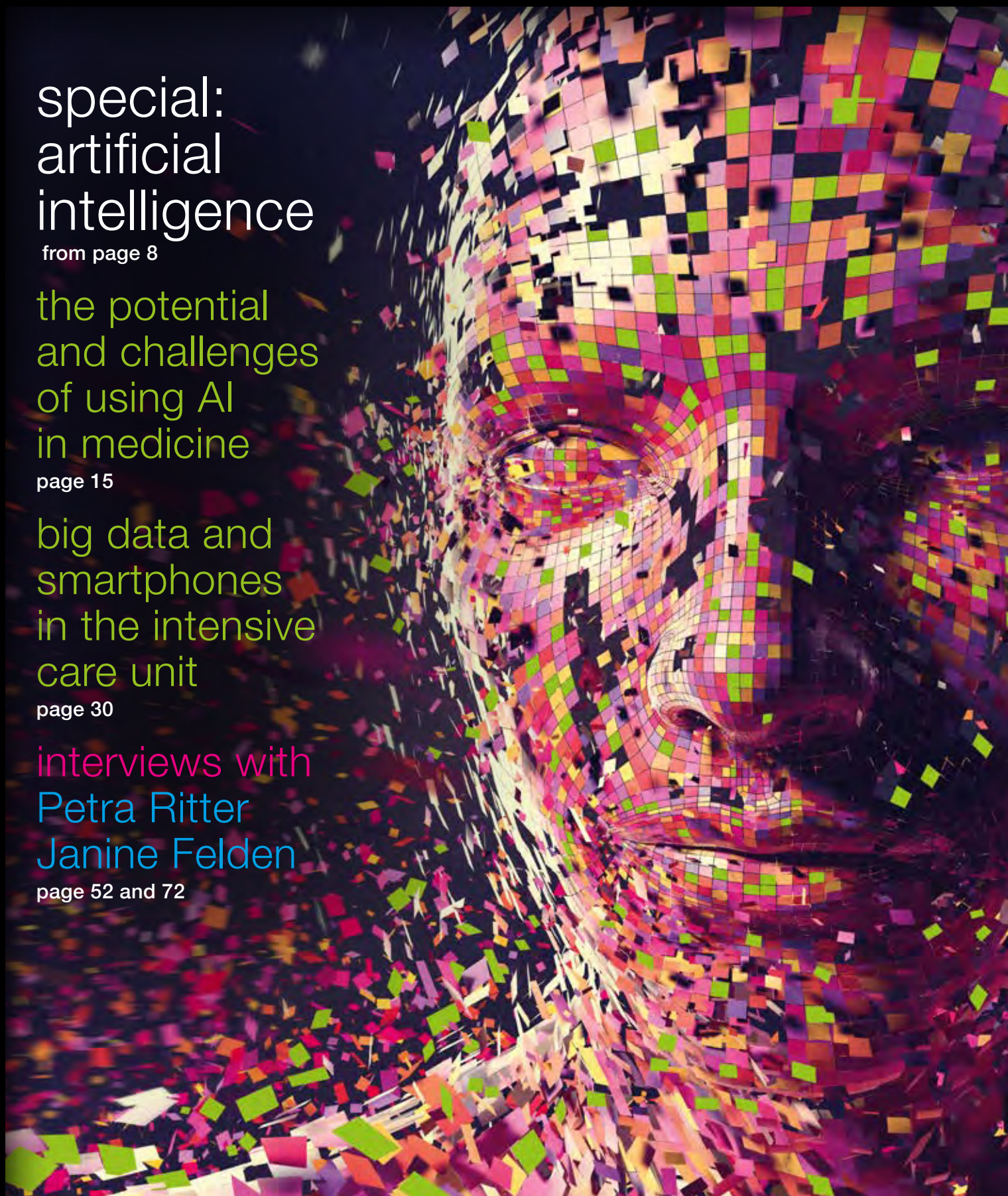
INTERNATIONAL EDITION

gesundhyte.de

THE MAGAZINE FOR DIGITAL HEALTH IN GERMANY

ISSUE 13 JULY 2021



special: artificial intelligence

from page 8

the potential and challenges of using AI in medicine

page 15

big data and smartphones in the intensive care unit

page 30

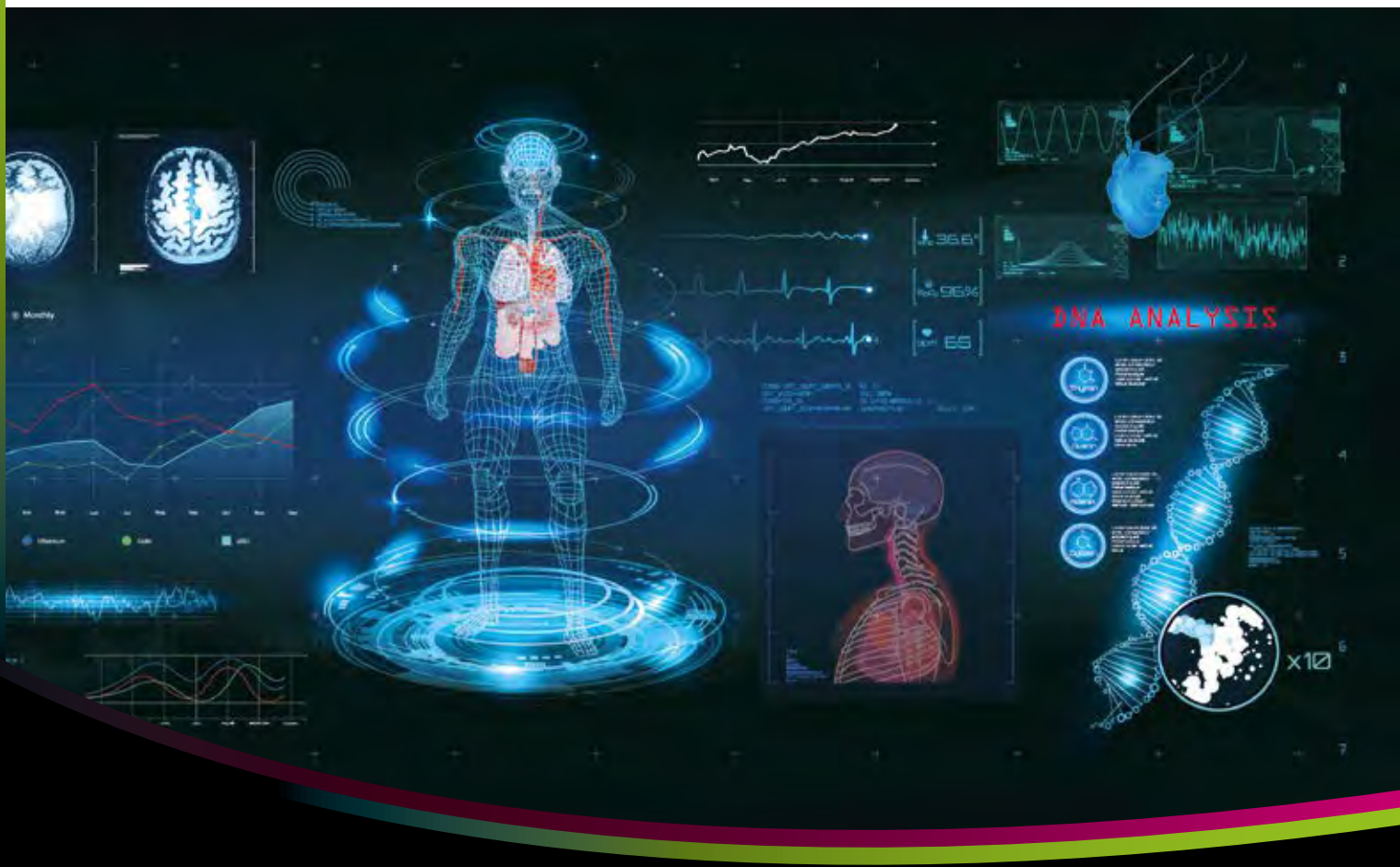
interviews with Petra Ritter Janine Felden

page 52 and 72

ISSN 2748-3533

SPONSORED BY THE

Federal Ministry
of Education
and Research



gesundhyte.de

is dedicated to finding solutions for today's most important challenges in health care focusing on Digital Health and Systems Medicine, two young disciplines that are considered to form the medicine of the future. The key is the combination of laboratory research data of different kinds with real-world data – from bench to bed-site. Innovative technologies and methods will pave the way for more precise predictions and personalised therapy. Already today, this approach is successfully used in Oncology and will be extended to other diseases in the future. Here, the establishment of robust and standardized IT infrastructure plays a major role to allow for the secure exchange of patient data with research teams. Read the magazine gesundhyte.de to find out how this innovative branch of science works to provide solutions for our current and future challenges in medicine.

Cover photo: iStock © peepo

Photo: AdobeStock_© SergeyBitos

greetings

Dear Reader,

The COVID 19 pandemic has presented the world with major challenges. However, it also has led us to realize the relevance of research. Due to the joint efforts of funders, the international research community and industry, our understanding of the disease, and the development of new diagnostics, therapeutics, and vaccinations have advanced at an unparalleled pace in the history of health research.

At the same time, there are still many unsolved research questions, not just concerning COVID-19, but also other severe diseases like cancer or dementia. To tackle these challenges, we need advances in experimental methods and techniques – and we need new approaches to model and analyse the data we collect. Computational biology and artificial intelligence are already indispensable and will play an even more important role in the future to decipher hidden connections between genomes, molecular factors and the condition of patients as well as their reaction to treatments.

These methods, however, rely heavily on large, accurate, and high quality datasets, which in turn require reliable infrastructures and a close collaboration of research and healthcare. The process of collecting and sharing of personal health information has to be transparent and address potential questions regarding ethical, legal and social aspects. Together with the German Federal Ministries of Health and Economics, the Federal Ministry of Education and Research (BMBF) has launched the "Data for Health" Innovation Initiative in order to create the necessary framework.

Advancing the digital transformation of the life sciences and medicine is an important priority in the BMBF Health Research Agenda. We support the efforts of the research community through a variety of funding initiatives such as "Computational Life Sciences", "German Network for Bioinformatics Infrastructure de.NBI" and the "Medical Informatics Initiative".

The projects featured in the current issue of gesundhyte.de highlight the broad scope and potential of artificial intelligence in life sciences and medicine. I hope that, like myself, you will enjoy the stimulating insights.

Prof. Dr. Veronika von Messling
Director General – Life Science Division
German Federal Ministry of Education and Research

Photo: Prof. Dr. Veronika von Messling

greetings

Dear Reader,

Last year's Nobel Prize in Medicine and Chemistry was awarded for hepatitis research and for the discovery of the CRISPR-Cas9 genetic scissors. CRISPR-Cas9, initially a basic research discovery, will offer more and more possibilities in the future in treating diseases through genome editing. Scientific research is continuously changing our diagnostic and therapeutic options. Just as medicine is changing, systembiologie.de has evolved into gesundhyte.de to focus its systems biological lens even more on clinical applications, new preventive measures and patient therapies.

The current pandemic has brought the issues of infection prevention and control and the tracing of transmission chains, as well as questions about adequate staffing levels at hospitals and nursing homes, to the fore. At Charité, we consider the current COVID-19 pandemic not only to be a major challenge, but also an opportunity: To ensure faster information exchange and the sharing of research data, we need to better link university hospitals, digitally empower patients and companies, and accelerate both initiated and ongoing processes, building on the consortia of the Medical Informatics Initiative.

In addition to focusing on infections, we will be addressing other important issues in this decade: in particular, the digitalization of medicine, which includes both the introduction of the electronic patient record (EPR) and research-compatible patient records, as well as personalized patient care. Precision medicine is not only becoming more important in the field of oncology, but also for other diseases, and will become the medicine of the future. In this area, Artificial Intelligence (AI) methods are driving new technologies and therapies. Above all, computer-aided decision-making tools – including machine learning methods – are increasingly playing a role not only in research, but also in daily hospital practice, whether it's assisting in patient diagnosis and the selection of the most promising therapy or helping to optimize care. There are many possible applications – we just need to know how to use them. AI also holds enormous opportunities in the COVID-19 crisis, which we should seize upon without disregarding associated risks. You will be able to review all these topics in this special "Artificial Intelligence" issue of gesundhyte.de.

I hope you find it an interesting read.

Prof. Dr. Heyo K. Kroemer
Chief Executive Officer, Charité – Universitätsmedizin Berlin

Photo: Wiebke Peitz / Charité

foreword

Dear Reader,

Did you know that “time” is the most commonly used noun in the German language? In our fast-moving era, it is probably most often associated with the negative statement “I don’t have time.” But it is also used to indicate a process of change, such as in the descriptive phrase “in changing times.” In much the same way, the field of systems biology is also continually undergoing change, a process reflected in the twelve issues of systembiologie.de since it was first published ten years ago.

After a somewhat longer pause, we have decided to make a fresh start. Since biomedical research has long been inconceivable without digital support, digitalization is now finally also making its way into the healthcare sector. Many hopes are also focused on the field of Artificial Intelligence (AI), which is expected to lead to major and frequently disruptive changes in healthcare. The promises are great, and the expectations equally so. Some of these promises will be realized within short-term time frames, while others will be a long time in coming. The topic is exciting and omnipresent, and will be with us for some time to come.

After ten years, we have made the decision to change the title of the magazine – from systembiologie.de to gesundhyte.de. You can look at this issue from both sides, the back side having the old design, the front side the new one. The topics we have chosen, which as usual have been brought to life by a large number of committed authors, are in step with the times. Of course, in this turbulent age, no medical magazine could possibly overlook the COVID-19 pandemic, and this issue also examines topics relating to the pandemic from an AI perspective. Even in the current COVID-19 situation, it is clear that digitalization is making a significant contribution to tackling the pandemic.

A new beginning is always associated with change. To this end, the editorial team has regrouped and we are pleased about the numerous new members who have given this magazine vital new momentum. The co-publisher is now the Berlin Institute of Health, which is itself undergoing an exciting process of transformation. But some things have not changed – we remain committed to reporting that is entertaining, but also always professionally oriented.

As I’m sure you’ve been counting, the word “time” appears ten times in my foreword. Even in these difficult times, I hope you will have the opportunity to read and enjoy this exciting magazine at your leisure.

Yours,

Roland Eils
Editor-in-Chief, gesundhyte.de

Photo: David Ausserhofer / BIH

index

AI and ethics in medical research

AI-based research in medicine and its ethical challenges

by Olga Levina

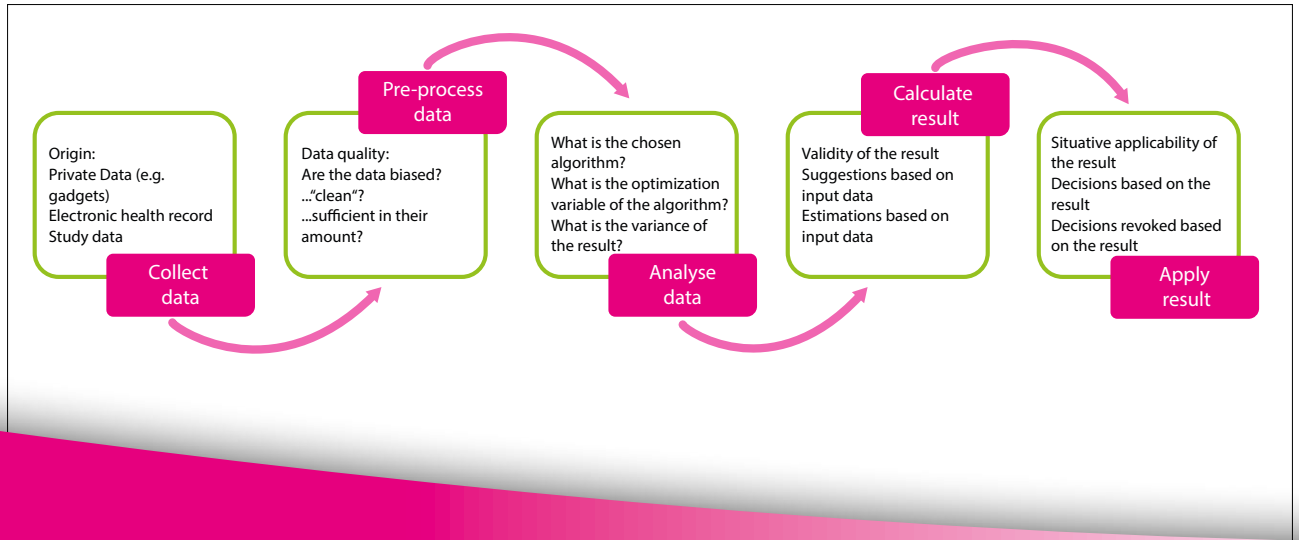
The issue of ethics in the context of artificial intelligence (AI) technologies is currently the subject of intense discussion and debate in the scientific community as well as among the public and policy-makers. The technologies in question can mostly be classified as so-called weak AI. These are information technology systems that focus on solving a specific problem using machine learning (ML) approaches. These are algorithms that are trained to target a specific problem domain. Hence, they can determine patterns in the data and contextually integrate them into relevant processes. ML-based technologies can identify patterns from large amounts of data and are thus well suited for structuring and analyzing medical data, which are often complex. Yet when using them in a real-world research environment, it is important to consider a number of aspects that may directly affect the quality of their output.

Machine learning systems in medical research

Using ML-based systems in medical research helps to realize several efficiency potentials. Among other things, patients can be assigned to specific studies more precisely or therapeutic agents can be identified and combined more efficiently. A ML-based system can evaluate complex clinical data and combine it with genetic information as well as information from medical studies in order to make statements about, for example, the effectiveness of drugs or therapies. Here, the ethical questions were already transformed into legislation, as it is important to ensure that all personal and sensitive data are adequately protected.

Current trends leading to the increasing use of digital and mobile applications and personal gadgets, such as smartwatches and personal fitness apps, open up the possibility for researchers to access a vast amount of digital personal behavioral data that was previously obtained mainly from self-reporting by patients or from observational studies. This development is crucial for medical research, since the amount and quality of data plays an important role in determining the results of ML-based systems. The more relevant and accurate

Figure 1: Societal issues in the data processing process (adapted from Levina, 2020)



Dr. Olga Levina is a research associate at the Berlin office of the FZI Research Center for Information Technology (Photo: Susanne Krautwald).

data used to train the system, the more accurate the generated output. Where the data comes from, how and when it was collected, and to what extent it relates to the research question are just some of the questions that are of relevance to derive reliable research statements. Unfortunately, when it comes to collecting and analyzing data from e.g. digital personal applications, it is currently not the researchers who dictate the format and content of – as well as access to – these data, but the IT companies that produce the apps.

Ethical challenges for using ML-based systems in medical research

In addition to the question of who has access to data, many of the ethical issues raised by the use of ML-based systems in medical research stem from the context of data-based systems, automation and trust in technologies. Consequently, these issues also become relevant for the research context as early as during the system's development. Figure 1 shows how a simplified data processing process works in practice during the design of an ML-based system.

From the very start, from data collection and the development and training of the algorithm, to the actual application of the results in the research process and beyond, such as when examining questions of responsibility, system developers and users must take socially relevant factors into account when designing a system that produces socially aware and trustworthy results.

Hence, the data analysis phase, see Figure 1, includes designing or choosing an algorithm to classify the outcome variables. Depending on such things as how the developers have defined the fault tolerance or the respective limits of to-be-identified classes, the question of the reliability of the results arises. The consequences of this at first glance mathematical decision, can be e.g. an incorrect recommendation by a decision support system to participating physicians to discharge patients who had not yet recovered (Caruana *et al.,* 2015). Or an advice for clinical treatment of patients despite that their recorded symptoms did not require it (Caruana *et al.,* 2015). Thus, although the calculation method is mathematically correct in its classification, it does not mean that the contextual requirements such as the interests of the treating physicians are being met.

In the result calculation phase, one key issue is transparency regarding decisions involved in the data processing. In order to be able to include e.g. a recommendation for a drug combination made by the ML-based system, researchers must be able to discern the basis for the recommendation, including what variables were included in the design. Without this information, the impact and applicability of the calculated outcome cannot be considered as robust. This therefore jeopardizes the validation of research results and their integration into medical knowledge.

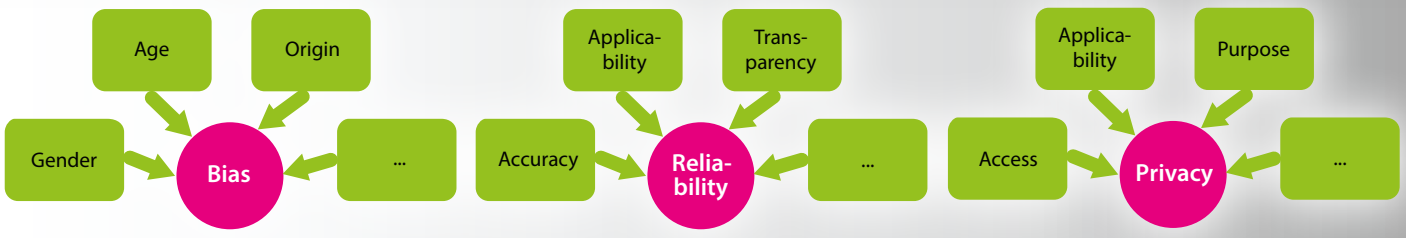


Figure 2: Selected ethical challenges in medical research (adapted from Levina, 2020)

Hence, data form the basis for the research results calculated by ML-based systems and thus for their future use in diagnostic systems, healthcare processes and therapeutic drugs. Accordingly, the quality and composition of data can have far-reaching consequences. Numerous studies have demonstrated that findings from medical research are often based on data that are poorly representative of the general population. For example, diagnoses and drugs have been developed for male patients that may be ineffective or even harmful for female patients (see, e.g., Popejoy and Fullerton, 2016). Moreover, collecting and analyzing data from geographic regions that have yet to undergo sufficient digitalization of their processes is particularly difficult and results in using the already trained ML-based systems in a different social and demographic context. The effects can be devastating. In sub-Saharan Africa, for example, women diagnosed with breast cancer are younger compared with their counterparts in Western countries. Consequently, systems trained to identify the disease at a more advanced stage and in older women cannot be used effectively in this setting (Black and Richmond, 2019). This bias in the data, i.e., the over- or underrepresentation of certain groups in the data set, means that advances in medical research are only applicable to certain populations. As data are removed from their context and considered a common good of the scientific community, a particular outlook on society and patients manifests itself in software code. Consequently, the use of the ML-based systems as well as the associated processes affect directly the current and future health of a large numbers of people.

Summary

AI technologies are already being used in many areas of medicine. While providing support in the research process, they create a number of ethical challenges that need to be recognized and mitigated, especially since these technologies can pose significant threats to patients' preferences as well as their safety, health and privacy. Yet current policies and ethical guidelines for the development of these systems lag behind the advances that AI technologies have made in healthcare delivery and health research. Although there are multiple efforts to engage in these ethical discussions, the medical community remains insufficiently informed of the ethical complexities that are implied by the expanding use of AI technologies in the research and treatment processes.

References:

Black, E., and Richmond, R. (2019). Improving early detection of breast cancer in sub-Saharan Africa: why mammography may not be the way forward. *Global Health* **15**, 3. https://doi.org/10.1186/s12992-018-0446-6

Caruana, R., *et al.* (2015). Intelligible models for healthcare, in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp1721-30.

Levina, O. (2020). A Research Commentary- Integrating Ethical Issues into the Data Process, in *WI2020 Community Tracks*. GITO Verlag, pp. 321–330. doi: 10.30844/wi_2020_z3-paper3.

Popejoy, AB., and Fullerton, SM. (2016). Genomics is failing on diversity. Nature 538:161.

Quartz (10 July 2017). If you're not a white male, artificial intelligence's use in healthcare could be dangerous.

The Society for Women's Health Research, The US FDA Office of Women's Health (2011) Dialogues on diversifying clinical trials.

Contact:

Dr. Olga Levina
Research associate
FZI Research Center for Information Technology
levina@fzi.de

https://url.fzi.de/levina

IEAI: at the forefront of shaping ethical AI

Generating global, egalitarian and interdisciplinary guidelines for the ethical development and implementation of AI

by Anastasia Aritzi and Caitlin Corrigan on behalf of the IEAI

Artificial intelligence (AI) is a powerful force that is transforming day-to-day life and interactions, environments and societies. The term encompasses a "collection of technologies and applications which process potentially very large and heterogeneous data sets using complex methods modelled on human intelligence to arrive at a result which may be used in automated applications."[1]

Billions of people use AI every day, mostly without noticing. Self-driving vehicles, digital assistants, chatbots, face recognition and personalized recommendations are just few examples of AI applications. AI is here to stay and will continue to transform society substantially. The most pressing question is how this transformation will take place and what the repercussions will be.

The Institute for Ethics in Artificial Intelligence (IEAI) seeks to be a leader in this process by exploring issues related to the use and impact of AI-based technologies. The IEAI, under the leadership of Professor Christoph Lütge, holder of the Peter Löscher Chair of Business Ethics at Technical University of Munich (TUM), forms part of TUM's Munich Center for Technology in Society (MCTS). Founded in 2019, the IEAI conducts independent, multidisciplinary research and celebrated its first anniversary last autumn.

The IEAI in brief

Imagine a scenario in which a company's AI-powered CV-screening software begins excluding people based on gender, sexuality, race, religion, marital status, pregnancy status, disability or age. Consider of a case in which decisions made by algorithms favour of the interests of cars' passengers at the expense of the safety of others (e.g. pedestrians or cyclists).

In both of these cases, in order to support the effective and safe advancement of the technologies, ethical concerns need to be adequately addressed. The example of CV screening software shows the potential for AI to alter or enhance discrimination in the workplace. Algorithmic specifications also can affect human

"AI cannot fly without ethics."
Prof. Christoph Lütge, Director of IEAI

Photo: Sonja Herpich

[1] German Data Ethics Commission - Recommendations of the Data Ethics Commission for the Federal Government's Strategy on Artificial Intelligence (9 October 2018)
https://www.bmi.bund.de/SharedDocs/downloads/EN/themen/it-digital-policy/recommendations-data-ethics-commission.pdf?__blob=publicationFile&v=3

safety when they are making decisions for vehicle movement. And issues such as accountability, transparency and violations of privacy come into question when we are talking about systems that use massive amount of potentially personalized data to make impactful decisions.

IEAI's inter-, multi-, and transdisciplinary research approach seeks to identify these challenges, deliver practical and action-oriented outcomes and develop tangible frameworks, methodologies, and algorithmic tools to provide AI developers and practitioners with a set of ethical best practices. Faculty members from two distinct disciplines, supported by highly skilled researchers, experts and doctoral students, work closely together towards the development of operational ethical frameworks in the field of AI.

Seven clusters represent the IEAI's current key research areas, with individual projects conducted within these clusters (See Table 1). These research projects seek to answer questions such as: How can AI best be used in workplace? Is it possible to integrate ethical behavior into autonomous vehicles? How might AI-driven marketing challenge consumer choice, autonomy and behavior? And what ethical and performance challenges or risks may result from the use of AI in health care related tasks?

Through an initial funding grant from the US company Facebook, support from TUM and the Bavarian Government and a growing group of external partners, the IEAI is able to conduct long-term, expansive and independent research. The Institute has an approach that promotes active collaboration between the technical, engineering and social sciences, while also actively courting interaction with a wide group of international stakeholders from academia, industry and civil society. This exhaustive approach enables the IEAI to comprehensively address a growing group of ethical challenges arising at the interface of technology and human values. It also aids in the development of thoroughly operational ethical frameworks in the field of AI. In its role as a platform for building cooperation, the IEAI organizes workshops, conferences and seminars to promote exchange between a wide range of important stakeholders (academia, civil society, government, industry etc.). To this end, the institute also provides leadership and participates in several networks related to AI ethics such as: AI4People, ITU Focus Group on AI for Autonomous & Assisted Driving, the Global AI Ethics Consortium and the Responsible AI Network-Africa. By joining forces with researchers and practitioners worldwide, the IEAI is able to address real-world challenges and contribute to the broader conversations and concerns surrounding ethics and AI on an international level.

Table 1: The IEAI key research areas

1. **AI AND THE FUTURE OF WORK**
   A Human Preference-aware Optimization System
2. **AI, MOBILITY AND SECURITY**
   ANDRE – AutoNomous DRiving Ethics
3. **AI, CHOICE AND AUTONOMY**
   Consumer Perception and Acceptance of AI-enabled Recommender System
4. **AI IN MEDICINE AND HEALTH CARE**
   Building Strategic Partnerships to Understand Ethics and the Use of AI to Manage Health Related Crises
   ETHAN – Ethical AI for Pandemic Management
   METHAD - Toward a MEdical ETHical ADvisor System for Ethical Decisions
   Public Trust in AI and the Ethical Implications: A Comparative Study of Governmental Use of AI During the Covid-19 Pandemic
   Rule of Law, Legitimacy and Effective COVID-19 Control Technologies
   The Ethics and Practice of AI Localism at a time of COVID-19 and beyond
5. **AI AND ONLINE BEHAVIOUR**
   Online Firestorms and Resentment Propagation on Social Media: Dynamics, Predictability, and Mitigation
   Online-Offline Spillovers-Potential Real-world Implications of Online Manipulation
   Personalized AI-based Interventions Against Online Norm Violations: Behavioural Effects and Ethical Implications
6. **AI, GOVERNANCE AND REGULATION**
   TrustMLRegulation - Managing Trust and Distrust in Machine Learning with Meaningful Regulation
7. **AI AND SUSTAINABILITY**
   Artificial Intelligence for Earth Observation (AI4EO)

For more information about the Projects visit https://ieai.mcts.tum.de/research/
Table's last update: January 2021

KI, ethics and the COVID-19 crisis

From early warning and contagion trend detection about the novel coronavirus spread, to individual contact tracing for infections and identification of effective treatments, AI is playing a multitude of roles during the COVID-19 crisis. In the process of managing the COVID-19 pandemic, many AI-based tools have been used (or are being developed) to gather and analyze relevant data, develop treatments, make medical decisions, track infected populations and manage quarantines and information dissemination. However, the use of some of these tools (e.g. contact-tracing apps, facial recognition, quarantine surveillance) has fuelled discussions around privacy, security and human rights and raised serious ethical questions about the tough trade-offs that might extend well beyond this crisis. Of course, there are always trade-offs, but these trade-offs need to be transparent, clearly identified and explained.

AI is a powerful tool that has the ability to strengthen and amplify efforts for managing a pandemic in many of the ways mentioned above, but it also has the potential to entrench and increase harms and discrimination if not used responsibly (e.g. inaccurate, insufficient or biased data can lead to inappropriate or discriminatory outcomes). What the COVID-19 crisis has made clear is the need for a coordinated, dedicated data and technology infrastructure and ecosystem for tackling dynamic societal and environmental threats. The crisis has also underlined the urgent need to discuss the ethical considerations in the use of AI in healthcare and public health contexts and develop operational ethical frameworks in the field of AI.

The IEAI's response to COVID-19 crisis

↗ Research Brief: Ethical Implications of the Use of AI to Manage the COVID-19 Outbreak

"Explicit and well-defined ethical guidelines can play an important role to manage health-related crises and prevent pandemics. However, practitioners must be able to trust in AI systems in order to use them effectively."[2]

Do we value public health and beneficence over *individual privacy* and *autonomy*? How does AI enabled decision-making for questions of who receives limited resources for healthcare reconcile with the need for justice and non-maleficence? In the case of a rush to create tools and put them into action, where does explicability rank in the order of importance?

[2] IEAI. (2020). Ethical Implications of the Use of AI to Manage the COVID-19 Outbreak. April 2020. https://ieai.mcts.tum.de/wp-content/uploads/2020/04/April-2020-IEAI-Research-Brief_Covid-19-FINAL.pdf

Figure 1: IEAI's human capital in numbers. For more information about the people working at IEAI visit https://ieai.mcts.tum.de/about-ieai/people
Figure's last update: January 2021

These are some of the questions that must be considered alongside the increasing development of AI tools to be used in healthcare crises. In its growing role as a resource for information on AI ethics, the IEAI published a Research Brief in Spring 2020. The Brief posed questions on the ethical considerations in the use of AI to manage health-related crises, outlined some of the current and potential uses for AI-based tools in managing pandemics and discussed the ethical implications of these efforts. Visit IEAI's website (Research and Publications section) and read the Research Brief.

↗ Launching the Global AI Ethics Consortium

"The time for analyzing how AI is deployed -whom it affects, how it affects them, what are its broader social and economic impacts- is now."[3]

The various uses of AI to manage pandemics, as well as the ethical challenges related to them (e.g. data privacy, data protection) require multidisciplinary and multi-stakeholder engagement, as well as international collaboration on developing AI governance. To this end, the IEAI led the way in the creation of the *Global AI Ethics Consortium (GAIEC)* in April

[3] Global AI Ethics Consortium. (2020). The Global AI Ethics Consortium on Ethics and the Use of Data and Artificial Intelligence in the Fight Against COVID-19 and other Pandemics- Statement of Purpose. https://ieai.mcts.tum.de/wp-content/uploads/2020/05/COVID-19-Statement-of-Purpose-FINAL-May-2020.pdf

The Institute for Ethics in Artificial Intelligence is an institution of the Technical University of Munich. (Photo: Andreas Heddergott/TUM).

2020. The GAIEC has an initial focus on promoting *Ethics and the Use of Data and Artificial Intelligence in the Fight Against COVID-19 and other Pandemics*. It brings together prominent members of academic institutions and research centers worldwide to promote trust in data and technology, maximize the potential of AI while limiting its harms, help to navigate current uncertainty and create ethical frameworks. The GAIEC aims to make its expertise available to other stakeholders, launch its own projects and create an accessible repository for all research results on ethics in artificial intelligence in the context of the COVID-19 crisis. As part of the GAIEC, the IEAI in cooperation with The Gov Lab at NYU released a free online course "AI Ethics – Global Perspectives" in February 2021.

↗ Call for multidisciplinary research related specifically to ethics and the use of AI to manage pandemics and health-related crises

"Academic research plays a fundamental role in designing tools and frameworks that can help governments, companies, NGOs and other actors navigate and arrive at sound and justifiable responses in the preparation and management of pandemics."

The various uses of AI to manage pandemics, as well as the ethical challenges related to them, require also multidisciplinary research. Therefore, the IEAI announced in Summer 2020 a call for proposals that seek to study and design research-based concrete and practical solutions for issues related specifically to ethics and the use of AI to manage pandemics and health-related crises. The selected research proposals were announced in August 2020 with the projects starting shortly thereafter (See Table 1 - AI in Medicine and Health Care).

What's next

"TRAIF 2021: Promoting a sustainable, inclusive and comprehensive framework for the use of AI that delivers global benefit."

The Responsible AI Forum 2021 (#TRAIF2021) aims to (1) bring together members of industry, civil society, government and academia to discuss the most relevant and pressing issues related to the responsible use of AI through shared stories, cutting edge research and practical applications and (2) encourage exchange between research and practice through productive discussion and demonstration. Organized by the IEAI, TRAIF 2021 is scheduled to take place this winter in Munich. Learn more and be part of it: responsibleaiforum.com.

IEAI

Contact:

Prof. Dr. Christoph Lütge
Director of the IEAI
INSTITUTE FOR ETHICS IN ARTIFICIAL INTELLIGENCE
Munich
ieai@mcts.tum.de

https://ieai.mcts.tum.de

the potential and challenges of using AI in medicine

Improving recovery chances, preventing diseases, empowering patients

by Klemens Budde and Karsten Hiltawsky
for the Working Group on Health Care, Medical Technology, Care of Plattform Lernende Systeme

From prevention and early diagnosis to patient-focused therapy and beyond, artificial intelligence (AI) and self-learning systems have the potential to facilitate better therapeutic outcomes and to improve healthcare in general. Yet the deployment of AI-based decision support systems in the healthcare system comes with its own challenges. It requires high IT security standards as well as the trust of medical professionals, care workers and patients. This article addresses the opportunities, challenges and requirements for the secure, safe, ethical and economically viable use of AI in healthcare and care provision.

While the current coronavirus pandemic highlights both the vulnerability and resilience of our healthcare system, the crisis also offers us the opportunity to leverage innovative technologies to increase the resilience of this system. Epidemiology in particular – i.e., the study of diseases on a population level – benefits from AI and machine learning methods that allow researchers to analyze large amounts of data efficiently in order to generate new insights. By facilitating improvements in disease prevention, early detection, and targeted treatment selection, AI can also contribute significantly to better and more personalized patient care now and in the near future. This article is based on a **report**, prepared by the Working Group on Health Care, Medical Technology, Care of **Plattform Lernende Systeme**, that summarizes significant opportunities and challenges of self learning systems in healthcare. Initiated by the German Federal Ministry of Education and Research (BMBF), the **Plattform Lernende Systeme** comprises focused working groups that bring together leading experts in AI and self-learning systems. The aim of the platform is to ensure that AI technologies are deployed and developed in the best interests of society.

Lernende Systeme
GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE

The potential of AI in medicine

A glimpse into the hospital of the future reveals the wide range of potential applications of self-learning systems: Using AI systems to interpret imaging results, physicians will be able to diagnose their patients with greater accuracy and precision. AI will increasingly allow clinicians and researchers to develop new approaches to prevention, study rare genetic diseases and discover previously unknown medical correlations. By using networked data, self-learning systems can recommend suitable preventive approaches or treatments and thus support physicians in their decision-making processes. AI-based analyses of medical imaging results, from radiography to nuclear medicine imaging, magnetic resonance imaging and ultrasound examinations of organ systems, will be more precise, faster, and more reliable than ever before. Yet because treatment decisions cannot be derived solely from AI algorithms and will always require human intelligence and agency, medical and care professionals cannot and should never be replaced by these techno-

logical innovations. Rather, AI-based decision support systems should provide a complementary yet important source of information on which to base treatment choices and other decisions. Furthermore, physicians will continue to be indispensable in their role as communicators who discuss diagnoses and treatment options with their patients in order to arrive at the best possible treatment decisions.

Care provision is another area in which the interaction of humans and technology offers great potential. AI-driven voice recognition could help care workers with routine tasks such as documentation, giving them more time to provide personal attention. Soon, robot assistants and AI-based technologies (e.g., exoskeletons) may also allow us to live independent lives well into our old age. The better treatment options and improved chances of recovery made possible by medical AI applications are not only beneficial to those who are ill or in need of care. Self-learning systems hold particularly high potential for disease prevention in healthy patients and early diagnosis. Smartphone apps and wearables, for example, record and analyze their owners' personal health data with privacy-preserving methods. The results are used to support healthier lifestyle choices and detect early symptoms of illness.

Sharing and protecting health data

To ensure that self-learning systems in healthcare are used for the good of patients, certain prerequisites need to be put in place. This includes the creation of a health database or distributed health databases, as AI systems are only as good as the data available to them. In order for self-learning processes to perform, they need sufficient volumes of diversified, usable data. As a rule, an AI system can become more reliable and produces more stable results with each data augmentation. This rule also applies to AI applications in healthcare. Moreover, the system should combine data from multiple sources so as to generate applicable analyses with high validity. The development, deployment and acceptance of AI-based solutions in the healthcare system therefore requires representative, structured and managed health databases. Personalized health data collected during the provision of standard healthcare services should be made available continuously to machine learning applications. These data could then be used, for example, to support both outpatient and inpatient care. Patient data and patient-generated health data pose questions concerning the usage and storage of such data as well as IT security in AI systems. It also raises concerns regarding privacy and the fear of the "transparent patient." For enough people to consent to the use of their data in a linked database and for the deployment of AI in medicine to be accepted by society, these questions and concerns need to be addressed and patients and care recipients must be given access to and control over their health data.

Building skills in medical training and care provision

The new technologies will change the job requirements and daily work routines of medical and care professionals. To use AI-based solutions in diagnostics or treatment selection, medical practitioners need basic knowledge of machine learning and information technology (AI literacy). The doctor-patient relationship will also change as algorithms support decision-making processes and patients take a more active role in col-

Publication "Self-Learning Systems in the Healthcare System"
The paper is based on a report by the Working Group on Health Care, Medical Technology, Care of Plattform Lernende Systeme. The expert report illustrates the potential benefits of AI technologies through research examples and an application scenario on lung cancer.

Lernende Systeme – Germany's Platform for Artificial Intelligence (ed.) "Self-Learning Systems in the Healthcare System – A Report by the Working Group on Health Care, Medical Technology, Care," Munich 2019. The full German version is available online at: www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG6_Bericht_23062019.pdf

An executive summary in English is available at: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen_EN/AG6_Executive_Summary_final_200206.pdf

Exploiting the potential of AI in medicine (Photo: iStock © metamorworks)

lecting and analyzing data. Realizing the potential of AI in the healthcare system requires skill building in medical training and care provision as well as research on human-machine interaction. Clinical flagship projects can serve as role models for translating research insights into practice and testing their implementation and benefits. Advances in healthcare AI will also require investments from companies in the medical technology, pharmaceutical, hardware and software sectors.

Bringing AI innovations to patients and care recipients

Introducing AI in the healthcare system is a far-reaching innovation process. It requires careful integration of all relevant stakeholders, including patient associations, employee representative bodies and ethics committees. There are also regulatory issues to be addressed. These include setting up liability frameworks for the use of AI innovations by medical and care professionals. Regulatory approval of AI decision support systems poses specific challenges as these systems may continuously evolve by learning from new data. Existing rules for clinical trials and medical device approval fail to take this evolutionary potential of AI-based systems into account. Thus, new rules for self-learning medical devices need to be developed in order to provide legal certainty for manufacturers.

However, the benefits for patient and care recipients must be at the heart of any technological advancement. One great challenge in this context is the transparent presentation of results. For even if the system demonstrably delivers good results, it must be possible to establish why a particular result is achieved (causality). Replicability will also be a key criterion for further dissemination and use, because both physicians and patients will want to know which medical findings and data results are based on, especially with regard to potentially erroneous primary data. In order to advance AI innovation in healthcare and ensure it is used for the benefit of all patients, relevant research at university hospitals and other research institutions needs to be strengthened. In addition, we need companies that will develop the technologies into market-ready solutions so that hospitals, doctors' offices and care facilities can deploy safe and useful AI innovations. A precondition for investment in the development of AI medical products is potential reimbursement, which requires a clear legal framework that enables and regulates the use and development of AI applications in healthcare. It also requires regulatory and technical frameworks that will render the results of AI-based innovations transparent, measurable and comparable. This includes the regulation and funding of clinical trials that validly demonstrate the safety, efficacy and usefulness of machine learning-based medical devices.

Outlook: The importance of further ethical debates

The use of AI-based solutions also requires a critical debate of relevant ethical issues as well as public acceptance of artificial intelligence in medicine. Only if effective data protection and compliance with social norms and values are guaranteed will people trust AI. Using AI raises numerous ethical issues, and even more so in the healthcare system, which should give highest priority to the individual's physical and psychological integrity and which deals with most sensitive data. A broad public debate about technological innovations is therefore crucial. Patient associations and ethics committees should be involved in the discussion about ethical standards and in policymaking bodies early on. There will continue to be ethical dilemmas, and the answers to some questions are still missing as the effects of these technological innovations cannot be foreseen in their entirety. To date, important legal frameworks for the use of health data for AI are lacking; they require a careful balancing of interests between medical benefits and the protection of personality rights. The introduction of AI in the healthcare system may also lead to new conflicts of interest among different stakeholders (e.g., the scientific community, the private sector, start-ups, pharmaceutical companies, medical technology companies, charity organizations and health insurers). It also remains unclear how AI may impact the decision-making autonomy of physicians, care workers and patients. Basing judgments and decisions on ethical criteria is an integral part of everyday medical practice and care provision. Self-learning systems will not be able to take over this task, which is why AI will never be more than effective support for human action. Appropriate regulation by policymakers must take this into account.

Contact:

Prof. Dr. med. Klemens Budde is Co-Head of the Working Group on Health Care, Medical Technology, Care of Plattform Lernende Systeme and Deputy Medical Director of the Medical Department, Division of Nephrology and Internal Intensive Care Medicine at Charité – Universitätsmedizin Berlin.
klemens.budde@charite.de

https://nephrologie-intensivmedizin.charite.de/metas/person/person/address_detail/budde-1/

Dr.-Ing. Dr. med. Karsten Hiltawsky is Co-Head of the Working Group on Health Care, Medical Technology, Care of Plattform Lernende Systeme and since 2012 has worked at Drägerwerk AG & Co. KGaA, where he is currently Head of Technology & Intellectual Property.
karsten.hiltawsky@draeger.com

About Plattform Lernende Systeme

Designing self-learning systems for the benefit of society is the goal pursued by the Plattform Lernende Systeme which was launched by the Federal Ministry of Education and Research (BMBF) in 2017 at the suggestion of acatech. The members of the platform are organized into working groups and a steering committee which consolidate the current state of knowledge about self-learning systems and Artificial Intelligence. They point out developments in industry and society, analyze the skills which will be needed in the future and use real application scenarios to demonstrate the benefit of self-learning systems. A Managing Office at acatech coordinates the work of the platform.

Photos: Draegerwerk AG Co. KGAA and Wiebke Peitz

"doctor, do you understand me?"

How learning systems help to understand medical terminology and which role clinical guidelines play in this context

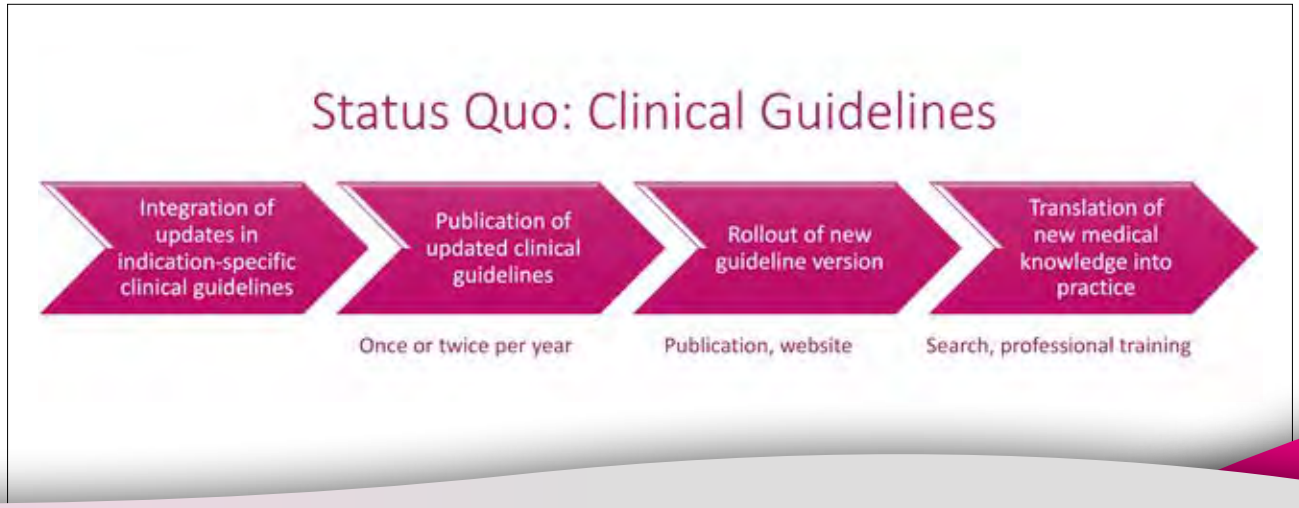
by Florian Borchert, Christina Lohr, Luise Modersohn, Udo Hahn, Thomas Langer, Gregor Wenzel, Markus Follmann, and Matthieu-P. Schapranow

Today, medical knowledge relevant for the treatment of cancer patients is provided in the form of so-called clinical guidelines. Their content is arranged according to indication and maintained – often through manual research – by qualified experts of the national oncology societies. As soon as new findings are made available, existing clinical guidelines are updated and new versions are published.

Dear readers, we would like to take you on a journey into the not-too-distant future, when the health system is supported by digital solutions in many areas. Even the treatment of cancer patients benefits from them, for example, through the use of learning systems to help amongst others with the mandatory clinical documentation, support the decision making process and predicting the probability for clinical endpoints. Figure 1 shows today's process of creating clinical guidelines whilst Figure 2 shows how, in future, evidence-based oncology guidelines will be put into clinical practice. Thanks to the use of learning systems based on latest Artificial Intelligence (AI) algorithms it will happen faster and more frequently than today possible.

The exchange of medical data between hospitals in Germany happens via the digital infrastructures established by the Medical Informatics Initiative. Hospitals and clinics are more than mere data repositories; they are also sources of new medical knowledge. They provide evidence-based insight into the success of therapies on the basis of concrete cases, and are therefore important partners in the evaluation of clinical practice guidelines. AI algorithms enable the automatic analysis of hundreds of thousands of patient histories in a blink of an eye. Experts from oncology societies regularly evaluate the results of the AI algorithms and decide to what extent existing guidelines need to be updated. Machine-readable versions of the clinical guidelines are becoming increasingly important, for example, to support knowledge transfer and quality assurance conducted by clinical information systems. The automatic processing of human language – also known as natural language processing (NLP) – will be an important technical aspect in future.

Figure 1: Today's linear clinical guideline development process (simplified representation)



Source: Matthieu-P. Schapranow

Figure2: Learning AI-based Clinical Guideline System. Iterative process for guideline generation using a learning AI-based clinical guideline system (Source: Schapranow, based on www.flaticon.com/packs/medical-icons and https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/5_Publikationen/Ministerium/Flyer_Poster_Etc/Unser_Gesundheitssystem_Schaubild_2020.pdf).

Machine learning for the processing of biomedical texts

The processing of human language by computers has always been a central goal of AI. However, human language is a complex problem for machine processing for many reasons, for example, it can be ambiguous, context-dependent and often only interpretable through implicit information. In specific areas, like medicine, additional challenges arise, such as the distinct technical vocabulary, the frequent use of abbreviations and the highly condensed linguistic style that differs significantly from everyday language.

Machine learning can also help to achieve considerable progress in the research area of NLP. The currently most successful procedures for many tasks, for example, the extraction of medically relevant information from texts, are based upon a process called supervised learning. For supervised learning, human annotators add content-specific metadata manually to available collections of actual texts (so-called text corpora) to provide the algorithms with suitable input from which they can learn. The availability of large, richly annotated text corpora in combination with machine learning methods based on artificial neural networks (deep learning) is a major driving force in current NLP research (Wu *et al.*, 2020). So far, NLP research has been dominated by English-language texts, as English-language text corpora are more widely available than any other language. However, this focus on English-language documents is problematic. For example, language analysis learning models trained solely on English texts are generally not applicable to texts in other languages.

The lack of NLP models specific to individual languages presents a real obstacle to the automatic analysis of clinical guidelines, because they are usually published in the respective national language. This problem is particularly acute in Germany, as almost no other medical texts are available in German for NLP machine learning models. Specialist texts are usually published in English-language journals or are subject to copyright restrictions. Furthermore, clinical documents like discharge letters and pathology reports are only accessible under very high data protection requirements and are often not available to researchers in the required de-identified and publicly accessible form.

GGPONC: A German-language text corpus based on clinical practice guidelines

German-language NLP research can now really take off thanks to partners in the HiGHmed and SMITH consortia of the BMBF Medical Informatics Initiative. With GGPONC (Borchert *et al.*, 2020), the Hasso Plattner Institute for Digital Engineering at the University of Potsdam, the JULIE Lab at Friedrich Schiller University Jena and the German Guideline Program in Oncology

have created a German-language medical text corpus based on clinical guidelines for the very first time, and were able to make it accessible to researchers in a very short space of time at the following URL:

www.leitlinienprogramm-onkologie.de/projekte/ggponc-english/

The German Cancer Society's clinical guidelines are an important data source for NLP research. They provide a perfect basis for the creation of a German-language medical text corpus that is completely free from personal health information and thus has no data protection restrictions.

The guideline recommendations and background texts, which form the basis of the new German Guideline Program in Oncology app, are provided by the German Cancer Society itself. The first version of the corpus contains over 1.3 million tokens from 25 guidelines, and is therefore one of the most comprehensive German-language clinical text corpora to date – and the only one that is also publicly accessible.

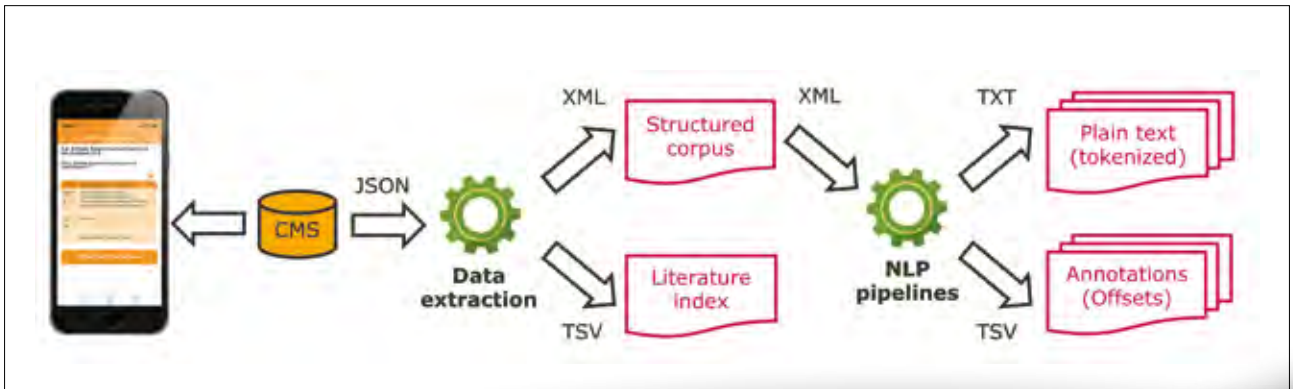
Figure 3 shows how, once the text resource has been created, medically relevant information types are automatically identified in the guideline texts by NLP techniques. German-language UMLS and TNM terminology form part of the basis of this process. A selected subset of these automatically identified information types were validated by medical experts and are now provided as a "gold standard" for further work. Also, by mapping the information types found onto other ontologies, such as SNOMED CT, it becomes possible to enrich the texts with semantic information to make them more machine-readable.

How can machine learning systems assist clinicians?

Going forward, the question is in what form clinicians and medical experts can be involved at an early stage in the development of NLP systems. The dependency of supervised learning on manually annotated data is one limitation of such approaches. It also seems an inefficient use of expert knowledge that professionals are incorporated into such systems purely to carry out time-consuming annotations.

One potential solution would be to adapt models from other areas to the context of medicine (transfer learning). This would make it possible to train high-quality word representations in the context of deep learning (word embeddings) using public data like our medical text corpus GGPONC. Such an approach would enable the development of IT systems in a clinical context with less training data required from the target area.

Figure 3: Pipeline for the automated creation of the annotated GGPONC corpus using data from the German Guideline Program in Oncology's content management system



Source: Florian Borchert

In addition, various formats allow also to have humans in the loop during development. For example, a dedicated NLP system for information extraction could be directly integrated into the content management system, e.g. to support automatic keyword extraction in the course of guideline creation and update. The human editor can immediately give feedback to the model predictions to allow continuous improvement of the model. Also promising are novel approaches known as "weak supervision", which enable NLP systems to be trained with less human involvement. Instead of making small-scale annotations, experts in a particular area can encode their knowledge in the form of heuristics, metadata or by accessing existing curated knowledge databases, and thus automatically generate annotations for a large set of texts (Ratner, Hancock & Ré 2019). Researchers are currently working out what the interfaces for this sort of interaction with a learning system might look like.

Project profile:

The German Guideline Program in Oncology NLP Corpus (GGPONC) is a collaborative project that was started in 2019 by the BMBF Medical Informatics Initiative, represented by the Digital Health Center of the Hasso Plattner Institute and the JULIE Lab of Friedrich Schiller University Jena, together with the German Cancer Society.

Referenzen:

Seufferlein, T., Kopp, I., Post, S., Jonat, W., Kreienberg, R., Nothacker, M., Marks, A., Nettekoven, R., Langer, T., Follmann, M., Bamberg, M. (2019). Onkologische Leitlinien – Herausforderungen und zukünftige Entwicklungen. Forum. 34(3) : 277–283.

Borchert, F., Lohr, C., Modersohn, L., Langer, T., Follmann, M., Sachs, J. P., Hahn, U., & Schapranow, M.-P. (2020). GGPONC: A Corpus of German Medical Text with Rich Metadata Based on Clinical Practice Guidelines. ArXiv:2007.06400 [Cs]. http://arxiv.org/abs/2007.06400

Cohen, K. B., and Demner-Fushman, D. (2014). Biomedical Natural Language Processing (Vol. 11). John Benjamins Publishing Company.

Ratner, A.J., Hancock, B., and Ré, C. (2019). The Role of Massively Multi-Task and Weak Supervision in Software 2.0. CIDR.

Wu, S. T. *et al.* (2020). Deep Learning in Clinical Natural Language Processing: A Methodical Review in Journal of the American Medical Informatics Association, 27, 457–470.

Contact:

Florian Borchert, M.Sc.
HPI Digital Health Center,
Hasso Plattner Institute,
University of Potsdam
florian.borchert@hpi.de

https://hpi.de/digital-health-center/members/working-group-in-memory-computing-for-digital-health/florian-borchert.html

Dr.-Ing. Matthieu-P. Schapranow
Working Group Leader
"In-Memory Computing for Digital Health"
Scientific Manager Digital Health Innovations
HPI Digital Health Center,
Hasso Plattner Institute,
University of Potsdam
schapranow@hpi.de

https://hpi.de/digital-health-center/members/working-group-in-memory-computing-for-digital-health/dr-ing-matthieu-p-schapranow.html

Photos: Schapranow and Borchert

precise diagnostics with artificial intelligence

Company profile: Aignostics GmbH

by Stefanie Seltmann

The usual way to determine if a suspicious lump is cancerous involves microscopic examination in a pathology laboratory. To ease the workload of pathologists who face a rising number of cancer cases and increasingly finer molecular details and to enable tailored treatments, scientists at the Institute of Pathology of Charité – Universitätsmedizin Berlin have teamed up with colleagues at TU Berlin to develop a digital image analysis system that uses Artificial Intelligence (AI) to evaluate microscopic images. The Digital Health Accelerator (DHA) of the Berlin Institute of Health (BIH) is helping the scientists to make the long-term research project a translational success. In April 2020 the scientists founded the Aignostics GmbH.

AIGNOSTICS

Professor Frederick Klauschen studied physics and medicine, "a good combination for pushing forward the digitalization of healthcare," says the Deputy Director of Charité's Institute of Pathology on Campus Charité Mitte. With ever more samples from ever more patients, digital assistance systems can help to avoid mistakes," Klauschen explains. "And humans are not as good at determining what percentage of tissue is affected by cancer or what percentage of tumor cells contain a certain therapeutically relevant receptor. Here our 'digital colleague' can provide valuable help, because it is both faster and more precise when it comes to quantitative analysis."

In order to train the "digital colleague" in diagnostic pathology, the team led by Klauschen collaborated with lots of human colleagues from various university hospitals. They mapped the pathological changes on thousands of digital microscopic images of tissue sections. The pathologists "fed" the software with these findings, so that it could "learn" to distinguish, for example, tumor tissue from healthy tissue. "We generated these so-called annotations for different diseases," explains Klauschen. The scientists have so far trained the software to reliably identify lung, breast and colon cancer as well as immune cells in tumor tissue and various tumor markers. It can also be used to analyze infections as well as degenerative, connective tissue and autoimmune diseases, "anything that causes visible changes in tissue."

Explainable Artificial Intelligence

"The system we developed also reveals how the AI makes its decisions," explains Professor Klaus-Robert Müller, co-founder of Aignostics and Chair of the Machine Learning Department at TU Berlin. "The software creates so-called heat maps, which show precisely which cells or image areas were decisive for the algorithm's classification of the tissue as cancerous or non-cancerous." By using these heat maps the pathologists can assess whether the AI analysis is plausible and correct. But the technology also opens up entirely new possibilities for research. If, for example, one trains the AI with the positive and negative courses of a particular therapy, the resulting heat

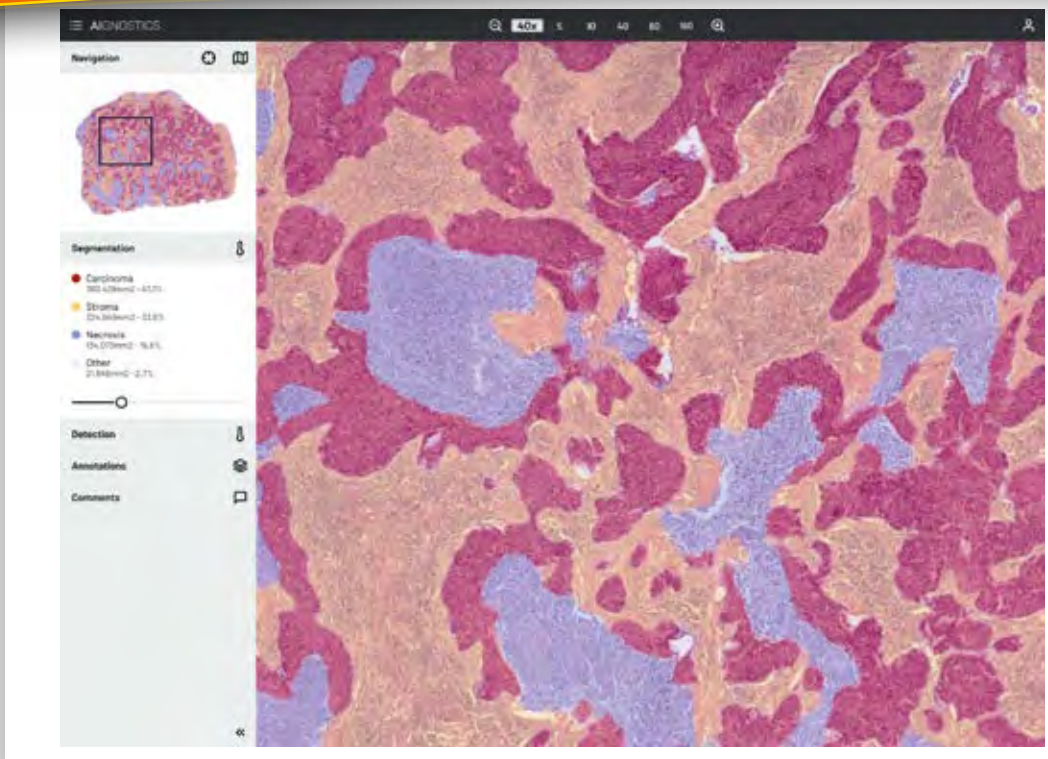


Figure 1: Aignostics Portal (Source: Aignostics GmbH).

maps could enable pathologists to discover new characteristics ("biomarkers") that predict therapeutic success. "We call this approach 'explainable AI,'" says Müller.

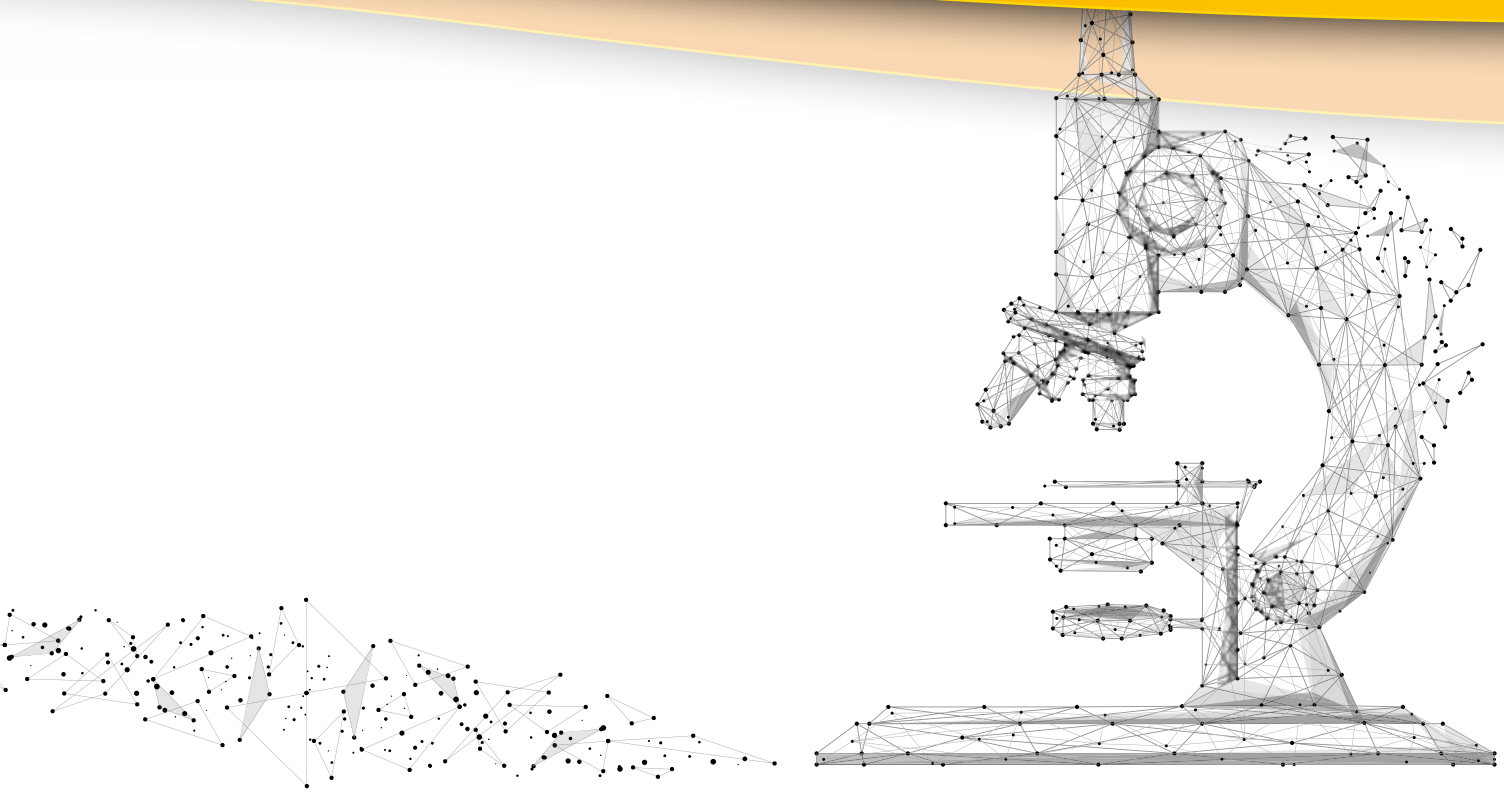
In the field of routine diagnostics, the AI system promises not only to save time and help avoid mistakes, but also to pave the way for personalized medicine. Therapeutic decisions increasingly require the precise detection and quantification of certain characteristics in tissue samples. The software developed by the scientists is already being used in diagnostics at Charité's Institute of Pathology. Other institutes and clinics plan to follow suit; the software is currently in the process of being certified for broad use.

Other applications include research and drug development

"Our software is also suitable for the expensive and lengthy approval process for new drugs," explains Klauschen. "If, for example, a pharmaceutical company develops a new active substance, the company must carry out extensive clinical trials – often in several countries – to show that the drug works in both animals and humans and has an acceptable side-effect profile. Pathological studies are an important component of this research. And it is precisely here that our software can help." The first collaborations on the development of active substances are already underway with various manufacturers. Klauschen adds: "Our aim is also to develop so-called companion diagnostics." This refers to the use of a targeted drug in combination with a suitable target molecule, for example, to treat a tumor: The physician can only prescribe the drug if a suitable target molecule is present. "If our software could help detect the target molecule faster and more accurately, this would be great for both patients and manufacturers," notes Klauschen.

BIH Digital Health Accelerator supports the development of digital solutions

It's a long way from the initial idea to a successful spin-off. That is why the BIH offers help at exactly this interface between research and application through its Digital Health Accelerator program. "The BIH is supporting innovators from Charité and the Max Delbrück Center to develop their concepts into digital products and to translate these into clinical practice. In addition to a licensing arrangement or an industrial cooperation, this can include a spin-off and thus create jobs," explains Tim Huse, head of the Digital Health Accelerator at BIH Innovations. "The focus in this project was mainly on rounding out the team, developing the business strategy and the product, and preparing for the spin-off." The support provided by the Digital Health Accelerator program and the BIH Innovations team ranges from financial aid to coaching and mentoring by experts and access to networks of talent,



Source: shutterstock / anttoniart

development partners, industry players and investors, to the provision of a co-working space for innovation and translation activities and in-depth guidance for spin-offs.

Research and routine diagnostics

The Aignostics team currently consists of 20 employees as well as a network of over 20 pathologists and research partners, who are helping with development and continuously testing the software. “But we expect to be able to significantly increase the size of our team in 2021,” says CEO Viktor Matyas. “We will then be in a position to expand the current applications and develop new ones,” adds Dr. Maximilian Alber, the company’s CTO. What is especially important to Klauschen is the close proximity to routine clinical practice: “The development of our solution is clearly geared to meeting the needs of routine diagnostics and clinical medicine. Since I use the software on a daily basis, we can immediately identify ways to improve it. We hope this is the best way to help patients.”

About the company:

Aignostics is a spin-off from Charité – Universitätsmedizin Berlin that develops AI-based pathology solutions. The groundwork for Aignostics was laid by a team of physicians led by Prof. Dr. Frederick Klauschen, together with researchers from the Fraunhofer-Gesellschaft and TU Berlin led by Prof. Dr. Klaus-Robert Müller, who filed their first patent for the pathology software in 2011. The interdisciplinary team enrolled in the Digital Health Accelerator (DHA) program of the Berlin Institute of Health (BIH) in early 2018 before spinning off in early 2020. A particular focus is on overcoming the much-discussed “black box” problem of AI in pathology, for which Aignostics has developed a proprietary “Explainable AI” platform.

Contact:

Prof. Dr. Frederick Klauschen
Lead Medical Adviser | Aignostics GmbH, Berlin
info@aignostics.com

www.aignostics.com

Photo: Frederick Klauschen

breathing new life into old data

AI makes it possible to extract new information from existing data

by Jakob Simeth, Marian Schön, Michael Huttner, Paul Heinrich, Michael Altenbuchinger, and Rainer Spang

Tissues are made up of many different cells with a wide variety of functions. Modern single-cell RNA expression data have shown us for the first time what the individual cells in a tissue do and how they change in a tumor. Yet these data are new and information on clinical courses is still needed – for these take time. At the same time, however, we have access to extensive collections of (bulk) tissue gene expression profiles that contain clinical course information, but these are not resolved at the cellular level. In our TissueResolver project, we are developing tools that make use of single-cell data to learn how to reconstruct the tissue cell composition and cell type-specific expression from various bulk profiles.

Single-cell RNA sequencing (scRNA-seq) is developing rapidly: When single-cell sequencing emerged a few years ago, only a few hundred cells could be isolated from a tissue and then sequenced. But new microdroplet technologies allow the simultaneous sequencing of millions of cells in a single run. The technology can provide information about the composition and function of a tissue. As with flow cytometry, it is possible to assign cells to specific types and subtypes and count their numbers in the tissue. This already reveals a good deal about the condition of the tissue; for example, one can find out whether the number of T cells infiltrating a tumor will influence mortality, or whether the number has changed following immunotherapy. In addition, tissue scRNA-seq analysis allows the gene expression of only a subset of all cells to be analyzed separately. This is quite relevant, for if an upregulation of cell cycle genes is observed in a bulk profile, this indicates proliferating cells. However, it makes a big difference whether the tumor cells are growing or whether our immune system is merely producing more T cells to fight the tumor.

Such questions can be answered by modern scRNA-seq methods. There is only one catch: This technology is still very new and very expensive. So far, little data have been available. On the other hand, there are thousands and thousands of bulk tissue data sets, which contain sequence data from many different tissues and tumors. These data have been made freely and publicly available over the last two decades. It will be about a decade before scRNA-seq data are available on a similar scale.

Yet the usefulness of bulk gene expression is limited. Cell type-specific signals cannot be easily filtered out from the overlapping expression of millions of cells. Much in the same way, it is difficult for us humans to filter out individual voices from the babble of voices in a large crowd. Thus, specialized software tools are needed to reliably separate cell type-specific signals, such as those from specific immune cells, from the tissue expression clutter. In our project, we aim to fill this gap and make such tools freely available to the scientific community.

Single-cell data are the key to understanding tissue data

To resolve existing tissue data *in silico* on a cellular basis, we use the few single-cell data sets that already exist. The basic idea is simple – we combine the gene expression of cells whose type we know so that they are as similar as possible to the simulated tissue bulk of a known composition. This is repeated

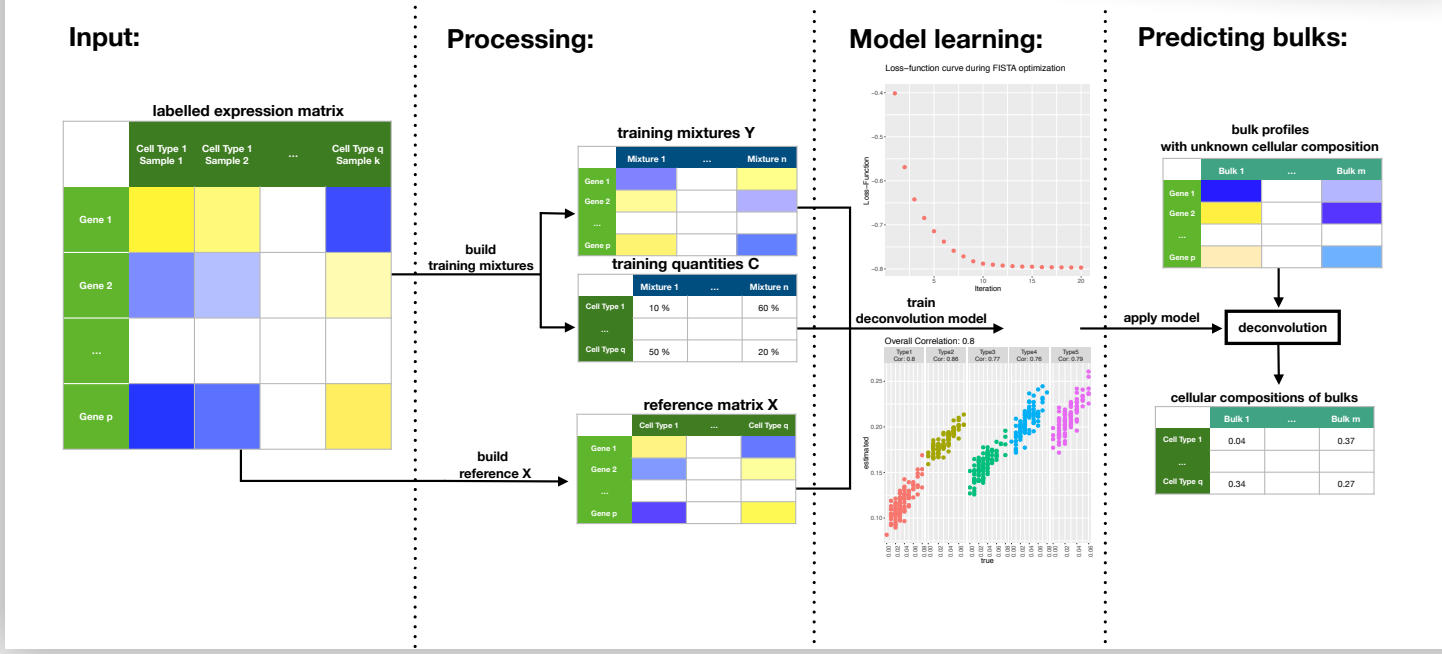


Figure 1: Illustration of our deconvolution process. We use cell-type labeled single-cell data both to compute a signature matrix and simulate mixtures (artificial bulks) of known cellular composition. Using machine learning, we learn weights of individual genes in such a way that the known composition of the simulated bulks can be predicted best. This model can then be used to analyze the cellular composition of unknown tissues (Source: Marian Schön / Schön *et al., 2020*).

for many cells and tissues until we know how single-cell data must be combined to be "similar" to a tissue. With this knowledge, we can then reverse the problem and look directly at the tissue expression of an unknown composition and estimate the proportions of particular cell types within it. Initially, we are interested in finding out what cells are present in a tissue and in what quantity. This question is answered in a process called digital tissue deconvolution (DTD), and for this to happen, it is important to capture the characteristic features in the gene expression of a cell type. We determine this cell type "signature" using machine learning methods. We have published a first algorithm for this (Görtler *et al.*, 2020) and made the associated software publicly available (Schön *et al.*, 2020).

Tissue deconvolution and the role of machine learning

DTD accuracy depends on the design of the signature matrix. In the same way that our brain can focus on certain voices that it knows, using their characteristics, and block out other voices, we need to select those genes that are particularly characteristic of a cell type and block out others that code for more general tasks in cells. No information about the cell type can be expected from these. For example, the genetic programs for cell division must be retrievable by all cells, so a high expression of these genes does not provide information about which cell type is involved. So some caution must be exercised in selecting such signature genes. On the one hand, too few genes should not be selected, because deconvolution is then prone to measurement inaccuracies and technical artifacts. On the other hand, the results will also be inaccurate if too many genes are included, because then the weak signals from small cell populations will not be visible. The right balance can be found through machine learning. We assign weights to genes, which we have learned from the data itself, i.e., we optimize these values so that the training and validation data of a known composition can best be deconvolved into their cell contents. This means that only those genes are selected that are particularly informative about the cell type. This leads to good deconvolution performance while significantly improving the detection of rare cells (see Figure 1, Goertler *et al.*, 2020).

Virtual tissue models

We go one step further in the design of the signature matrix. As we have data from many single cells, we never know *a priori* what we will find in a tissue. We therefore develop algorithms that select from suitable single-cell data those which, when

added, are as similar as possible to the tissue expression to be analyzed. In this way, we use the full spectrum of existing cells in humans and do not even have to assign them to a known cell type. We refer to the result as a "virtual tissue model": a limited number of representative cells that are assigned a weight factor that roughly corresponds to the frequency of similar cells in the analyzed tissue. This ensemble of cells is thus close to what would otherwise be obtained using state-of-the-art, high-throughput scRNA-seq technologies.

So it is possible to compare two tissues across their models by applying a dimensional reduction technique in order to visualize the individual cells and label the identified cells of both tissue models (see Figure 2). If one then further assigns a type to each cell – either from the original data set or through appropriate clustering – differences become apparent beyond what might be explained by cellular composition alone. For example, one can examine how B cells in tumor tissue differ from those in the peripheral bloodstream, even when single-cell measurements from the tissues are not available.

Cell type-specific expression from virtual tissues

For decades, we have been comparing gene expression in tissues. Yet we never know whether differences are due to gene regulation or merely reflect a different cellular composition in the tissues. Even if they are due to gene regulation, we do not know in which of the tissue cells this regulation occurred. Proliferation signals may originate from tumor cells or indicate an immune response in the tumor. We want to visualize this difference in retrospect. To do this, we need to disentangle the expression patterns of individual cells in the tissue – and here too, we use single-cell data and tissue models to learn how to do this. This trick allows us to answer entirely new questions with the benefit of hindsight. How does the gene expression of T cells from tumor patients with a good prognosis differ from those with a poor prognosis? How strongly do B cells express certain genes when they infiltrate the tumor? Do the tumor cells or the immune cells fighting that tumor generate the proliferation signals?

Tissue normalization enables a better view of the tumor

Often, of course, single-cell data are not available for the same cells that are found in the tissue. For example, if only immune cells are available in the single-cell reference and the tissue to be analyzed comes from melanoma, only the immune cell portion of the tissue can be explained in detail. This may seem like a disadvantage at first, but it actually allows us to better understand the tumor. By removing the expression of the cells present from the overall expression, a clearer picture emerges. For example, we can explain much of the data variation that is simply due to the random choice of where the tissue was taken from in the tumor. By removing this easily explained variation in expression, we create a level playing field and can better compare the data that remains.

Breathing new life into old data

Sequenced tissues are publicly and widely available and the more single-cell measurements are accessible, the more we can learn – with our new tools – from this old data. This is especially important because much of these data were collected years ago and thus clinical follow-up information is already available, meaning the data can be linked to a prognosis. We anticipate that it will be another decade before similar conclusions are available for single-cell measurements. Nor is a replacement for tissue measurements foreseeable in clinical practice and personalized medicine of the future. Single-cell diagnostics are simply too costly to be clinically routine.

It is therefore all the more important to exploit the existing treasure trove of data and to develop tools that are helpful in this respect. In our project, we are dedicated to the development of such software and publish it freely, making it available to everyone. A start has been made with our tool for tissue deconvolution (DTD), which is available as a package for the programming language "R" (Schön *et al.*, https://github.com/spang-lab/DTD). In parallel, readily trained models can also be used in a simple way with a web tool (https://dtd.spang-lab.de). We will continue to expand these tools over time, so that our new algorithms and developments can soon be deployed and valuable existing data sets can be analyzed further and in more depth. We hope this work will benefit a broad community of scientists – from tumor biology, immunology and many other areas – and help them gain new insights.

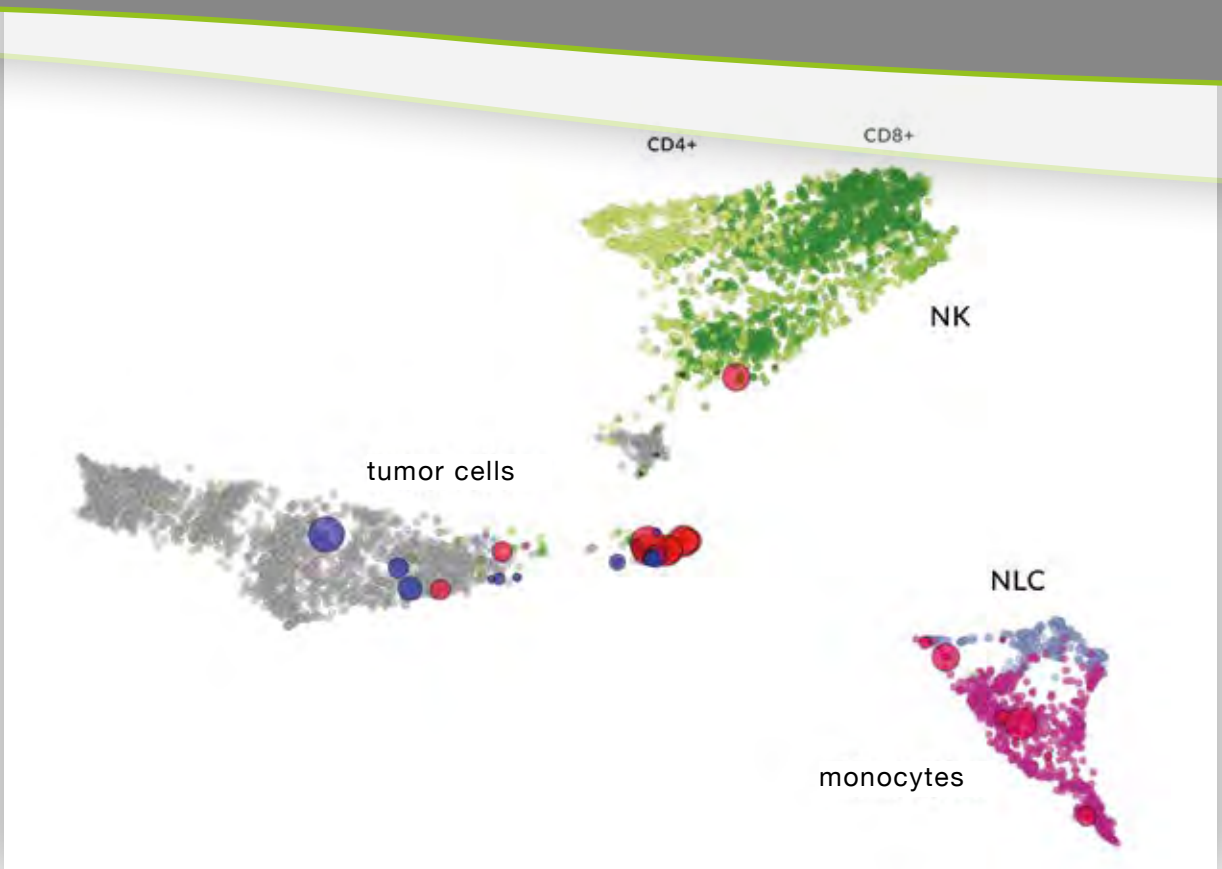


Figure 2: Illustration of single-cell profiles and bulk tissues from patients with chronic lymphocytic leukemia (CLL). Small dots represent single cells, colored by their cell type (green cluster: immune cells, purple cluster: monocytes, gray clusters: tumor cells). Using these single-cell profiles two bulk tissues were analysed and representative cells were chosen that resemble the two bulks best (blue and red circles, diameter corresponds to the frequencies of these cells). The two samples differ substantially: While the tumor of the first (blue) patient can be found predominantly in the left gray cluster of the tumor, the other patient's tumor (red) is more similar to the (also malignant) cells in the center. Differences can also be observed in the frequencies and distributions of immune cells and monocytes (Source: Jakob Simeth).

Project profile:

TissueResolver is a project funded by the German Federal Ministry of Education and Research (Grant No. 031L0173). It was launched in 2019 at the Institute of Statistical Bioinformatics of the University of Regensburg, with the goal of developing algorithms and providing interactive tools to better investigate bulk tissue data.

References:

Schön, M., Simeth, J., Heinrich, P., Görtler, F., Solbrig S., Wettig T., Oefner, P.J., Altenbuchinger, M., and Spang, R. (2020). DTD: An R Package for Digital Tissue Deconvolution. Journal of Computational Biology, 386-389.

Görtler, F., Schön, M., Simeth, J., Solbrig S., Wettig T., Oefner, P.J., Spang, R., and Altenbuchinger, M. (2020). Loss-Function Learning for Digital Tissue Deconvolution. Journal of Computational Biology, 342-355.

Contact:

Prof. Dr. Rainer Spang
Statistical Bioinformatics
Institute for Functional Genomics
University of Regensburg
rainer.spang@ur.de
www.spang-lab.de

Jakob Simeth
Statistical Bioinformatics
Institute for Functional Genomics
University of Regensburg
jakob.simeth@ur.de

Photos: Rainer Spang and Jakob Simeth

big data and smartphones in the intensive care unit

How smartphones and Artificial Intelligence help treating ventilated patients

by Gernot Marx, Sebastian Fritsch, Johannes Bickenbach, Julian Kunze, Oliver Maaßen, Saskia Deffge, Silke Haferkamp, and Andreas Schuppert

When the first cases of a new lung disease appeared in Wuhan, China in January 2020, hardly anyone foresaw what the world would soon face. It became clear early on that the majority of patients who became particularly ill or even died were suffering from acute respiratory distress syndrome (ARDS). The clinical picture is by no means new to intensive care physicians. Even apart from COVID-19, it kills about 40 % of all patients. But although there are actually clear criteria for diagnosis, it is still identified all too rarely. This challenge is being addressed by the clinical use case "Algorithmic Surveillance in Intensive Care" (ASIC) under the auspices of the SMITH consortium.

ARDS and ventilation

To understand ARDS, one has to imagine the lungs as a sponge, in which air-filled spaces alternate with fine tissue structures. If damage occurs to the lungs, e.g. due to pneumonia, fluid flows into the tissue. The sponge fills up, creating what is known as "pulmonary edema." As a result, air-filled spaces fill with fluid, tissue structures swell and gas exchange deteriorates. The lungs become "wet and heavy," significantly reducing the mobility of the tissue. Often, the lungs can then no longer meet the body's oxygen requirements. At the same time, the patient's muscular breathing effort increases to such an extent that old or weakened patients, in particular, are no longer able to muster the effort.

In the end, both of these factors often necessitate mechanical ventilation, in which oxygen-enriched air is introduced into the lungs via a breathing tube using positive pressure. After the end of "exhalation," positive pressure remains on the lung tissue – known as positive end-expiratory pressure (PEEP) – in order to "push" fluid out of the lung tissue and ventilate it more thoroughly. Mechanically ventilated patients with advanced ARDS are placed in prone position with the same goal. This enables areas of the lung that have closed up in the supine position due to the weight of the wet lungs to move upwards and open up again. Often, this can improve gas exchange in the lungs.

Mechanical ventilation is frequently life-saving. However, it also has downsides. The lungs can be damaged due to high pressure and/or excessively large volumes that are generated by the ventilator. The tissue of the inelastic lung can be overstretched and injured by the forces acting upon it. This, in turn, sets inflammatory processes in motion that further damage the lungs. A vicious cycle of ventilation, further pulmonary deterioration, and renewed ventilation intensification ensues. The core goal of therapy is therefore to ensure that ventilation

Figure 1: The ASIC App supports the treating intensive care physicians in the diagnosis and therapy of ARDS (Photo: University Hospital RWTH Aachen).

Figure 2: A significant proportion of ARDS disease is not detected or is detected too late
(Photo: University Hospital RWTH Aachen).

no further damages the lungs, i.e., with smaller breathing volumes and low peak pressures. Unfortunately, despite advances in therapy, 40 % of all ARDS patients still die from their illness (Bellani *et al.*, 2016).

The "Berlin definition"

ARDS was first described in 1967 and has been redefined several times since then. The current version from 2012 was named after the meeting venue of the expert group and is therefore known as the "Berlin definition" (ARDS Definition Task Force, 2012). An essential part of the definition is what is known as the Horowitz index. This expresses the ability of the lungs to deliver oxygen to the blood and is calculated from the oxygen partial pressure in the arterial blood and the oxygen percentage in the inhaled air. It also determines the severity of ARDS. Other criteria for an ARDS diagnosis include the course of the illness over time, X-ray changes and the exclusion of other conditions, such as hyperhydration or heart failure.

Unfortunately, a large number of ARDS cases are not recognized or are only recognized too late. In a large observational study involving 29,000 patients in 439 intensive care units (ICUs) in 50 countries over four weeks, barely half of all mild ARDS cases were recognized (Bellani *et al.*, 2016). Thus, it must be assumed that an undiagnosed illness is also not treated correctly.

The ASIC app

This is where the ASIC app comes in. It is used by intensive care physicians at the university hospitals of the SMITH consortium with on-duty smartphones. The app communicates with the patient data management system (PDMS) and regularly verifies whether the Horowitz index falls below the threshold for ARDS. If it does, the physician is notified via push message of the potential presence of ARDS in one of his or her patients and is prompted to evaluate the remaining diagnostic criteria. If the criteria are met, the app displays the most important recommendations for ARDS therapy from the corresponding guideline. The physician can thus make a diagnosis much earlier, review the previous therapeutic steps and, if necessary, make evidence-based adjustments. Before being used at the patient's bedside, the ASIC app underwent the CE conformity procedure and was certified as a Class I medical device.

COVID-19

The fact that ARDS is a major problem in COVID-19 therapy became clear early on. After 20 million patients, it already became obvious that ARDS could have a massive negative impact on the treatment outcome. How many SARS CoV-2-infected patients actually develop ARDS is not yet clear due to the dynamic situation and highly fluctuating numbers. At the same time, some authors are already suggesting that ARDS in COVID-19 is different from "conventional" ARDS (Marini and Gattinoni, 2020). According to the authors, some patients initially show marked gas exchange dysfunction but only minimal pulmonary edema. The lung tissue thus remains rather elastic. Since established treatment recommendations assume a "wet" lung with marked pulmonary edema, recommendations could not only be unbeneficial for patients with COVID-19 ARDS, but may even be harmful. However, to investigate this theory on the basis of clinical trials, data from a large number of patients would be needed. ICU patients are well suited for this purpose, as approximately 1,200 data points are saved per day per

Figure 3: The authors (Photo: University Hospital RWTH Aachen)

patient. Unfortunately, most of the systems used in the ICUs of German university hospitals are not compatible with each other. Likewise, the number of COVID-19 patients requiring intensive care at a German university hospital is likely to be rather too low for the above-mentioned statistical evaluations. So it is a major challenge to create a large pool of anonymized treatment data from ICU patients that is also suitable for the application of new data science methods. This is the second important field in which ASIC plans to bring about innovation, thereby improving patient care.

Data science and data integration centers

Artificial intelligence, big data analysis, machine learning and data science are all considered to be of key importance for the future of medicine, as they make it possible to extract good predictions from data, even in highly complex contexts that are not fully understood. For example, it is already possible to predict the occurrence of septic shock several hours in advance, based on small changes in vital signs that are imperceptible to humans (Ghalati *et al.*, 2019). In addition, systems medicine models, known as virtual patients, can describe the development of complex illnesses on the basis of medically established processes. Developing such models requires algorithms adapted to the specific situation in combination with large databases. However, what is currently the most widely-used database for ICU patients, the MIMIC-III database, contains only U.S. patients. Its data can only be transferred to German patients to a limited extent. Therefore, the goal of the ASIC use case is to combine the data of ICU patients from all participating university hospitals and make them usable for research. To this end, a number of challenges need to be solved, which are being addressed within the framework of ASIC. Data integration centers (DIC) have been created at each participating site. Such a DIC extracts data from primary care systems, depersonalizes them, and converts them to an interoperable format so that the data can be merged. Analysis of this data can then identify differences between specific patient groups, such as those with COVID-19 ARDS and conventional ARDS. The most successful therapeutic strategies can then be identified and future therapies adapted accordingly. Using this nascent database, ASIC will be able to develop an early warning system for ARDS, analogous to the existing system for septic shock, as well as a virtual patient model for ARDS. The latter can predict the effect of certain therapeutic measures under specific conditions without having to verify this with the effort of costly clinical studies and patient experiments.

Data protection

In order to implement these ambitious projects, a special emphasis has been placed on a data protection-compliant approach. For example, the use of medical data had to take into account not only the overarching requirements, such as the EU's General Data Protection Regulation (GDPR), but also differing legislation from six German states. To this end, several legal opinions were ob-

tained, a comprehensive data protection concept was drawn up by external experts, and a dialog was conducted with the data protection officers of all participating university hospitals. The ethics committees of participating hospitals were also consulted and ultimately gave their approval.

Outlook

ASIC is a use case of the SMITH consortium. It exemplifies the role played by the DIC of the Medical Informatics Initiative (MI-I) in research and the improvement of care (see also the article **The Data Integration Center – A Hub for Medical Research and Care Data**, page 84). The focus on the clinical picture of ARDS is primarily intended to demonstrate the functionality of the methods used. If proven successful in the project, these can be extended to other disease patterns, such as sepsis or acute kidney failure, with manageable effort. Until then, the app will still be warning physicians on ICUs about possible cases of ARDS.

About the SMITH consortium:

The SMITH consortium and the clinical use case ASIC are funded by the Federal Ministry of Education and Research (BMBF) as part of the MI-I. Ten German university hospitals and other partners from research and industry have joined forces in the SMITH consortium to realize the interoperable exchange and use of health-related data from research and patient care across sites (Winter et al., 2018). For this purpose, DIC are being established at the university hospitals and connected through a technical network. At the same time, innovative IT solutions will be developed in two clinical use cases and in one methodological use case to demonstrate the possibilities and potential of modern digital services as well as the functionality and benefits of the new infrastructure.

SMITH consortium partners
(www.smith.care/konsortium)
Leipzig University
University of Leipzig Hospitals and Clinics
Friedrich Schiller University Jena
Jena University Hosptal
University Hospital RWTH Aachen
RWTH Aachen University
Fraunhofer Institute for Software and Systems Engineering ISST
Bayer AG Leverkusen
Maerz Internetwork Services AG
Averbis GmbH
ID Information & Documentation in Health Care GmbH & Co. KGaA
Juelich Research Center
University Hospital Halle (Saale)
University Hospital Bonn
University Medical Center Hamburg-Eppendorf
Essen University Hospital
University Medical Center Rostock
Duesseldorf University Hospital
University Hospitals of the Ruhr University of Bochum

References:

Bellani, G., Laffey, J.G., Pham, T., Fan, E., Brochard, L., Esteban, A., Gattinoni, L., van Haren, F., Larsson, A., McAuley, D.F., *et al.* (2016). Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. Journal of the American Medical Association (JAMA) 315, 788-800.

Ghalati, P.F., Samal, S.S., Bhat, J.S., Deisz, R., Marx, G., and Schuppert, A. (2019). Critical Transitions in Intensive Care Units: A Sepsis Case Study. Scientific reports 9, 12888.

Marini, J.J., and Gattinoni, L. (2020). Management of COVID-19 Respiratory Distress. JAMA 323, 2329-2330.

The ARDS Definition Task Force (2012). Acute Respiratory Distress Syndrome: The Berlin Definition. JAMA 307, 2526-2533.

Winter, A., Staubert, S., Ammon, D., Aiche, S., Beyan, O., Bischoff, V., Daumke, P., Decker, S., Funkat, G., Gewehr, J.E., *et al.* (2018). Smart Medical Information Technology for Healthcare (SMITH). Methods of information in medicine 57, e92-e105.

Contact:

Prof. Dr. Gernot Marx, FRCA
Head of the Department for Intensive Care Medicine
University Hospital RWTH Aachen
gmarx@ukaachen.de

www.smith.care

#nCoVStats

How data science is helping to understand the coronavirus pandemic

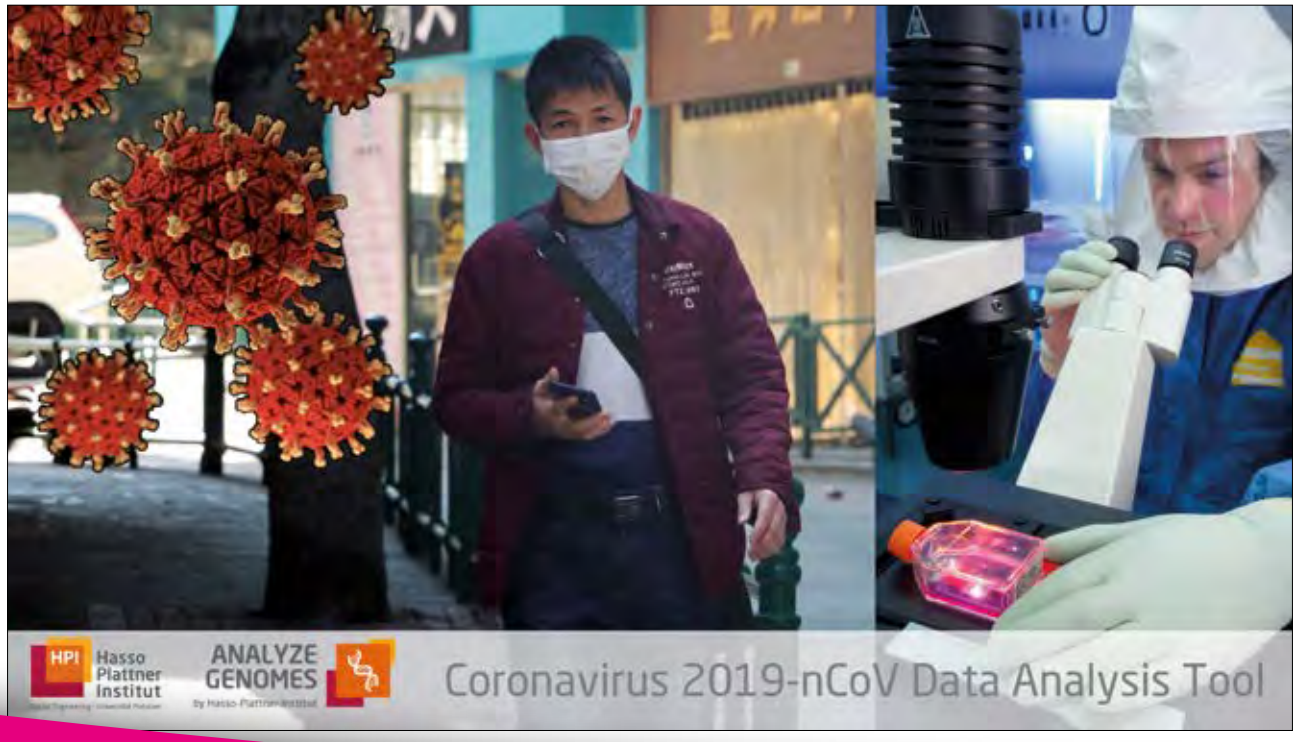
by Matthieu-P. Schapranow

The pandemic triggered by the coronavirus SARS-CoV-2 outbreak in late 2019 continues to define the lives of us all still today. In the fight against the virus, public data have become important for experts and the general public as a source of independent information. This is where a group of researchers at the Hasso Plattner Institute for Digital Engineering (HPI) at the University of Potsdam comes into play. They provide free graphical data exploration tools on the website https://we.analyzegenomes.com, including daily situational reports on the coronavirus pandemic using the Twitter hashtag #nCoVStats to all interested people worldwide.

Containment of the virus through systematic identification of suspected cases

The era of borderless globalization offers numerous advantages. We can travel wherever we want, and even the most exotic products can be ordered quickly via the internet and arrive just a few days later in our homes. But viruses also travel around the globe faster than ever before. This process is aided by countless flight connections that link even the most remote places on the planet with large cities in just a few hours.

At HPI, we build on our experience in researching other epidemics. For example, HPI researchers worked together with international scientists during the 2014 Ebola epidemic outbreak in West Africa. Contact tracing was and still is one of the most important containment strategies against the virus. It involves isolating all potential contacts of confirmed infected persons over the period of incubation and regularly check back with them with regards to disease-specific symptoms. The risk of infecting others be only be reduced by consistently identifying contact persons and isolating them as soon as possible.



Source: CDC, Macau Photo Agency, M. Schapranow

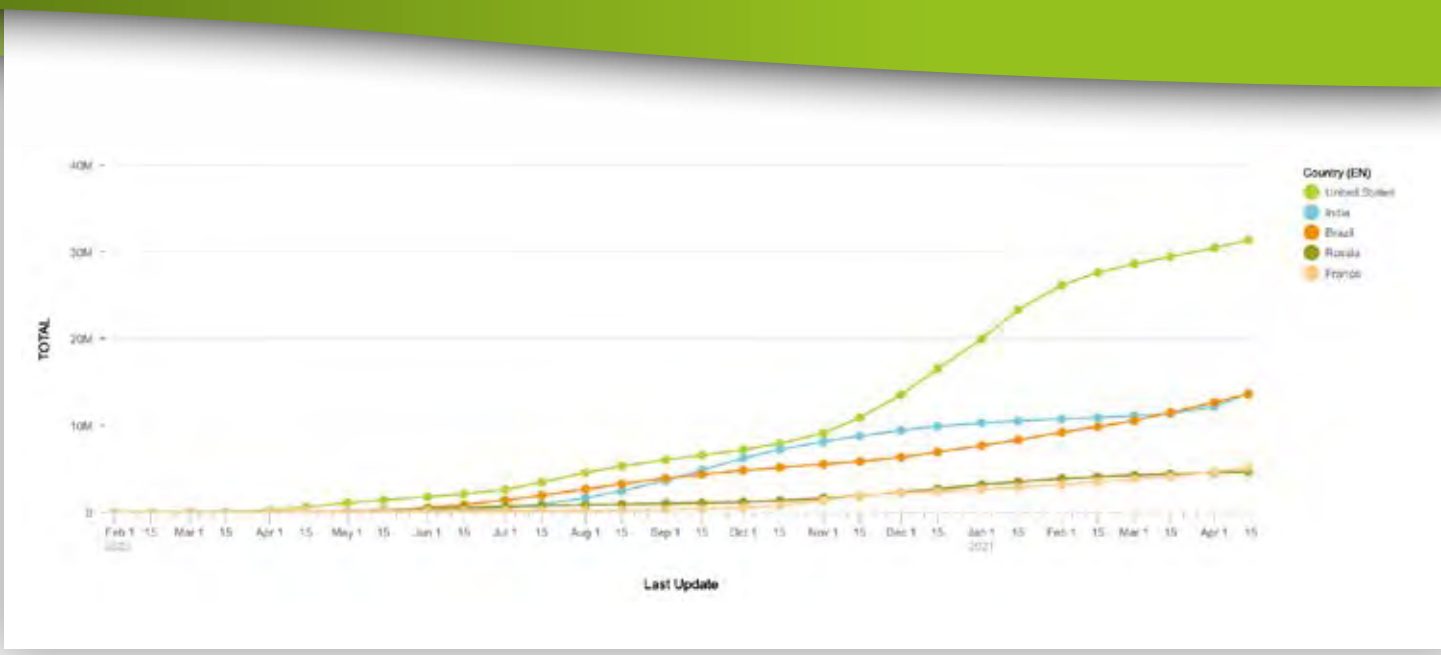


Figure 1: Trends in infection rates in the top 5 countries (as of April 14, 2021, source: https://we.analyzegenomes.com/apps/ncovstats/).

Contact tracing is also a key to success in the current coronavirus pandemic. Particularly during the first cases in Germany, personal interviews were used very successfully to reconstruct a list of people with whom confirmed infected persons had been in contact over the past days. With more and more contact tracing carried out, it became clear that a more and more qualified resources were bound by it. It had already become apparent during the Ebola epidemic in 2014 that qualified contact tracing personnel would quickly become scarce. Therefore, a contact tracing app had already been developed at HPI together with an international team of scientists and evaluated with experts in Nigeria. With the help of the app, non-medical personnel were also able to support the contact tracing process after participating in a short training session. In 2020, HPI again supported the development of what has become known as the CovApp, which helps to record relevant symptoms in suspected cases.

Once again, the use of digital applications demonstrates how, in times of scarce resources, apps can help to use limited resources more effectively, thus enabling medical professionals to devote their attention to emergencies.

The basis for sound decision-making: Access to always up-to-date data

In addition to contact tracing data, treatment data from hospitals are an important source for decision-making. Hospitals provide high-quality data, as it comes from validated laboratory tests. For example, it can provide precise information on the number of new patients, details of the virus subtypes as well as the number of recovered patients and the number of deaths. However, such key figures are collected in a decentralized manner using individual IT systems. There is no central register which combines all the decentral data without delay, so far. However, many important decisions are based on up-to-date nationwide data. For example, epidemiologists use current data on infected persons per region to make estimates on the spread of the disease, as well as appropriate recommendations, e.g. for business closure or contact restrictions. Figure 1 shows the development of infection for the top five countries worldwide. One can easily see that the United States lead the top 5 board with 30M+ reported cases, but returned to a linear growth since January 2021 after the start of the national vaccination program. Brazil and India follow with almost 14M each, but both countries show recently an upward trend in reported cases. France and Russia report both about 5M cases and a linear growth over the past weeks. Due to the different reporting structures in the healthcare and political system, we expect a higher number of unreported cases for some of the countries.

At HPI, researchers recognized the gravity of the situation early on. As early as January 2020, just weeks after the first confirmed coronavirus case, they began identifying available international data sources with case numbers on SARS-CoV-2. Since at that time the center of the epidemic was in China, we focused on Chinese data sources first. The global case counts were integrated into an in-memory database at HPI. This database technology, which is being researched at HPI, makes it possible to analyze even huge amounts of data at a blink of an eye using any cri-

teria. Among other things, the database stores the currently reported case numbers of healthy, ill and deceased persons per country or region together with time stamps. So-called crawler programs are used to automatically identify updated data and to integrate them into the HPI database. In this way, a longitudinal database of worldwide case numbers was created, which now comprises around 90,000 entries for almost 600 countries and regions worldwide.

Visualization helps to identify correlations

After compiling updated data, the next step is to analyze them. The data can be used to make statements about the current global situation, but they can also be analyzed retrospectively, e.g., to identify trends in individual countries or regions. Here, software systems are used that support the exploration of large amounts of data through interactive visualizations. Figure 2 shows an example that compares the reported case numbers from April 20 and July 20, 2020 for each country using circular diagrams. One can quickly see how much the case numbers have increased already within the three-month period, especially in North and South America, but also in parts of Europe and Russia. They far exceed the number of cases in China, the country of origin, which is visualized as a relatively tiny spot in the diagram.

Analysis challenges: Varying quality of data

Simultaneously, one can also see that the African continent, for example, has reported low case numbers. But is this really the case? Here we encounter another challenge: quality of reported data. Even if we have access to reported data from nearly every country, we have no influence on the quality and the way data was collected. This is not just a question of the accuracy of the reported numbers, but also of the definitions and assumptions used per country. For example, what criteria are used to decide whether a suspected case is reported as confirmed or not? At the beginning of 2020, medical centers lacked capacity for comprehensive testing of suspected cases. Instead of a PCR test for viral RNA, other indicators, such as CT images of the lungs, were used to determine cases. However, this different approach means that the reported numbers are subject to different measurement errors per country.

In African countries with less well-established healthcare systems, systematic testing of suspected COVID-19 cases is extremely difficult. Documenting suspected cases and acquiring data from regional medical centers also present logistical hurdles for governments. Based on experience from previous epidemics, it can therefore be assumed that the publicly reported figures represent only a fraction of the expected cases. Compounding this situation is the fact that fortunately only a relatively small proportion of infected persons develop severe symptoms requiring hospitalization.

Accurate forecasts through artificial intelligence

In Germany, we also know that many infected people sometimes show only mild or no symptoms at all, which means they are not registered by a visit to the doctor. To account for this error in national figures, residents were surveyed and tested across Germany's regions that are considered coronavirus hotspots. It is hoped that these regional studies will provide a more precise forecast of the real number of cases in Germany and a better understanding of how the virus is transmitted. HPI researchers also use the database of collected COVID-19 data as a basis for forecasts. Machine learning and artificial intelligence methods are used, for example, to forecast case numbers for other countries based on developments in China or to evaluate the effectiveness of measures that have been taken in other countries.

We have already been able to learn important things from the pandemic: The sooner we can access up-to-date nationwide data, the sooner appropriate measures can be taken to contain the pandemic. Established clinical processes, systematic testing of suspected cases, a central registry for recording suspected cases, and appropriate software tools for interactive and flexible analysis of data are just a few of the components that will prepare our healthcare system for future pandemics. This is also why we should build up these capabilities systematically now, and not simply shelve the findings from the current coronavirus pandemic.

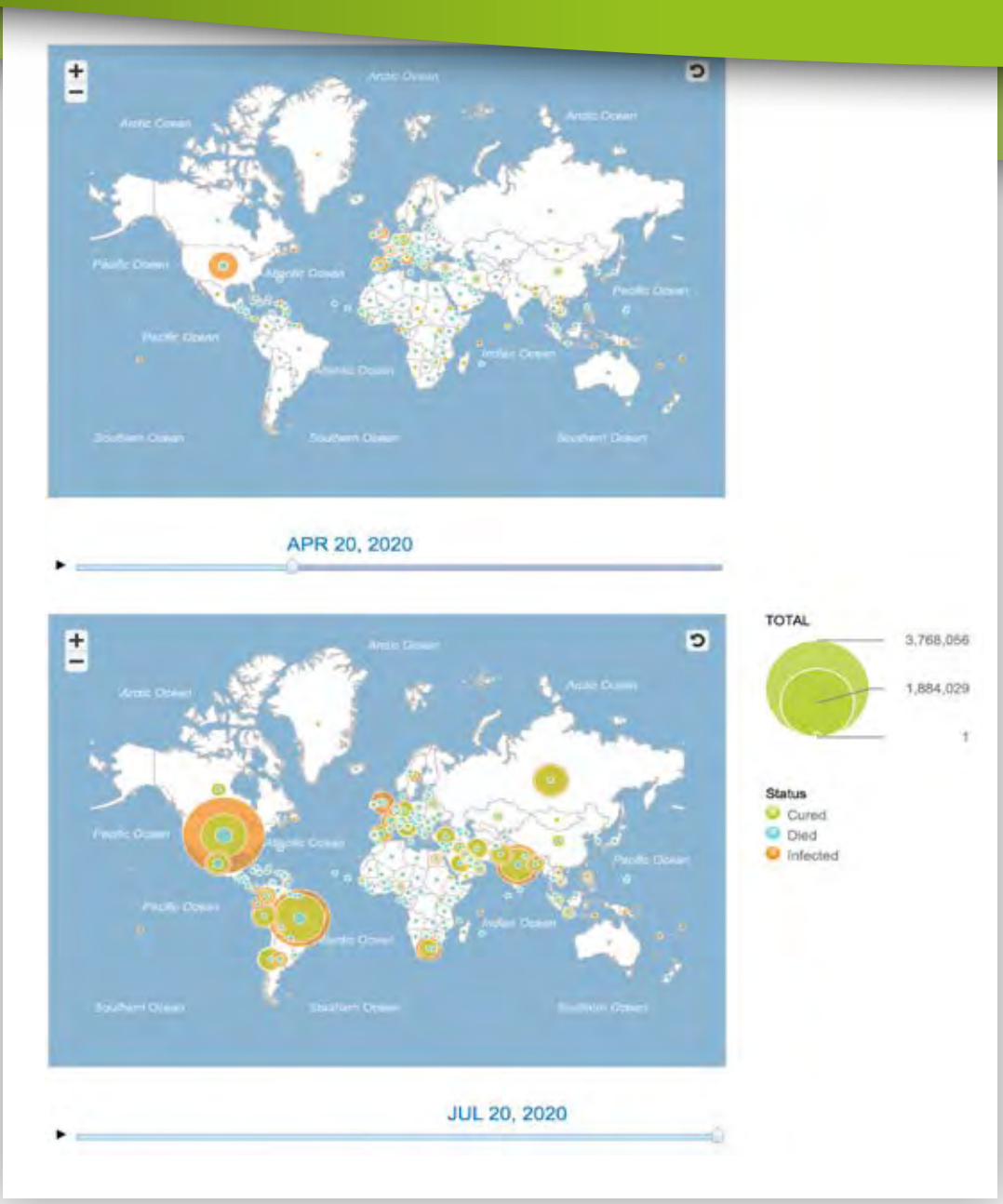


Figure 2: Visualization of global COVID-19 case numberes on a world map, top April 20, bottom July 20, 2020. Circular area reflects absolute numbers (source: https://we.analyzegenomes.com/apps/ncovstats/).

Project profile:

The **#nCoVStats** project was launched in early 2020 at the Hasso Plattner Institute's Digital Health Center in Potsdam. Its aim is to make up-to-date figures on the spread of coronavirus and graphical data exploration tools available free of charge via the internet to citizens, policymakers, journalists and scientists.

Contact:

Dr.-Ing. Matthieu-P. Schapranow
Working Group Leader
"In-Memory Computing for Digital Health"
Scientific Manager Digital Health Innovations
HPI Digital Health Center,
Hasso Plattner Institute,
University of Potsdam
schapranow@hpi.de

https://hpi.de/digital-health-center/members/working-group-in-memory-computing-for-digital-health/dr-ing-matthieu-p-schapranow.html

Photo: Schapranow

the chronic kidney disease nephrologist's app

Personalized systems medicine for chronic kidney disease patients

by Ulla T. Schultheiß, Johannes Raffler, Sahar Ghasemi, Robin Kosch, Michael Altenbuchinger, and Helena U. Zacharias

The kidney plays a key role in regulating numerous systemic processes in the human body. Chronic kidney disease, one of the leading causes of death worldwide, is a complex disease characterized by a highly variable disease progression and multiple comorbidities significantly differing between affected individuals. These facts complicate not only the prediction of adverse events and individual disease progression, but also treatment planning and medication management. The e:Med junior research alliance CKDNapp has set itself the goal of developing an app for nephrologists to support them in providing personalized treatment for chronic kidney disease (CKD) patients.

Transparent decision support for practicing nephrologists

The current condition and the expected disease progression of a CKD patient is dependent on numerous factors, including demographics, disease history, lifestyle, and medication. Traditional and novel biomarkers can also be used to provide further insights into disease progression. In order to provide optimal care for affected patients, the treating physician has to jointly evaluate and integrate all of these distinctive and complex data based on medical knowledge. This is the only way to tailor a disease treatment to the individual patient – i.e., to personalize patient treatment.

Figure1: (A) Schematic workflow of the development and application, and (B) detailed features of CKDNapp

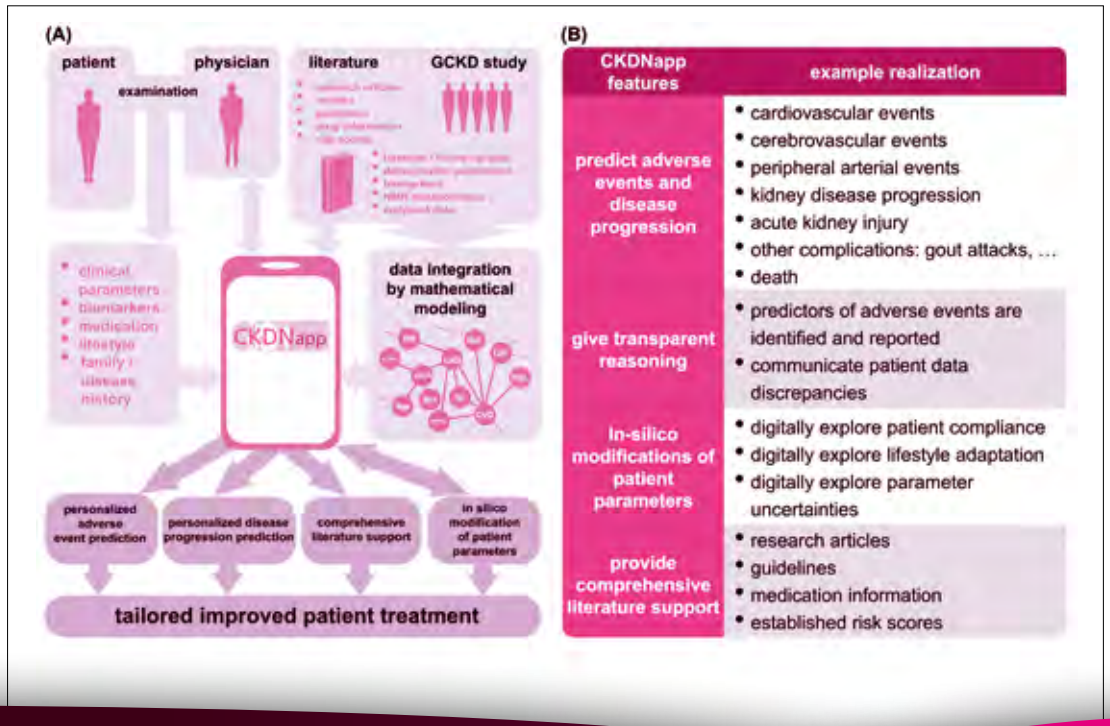


Image: Michael Altenbuchinger and Helena U. Zacharias.

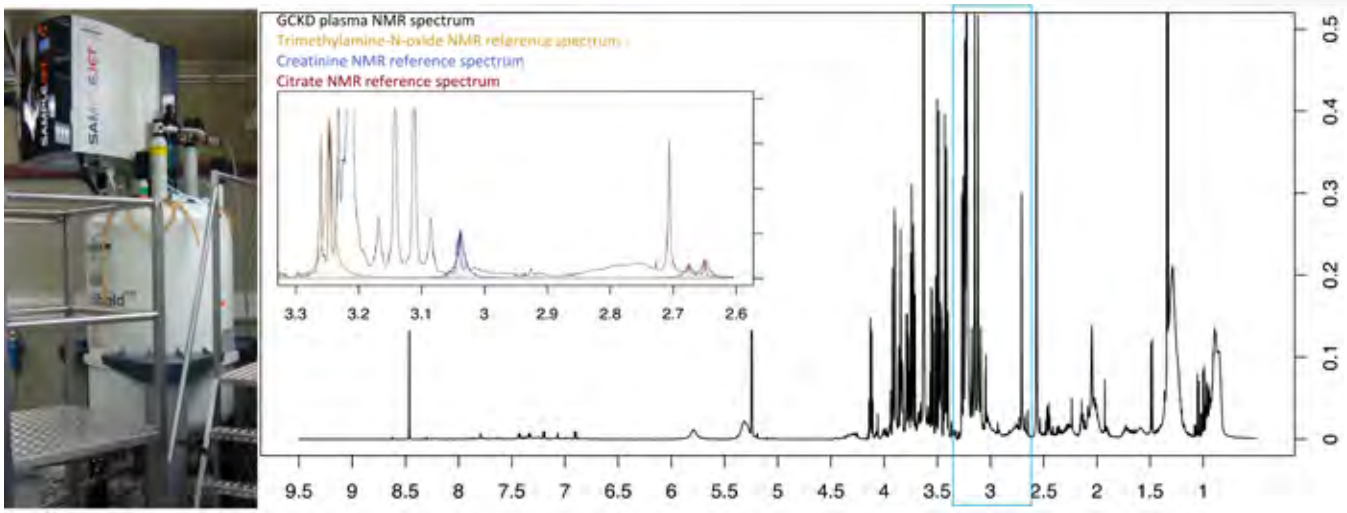


Figure 2: The NMR spectrometer used in this project (left) provides high-resolution NMR spectra of the investigated GCKD plasma samples (right). Signals in the spectrum reflect specific metabolites that can be identified based on their signal position on the x-axis. The signal height on the y-axis reflects the concentration of a metabolite in the investigated plasma sample. The metabolites, e.g. creatinine, can be identified by comparison with reference spectra (see magnification of the spectrum section marked in light blue on the left).
(Source: Helena U. Zacharias and Wolfram Gronwald).

The e:Med junior research alliance CKDNapp, coordinated by Helena Zacharias, is developing the CKD Nephrologist's App (CKDNapp for short) to support practicing nephrologists in the complex process of data integration – and thus in personalized treatment of CKD patients. When examining a patient, the extensive patient data collected by the attending physician can be fed into CKDNapp (Figure 1). The app itself is built on two complementary pillars: (1) comprehensive mathematical models to aid with diagnosis and predictions – for example, for the personalized prediction of cardiovascular events, end-stage kidney failure, or patient death – and (2) an extensive collection of established risk predictors taken from the literature.

Most importantly, CKDNapp will identify and report all relevant predictors of these patient events to the nephrologist, who will always be fully informed and in control when making decisions. In addition, the attending physician will be able to adjust patient parameters, change diagnoses or check on disease progression virtually in the app. Making virtual changes to a patient's lifestyle and letting CKDNapp predict a potential reduction in risk for that patient due to these changes, could, for example, increase the patient's motivation to actually implement such changes.

Small molecules become large biomarkers

The core of CKDNapp consists of different mathematical models for refining personalized diagnoses and predicting adverse health events and disease progression. The e:Med junior research alliance CKDNapp will calculate models on the basis of comprehensive demographic, phenotypic, clinical endpoint, and metabolomics data obtained from the German Chronic Kidney Disease (GCKD) study (https://www.gckd.org), one of the world's largest prospective CKD observational studies, initiated in 2010 and including > 5,000 patients and a follow-up period of ten years and more. The scientists are particularly excited about the metabolomics data obtained from plasma samples from the GCKD study.

But why is the metabolism of chronic kidney disease patients even of interest? Metabolites – small organic molecules such as amino acids or sugars – are metabolic intermediates and/or end products and can be found in all body fluids and tissues. The kidney is one of the main regulators of the body's metabolism, as it filters metabolic end products out of the blood and/or secretes them into the urine. If this important mechanism is disrupted by kidney impairment, metabolic dysregulations may follow. This, in turn, is reflected in an altered metabolite composition in the blood and/or urine. Metabolites can therefore serve as important biomarkers for a patient's current physiological state, but also for predicting future health events.

Obtaining a meaningful picture of human metabolism requires the analysis of as many metabolites as possible. The junior research alliance **CKDNapp** is therefore pursuing a hypoth-

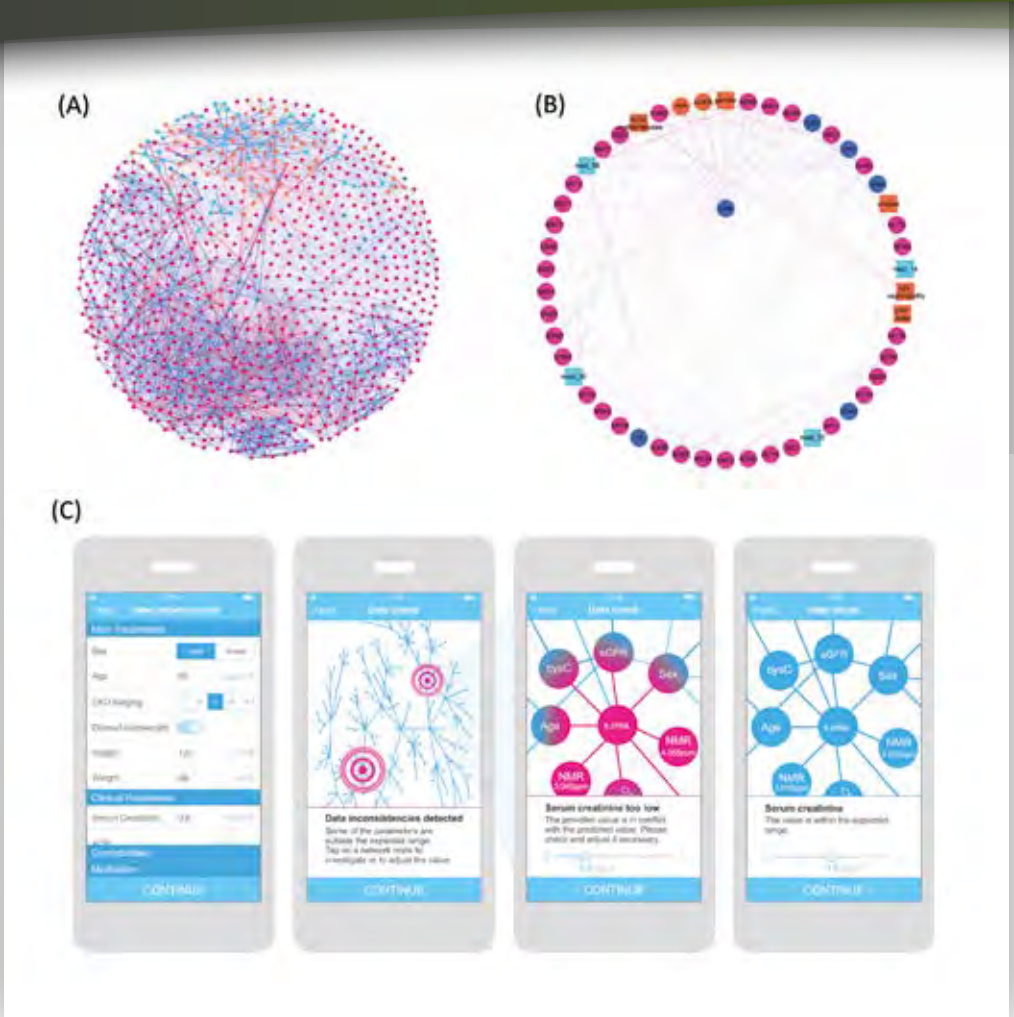


Figure 3: The junior research alliance "CKDNapp" uses newly developed algorithms to calculate a mathematical CKD model from the GCKD data that reveals all direct dependencies between individual patient parameters. (B) This model allows, for example, to reveal all patient parameters directly related to a patient's serum creatinine level. (C) Patient data uploaded to CKDNapp by the nephrologist is checked for data inconsistencies. Here, the app compares the read-in data with the mathematical model and alerts the user that the provided serum creatinine value may be too low. The physician can now "virtually" change this parameter to resolve the data inconsistency. (Image: (A) taken from Altenbuchinger *et al.*, 2019. (B) Helena U. Zacharias. (C) Johannes Raffler, Michael Altenbuchinger, and Helena U. Zacharias.)

esis-free approach to measure metabolites, which allows the determination of all detectable metabolites in a sample without prior selection (non-targeted metabolomics). The nuclear magnetic resonance (NMR) spectroscopy technique used in this project enables such a non-targeted approach to metabolomics measurements (Figure 2). During the course of the project, these metabolomics measurements will be repeated on all plasma samples from a follow-up visit, approximately two years after the baseline visit to enrich the CKDNapp models with time-course data. A wide variety of data on health events such as myocardial infarctions, kidney failure, and strokes are being continuously collected and processed for further data analysis.

The extensive patient information made available by the GCKD study, which currently contains more than 15,100 different data points, is thus a valuable treasure just waiting to be mined.

Computer-aided data integration unravels highly complex systems biology relationships in the human body

The integration of all these patient parameters is performed using state-of-the-art machine learning methods. The junior research alliance CKDNapp is, for instance, developing new mathematical models that enable the precise prediction of kidney failure with replacement therapy (Zacharias *et al.*, 2019).

Important patient information such as medical complications, comorbidities, demographic factors, the administration of medication, laboratory parameters, and metabolomics data are all deeply interwoven. This leads to complex dependencies between variables that can complicate the interpretation of these mathematical models. However, CKDNapp will be able to utilize these complex association patterns and map them into a consistent mathematical model to provide a detailed picture

Figure 4: Organization of the junior research alliance "CKDNapp" (Photos: Ulla T. Schultheiß, Helena U. Zacharias, Sahar Ghasemi, Michael Altenbuchinger, Robin Kosch, Johannes Raffler).

of chronic kidney disease (Figure 3A). Such mathematical models, "trained" on GCKD study data using newly developed algorithms, allow scientists to find out which patient parameters are directly interdependent (Altenbuchinger *et al.*, 2019, 2020). For example, serum creatinine levels change with patient age and are also dependent on different metabolites, drugs, and kidney disorders (Figure 3B). Comprehensive modeling of these complex parameter relationships can provide new insight into important pathomechanisms underlying chronic kidney disease.

In particular, CKDNapp will be able to use this model to estimate the correlation between new patient data and the model, thereby detecting possible discrepancies. Imagine, for example, that a patient's serum creatinine level was measured incorrectly. The nephrologist first loads all patient data onto CKDNapp and lets the app perform an automatic data check (Figure 3C). CKDNapp detects an inconsistency between the uploaded patient data and the mathematical CKD model, and tells the nephrologist that the entered serum creatinine value seems too low. The physician now has the option to change this value in the app until the data inconsistency is resolved. Thus, CKDNapp highlights potential data inconsistencies, reduces potential prediction errors and improves treatment recommendations. In addition, this app feature can be used to show patients how lifestyle changes could have a positive impact on their prognosis. Since the mathematical model trained on the GCKD study data has "learned" all dependencies between individual patient parameters simultaneously, CKDNapp will be able to highlight such data inconsistencies without being guided by a specific hypothesis.

CKDNapp transfers metabolic biomarker signatures into clinical practice

CKDNapp will be made available free of charge as a user-friendly software with intuitive interfaces. Even during the design process, the junior research alliance CKDNapp is testing both the predictive power of the newly developed mathematical models and the applicability of the app itself in a controlled clinical setting.

But can the metabolic CKDNapp models developed on the basis of highly specialized NMR spectroscopy measurements even be applied in clinical practice? This is a question that the CKDNapp junior research alliance intends to answer before the app is launched. Currently, NMR spectroscopy is barely ever used in routine laboratory diagnostics, whereas the use of mass spectrometry is quite common. The junior research alliance will first test whether the NMR-based metabolic CKDNapp models can be applied to mass spectrometry data without losing significant predictive power. This process relies on a newly developed machine learning algorithm called

zero-sum regression (Altenbuchinger *et al*., 2017a), which has already been used to successfully transfer proteomic biomarker signatures from one analytical measurement platform to another (Altenbuchinger *et al*., 2017b). The CKDNapp project will conduct detailed investigations to test whether such a platform transfer is also possible for metabolic biomarker signatures. Ultimately, it should also be possible to measure the new metabolic CKDNapp biomarkers using standardized analytical methods, so as to enable broad and cost-optimized use of the app in routine clinical diagnostics. This, too, will be assessed by the junior research alliance CKDNapp, thus bringing systems medicine into the clinic.

Project profile:

The e:Med junior research alliance "CKDNapp: A toolbox for monitoring and tailoring treatment of chronic kidney disease patients – a personalized systems medicine approach" has been supported by the German Federal Ministry of Education and Research (BMBF) since 2019. The funding will run for five years. The interdisciplinary team comprises the research groups of four junior group leaders who devote themselves to the following subprojects (Figure 4):

- **SP 1:** Dataset consolidation, generation of clinical input variables, and medical interpretation and validation (Ulla T. Schultheiß, University Medical Center Freiburg)
- **SP 2:** Mathematical modeling for CKDNapp (Helena U. Zacharias, University Medical Center Schleswig-Holstein and Kiel University)
- **SP 3:** Algorithmic foundation of CKDNapp models (Michael Altenbuchinger, University of Hohenheim)
- **SP4:** CKDNapp application and web service development (Johannes Raffler, Helmholtz Center Munich)

Further information: https://www.ckdn.app

References:

Altenbuchinger, M., Rehberg, T., Zacharias, H.U., Stämmler, F., Dettmer, K., Weber, D., Hiergeist, A., Gessner, A., Holler, E., Oefner, P.J., *et al.* (2017a). Reference point insensitive molecular data analysis. Bioinformatics.

Altenbuchinger, M., Schwarzfischer, P., Rehberg, T., Reinders, J., Kohler, C.W., Gronwald, W., Richter, J., Szczepanowski, M., Masqué-Soler, N., Klapper, W., Oefner, P.J., Spang, R. (2017b). Molecular signatures that can be transferred across different omics platforms. Bioinformatics.

Altenbuchinger, M., Zacharias, H.U., Solbrig, S., Schäfer, A., Büyüközkan, M., Schultheiß, U.T., Kotsis, F., Köttgen, A., Spang, R., Oefner, P.J., *et al.* (2019). A multi-source data integration approach reveals novel associations between metabolites and renal outcomes in the German Chronic Kidney Disease study. Sci. Rep.

Altenbuchinger, M., Weihs, A., Quackenbush, J., Grabe, H.J., and Zacharias, H.U. (2020). Gaussian and Mixed Graphical Models as (multi-)omics data analysis tools. Biochim. Biophys. Acta - Gene Regul. Mech.

Zacharias, H.U., Altenbuchinger, M., Schultheiss, U.T., Samol, C., Kotsis, F., Poguntke, I., Sekula, P., Krumsiek, J., Köttgen, A., Spang, R., *et al.* (2019). A Novel Metabolic Signature to Predict the Requirement of Dialysis or Renal Transplantation in Patients with Chronic Kidney Disease. J. Proteome Res.

Contact:

Prof. Dr. Helena U. Zacharias
Scientific coordinator of the e:Med junior research alliance "CKDNapp"
Associate Professor "Clinical Metabolomics"
University Medical Center Schleswig-Holstein and Kiel University
Internal Medicine I/Institute of Clinical Molecular Biology
Kiel
h.zacharias@ikmb.uni-kiel.de

Photo: Zacharias

alternative splicing – a systems medicine perspective

How does alternative splicing impact disease development?

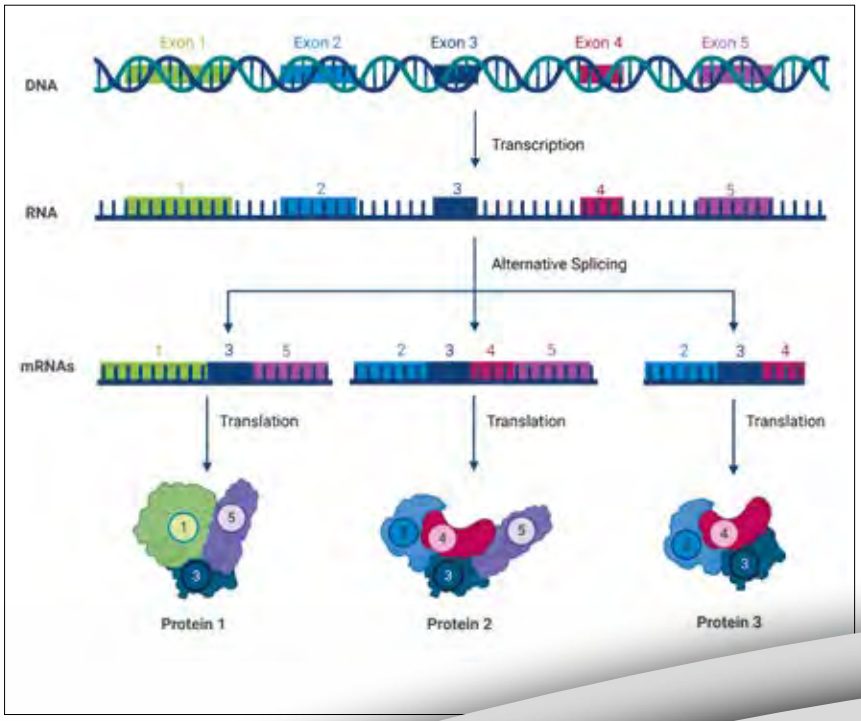
by Tim Kacprowski, Nina Kerstin Wenke, Sabine Ameling, Kristin Wenzel, Olga V. Kalinina, and Markus List on behalf of the entire Sys_CARE consortium

Alternative splicing (AS) is a biological mechanism by which different transcripts and ultimately different proteins (isoforms) can be formed from the same gene. Current studies based on molecular data (so-called omics data) have made little distinction between various isoforms. A systematic analysis of the association between alterations in AS and particular diseases is therefore still lacking. Understanding the far-reaching effects of AS requires an integrative systems medicine approach. In following this approach, **Sys_CARE** strives to investigate the significance of AS for the onset and progression of diseases and develop the bioinformatics tools required for such an investigation.

Alternative splicing in dilated cardiomyopathy (DCM) and hypertensive nephrosclerosis (HN)

AS leads to different protein isoforms which differ in structure and function (Figure 1) and therefore act differently in protein-protein interactions and complex signaling pathways. So far, even molecular-based (omics) datasets rarely distinguish between individual isoforms. A detailed exploration of the various effects AS has on disease development and progression is therefore still missing. Olga V. Kalinina, Professor at the Helmholtz Institute for Pharmaceutical Research Saarland (HIPS) and the Helmholtz Centre for Infection Research (HZI), emphasizes: *"If we want to understand the molecular mechanisms of disease development, we need to consider not only gene expression but also protein structures and their potential impact on function and protein-protein interactions when we integrate multi-omics data."*

Figure 1: Alternative splicing



Due to alternative splicing, a single gene can give rise to a diverse set of transcripts. Differences in exon usage lead to changes in the composition and sequence of the mRNA resulting in structural and functional differences on the protein level (isoforms) after translation (created with BioRender.com).

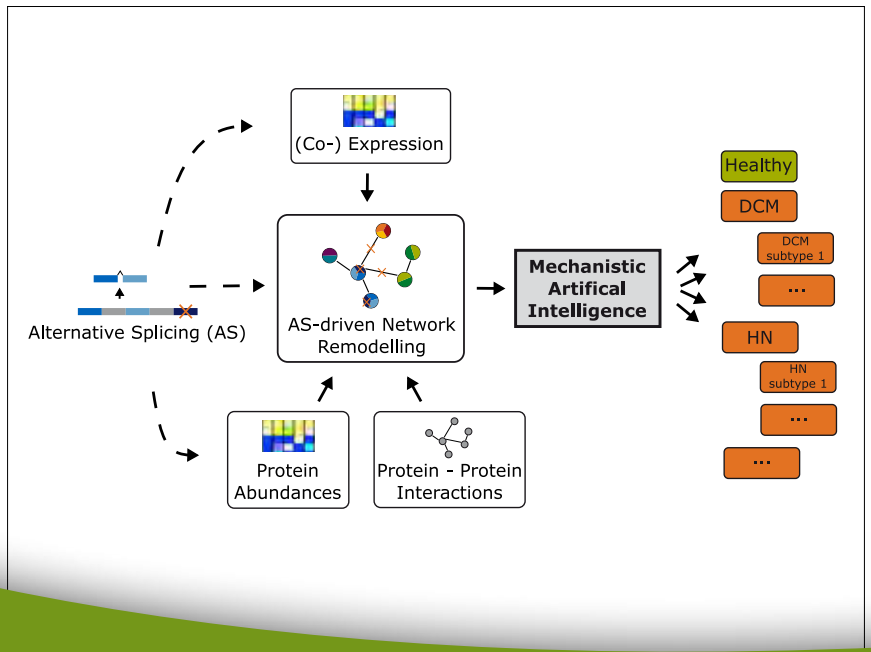
Source: Sys_CARE consortium

The Sys_CARE project focuses on two disease patterns in which AS is thought to play a role. Dilated cardiomyopathy (DCM) is a serious cardiac disease characterized by the presence of left ventricular (LV) or biventricular dilatation and systolic dysfunction. DCM is the most common form of cardiomyopathy as well as the most common cause of non-ischemic heart failure and end-stage heart transplantation. Both genetic and non-genetic causes contribute to DCM. An increase in the number of mutations in specific genes, such as titin, laminin or myosin, is known to be associated with the onset of DCM. *"In addition to genetic causes, inflammatory processes play a role in DCM development. The molecular mechanisms of the pathogenesis of this common myocardial disease are not fully understood and need to be studied in more detail,"* says Stephan B. Felix, Professor at the University Medicine Greifswald. While AS has been shown to play a role in the development of heart diseases (Beqqali 2018), no causal relationship between AS and DCM has been specifically demonstrated to date. Likewise, the disease mechanisms of hypertensive nephrosclerosis (HN), the second condition investigated in our study, are not fully understood. The role of AS in HN has not been investigated yet. However, AS for cell-typical genes such as Wilm's tumor protein 1 (WT-1) have been observed in kidney-specific cells called podocytes, resulting in different isoforms (Barbaux *et al.*, 1997).

"These examples show the need for a systematic analysis of the involvement of AS in cardiac and renal diseases," affirms Uwe Völker, Professor at the University Medicine Greifswald. So far, data are based predominantly on gene expression array technologies, which do not allow for an identification of novel gene products. Moreover, most transcriptome-wide AS studies investigating cardiomyopathies use mouse models. Although AS has been detected in animal models of kidney disease, transcriptome-wide AS analyses have been limited to patients with genetic kidney disease and have not been applied to the more common, non-genetic kidney diseases such as HN. Consequently, there is a pressing need for systems medicine studies using larger cohorts and highly sensitive sequencing methods.

"Dilated cardiomyopathy and hypertensive nephrosclerosis are two heterogeneous diseases currently defined phenotypically, i.e., by their symptoms. The systems medicine paradigm, however, calls for a mechanistic characterization of complex diseases that takes into account the molecular processes and metabolic pathways involved," says Jan Baumbach, Professor at the Technical University of Munich (TUM). A mechanistic understanding of a disease and its development allows for targeted therapies that address the cause of the disease instead of just fighting the symptoms. To this end, Sys_CARE applies innovative network analysis methods

Figure 2: Systems medicine interpretation of alternative splicing



Source: Sys_CARE consortium

Sys_CARE integrates data of multiple omics levels to provide a complementary view on alternative splicing. Gene expression data, protein abundances and protein-protein interaction data are integrated in a heterogeneous network and allows for studying the effect of AS on a network-level. These networks are then used as features in an artificial intelligence (AI) to detect mechanistic patterns. This means that our approach is not limited to classification but it also extracts molecular mechanisms which can explain diseases, supports diagnosis and disease subtyping.

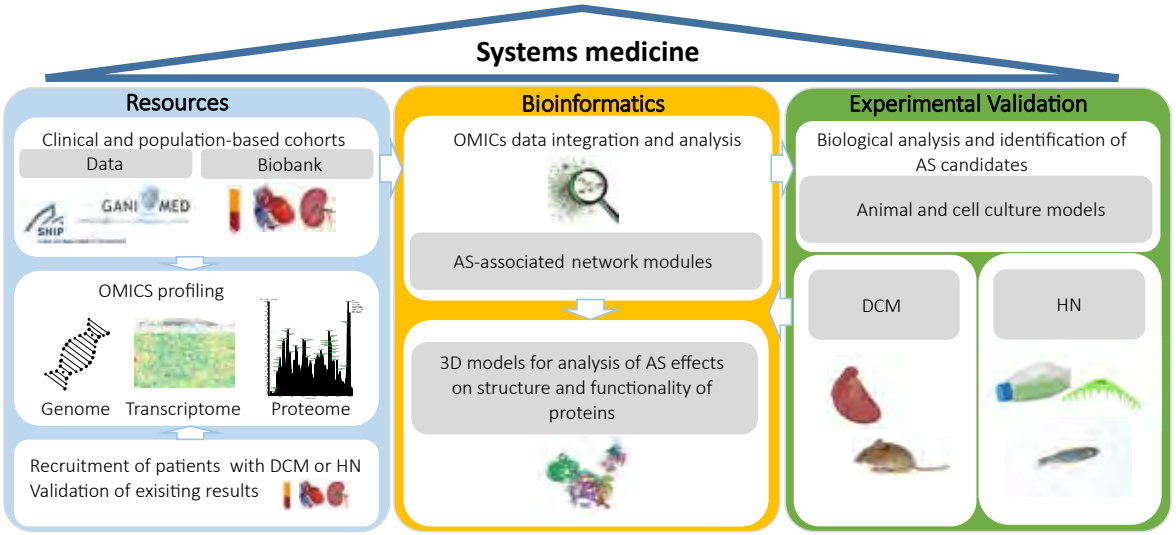


Figure 3: Systems medicine for the detection and functional characterization of alternative splicing. In the Sys_CARE consortium, an interdisciplinary team of clinical and basic researchers as well as bioinformaticians works closely together to make data integration, analysis of AS variants and their analysis on a network level as well as 3D modelling of proteins feasible. The overall aim of the project is to use human omics data to gain a better understanding of AS variants and their functional relevance in diseases such as DCM and HN, ultimately leading to new insights that can be exploited for clinical application. (Source: Sys_CARE consortium)

to identify not only protein isoforms affected by AS, but also their impact on protein-protein interactions and functional networks with heterogeneous multi-omics data (Figure 2). This may lead to the identification of disease subtypes for which optimal therapies can be developed following the concept of personalized medicine.

Sys_CARE – innovative, interdisciplinary and translational approaches in systems medicine

The Sys_CARE consortium's systems medicine research is based on clinical and population-based cohorts and uses existing as well as newly generated multi-omics data. The research focuses on the influence of AS on the formation of protein complexes and on regulatory mechanisms. Sys_CARE is specifically investigating changes in molecular interaction networks that influence the development and progression of diseases. The new bioinformatics methods that the consortium is developing should allow researchers to integrate gene expression data into these networks to clarify how AS impacts cellular mechanisms and regulatory programs.

Alongside this research, Sys_CARE is tracking AS events in time-course studies using *in vivo* and *in vitro* mouse and zebrafish models of DCM and HN, respectively. This offers an opportunity to screen for AS events during disease progression, allowing for an identification of dynamic AS mechanisms. The results are used in turn to iteratively improve the network analysis methods (Figure 3).

Another key step is to confirm AS events by identifying specific protein isoforms or peptides suitable for clinical diagnosis. A comparison of blood-based and biopsy-based AS events should reveal the extent to which disease-associated AS events found in blood reflect mechanistic processes in cardiac or renal tissue. These may lead to biomarkers for clinical diagnosis that could be directly measurable in the patient's blood without invasive procedures to obtain tissue biopsies (biopsy specimens).

First Sys_CARE results

To allow researchers to investigate the influence of AS at the network level, the web-based application DIGGER (www.exbio.wzw.tum.de/digger , Louadi *et al.*, 2020) was developed as part of Sys_CARE. Previous tools and databases are mostly limited to protein-protein interactions, therefore neglecting the influence of AS, where the protein domain required for an interaction may not be part of a protein isoform. To allow for more detailed analyses, DIGGER maps such interactions at the

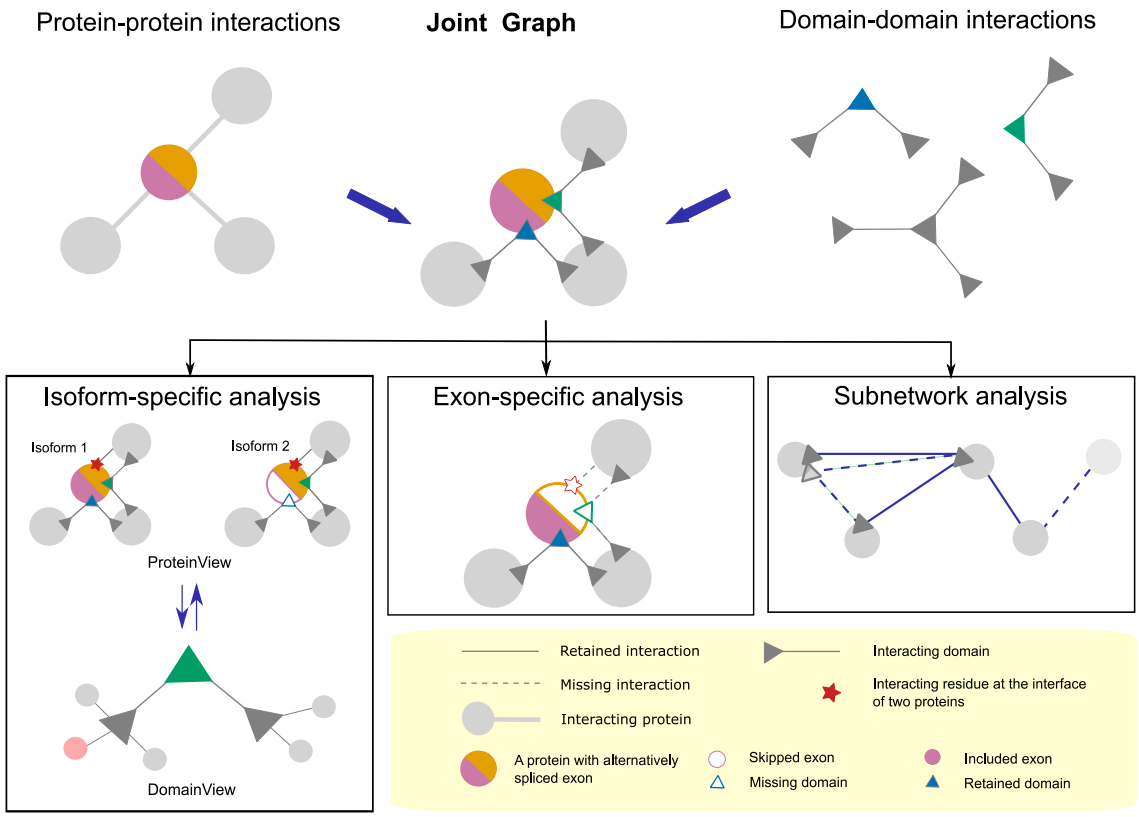


Figure 4: DIGGER overview: protein-protein interactions are integrated with domain-domain interactions to obtain a structurally augmented view on the interactions of each gene. These integrated data allow for comparing the comparison of different protein isoforms (left), understanding the effect of exon skipping (middle), or the analysis of AS on the network-level (right). (Source: https://www.exbio.wzw.tum.de/digger/)

level of individual protein domains and the underlying exon structures (Figure 4). To this end, the app integrates information on protein-protein interactions and domain-domain interactions in a joint network. In addition, DIGGER reveals interacting residues, i.e., amino acids found at the interface of two interacting proteins. These residues were identified using experimental X-ray crystallography methods and nuclear magnetic resonance images of interacting protein structures, providing complementary evidence of possible protein interactions. DIGGER allows users to search for individual or groups of exons or isoforms to visualize and explore their interactions under the influence of AS.

Translational potential of Sys_CARE for precision medicine

The bioinformatics methods developed in Sys_CARE will advance studies of AS in disease and have a lasting impact on the field of systems medicine. We expect the insights gained in Sys_CARE regarding the role of AS in DCM and HN to contribute to the advancement of precision medicine. The translational potential of the basic research done here is demonstrated by recent clinical trials using antisense oligonucleotides (AONs) as a therapeutic tool to effectively control splicing (Spitali and Aartsma-Rus 2012). This suggests that effective treatment of diseases such as DCM and HN, following a mechanistic understanding of the role of AS, will soon be possible.

Project profile:

Sys_CARE is funded by the BMBF as part of e:Med Systems Medicine program and brings together expertise in basic medical research (Prof. Dr. Uwe Völker, Prof. Dr. Karlhans Endlich and Prof. Dr. Nicole Endlich of the University Medicine Greifswald), functional protein modeling (Prof. Dr. Olga V. Kalinina of the Helmholtz Institute for Pharmaceutical

Research Saarland and the Helmholtz Centre for Infection Research), multi-omics network analysis (Prof. Dr. Jan Baumbach, Dr. Markus List and Dr. Tim Kacprowski of the Technical University of Munich) and clinical practice (Prof. Dr. Stephan B. Felix and Prof. Dr. Sylvia Stracke of the University Medicine Greifswald). The overall goal of Sys_CARE is to investigate the significance of AS for the onset and progression of diseases and to develop the bioinformatics methods required for such an investigation.

The key subprojects are:

SP 1: Clinical Omics Profiling
(Prof. Dr. Uwe Völker)

SP 2: Modeling of Protein Structure and Function
(Prof. Dr. Olga V. Kalinina)

SP 3: AS-Associated Network Modules
(Prof. Dr. Jan Baumbach)

SP 4: Experimental Studies of Disease-Associated AS Events
(Prof. Dr. Stephan B. Felix)

References:

Barbaux, S., Niaudet, P., Gubler, M.C., Grünfeld, J.P., Jaubert, F., Kuttenn, F., Fékété C.N., *et al.* (1997). „Donor Splice-Site Mutations in WT1 Are Responsible for Frasier Syndrome." *Nature Genetics* 17 (4): 467–70.

Beqqali, A. (2018). „Alternative Splicing in Cardiomyopathy." *Biophysical Reviews* 10 (4): 1061–71.

Louadi, Z., Yuan, K., Gress, A., Tsoy, O., Kalinina, O.V., Baumbach, J., *et. al.* (2020). „DIGGER: exploring the functional role of alternative splicing in protein interactions." *Nucleic Acids Research.* 10.1093/nar/gkaa768.

Spitali, P., and Aartsma-Rus, A. (2012). „Splice Modulating Therapies for Human Disease." *Cell* 148 (6): 1085–88.

Contact:

Prof. Dr. Jan Baumbach
Chair of Computational Systems Biology
University of Hamburg (UH), Hamburg
jan.baumbach@uni-hamburg.de
www.cosy.bio

Prof. Dr. Uwe Völker
Interfaculty Institute for Genetics and Functional Genomics,
Department of Functional Genomics
University Medicine Greifswald
voelker@uni-greifswald.de
https://biologie.uni-greifswald.de/struktur/institute/interfakultaeres-institut-fuer-genetik-und-funktionelle-genomforschung/department-of-functional-genomics

Prof. Dr. Olga V. Kalinina
Helmholtz Institute for Pharmaceutical Research Saarland (HIPS) / Helmholtz Center for Infection Research (HZI),
Saarland University, Saarbrücken
Olga.kalinina@helmholtz-hips.de
https://www.helmholtz-hzi.de/en/research/research-topics/anti-infectives/drug-bioinformatics

Prof. Dr. Stephan B. Felix
Department of Internal Medicine B,
University Medicine Greifswald
felix@uni-greifswald.de
http://www2.medizin.uni-greifswald.de/inn_b/

Prof. Dr. Karlhans Endlich
Institute of Anatomy and Cell Biology,
University Medicine Greifswald
Karlhans.endlich@uni-greifswald.de
http://www2.medizin.uni-greifswald.de/anatomie

bioinformatics and systems cardiology

An institutional profile of the Klaus Tschira Institute for Integrative Cardiology

by Tobias Jakobi and Christoph Dieterich

The Klaus Tschira Institute for Integrative Cardiology focuses on three thematic areas. First, RNA maturation and processing, whereby a strict control of RNA biology is particularly crucial in heart development and physiology. Our laboratory has succeeded in publishing numerous software solutions aimed at investigating the complex world of RNA. Second, we have established the field of systems cardiology for *in vitro* and *in vivo* models of heart failure. Third, we are building a bridge into the field of clinical data science through the HiGHmed consortium, which is part of the Medical Informatics Initiative. Here, it is worth mentioning in particular our AI work in the field of unstructured German texts from cardiological settings.

Klaus Tschira Stiftung gemeinnützige GmbH

KTS

Klaus Tschira Foundation

The Klaus Tschira Institute for Computational Cardiology was founded in September 2015 with the support of the Klaus Tschira Foundation and is headed by Prof. Dr. Christoph Dieterich. Our focus in bioinformatics is on processing genetic information from DNA to proteins. This has often been seen as a straightforward process – one in which RNA only represents an intermediate product. Yet this picture does not do justice to the role of RNA, because RNA is actually an interactive and dynamic information carrier that fulfills a variety of functions. The stability and translational efficiency of RNA are controlled by its secondary structure as well as by interactions with RNA binding proteins and non-coding RNAs such as microRNA or long non-coding RNA (lncRNA). Co- and post-transcriptional processes, such as RNA modifications, can also alter RNA molecules at the base-pair level and thus influence the final protein sequence even after transcription. With the rediscovery of the class of circular RNAs (circRNAs), another, still largely unexplored, group of RNA molecules has also found its way into the realm of non-coding RNAs. The interplay of all these parts in a large interaction network is now known as post-transcriptional gene regulation and controls numerous processes in human cells. Our work traditionally starts out with specific questions or observations from RNA biomedicine.

A possible question might be:

> " Heart muscle cells grow not only through exercise, but also as the result of pathological influences such as hypertension. But why do the long-term effects differ significantly at the molecular and clinical level?"

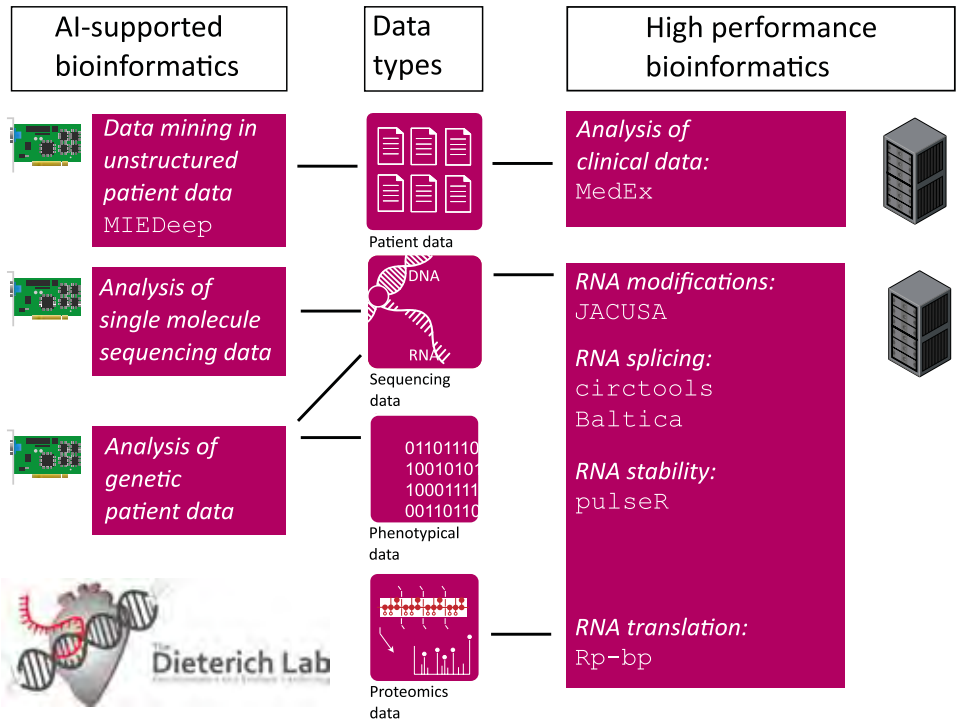


Figure 1: Software Overview. New methods based on artificial intelligence are entering various fields of bioinformatics and medical informatics (left), even though classic bioinformatics tools still dominate a large part of the workflows (right) (Source: Tobias Jakobi).

Open-source software tools for the scientific community

We typically work with our experimental partners to develop hypotheses, which we then test using established bioinformatics and statistical methods as well as software and procedures that we have developed ourselves. Newly developed software tools are made available to the scientific community in an open-source format and are continuously further enhanced.

For example, the research group has developed software that is able to recognize modified RNA base pairs from sequencing data (Piechotta *et.al,* 2017). Other specialized software solutions for RNA splicing are Baltica and circtools, the latter of which is especially useful for circular RNA (Jakobi *et al.,* 2019). The software has been implemented to cover the entire work flow from quality analysis of raw data to detection and reconstruction of circular RNAs to the design of molecular genetic primer sequences for validation experiments.

The stability of RNA is a critical factor for many regulatory functions. In many cases, the availability of RNA templates is rapidly regulated by decay processes or synthesis depending on the context. With PulseR and further theoretical work, the research group has developed a tool for analyzing RNA metabolic kinetics from RNA sequencing data (Uvarovskii *et al.,* 2019).

Yet in other cases, it is important to know which RNAs are actually translated into proteins and how the translation of the proteins compares to the transcription of the RNA. Ribosome profiling using high-throughput sequencing (Ribo-seq) is a promising new technique for characterizing ribosome distribution on RNA with base-pair resolution. The ribosome is responsible for the translation of mRNA into proteins, so that information on its occupancy provides a detailed view of ribosome density and position, which could be used, among other things, to discover newly translated open reading frames (ORFs). A Bayesian approach to predict ORFs from ribosome profiles has been implemented in the software Rp-Bp (Malone *et al.,* 2017).

Figure 2: Group photo. From left to right: Maja Bencun, Aljoscha Kindermann, Thiago Britto Borges, Isabel Naarmann-de Vries, Etienne Boileau, Tami Liebfried, Jessica Eschenbach, Christoph Dieterich, Magdalena Smieszek, Phillip Richter-Pechanski, Qi Wang,and Tobias Jakobi (Photo: Tobias Jakobi)

Systems cardiology requires adequate and specialized hardware

Quantitative systems cardiology is characterized by immense amounts of data that are no longer manageable on ordinary workstation computers. For this purpose, the research group maintains its own network of high-performance computers, which are capable of analyzing even extensive experimental data sets in a short time. The computer cluster currently consists of 26 dedicated computing nodes with a main memory of up to one terabyte, which is needed for such tasks as genome assembly or the parallel analysis of large OMICS data sets.

In addition, the computer cluster has been equipped with a dedicated server that accommodates NVIDIA graphics processing units (GPUs). This special hardware is derived from 3D graphics cards for computer games, which have become increasingly powerful in recent years and are predestined to process machine learning and AI tasks due to their highly parallel architecture. The special system is used for a variety of tasks, ranging from extracting the sequence of base pairs from raw sequencing data to text mining in medical documents to the analysis of patient genomes from molecular genetic data.

Software solutions for structured and non-structured patient data

In clinical practice, large amounts of data from a wide variety of areas are routinely generated. Our software, the Medical Data Explorer (MedEx) (Kindermann *et al.,* 2019), is an intuitive, web-based solution with options for easy data import. We combine a modern dynamic web interface with an in-memory database solution for near real-time responsiveness. MedEx offers various visualization options in order to provide a simple overview of the loaded data, and to generate hypotheses and perform elementary analyses.

In the clinical setting, much treatment-relevant information is still recorded in the form of unstructured German texts. A typical example is the discharge letter, which is intended as a transfer document for communication between physicians. Our

Medical Information Extraction using deep learning (MIEdeep) project aims to make this data source usable for extracting information. For this purpose, innovative approaches from the fields of deep neural networks (deep learning) and natural language processing (NLP) are used. We combine approaches of machine learning for data preparation, creation of training data and information extraction with a modern graphical user interface suitable for use in a clinical setting.

References:

Piechotta, M., Wyler, E., Ohler, U., Landthaler, M., Dieterich, C. (2017). JACUSA: site-specific identification of RNA editing events from replicate sequencing data. *BMC Bioinformatics*, 18, 7, doi:10.1186/s12859-016-1432-8.

Jakobi, T., Uvarovskii, A., Dieterich, C. (2019). circtools – a one-stop software solution for circular RNA research. *Bioinformatics*, 35, 2326–2328, doi:10.1093/bioinformatics/bty948.

Uvarovskii, A., Vries, I.S.N., Dieterich, C. (2019). On the optimal design of metabolic RNA labeling experiments. *PLOS Computational Biology*, 15, e1007252, doi:10.1371/journal.pcbi.1007252.

Malone, B., Atanassov, I., Aeschimann, F., Li, X., Großhans, H., Dieterich, C. (2017). Bayesian prediction of RNA translation from ribosome profiling. *Nucleic Acids Res*, 45, 2960–2972, doi:10.1093/nar/gkw1350.

Kindermann, A., Stepanova, E., Hund, H., Geis, N., Malone, B., Dieterich, C. (2019). MedEx - Data Analytics for Medical Domain Experts in Real-Time. *Stud Health Technol Inform*, 267, 142–149, doi:10.3233/SHTI190818.

Contact:

Prof. Dr. rer. nat. Christoph Dieterich
Professorship for Bioinformatics and Systems Cardiology
INTERNAL MEDICINE III: HEART, VASCULAR AND LUNG
Heidelberg University Hospital &
Klaus Tschira Institute for Computational Cardiology
christoph.dieterich@uni-heidelberg.de

https://dieterichlab.org

Dr. rer. nat. Tobias Jakobi,
Postdoctoral Fellow
INTERNAL MEDICINE III: HEART, VASCULAR AND LUNG
Heidelberg University Hospital &
Klaus Tschira Institute for Computational Cardiology
tobias.jakobi@med.uni-heidelberg.de

https://tobias.jako.bi

the virtual brain

Interview with Petra Ritter, BIH Johanna Quandt Professor for Brain Simulation at the Berlin Institute of Health (BIH) and Charité

Petra Ritter works in a modern office at Robert-Koch-Platz on Campus Charité Mitte. She is BIH Johanna Quandt Professor for Brain Simulation at the Berlin Institute of Health (BIH) at Charité. Ritter also leads an international project called The Virtual Brain, which uses computer simulation for a better understanding of the brain. Stefanie Seltmann spoke with her about the project.

gesundhyte.de: *Prof. Ritter, what exactly is The Virtual Brain?*

Prof. Dr. med. Petra Ritter: The Virtual Brain is a brain simulation platform developed by our international consortium that has been publicly available since 2012. Any scientist can use it to construct a patient-specific brain and simulate its activity on the computer.

Brain examinations are usually conducted using EEG, MRI or PET scans, from which researchers can draw conclusions about brain activity or even individual nerve cells (neurons). Do you feed this information into the computer?

Yes, we obtain this information from individual subjects and integrate it into mathematical models of their brains. But we start by simplifying things quite a bit, as we want to create a model for a brain with particular characteristics that are key to whatever question we are interested in at that time.

A digital interactive brain atlas is available in many languages at https://www.brainsimulation.org/atlasweb/

Source: https://www.brainsimulation.org/bsw/zwei/research-visualization

Prof. Dr. med. Petra Ritter (Photo: David Ausserhofer)

A brain has one hundred billion neurons that are connected to one another with yet more synapses. You can't possibly reproduce them all...

No, there is no supercomputer in the world that could cope with this huge amount of interacting elements. That's why we have to scale it down. Many neurons are synchronized and interconnected. Just like when several birds flock in the same direction, you can determine the path of one individual by looking at the path of the flock as a whole. We do the same with neurons. By bundling them together, we can significantly reduce the complexity of the brain.

> "There is no supercomputer in the world that could cope with this huge amount of interacting elements."

Let's take a look at a real-life example. What do you see in Alzheimer's patients with your model?

A PET scan, for example, shows us that patients with Alzheimer's disease have beta-amyloid deposits in many areas of the brain. We also know from animal and cellular studies that these protein deposits seem to disrupt the function of inhibitory neurons in their vicinity. We can take these two pieces of knowledge and build them into our model. So, we create the spatial distribution of the protein deposits from the PET data, and a small mathematical model that translates the presence of this protein into the reduced activity of nearby inhibitory cells. If we put this together and start a simulation of the patient's brain, we see that just by taking the protein deposits into account, we get altered EEG signals in the simulation – in other words, things slow down. We see this same "slowing down" in real brain data.

The hope would be that, in line with the BIH's motto "Turning Research into Health," you could then go on to simulate which procedure or drug treatment might provide the most help?

A multicenter clinical trial on epilepsy is already underway in France. Thirty percent of epilepsy patients do not respond to medication, which means their only option is neurosurgery. In preparation for surgery, electrodes are placed on or in the brain. It is not possible, however, to place these electrodes everywhere, so we end up with holes in this map of brain activity. Our model can take all the available information and

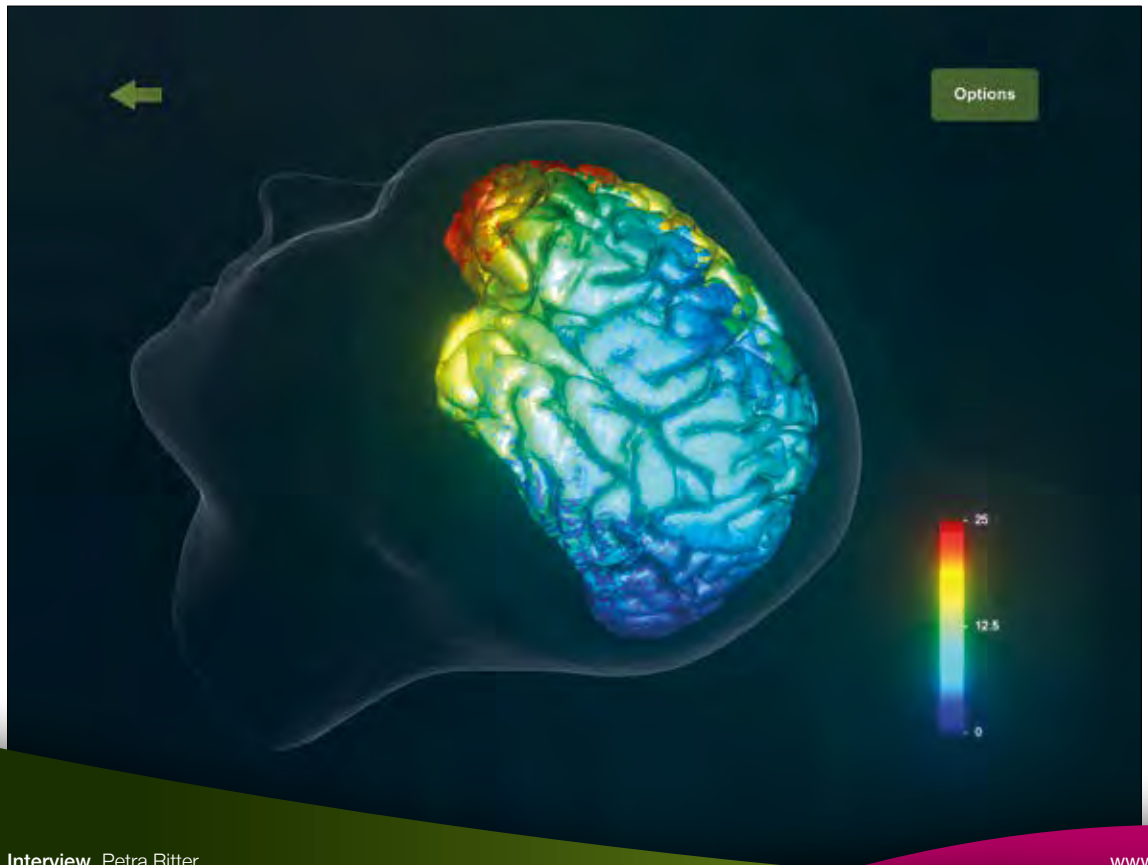
then simulate the activity in those places where information is missing. First indications have been published which suggest that the model´s prediction regarding the epileptic zone, where the pathological activity is starting and is spreading from may be better than conclusions from neurologists just looking at the data and evaluating it. Another study is now starting in France with four hundred patients. Half will undergo surgery using as input information gleaned from The Virtual Brain, the other half with the input of classical information – i.e., the human team of experts looking at and drawing conclusions from the data. In less than four years, we will be able to say whether the computer model actually provides better predictions and results in a greater reduction in the number of epileptic seizures after the operation than the conventional approach.

Could this type of brain model also be used to predict the effect of medication?

"The hope is that in the future it will be possible to use the computer model to determine which drug will be most effective for a specific patient."

We have actually already demonstrated this on the Alzheimer's brain. Memantine is a common Alzheimer's drug, which often produces at least a temporary improvement in the cognitive performance of patients. We were able to show on the

This image shows an EEG projected onto the cortical surface of the brain



Source: https://www.brainsimulation.org/bsw/zwei/research-visualization

computer that the brains of Alzheimer's patients actually display normalized function following the virtual administration of this drug. The hope is that in the future it will be possible to use the computer model to determine which drug will be most effective for a specific patient.

Would it even be possible to conduct a first clinical trial on your model?

Yes, we really hope that in the future, before certain procedures are performed on animals or human patients, they can first be tested on computer brain models – and that these tests will enable us to recognize things like side effects in advance.

What role does artificial intelligence play in your project?

It plays a very important role. On the one hand, artificial intelligence learns from our models of biological networks and their ability to produce cognition, and on the other hand, we use artificial intelligence to build better models and make better predictions for the patients.

So, artificial intelligence helps human intelligence to understand natural intelligence?

You could certainly say that, yes.

Thank you for your time, Prof. Ritter.

My pleasure.

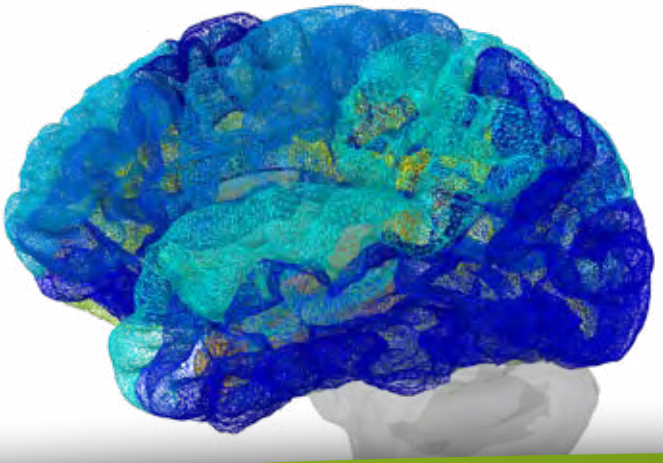
Stefanie Seltmann conducted the interview.

An app for mobile devices is also being developed in the lab of Professor Ritter (Source: https://www.brainsimulation.org/bsw/zwei/research-visualization#).

Contact:

Prof. Dr. med. Petra Ritter
Director of Brain Simulation Section (CCM)
Charité – Universitätsmedizin Berlin
petra.ritter@charite.de

www.thevirtualbrain.org
https://brainsimulation.charite.de/en
www.brainsimulation.org
https://virtualbraincloud-2020.eu



Source: https://www.thevirtualbrain.org/tvb/zwei/brainsimulator-screens

WWW.THEVIRTUALBRAIN.ORG

model exchange for regulatory genomics (MERGE)

Fostering the sharing and reuse of predictive models in genomics through software standardization

by Julien Gagneur, Oliver Stegle, and Michael J. Ziller

Artificial intelligence is transforming our ability to decipher individual genomes. Novel models in regulatory genomics, which predict the effect of DNA variations on cellular processes, are critical for elucidating the mechanisms of disease-causing variants and mutations in tumours. Despite the promise of this field, there is a lack of coherent software infrastructure, workflows, and interfaces to deploy these new technologies at scale. To address this issue, **MERGE** will establish a platform for the exchange, benchmark and extension of regulatory genomics models. We will apply these models to understand the molecular basis of diseases such as rare diseases, cancer, and psychiatric illnesses.

Regulatory genomics

Over the last 10 years, the biomedical community has built up a tremendous resource of genomic and molecular data which provides the basis for personalized genome interpretation. Pertinent examples include data generated from international efforts such as ENCODE, Roadmap Epigenome, Blueprint, TCGA, FANTOM5, and GTEx, which have generated comprehensive maps of the genetic, epigenetic (e.g. (Ziller *et al.* 2015)) and transcriptional state across a wide range of cell types, individuals and diseases. Collectively, these rich resources hold great promise to allow for decoding functional roles of disease-associated genetic changes in coding and non-coding regions of the genome. While variants in coding sequences are frequently associated with gain or loss of function mutations, the situation is more subtle in the case of non-coding genetic variants. In particular, these variants might act on distinct layers of gene regulation, altering three-dimensional genome architecture, transcription levels, RNA splicing, RNA stability, or translation frequency.

Multiple strategies have been used to understand the role of non-coding sequences in gene regulation. One strategy is to identify genetic variants associated with gene regulatory changes using large cohorts with matched molecular profiles (e.g. gene expression profiles) and genetic data. However, such association-based approaches provide little mechanistic insights into the molecular basis of specific genetic variants. Moreover, due to power limitations, they fall short of providing insights into rare genetic variants, such as those arising in individual tumours or implicated in rare diseases. A second strategy is based on biophysical models of gene regulatory mechanisms, modelling, for instance, the binding affinity of transcription factors to DNA sequence. However, biophysical models fall short in realistically capturing the actual complexity of *in vivo* gene regulatory mechanisms.

The deep learning revolution

A third strategy, lying conceptually between biophysical modelling and association-based approaches, leverages artificial intelligence techniques. Computational biologists have shown how genomics data can be cast to data representations encountered in computer vision and natural language

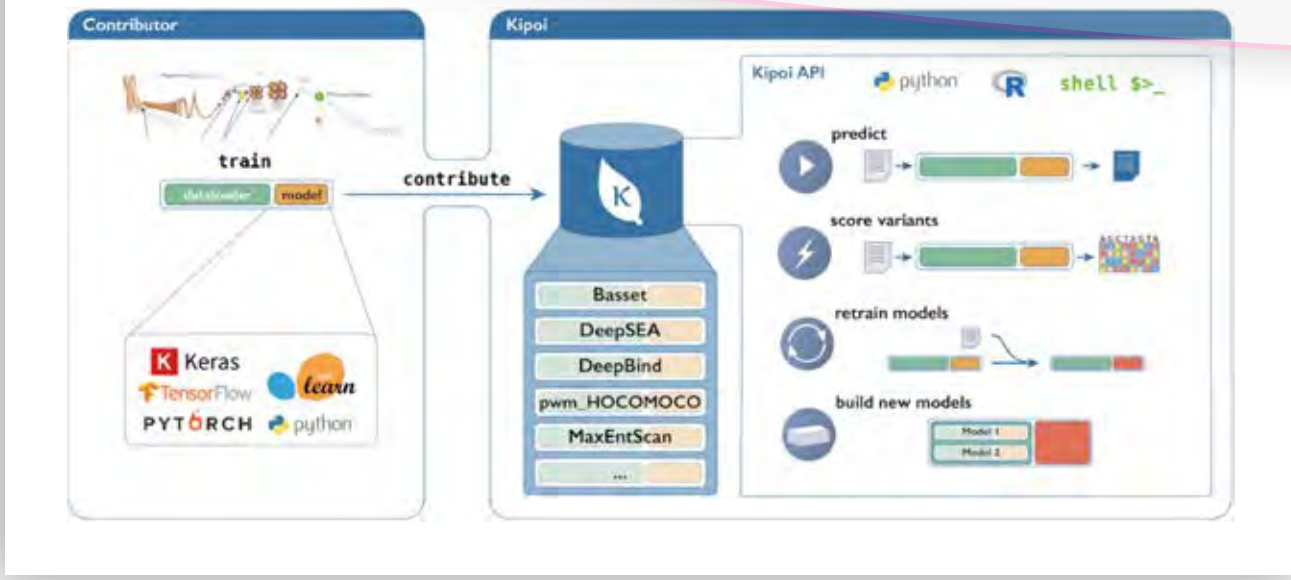


Figure 1: Concept of Kipoi. Standardized interface for regulatory genomics models enable model sharing, their use for new data and model composition. Kipoi facilitates downstream analytics of models, including predictions. The model repository is available at https://kipoi.org/ (Source: Roman Kreuzhuber, Kipoi team).

processing. This allowed leveraging techniques from deep learning, a recent branch of Artificial Intelligence, to model relationships between genotype and molecular phenotypes. Deep learning models are flexible classes of mathematical functions that are fitted in an end-to-end fashion to predict an output, say RNA abundance, from an input, say DNA sequence. Unlike genetic association models, deep learning models can be applied to genetic variants that were not seen in the training data. Therefore, they can generalize to rare variants and de novo mutations, which is relevant for rare diseases, or to somatic mutations unique to a specific tumour. Moreover, in contrast to biophysical models, deep learning models are not biased by strong modelling assumptions that may be violated *in vivo*. Furthermore, effective software (deep learning framework such as pytorch, tensorflow, or keras) and hardware infrastructure (Graphical Processing Units) have been developed that enable training complex models on massive amounts of data such as those provided by large international genomics projects. Deep learning models have become the new state-of-the-art predictive models for a wide range of gene regulatory layers (Eraslan *et al.*, 2019) including transcription factor binding sites, chromatin modifications (e.g. (Angermueller *et al.*, 2017)), DNA contact maps, gene expression, splicing (e.g. (Cheng *et al.*, 2019)), RNA degradation, and translation.

Share the models!

While there is a large number of publications that describe models for key tasks in regulatory genomics, their potential is currently underutilized. A major reason for this is the lack of a coherent framework to share and exchange models in the community. The exchange of such models has been limited due to:

- **Lack of unified model interfaces for handling genomic data**
- **Heterogeneous machine learning frameworks, and software dependencies**
- **Need for interpretability of machine learning models to gain biological insights**
- **Lack of comparability of methods, i.e. for objective benchmark**
- **High complexity for reusing and applying existing models to new data**
- **High entry barrier for bioinformaticians who are not machine learning experts.**

To address these challenges, we are developing the programmatic standard and repository Kipoi (Avsec *et al.*, 2019). Kipoi is a so-called model zoo, i.e. a repository hosting hundreds of trained models (Figure 1). Our solution addresses develop-

ers, who can deposit their models and derive new models based on existing building blocks. However, the primary focus is on users. Kipoi models come with code to preprocess and load input data in major file formats, which means that existing models can be applied to new data in a few lines of code, via python, R or via the command line. Models can also be retrained and fine-tuned to new datasets and tasks. Furthermore, because Kipoi models are standardized, it is now possible to easily benchmark models from heterogeneous origins, and to build new models based on existing ones, thereby enabling genuine compositional modelling. These model standards also allow for generic routines for the downstream analysis and the interpretation of Kipoi models. Major machine learning frameworks are already supported by Kipoi (Tensorflow, Pytorch, Keras, and Scikit-learn). Altogether, Kipoi applies the FAIR sharing principle (Findable, Accessible, Interoperable, and Reusable), which have been defined for data, to trained machine learning models.

The MERGE work packages

MERGE is a collaborative project between the groups of Julien Gagneur, TUM, Oliver Stegle, DKFZ, who have both initiated Kipoi, and Michael Ziller from the MPI Psychiatry. MERGE will further develop Kipoi by extending the functionality of the repository and the performance of the models it contains (Figure 2). We will develop generic algorithms to improve Kipoi's usage in cloud environments and interpretability of the models. We will leverage on the de.NBI cloud instances for test and deployment on the cloud. We will extend the model repository with models of transcriptional enhancer, splicing and translation fitted on experimental data probing natural genetic variations, as well as high-throughput genetic perturbation assays. Furthermore, the Kipoi model zoo will be used to interpret cancer genomes, genetic variants associated with common diseases and rare diseases. The team members have on-going collaborations on these three disease areas and support for continuous collaboration in the context for MERGE at the German Cancer Research Center in Heidelberg, the MPI Psychiatry, the German Heart Center, and the Institute of Human Genetics of the TUM.

Outlook

Models and approaches developed in MERGE will be critical to fully leverage the exponential growth of human genome data that are expected to be generated in the coming years. Initiatives such as the European '1+ Million Genomes' Initiative (https://www.pubaffairsbruxelles.eu/germany-joins-the-1-million-genomes-initiative-eu-commission-press/), which is also signed by Germany, could profit from these advances and future developments. In parallel to novel AI technology and tools, it will be critical to foster bridges to data infrastructures that hold these key data resources. The German Human Genome Phenome Archive (http://www.ghga.de) is an example of an emerging infrastructure for human omics data, which is currently being developed within the framework of the National Research Data Infrastructure (NFDI). Advances derived from projects such as MERGE will also be connected to the national Bioinformatics infrastructure initiative de.NBI (https://www.denbi.de/). More broadly, the integration of AI methods and tools with data infrastructures is an important future challenge.

Key data:

Project: MERGE (Model Exchange for Regulatory Genomics)
Funding Scheme: CompLS – Computational Life Sciences (https://www.gesundheitsforschung-bmbf.de/de/compls-computational-life-sciences-9161.php)
Partners: Julien Gagneur (Technical University of Munich, Munich), Oliver Stegle (Deutsche Krebsforschungszentrum, Heidelberg), Michael Ziller (Max-Planck-Institute for Psychiatry, Munich)

References:

Angermueller, C., Lee, H.J., Reik, W., and Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. *18*, 67.

Avsec, Ž., Kreuzhuber, R., Israeli, J., Xu, N., Cheng, J., Shrikumar, A., Banerjee, A., Kim, D.S., Beier, T., Urban, L., *et al.* (2019). The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. Nat. Biotechnol. *37*, 592–600.

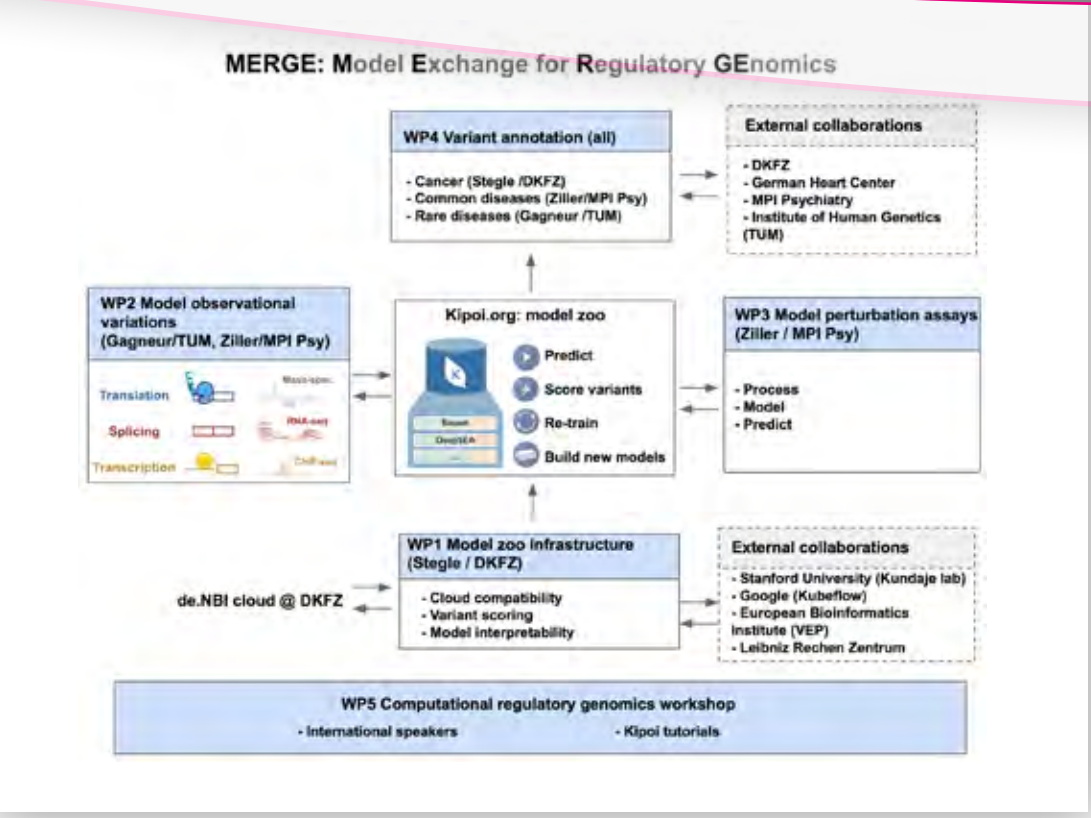


Figure 2: Approach and overview of the MERGE project. Central to the plans are extensions of Kipoi, to (1) integrate cloud components into Kipoi, (2) extend the functionality of Kipoi for modelling molecular consequences of natural variation from observational data, (3) the effective use of perturbation assays and (4) the use of models for the interpretation of human genetic variation. MERGE is anchored in a strong network of collaborations (Source: Julien Gagneur).

Cheng, J., Nguyen, T.Y.D., Cygan, K.J., Çelik, M.H., Fairbrother, W.G., Avsec, Ž., and Gagneur, J. (2019). MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome Biol. *20*, 48.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet. *20*, 389–403.

Ziller M.J., Reuven E., Yaffe Y., Donaghey J., Pop R., Mallard W., Issner R., Gifford C.A., Goren A., Xing J., Gu H., Cacchiarelli D., Tsankov A., Epstein C., Rinn J.L., Mikkelsen T.S., Kohlbacher O., Gnirke A., Bernstein B.E., Elkabetz Y., Meissner A. (2015), Dissecting neural differentiation regulatory networks through epigenetic footprinting. Nature *518*, 355–359.

Contact:

Prof. Dr. Julien Gagneur
Computational Molecular Medicine
Technical University of Munich
gagneur@in.tum.de

https://www.in.tum.de/gagneurlab

Prof. Dr. Oliver Stegle
Computational Genomics and Systems Genetics (B260)
German Cancer Research Center, Heidelberg
& European Molecular Laboratory Heidelberg
o.stegle@dkfz.de

https://www.dkfz.de/en/bioinformatik-genomik-systemgenetik/index.php

Prof. Dr. Michael Ziller
Functional Genomics in Psychiatry
University of Münster &
Genomics of Complex Diseases
Max Planck Institute of Psychiatry
Munich
michael_ziller@psych.mpg.de

https://www.psych.mpg.de/2165045/ziller

Photos: Gagneur, Stegle, Ziller

"bench to bedside": systems pharmacology in practice

How computer-based methods are driving the digital transformation of pharmaceutical research and therapeutics

Company profile: esqLABS GmbH

by Stephan Schaller

The pharmaceutical industry is expected to offer more cost-effective treatments for complex disease patterns. Delivering on this expectation is increasingly becoming a challenge. Fast prototyping and personalized medicine are two solutions that enable shorter research and development (R&D) times and help to determine the optimal course of treatment for individual patients. A crucial prerequisite for these approaches is a more rigorous adoption of data processing and analysis methods in pharmaceutical R&D and clinical practice. Here, software solutions for analyzing (clinical) data can help informed decisions be made more quickly.

esqLABS
WE EMPOWER HEALTH CARE

Fast prototyping in pharmaceutical R&D

Researching new medicines is a lengthy, highly complex, and cost-intensive scientific process.

For several years now, within the pharmaceutical industry, the rapid rise in the amount of data generated during R&D has fostered a growing appreciation – and a significantly increased need – for using computer models to analyze and process data (EFPIA MID3 Workgroup *et al.*, 2016). Although highly specialized individual solutions currently exist for the various segments along the value chain (Figure 1), they do not guarantee knowledge transfer and transparency.

And yet it is precisely this transfer of knowledge and expertise throughout the various research and development phases and the successful demonstration of a drug's efficacy in clinical trials, which are the key challenges for R&D in the pharmaceutical industry (Goldblatt and Lee, 2010).

esqLABS has recognized the value in using physiologically-based quantitative systems pharmacology (PB-QSP) model platforms that track drug distribution and efficacy in the body. The architecture of these models enables integrating a variety of different data from cell cultures, animals, individual patients, and patient populations generated in pharmaceutical research (Figure 1). Combining such data in a single platform promotes knowledge transfer and retention throughout the entire R&D process, resulting in quality and efficiency improvements and making the platforms more reliable than previous technologies in their prediction of clinical applications (Jones *et al.*, 2006).

With these models, esqLABS can investigate a wide variety of extrapolation scenarios in drug development using translational modeling approaches to integrate research data. These scenarios include predicting drug effects in humans based on

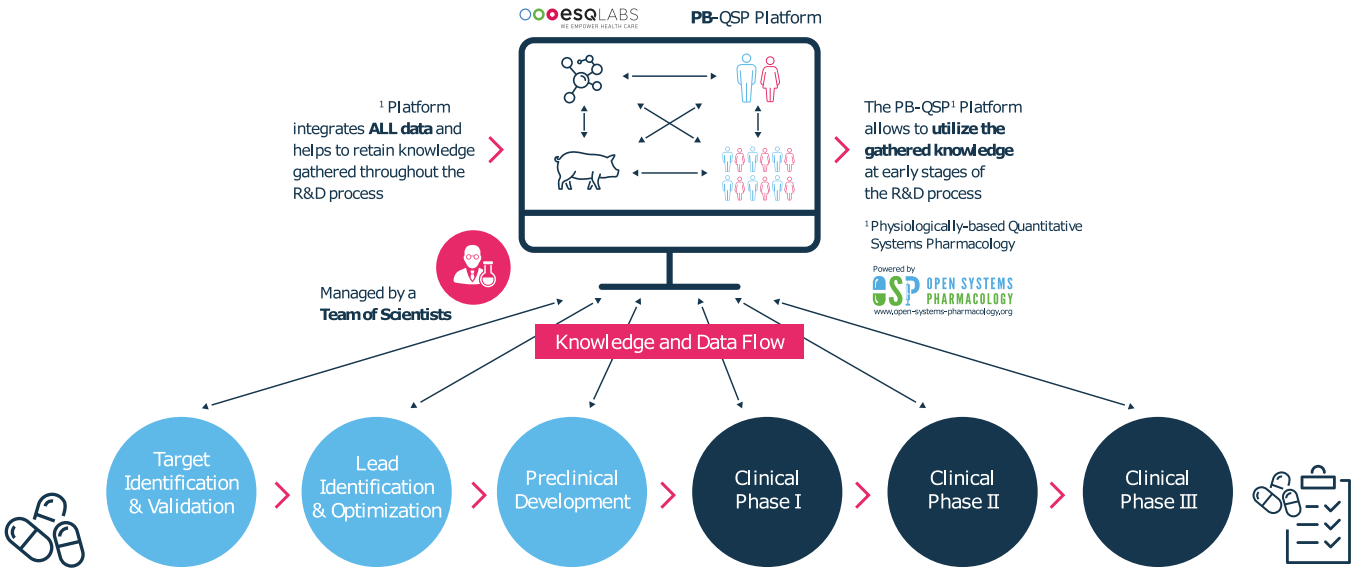


Figure 1: Use of computer models in the research & development process in the pharmaceutical industry. The "state-of-the-art" approach of the "patchwork" of QSAR, systems biology, PK/PD, conventional QSP platforms, and popPK/PD models is compared to the "one platform for all" approach of PBPK/PD QSP platforms from esqLABS (Source: esqLABS GmbH).

in vitro and animal studies, predicting drug interactions, and predicting the effects of age, disease, and genetics on a drug's distribution and efficacy.

The computer models contained in the platform are adapted to their specific R&D processes in cooperation with pharmaceutical companies and help with simulations in decision-making in drug development, shortening R&D cycles and thus harboring considerable cost-saving potential.

Personalized medicine

The use of computer models of diseases to calculate the optimal drug dose for individual patients within so-called therapeutic decision support (TDS) systems for use in hospitals holds promise to enable the paradigm shift from population-based medicine to personalized medicine.

The PB-QSP platforms developed by esqLABS and the resulting solutions make it possible to calculate in advance how much medication to dispense to individual patients and what effect this personalized dosage would have. Such technologies help avoid errors in communication and dosage calculation while also taking pressure off medical staff and increasing process efficiency.

TDS systems are of particular value for therapies with a narrow therapeutic window, as even slight deviations in the dosage of these drugs can have life-threatening consequences (Blix *et al.*, 2010; FDA).

The platform, therefore, helps with the precise dosing of complex medications. Consequently, the development of personalized treatments for individual patients also leads to better treatment for patients overall (Stern *et al.*, 2016). In addition to the direct benefits of the therapy, TDS systems can also help avoid follow-up costs incurred through incorrect dosing, such as an overdose, and their subsequent treatment, which can include hospitalization (Blix *et al.*, 2010).

Simulation models for research and therapy

To implement "fast prototyping" and personalized medicine, complex biological relationships need to be contextualized with mathematical models. Here, esqLABS is developing solutions which account for future requirements in (medical) sci-

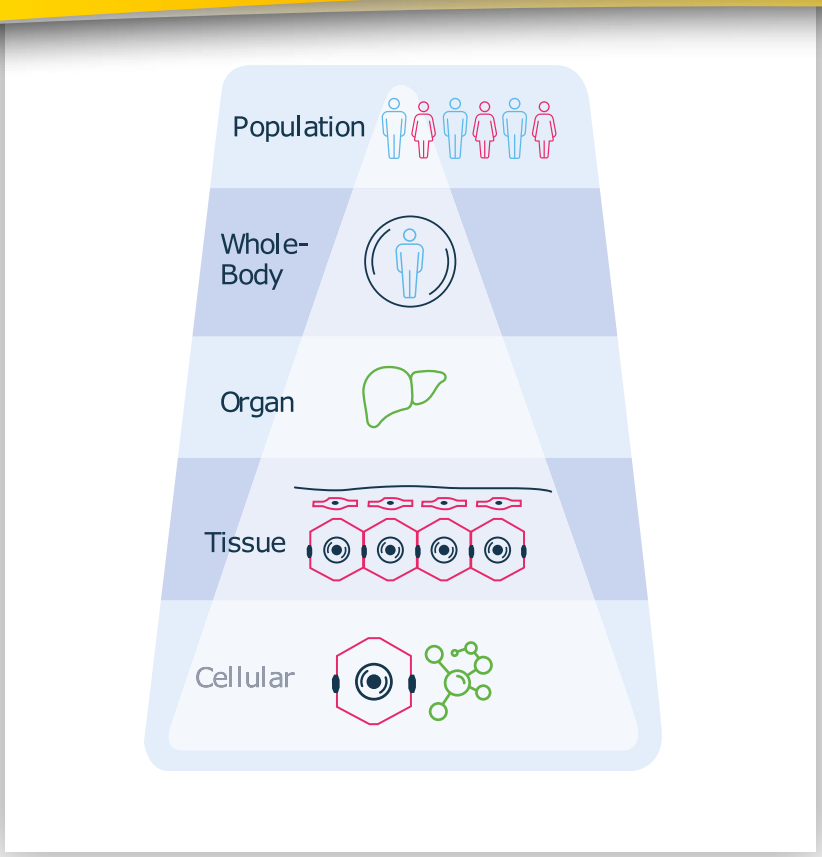


Figure 2: Multi-scale modeling and simulation. The Software Platform Open Systems Pharmacology Suite was developed to model and simulate biological processes focusing on pharmacokinetics, pharmacodynamics, and disease progression (including biochemical reaction networks). As a result, the platform enables the combination of several organizational and physiological scales, from cellular processes to populations (Source: esqLABS GmbH).

ence and industry. For their technology development and services, esqLABS GmbH utilizes the open-source Open Systems Pharmacology software platform (www.open-systems-pharmacology.org), and invests in technology transfer to apply these in personalized medicine.

One example is the further development of the Open Systems Pharmacology Suite (OSPS), an open-source software program that helps integrate and analyze complex biomedical data. The integration of complex disease processes, right down to cellular interactions, and the simulation of a wide variety of application scenarios, up to population level (Figure 2), can make model management and quality assurance quite challenging. A modularization concept and an automated qualification process have been derived to solve this challenge.

With its work, esqLABS integrates data and knowledge streams from all R&D phases within one platform using a generic multi-scale computer model. Such an approach, if rigorously applied, can increase the reliability and efficiency in decision-making in pharmaceutical research and drug therapy.

About the company:

Founded: June 2017

Number of employees: 8

Activities: Services, consulting, and software- and tool development in the field of computer-based modeling and simulation of drug distribution, action, and disease processes

- **PBPK-based analyses of systemic drug distribution (pharmacokinetics, PK)**
  - Age dependency in PK (pediatric analyses)
  - Dependency of PK on co-medications (DDI)
  - Pharmacogenomic dependency of PK
  - Dependency of PK on organ damage (liver/kidney)
  - (Oral) drug absorption, formulation development, and virtual bioequivalence analysis
- **Quantitative systems pharmacology (QSP) analysis of drug action (pharmacodynamics, PD) on different diseases**
  - Translational analysis of drug exposure & response profiles
  - Analysis of optimal personalized medication

Computer-based methods are driving the digital transformation of pharmaceutical research and therapeutics (Source: esqLABS GmbH).

References:

Blix, H.S., Viktil, K.K., Moger, T.A., and Reikvam, A. (2010). Drugs with narrow therapeutic index as indicators in the risk management of hospitalised patients. Pharm. Pract. 8, 50–55.

Coorevits, P., Sundgren, M., Klein, G.O., Bahr, A., Claerhout, B., Daniel, C., Dugas, M., Dupont, D., Schmidt, A., Singleton, P., *et al.* (2013). Electronic health records: new opportunities for clinical research. J. Intern. Med. 274, 547–560.

EFPIA MID3 Workgroup, Marshall, S.F., Burghaus, R., Cosson, V., Cheung, S.Y.A., Chenel, M., DellaPasqua, O., Frey, N., Hamrén, B., Harnisch, L., *et al.* (2016). Good Practices in Model-Informed Drug Discovery and Development: Practice, Application, and Documentation. CPT Pharmacomet. Syst. Pharmacol. 5, 93–122.

FDA, C. for D.E. and Generic Drug User Fee Amendments of 2012 - Narrow Therapeutic Index Drugs.

Goldblatt, E.M., and Lee, W.-H. (2010). From bench to bedside: the growing use of translational research in cancer medicine. Am. J. Transl. Res. 2, 1–18.

Jones, H.M., Parrott, N., Jorga, K., and Lavé, T. (2006). A novel strategy for physiologically based predictions of human pharmacokinetics. Clin. Pharmacokinet. 45, 511–542.

Stern, A.M., Schurdak, M.E., Bahar, I., Berg, J.M., and Taylor, D.L. (2016). A Perspective on Implementing a Quantitative Systems Pharmacology Platform for Drug Discovery and the Advancement of Personalized Medicine. J. Biomol. Screen. 21, 521–534.

Contact:

Dr. Stephan Schaller
Managing Director
esqLABS GmbH
Saterland, Germany
Stephan.schaller@esqlabs.com

www.esqlabs.com

Photo: Schaller

robots in the operating room

Robot-assisted surgery: from gimmick to standard

by Klaus-Peter Jünemann

In 1961, General Motors installed the first robots in the automotive industry. Today, virtually all industrial assembly lines are dominated by robotic arms, as recurring processes and work flows are handled better and more precisely by machines than humans. Robots today also work as "cobots" in interaction with humans. Digitilization and human-machine interaction are the key technologies of the 21st century. In industry, telecommunications, commerce – in nearly every area of life – these technologies are already taken for granted. But what about robots in the operating room? Are we willing to entrust our lives to a machine?

Why we need robot-assisted surgery

Robot-assisted surgery (RAS) already exists. The question is, are today's surgeons who back this technology too lazy or simply too poorly trained to perform conventional open or laparoscopic surgery? Are robots in the OR just a "nice to have," a gimmick for the surgeon? What is the benefit for the patient? Why should patients be willing to undergo surgery that is assisted by a robot?

For a better understanding, robots in medicine are telemanipulators based on the master-slave principle, i.e., the movements of the surgical instruments are controlled by the physician. The learning curve is shorter than for conventional laparoscopy, which means that robot-assisted surgery can be learned much faster.

Prof. Klaus-Peter Jünemann, MD
Professor and Speaker of the interdisciplinary "Kurt-Semm-Zentrum of Laparoscopic and Robot-Assisted Surgery" at the University Hospital Schleswig-Holstein, Campus Kiel (Photo: Felix Prell)

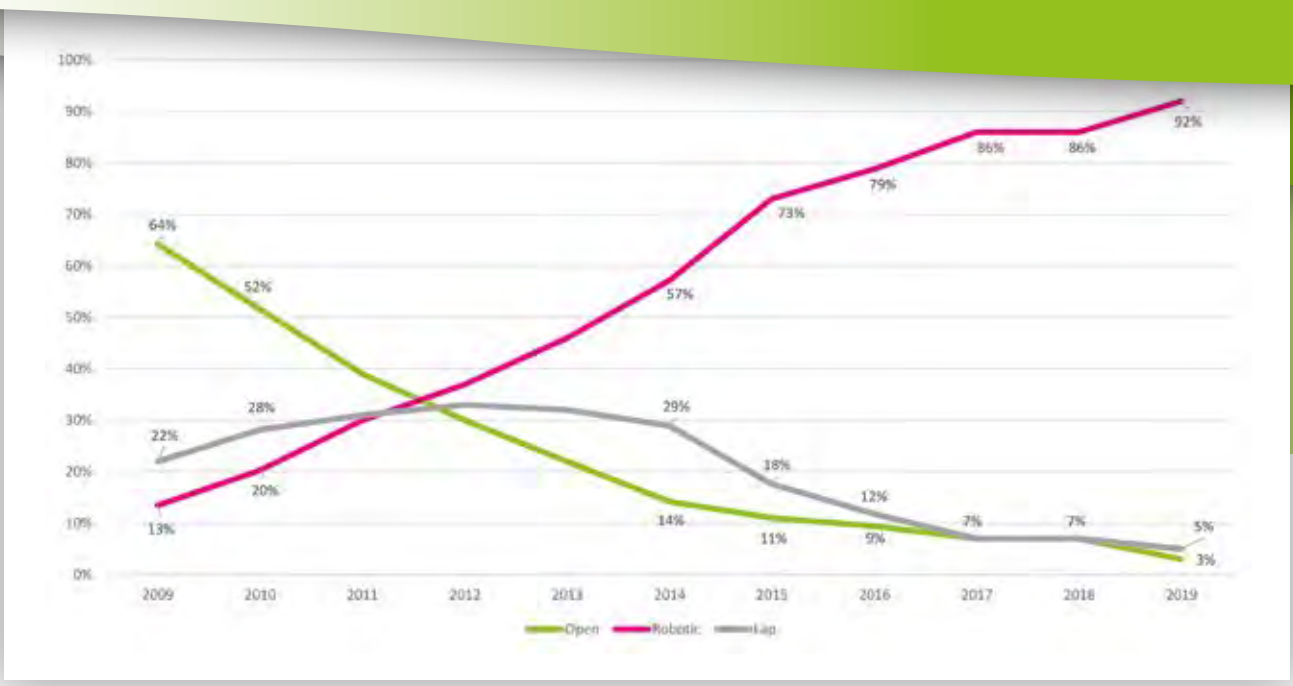


Figure 1: Hypothetical Evolution of Prostatectomies in the UK 2009 to 2019 (Source: Simmonds, 2020).

Unlike the laparoscopist, who stands hunched over the patient, the surgeon in robot-assisted procedures sits comfortably at a console and controls the endoscopic instruments without tremor via controllers or finger straps. Like the human wrist, the Endowrist® instruments of the market-leading da Vinci® system (by U.S. company Intuitive) offer seven degrees of freedom. The insufflated (i.e., filled with CO_2 gas) surgical area offers optimal room for movement. The surgical area is magnified ten times for the surgeon, in high resolution and in 3D, so that sensitive structures, such as blood vessels and nerve tracts, can be optimally protected.

RAS combines the advantages of the maximum leeway/maneuverability of open surgery with the minimally invasive approach of laparoscopic surgery. Studies have shown that oncologic results (e.g., in renal, prostate, bladder, colon, pancreatic and esophageal surgery) are at least equivalent.

The main advantage for the patient is the low complication rate (50 % less, compared to open procedures) and a faster healing time. Patients are easily convinced by the low percentage of wound healing complications due to minimally invasive approaches and the ability to get back on one's feet again. Moreover, there is a lower risk of the necessity for intensive care or reintervention. A tentative advantage of RAS is also emerging in functional outcomes, such as the preservation of continence and erectile function in radical prostatectomy. Robot-assisted radical prostatectomy is displacing not only open surgery, but also the laparoscopic alternative, as can be seen in the United Kingdom, where 92 % of all prostatectomies were done with assistance from the robotic system in 2019 (Figure 1).

This move from traditional to robot-assisted surgery is not unique to urology, as has been confirmed by a recent study on 169,000 general surgery patients in the Michigan area. While the first five years after the implementation of robot-assisted surgery were characterized by an increase – not only in robot-assisted procedures, but initially also in laparoscopic procedures over open surgery – the last two years also confirm a decrease in laparoscopic procedures in favor of robotic procedures (Figure 3).

Why conventional surgery will become obsolete

The competition introduced by robot-assisted surgery will continue, and is already affecting thoracic, maxillofacial and pediatric surgery, in addition to urology, general surgery and gynecology. RAS is a disruptive innovation that will almost completely replace conventional surgery. This process is unstoppable and inevitable. Unlike conventional procedures, RAS is a digital procedure, i.e., it offers capacity for innovation and is forward-looking. In the future, for example, preoperative imaging will be integrated into the operative image (augmented reality). Colored highlights of tumor areas and sensitive structures will guide the surgeon, and intelligent safety concepts will confine the space where the instruments are allowed to work and which structures must be given particular attention. In addition, digital recording of surgical procedures enables the implementation of artificial intelligence and expert systems. Other potential improvements can be found in the enhancement of surgical camera systems and the scope of visualization of the surgical area.

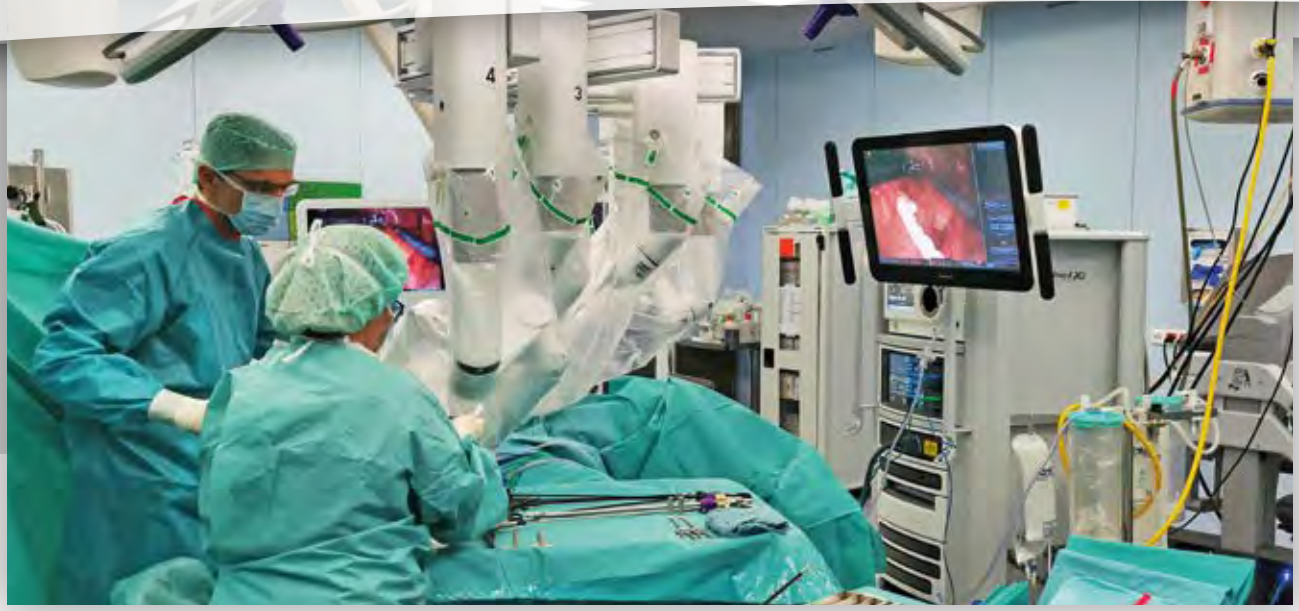


Figure 2: State of the Art – da Vinci surgery with the Xi-system (Source: F. Prell/G. Böhler).

The digitization of surgery opens up nearly unlimited optimization and expansion possibilities, right up to first initiatives in "autonomous surgery." Surgical training will also change and become more digital, moving away from animal models and cadaver studies to isolated perfused organ systems and increasingly detailed simulations. The determining factors in the transition to digital surgery will come from the outside, in particular the new disciplines of artificial intelligence resp. virtual/mixed reality, opening up new dimensions in surgical possibilities.

Why robot-assisted surgery is underrepresented in Germany

With the 150 da Vinci systems implemented in Germany to date, as many as 250 procedures are performed per system annually. This represents only 37,500 robot-assisted operations out of approximately 280,000 potentially robotic standard procedures carried out annually in urology, gynecology and general surgery, or only 13 %. The situation in terms of system distribution is similar. In the USA, the home to the chief manufacturer of robotic surgical systems, one system is used for approx. 95,000 people, while in Germany this figure is approx. 547,000. In some other European countries, robot-assisted surgery is already much more widespread than in Germany (e.g., in Sweden, with 280,000 citizens per system).

The potential represented by robot-assisted surgery is thus underexploited in Germany. The reasons for this are found first and foremost in the current funding shortfall for robot-assisted procedures due to the applicable case rates for reimbursement which cover only the costs of open surgery. In addition, the U.S. company Intuitive Surgical, manufacturer of the da Vinci® systems, as been the only established supplier of robotic surgery for a long time, while numerous competing surgical systems are on the verge of being launched. Robotic surgery is also becoming European, the leading representatives being Avateramedical (Germany), Distalmotion (Switzerland), Cambridge Medical Robotics (UK) and TransEnterix (Italy/USA). The new systems compete will fuel innovation in the field, while also occupying different niches.

Why robot-assisted surgery can help soften the side effects of the pandemic

The current COVID-19 crisis vividly illustrates the vulnerability of healthcare systems and demonstrates to what extent the economy depends on the performance of the former. The COVID-19 pandemic also has a serious and direct impact on clinical care and surgical medicine. In order to maintain sufficient capacity for COVID-19 cases, elective procedures are being postponed indefinitely, and beds and staff are being postponed withdrawn from regular interventions. These and additional bed capacities are kept as a standby buffer at great expense, regardless of whether they are actually needed or not. At the same time, the pandemic has significantly increased the risk of infection for clinical staff and caused higher demands for surgery as a result of backlogged procedures and treatments for post-coronavirus effects on the lung and heart.

Coping with this and future pandemics will require increased efficiency in clinical medicine while regular operations and bed capacities are largely maintained. By dramatically shortening the length of stay and rate of complications, RAS can provide new impulses, as inpatient beds will be used for shorter periods and intensive care beds will be used less frequently – accompanied by the respective benefits of staff reduction and faster recovery for patients. At the same time, robot-assisted surgery offers all the advantages of contact-free operation via consoles, providing literally a "non-touch" surgery approach.

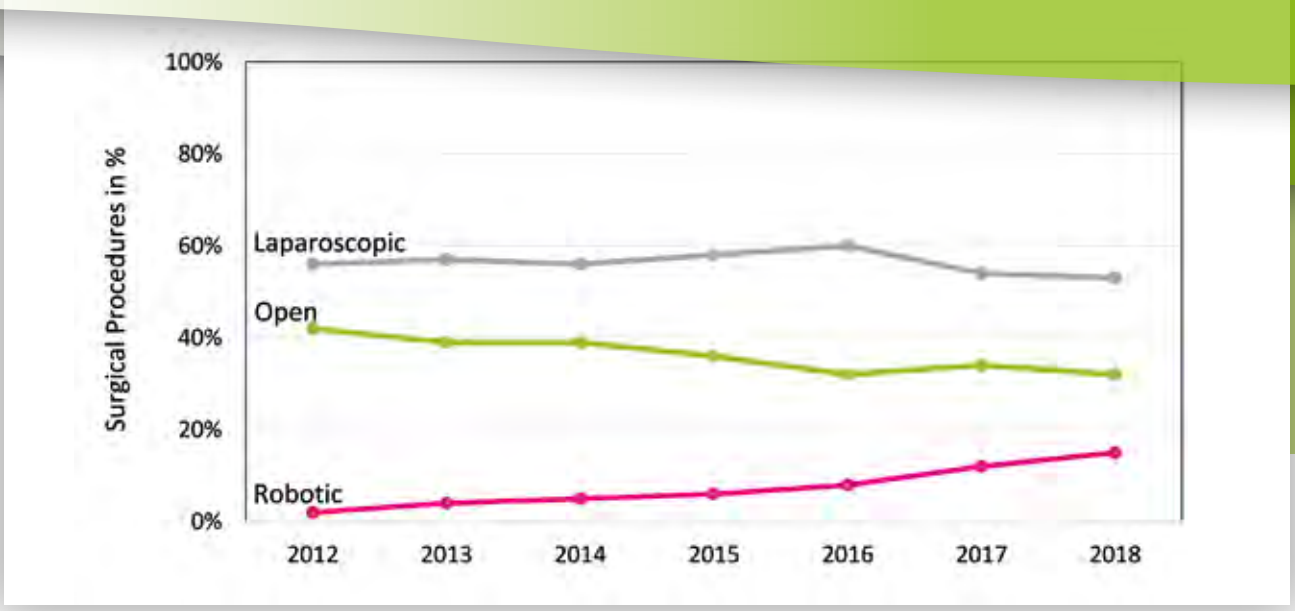


Figure 3: Temporal Trends in the Proportional Use of Robotic, Laparoscopic and Open Surgery (Source: Sheetz *et al.*, 2020).

In addition, the minimally invasive approach reduces the spread of potentially infectious aerosols from the patient's body. Consequently, robotic surgery is not only the key to a sustainable increase in healthcare efficiency, it also reduces the risk of infection in the operating room. In this way, the COVID-19 crisis may prove to be a catalyst for faster adoption of robot-assisted procedures.

Why robot-assisted surgery represents an opportunity for Germany

In the future, Germany can no longer blindly rely on its international reputation for high quality cars. The coronavirus crisis has gained Germany novel recognition and admiration in sophisticated medical technology. In order to assert itself as an innovation hub in high-end medical technologies, Germany must assume a leading role in robotics, artificial intelligence and digital innovations.

Robot-assisted surgery is a key technology and will also expand to diagnostic procedures, i. e. robot-guided biopsies. Without this technology, innovative procedures from the field of AI and augmented reality cannot unfold their benefits in surgery and diagnostics. There is a nearly infinite scope for making surgical therapies – in cancer, for example – less invasive and more precise with the help of digital technologies. We are only just at the beginning. In view of the coronavirus experience, high-end medical technologies will boost in prestige and have the potential to become the new flagship of the German and European economy. A keen political strategy is needed to pave the way.

References:

Stolzenburg, J.U., Kyriazis, I., Fahlenbrach, C., Gilfrich, C., Günster, C., Jeschke, E., Popken, G., Weißbach, L., von Zastrow, C., Leicht, H. (2016). National trends and differences in morbidity among surgical approaches for radical prostatectomy in Germany. World J Urol 34(11), 1515-1520. Epub 2016 Mar 24.

Cheng, C.L., and Rezac, C. (2018). The role of robotics in colorectal surgery. BMJ; 360: j5304. doi: 10.1136/bmj.j530.

Simmonds, Ch., BS, PGCE (2020): The Growth of Robotic Surgery. mimic Newsletter, Feb. 7, 2020

Sheetz K.H., Claflin J., Dimick, J.B. (2020). Trends in the Adoption of Robotic Surgery for Common Surgical Procedures. JAMA Netw Open. 2020 Jan3;3(1):e1918911.doi:10.1001/ jamanetworkopen.2019.18911.

Statistisches Bundesamt, Genesis-Online Datenbank, https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Krankenhaeuser/Publikationen/Downloads-Krankenhaeuser/operationen-prozeduren-5231401187015.html; Datenabfrage vom 16.04.2020 (Zahlen zu den häufigsten operativen Eingriffen in der Weichteilchirurgie)

Contact:

Almut Kalz
Project manager, Kurt-Semm-Center (Intersdisciplinary center for robot-assisted surgery) and Kurt-Semm-Academy (Surgical Training center)
University Hospital Schleswig-Holstein, Campus Kiel
almut.kalz@uksh.de
www.kurt-semm-zentrum.de

Prof. Klaus-Peter Jünemann, MD
Professor and Speaker, Kurt-Semm-Center (Intersdisciplinary center for robot-assisted surgery)
University Hospital Schleswig-Holstein, Campus Kiel

Federal Ministry
of Education
and Research

News from the BMBF

The dangerous dual role of the immune system

Our body's defence mechanisms are our most powerful weapon against viral pathogens, but the immune system can also pose a serious threat if it turns against healthy cells. A team of researchers in Berlin is using highly complex bioinformatic analyses to investigate the role of the immune system. The majority of people infected with SARS CoV 2 develop only mild symptoms similar to those of a common cold and usually recover within two weeks. However, a small minority of about ten to 15 percent of people become severely ill and develop life-threatening complications. The research team around mathematician Professor Roland Eils from the Berlin Institute of Health wants to find out what distinguishes these two groups of patients and what role our body's immune system plays in this context. The Federal Research Ministry is supporting their work through funding for de.NBI, the German Network for Bioinformatics Infrastructure.

"We had an inkling early on that the immune system of those infected has a decisive influence on the course of the disease," says Professor Eils. His research team used single-cell analysis to investigate their hunch.

Just like no two humans are the same, each cell responds differently to an infection. The scientists are looking into the cells' reaction by analysing how individual genes are regulated. Cells use gene regulation to respond to external factors. For example, cells can increase transcription of individual genes and consequently develop more receptors for interaction with other cells.

Single-cell analysis is an extremely complex process. At first, the researchers have to isolate thousands of cells from samples taken from the nose and throat of patients. Then they determine the activity of thousands of genes for each individual cell. This results in unfathomable amounts of data that researchers can only analyse with the help of extremely high-performing computers and specially designed bioinformatics tools.

Harmful alliance with the virus

We now know that SARS CoV 2 uses a receptor known as ACE2 to enter the cells of the mucous membrane of the nasopharynx and cause their infection. "Our research has shown that this receptor only occurs in a very small amount of cells altogether," says Professor Eils. This means that the viruses can initially infect only very few cells.

The researchers wanted to understand why the novel coronavirus is nevertheless so efficient at causing COVID 19. They took a closer look at the progression of the disease.

When the virus enters the body, infected cells alert the immune system, causing immune cells to release a specific factor called interferon gamma. This factor increases the immune response but also causes tissue cells to produce more of the ACE2 receptor. "The immune system thereby forms a harmful alliance with the virus," summarizes Eils. ACE2 levels in sick patients are about two to three-fold higher than in healthy persons. This increases the number of recep-

When the virus enters the body, infected cells alert the immune system.

Source: Adobe Stock / jijomathai

tors that the virus can use to enter cells in the body, which means that the virus can infect considerably more cells as the disease progresses.

Destruction of healthy lung cells

The immune system can play another similarly dangerous role. Researchers observed a hyper-inflammatory response in severely ill patients. “In these cases, immune cells kill both healthy lung cells and infected cells,” explains Professor Eils. “Patients with this kind of excessive immune response suffer from massive respiratory problems and usually require mechanical ventilation.” So far, the researchers cannot yet say what causes the immune system of some people to respond so dramatically and exacerbate the disease instead of fight it. “We need even more data to be able to understand this,” says Eils.

However, the researchers’ improved understanding of the hyperimmune response in severe cases of the disease is already offering some interesting therapeutic approaches.

“This is still a hypothesis but it would make sense to think that we could disrupt the interaction between tissue and immune cells in a targeted way,” believes Eils. To achieve this goal, his research team is proposing the use of drugs that are already in the process of obtaining a marketing authorization or have been approved to treat other viral diseases.

The team ensured that the publication of their research results and the underlying data are free of charge for the global community. They hope that clinical partners also in other countries will use their research findings to provide more efficient treatment for patients with severe cases of the disease.

Further information is available at:

bmbf.de/de/gefaehrliche-doppelrolle-des-immunsystems-12604.html

Smart use of data often thwarted by variety of data formats.

Source: Adobe Stock / ipopba

Data that helps to cure the sick

Three Federal Ministries that share a common goal: in September 2020, the Federal Ministries of Research, Health and Economic Affairs presented a joint roadmap for an innovation initiative under the motto “Data for Health” that will lay the foundation for the digital health care of the future.

Every medical examination and treatment generates medical data. This data plays an extremely important role in health research and holds the key to new findings about the causes of disease. At the same time, data provides an important basis for better health care, faster and more precise diagnosis and individualized patient treatment.

Creating the conditions for the smart use of data

Today, the use of many different data formats and standards often makes smart data use impossible. For example, medical reports are often written in free text while laboratory results come in the form of tables using different units of measurement. Even seemingly straightforward blood pressure readings are sometimes documented in different ways within the same hospital. Not even the smartest of software can make sense of such inconsistent data. This is why common rules and agreed international standards should be used in future to document digital health data.

Driving the harmonization and networking of data is a central objective of the innovation initiative "Data for Health: Roadmap for Improved Patient Care Through Health Research and Digital Technology". The initiative was jointly launched by the Federal Ministry of Education and Research, the Federal Ministry of Health and the Federal Ministry for Economic Affairs and Energy. The three ministries will formulate a joint strategy in cooperation with stakeholders from the sectors of health care, health research and the healthcare industry in order to achieve progress in this important area for the future.

Five priority areas of action in brief

1. Establishing and expanding structures for the digital networking of health care and health research
2. Enhancing the availability and quality of health data
3. Driving the development of innovative solutions for enhancing data security and data linkage
4. Embarking on the path to data-based medicine together
5. Developing prospects for future application at an early stage

Patient willingness to make their data available to research is another central prerequisite for the success of the initiative. This willingness depends on the people's confidence in the researchers and processes that play a role in collecting, processing and analysing their data.

The issue of data sovereignty is another key aspect. Who decides what data will be used for what purposes? There can only be one answer to this question: our citizens themselves have sovereignty over their data.

They decide whether and in what form their data can be used for research purposes now and in the future. A standardized sample text for obtaining patient consent has been developed and agreed by all German university hospitals in cooperation with the Federal and Länder data protection supervisory authorities. Another objective is to develop comprehensive information material, including on aspects of data privacy and data security, in order to gain public acceptance as well as patient trust.

Raising Germany's profile as a centre for health

In its High-Tech Strategy 2025, the Federal Government declares better links between research and health care one of twelve cross cutting missions. The goal is for Germany to become a global leader in the development and application of digital health innovations.

The "Data for Health" initiative of the three ministries involved is an important element of this mission as it is aimed at facilitating the development of novel technologies and areas of application at an early stage.

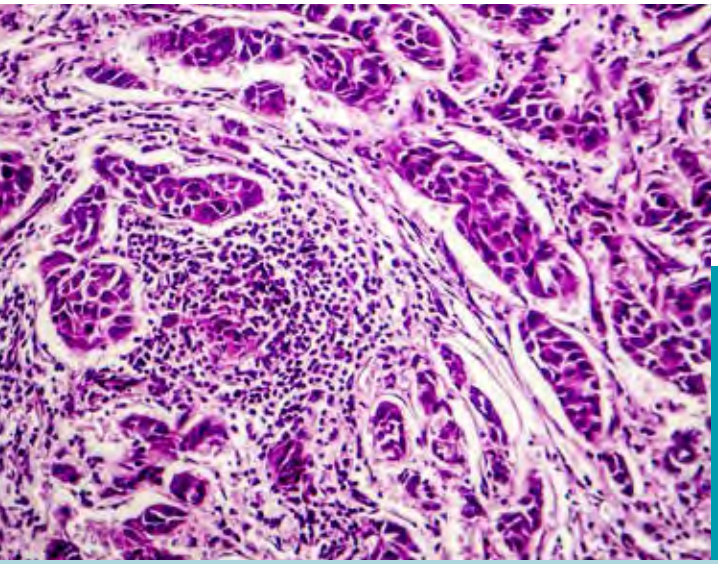
Further information is available at:

bmbf.de/de/daten-helfen-heilen-12503.html

Artificial intelligence to clean up molecular mess

Every tumour contains information that is crucial for the decision on the right kind of treatment. So far, however, it has been difficult to obtain this information. A new artificial intelligence (AI) tool is now being developed to solve this problem and advance cancer research by ten years.

Researchers refer to the material which provides them with all the important information as "tumour mess". The term "mess" is not meant in a negative sense here, but rather to illustrate how relevant information is currently mashed up in unmanageable amounts of data. Rainer Spang, professor of bioinformatics at the University of Regensburg, and his research team use AI to try and sort out as well as process this information. Should they succeed, their work would push cancer



Every tumour contains information that is crucial for the decision on the right kind of treatment.

Source: Adobe Stock/ Kateryna_Kon

research forward enormously. The Federal Ministry of Education and Research supports their research under the umbrella of the CompLS funding measure for computational life sciences.

Cancer researchers have been using measurements to determine what genes are active at what level of intensity in a specific tumour for 20 years. This helps clinicians to decide on the best choice of treatment. Every tumour is different and therefore only responds to certain specific drugs. Molecules inside tumour cells have to be broken apart in order to become accessible. This results in a chaotic molecular mix known to the research team at the University of Regensburg as a tumour mess. "However, since tumours do not only consist of tumour cells, this breaking apart of their structure results in a mixture of tumour cells as well as tissue and immune cells," explains Professor Spang. "This makes it difficult to assign the molecular information that we gain to the right cells."

Using AI to enhance treatment prediction

Tissue can, for example, contain information on the upregulation of genes responsible for cell division. The question here is where this signal comes from. If it comes from the tumour cells, the tumour will grow. But if it comes from the immune cells, this is a sign that the immune system is already fighting the tumour. For the doctors attending their patients, there is a huge difference between the two. Depending on the actual scenario, they will have to choose between completely different treatment plans. This is where the work of the research team around Professor Spang comes into the picture. They are developing an AI tool that assigns measurement data to individual cell types and therefore has the potential to significantly enhance treatment prediction.

While procedures which enable measurements for individual tumour cells have now been developed, practical experience with these procedures is still lacking. "We will have to wait and see how the disease develops in the cases that we have investigated," says Spang. This means that it is not yet possible to derive recommendations for treatment based on individual measurement data. By contrast, comparative data for the analysis of the tumour mess has been collated for decades, allowing researchers to draw conclusions about certain groups of patients. "It will take thousands of patients to obtain reliable data," concludes Spang. "And we'll have to wait and see how the disease develops over the course of the years." Professor Spang estimates that it could take up to 15 years before data obtained from single-cell analysis is ready to be harnessed.

"Push developments forward by at least ten years"

The bioinformaticians from Regensburg are planning to use their tool to obtain information on individual cells from measurements of the tumour mess. Although the relevant signals are indeed contained in the mishmash of tissue cells, they are also very much hidden. It is up to AI to filter these signals from the molecular jumble. "Achieving this goal could push the development of new methods forward by at least ten years to the benefit of cancer patients."

The new data which has been sorted by the research team could also provide previously unknown starting points for developing new treatments and drugs. Professor Spang cites the example of personalized immunotherapies that are being used more and more frequently in the fight against cancer. The practical application of this new approach could look like this: the AI tool establishes that specific immune cells have detected a tumour but are not attacking it because of a blockade. The attending doctor could now prescribe immunotherapy which would remove the blockade and trigger an immune response in the patient.

Further information is available at:

bmbf.de/de/kuenstliche-intelligenz-bringt-ordnung-ins-molekulare-chaos-9796.html

BMBF-Newsletter:

A monthly overview of important news from the BMBF
bmbf.de/newsletter

Stay up to date with the BMBF on Twitter, Facebook and Instagram.

twitter.com/BMBF_Bund

facebook.com/bmbf.de

instagram.com/bmbf.bund

bioinformatics: young women need role models

Interview with bioinformatician Janine Felden, PANGAEA group leader at the Alfred Wegener Institute in Bremerhaven

In the German Network for Bioinformatics Infrastructure, female group leaders are a distinct minority. Janine Felden is one of them. In this interview with gesundhyte.de, she talks about how she got into bioinformatics after starting out in another career, and what needs to change in order to interest more women in the field – one that fascinates and intrigues Felden to this day.

gesundhyte.de: *Unlike in biology, women are still the minority in bioinformatics. What is deterring them?*

Dr. Janine Felden: Sadly, when some people think of bioinformatics they still picture people sitting alone at their computers in darkened basements. That's an image that discourages young women in particular, as they usually want social interaction. If a woman has a fundamental interest in biology, she tends to opt for more hands-on work in the lab. But much of bioinformatics actually involves working with other people in a team. The subject lives off vibrant exchange with many different committed and communicative individuals. Bioinformaticians are anything but asocial recluses. Thankfully, during their biology studies some women do develop an interest in bioinformatics. Sometimes the question is just how they gain access to it. Many women get into it after embarking on another career, like I did.

How did you get into bioinformatics?

On a very circuitous route. First, I studied biology because I was fascinated by marine research – particularly microbial life in the sea. Ultimately, all life on our planet depends on microorganisms, as the Covid-19 pandemic has shown all too clearly. Unlike plants and animals, microorganisms cannot be distinguished by their external appearance alone. Only when their DNA is analyzed using bioinformatics tools can microorganisms and their characteristics be properly identified. So that's how I first came into contact with bioinformatics. I wanted to know what all those creatures in the sea were. And what they were up to. To answer such questions, we need bioinformatics. It provides the right tools and forms the basis of many different laboratory methods.

> "I am particularly fascinated by the wide range of questions that can be answered using bioinformatics tools."

What does your daily working life involve?

Contrary to the cliché, my work day involves plenty of interaction with real, live human beings. Where I work, most of the research groups are mixed gender. Men and women do tend to take a different approach, but at the end of the day the results are of the same quality. And it's actually beneficial to have different perspectives on things. My research group is primarily concerned with preparing and publishing scientific data for the PANGAEA database so that anyone who needs to can find, use and share those data. But our daily work is more than just organizing and integrating data. As a service platform within the German Network for Bioinformatics Infrastructure, we are in constant contact with our users. We steer and guide scientific projects from the initial idea to final publication, helping researchers with their data management and allowing DNA and environmental data to be linked.

Dr. Janine Felden (Photo: Felden).

What is it that fascinates you about bioinformatics?

I am particularly fascinated by the wide range of questions that can be answered using bioinformatics tools. It is bioinformatics that has shown just how many life forms exist all around us. With just a single sample of sea water we can answer several questions at once, such as "Who is there? What are they doing?" I've always been intrigued by the larger connections within and between ecosystems and I'm driven by the desire to understand them. But there are also exciting topics in medicine. These days, some cancer drugs are tailored to individual patients based on their DNA sequences, thus improving their survival chances. Without bioinformatics that wouldn't be possible. I think that's incredibly fascinating. It's also good to see that bioinformatics is now also driving other fields of science forward. One good example of that is the processing of vast amounts of data – known as "big data." Along with climate research, bioinformatics is leading this field and is now functioning as a driver for other disciplines.

How do you assess the future of bioinformatics?

In our information society, bioinformatics is a subject area that offers good prospects well into the future. This isn't just true of academic research, but also of private enterprise, where many exciting tasks await. Bioinformaticians are not overspecialized and can apply what they have learned to other fields of study. Handling exponentially increasing amounts of data and making them usable will require a good supply of well qualified bioinformaticians – whatever their gender.

What needs to change in order to encourage more women to work in bioinformatics?

It seems that we need to get rid of a few preconceived ideas. This is where it's important to have role models. Young women should be able to see that there certainly are women working in the field and that they can be very successful. It's also important for university programs to be more flexible. Women who start out working in the lab should have the opportunity to switch their focus to bioinformatics when they notice that laboratory work is not the only way to explore fascinating questions. And it can actually be easier to balance professional and family life in the digital world of bioinformatics, as it's possible to allocate your time more flexibly than when you work in a lab.

"Handling exponentially increasing amounts of data and making them usable will require a good supply of well qualified bioinformaticians – whatever their gender."

Some universities are offering special IT courses for women in order to get more of them interested in studying technology subjects. What do you think of that?

In general, I'm a fan of ensuring a good balance between men and women. A purely female environment simply doesn't re-

(Photo: iStock © metamorworks)

flect reality. So why should we have that during our studies? However, I do see that it is important to increase the attractiveness of bioinformatics for women and to remove the obstacles in their way. Growing up, girls often have different interests than boys. So it would probably be good to have initial learning options adapted to existing knowledge levels. It might therefore be useful to offer introductory courses for women. But that shouldn't turn into some kind of stigma. And why shouldn't young men who also lack prior knowledge not be able to benefit from such courses? Although young people today certainly have a greater affinity for technology than they did 20 years ago, it's simply not the case that all teenagers spends years programming their own little robots. For those who didn't, a low-threshold introductory course could help to arouse interest. If such things had been around in my day, maybe I would have actually studied bioinformatics.

Melanie Bergs and Gesa Terstiege conducted the interview.

Contact:

Dr. Janine Felden

PANGAEA Group Leader at the Computing Center

Alfred Wegener Institute – Helmholtz Center for Polar- and Marine Research, Bremerhaven

janine.felden@awi.de

www.awi.de/en/about-us/organisation/staff/janine-felden.html

PANGAEA database

PANGAEA – Data Publisher for Earth & Environmental Science – is a digital library system that focuses on data from earth system research and environmental science. It is jointly operated by MARUM – Center for Marine Environmental Sciences at the University of Bremen and the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research. The professional data management is provided as a service platform within the German Network for Bioinformatics Infrastructure (de.NBI), which the BMBF is supporting to the tune of €81 million through the end of 2021.

powering data analysis for COVID-19 research

The German Network for Bioinformatics Infrastructure offers an impressive range of bioinformatics tools and cloud computing resources

by Alfred Pühler, Irena Maus, Vera Ortseifen, and Andreas Tauch

For over a year now, the COVID-19 pandemic has kept the world in its grip. By collecting and evaluating information about the SARS-CoV-2, life science researchers are working to gain a better understanding of COVID-19 and develop vaccines and drugs to combat the pandemic. Extensive studies are yielding large amounts of research data that need to be analyzed in order to provide insights. The German Network for Bioinformatics Infrastructure (de.NBI) provides the expertise, computing resources and bioinformatics tools needed to perform such analyses. This article illustrates the network's considerable potential in the fight against COVID-19 and outlines the contributions of de.NBI members in German and European COVID-19 projects.

de.NBI: A network for analyzing life science research data

The German Network for Bioinformatics Infrastructure (www.denbi.de) was established in 2015 by the German Federal Ministry of Education and Research (BMBF) to provide life science researchers with an appropriate infrastructure for analyzing large amounts of data. Currently, the network consists of over 300 member scientists and 40 projects located across eight service centers. These service centers are thematically oriented and complement each other to cover various subdisciplines of life sciences, including services and expertise needed for data-based medicine.

The de.NBI infrastructure encompasses the areas of services, training and cloud computing (Figure 1). In the services area, a wide range of bioinformatics tools and databases are available to evaluate life science data. Closely linked to the services area is the training area, which educates the users of de.NBI services on how to employ the services and handle research data. The services area includes over 100 bioinformatics tools and workflows. Offering approximately 70 training courses per year, de.NBI has trained roughly 6,000 users to date.

de.NBI also plays an important role in the computing sector as an operator of its own de.NBI Cloud. Located at six academic cloud sites in Berlin, Bielefeld, Freiburg, Gießen, Heidelberg and Tübingen, the de.NBI Cloud has received extensive funding from the BMBF, thus allowing it to achieve considerable scale and perform analyses on large data volumes typical for the life sciences. The de.NBI Cloud is available free of charge to all life science researchers (see https://www.denbi.de/cloud). You can also learn more about the de.NBI Cloud in the next article on page 80.

Altogether, de.NBI provides an essential bioinformatics infrastructure in Germany. It also works closely with ELIXIR Europe, an intergovernmental organization that aims to create a single bioinformatics infrastructure across Europe, to which de.NBI is linked through the German ELIXIR node.

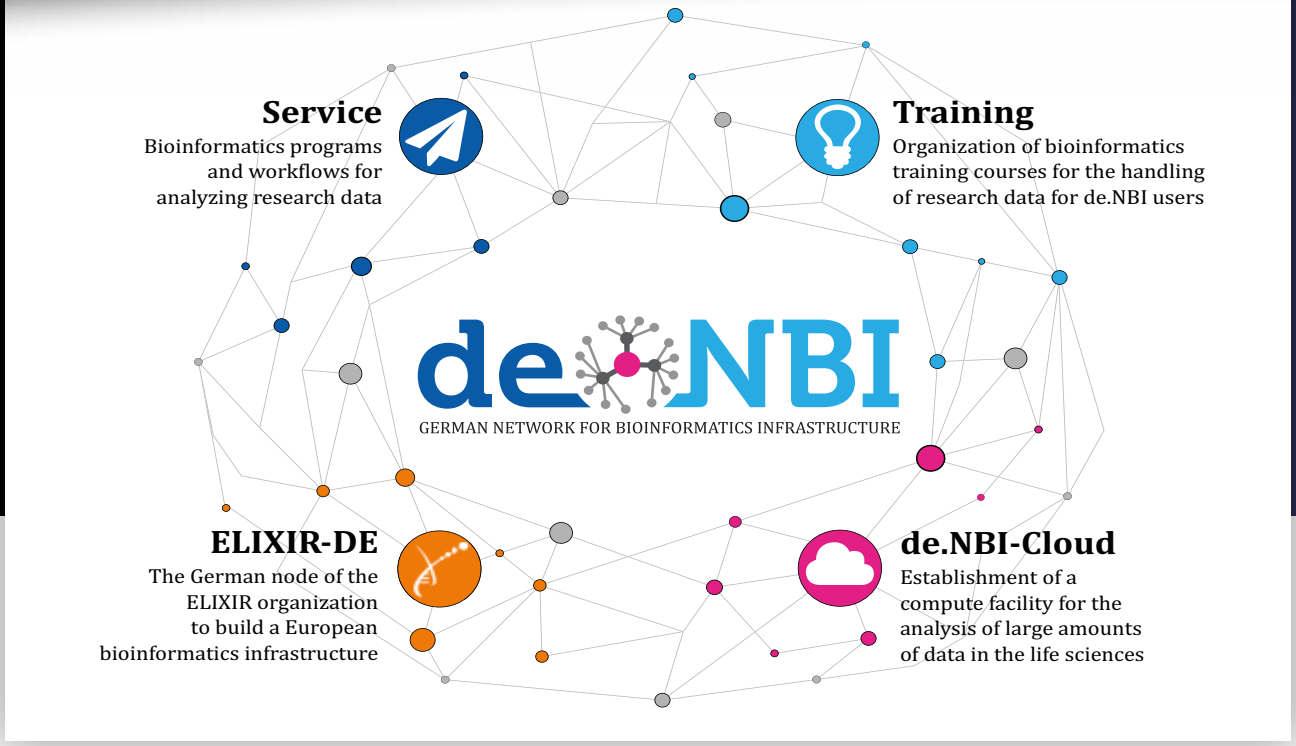


Figure 1: The main activities of the de.NBI network includes service and training as well as offering a cloud-based compute facility. Besides, the de.NBI network and its activities are also integrated into the European ELIXIR infrastructure (Source: de.NBI Administration Office).

Analyzing COVID-19 research data – an acid test for de.NBI

The current coronavirus pandemic is a major challenge for our society and therefore requires special attention from established scientific structures. It is important to advance molecular biological research on coronaviruses, especially SARS-CoV-2, and to investigate both the medical and epidemiological aspects of the infection process as well as COVID-19 disease progression. Efforts must also be made to develop drugs to treat patients with COVID-19 and to produce appropriate vaccines. All these research areas generate large amounts of data that only reveal their value after careful bioinformatics analysis. Big data analysis is one of de.NBI's greatest strength, and so the network is proving to be a real advantage for Germany in the coronavirus pandemic. The available de.NBI resources have been effectively and timely applied to the analysis of COVID-19 research data. de.NBI has thus passed the acid test of being able to meet current data analysis needs at short notice. To date, a total of 29 COVID-19 research projects with de.NBI involvement are existing and presented on the de.NBI website (https://www.denbi.de/covid-19).

de.NBI contributions to COVID-19 research projects

de.NBI supports COVID-19 research projects across six different categories (Figure 2). In total, there are 29 COVID-19-related research projects initiated by a wide range of research institutions in the de.NBI network (Table 1).

The category **"Sequence variants and geographical distribution of the SARS-CoV-2"** includes three projects focusing on sequence establishment and analysis of viral genomes. Researchers at the European Molecular Biology Laboratory (EMBL) in Heidelberg are using the genome analysis server GEAR to analyze SARS-CoV-2 chromatograms and mutations, while their colleagues at Charité – Universitätsmedizin Berlin and Heidelberg University have developed the online tool MapMyCorona. In Bielefeld, life scientists are decoding SARS-CoV-2 genomes by nanopore sequencing and subsequent analysis using de.NBI Cloud compute power.

The category **"Interactions of SARS-CoV-2 with human cells"** covers projects that focus on the interaction of the virus with human tissue cells or with the immune system in particular. The researchers' preferred method is single cell analysis of human cells. One joint project of the Heidelberg and Tübingen partners aims to analyze the relationships

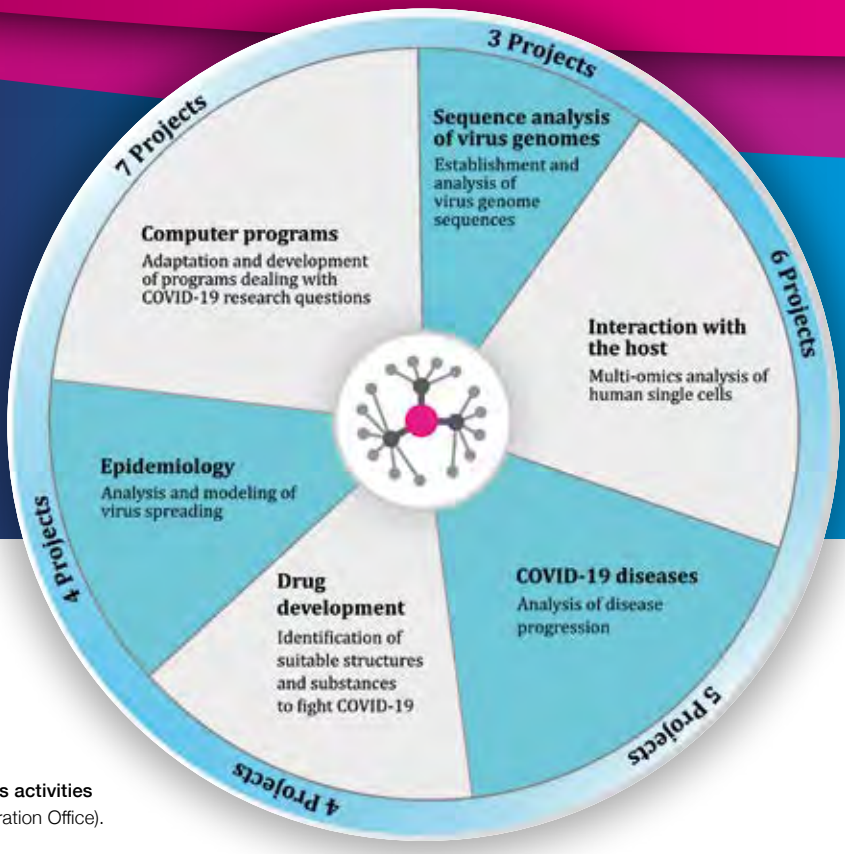


Figure 2: Six research rubrics that characterize de.NBI's activities in COVID-19 research projects (Source: de.NBI Administration Office).

between SARS-CoV-2 genomics and different courses of infection using the de.NBI Cloud. The category also includes three projects in Giessen studying lung diseases, in particular SARS-CoV-2 infections. The COVID-19 Omics Explorer Magellan, developed at Charité, provides tools for visualizing current COVID-19 research data to support more advanced analyses. Another project from Berlin focuses on cell response to infected tissue and analyzes the acquired omics data with various de.NBI applications. Moreover, in Berlin, researchers are analyzing the transcriptional events in single cells after infection of *in vitro* cell tissue with SARS-CoV-2.

In the category **"Analysis of the COVID-19 disease,"** three projects from Bochum are collecting information on COVID-19 disease progression and analyzing it applying special bioinformatics methods, including artificial intelligence. Researchers in Rostock are working on the development of disease maps and using text mining to facilitate drug selection. In Kiel, genome-wide association studies are being conducted to search for abnormalities in the genomes of COVID-19 sufferers.

Of particular relevance is the category **"Development of drugs for SARS-CoV-2 infections."** Researchers at the German Cancer Research Center (DKFZ) in Heidelberg are searching for substances with antiviral properties by testing whether they inhibit the transcription of the viral genome. At Heidelberg University, studies are being conducted on small interfering RNA fragments (siRNA) to discover whether they can prevent viral replication. Both research groups are using high-throughput microscopy techniques for their analyses. Researchers at the University of Freiburg are applying *in silico* methods to search for potential drugs. *In silico* methods are also being used by de.NBI colleagues in Hamburg in their search for substances capable of attaching to SARS-CoV-2 proteins.

The category **"Support for classical epidemiological studies"** plays a significant role in understanding virus spread. The projects in this category focus on modeling SARS-CoV-2 spread trajectories, with de.NBI supporting projects in Jena, Rostock and Heidelberg. The Jena project aims to model and predict the course of the SARS-CoV-2 pandemic. Also researchers in Rostock are using mathematical modeling to predict the pandemic's course. The de.NBI members in Heidelberg are employing the COPASI software to model SARS-CoV-2 infections.

Table 1: The 29 COVID-19 research projects within the de.NBI network

Rubric	Number of projects	Research Institute
Sequence variants and geographical distribution of the SARS-CoV-2	3	• The European Molecular Biology Laborarory (EMBL), Heidelberg • BIH at Charité, Berlin and Heidelberg University • Bielefeld University
Interactions of SARS-CoV-2 with human cells	6	• Justus Liebig University Gießen • BIH at Charité, Berlin and Heidelberg University • Heidelberg University • Eberhard Karls Universität Tübingen • Max Delbrück Centrum for Molekular Medicine, Berlin
Analysis of the COVID-19 disease	5	• Ruhr-University Bochum • Rostock University • Christian-Albrechts-Universität zu Kiel
Development of drugs for SARS-CoV-2 infections	4	• German Cancer Research Center (DKFZ), Heidelberg • Heidelberg University • University of Freiburg • Hamburg University
Support for classical epidemiological studies	4	• Fritz Lipmann Institute, Jena • Rostock University • Heidelberg University
Tool development to analyze COVID-19-related data	7	• Ruhr-University Bochum • University of Freiburg • Heidelberg Institute for Theoretical Studies • Bielefeld University • Eberhard Karls Universität Tübingen

"Tool development to analyze COVID-19-related data," is the last of the six research categories, comprises projects dedicated to the development of analytics software for COVID-19-related data, including software tools that support the operation of the de.NBI Cloud and the Galaxy platform. The Freiburg-based Galaxy platform in particular facilitates fast and uncomplicated SARS-CoV-2 data analyses for research question on coronaviruses (galaxyproject.eu). Most widely used, however, is the de.NBI Cloud, as it is able to perform extremely effective and rapid analyses of COVID-19 data as it accumulates.

de.NBI participation in European COVID-19 initiatives

On the international level, de.NBI is involved in European COVID-19 initiatives with international partners *via* the ELIXIR Germany node. Of particular note is the European COVID-19 Data Platform, which is part of an initiative of the European Bioinformatics Institute (EMBL-EBI), ELIXIR and the European Commission. A major part of this platform is the COVID-19 Data Portal (www.covid19dataportal.org) launched by EMBL-EBI on April 20, 2020. This new infrastructure aims to combine and continuously update relevant COVID-19 data sets and analytical tools.

According to the mission "Accelerating research through data sharing", the COVID-19 Data Portal offers researchers from all over the world a fast, open and sustainable way to share and analyze their data and provides insights on the novel coronavirus SARS-CoV-2. The portal gives access to a wide range of data beyond genomic sequences, including structural, expression, clinical and epidemiological data as well as an extensive collection of scientific literature.

National partners including the Technical University of Denmark, Hungary's Eötvös Loránd University and Germany's Heidelberg University Hospital coordinate closely with EMBL-EBI to provide the platform's infrastructure. Furthermore, de.NBI members are contributing significantly to the development and maintenance of the European COVID-19 Data Portal in cooperation with ELIXIR Europe.

Contact:

Prof. Dr. Alfred Pühler
Coordinator German Network for Bioinformatics Infrastructure
Center for Biotechnology (CeBiTec)
Bielefeld University
Bielefeld
puehler@cebitec.uni-bielefeld.de

www.denbi.de

Dr. Irena Maus
Communication and Outreach Officer
de.NBI Administration Office
irena.maus@cebitec.uni-bielefeld.de

Dr. Vera Ortseifen
Node Coordinator ELIXIR Germany
ELIXIR Germany Administration Office
vera@cebitec.uni-bielefeld.de

apl. Prof. Dr. Andreas Tauch
Head of Node of ELIXIR Germany
ELIXIR Germany Administration Office
tauch@cebitec.uni-bielefeld.de

Photos: Pühler, Maus, Ortseifen, Tauch

the de.NBI Cloud as an academic solution for life scientists

Cloud computing provides flexible, scalable computational and storage resources for big data applications

by Christian Lawerenz and Alexander Sczyrba

Advances in modern technology are presenting ever greater challenges in the life sciences. Sequencing and imaging techniques, for example, generate enormous amounts of data – which are growing at an increasingly rapid pace with each new generation of laboratory equipment. While today even small research labs can generate big data relatively easily, evaluating that data often leads to bottlenecks due to locally limited computing resources or the complex administration required for running large computer systems. Cloud computing offers a new approach: highly scalable software solutions that run on a high-performance infrastructure and can be leased as needed.

One cloud fits all

The Federal Ministry of Education and Research (BMBF) has launched the German Network for Bioinformatics Infrastructure (de.NBI) to solve the "big data problem" in life sciences. To give the life sciences access to computing and storage capacities as well as important de.NBI services, a cloud infrastructure was established.

The de.NBI Cloud is an academic, non-commercial cloud solution that aims to provide life science researchers in Germany with computational and storage resources free of charge. By supplying the relevant methods and reference data for the respective research area, de.NBI enables researchers to properly analyze their own data sets. The de.NBI Cloud infrastructure is tailored to the particular requirements of a wide range of analyses in the life sciences – for example with GPU clusters for use in the field of machine learning or image processing, or special high-memory computers for particularly memory-intensive projects. For the scientist, the location at which the analyses are actually carried out is no longer relevant. The use of cloud computing environments can therefore result in significant cost savings and eliminate the need to invest in local computer hardware – which is often only used sporadically by individual researchers.

In addition, the cloud approach allows analytical tools to be easily reused. Thanks to the virtualization methods employed, the software can often be rolled out to other systems with minimal effort. Here, the de.NBI Cloud staff offer their professional expertise – from the integration of smaller projects to help promote young researchers up to large-scale international projects with the participation of German partners. These projects include the ELIXIR network, an intergovernmental organization that brings together life science resources from all over Europe, and the European Open Science Cloud (EOSC).

de.NBI Cloud as a federation of German cloud sites

The de.NBI Cloud is a federation consisting of cloud sites at the Universities of Bielefeld, Freiburg, Giessen, Heidelberg and Tübingen. Further installations are located at the German Cancer Research Center (DKFZ) and the Berlin Institute of Health (BIH) at Charité, and another installation is being planned for the de.NBI Cloud's associated partner at the European Molecular Biology Laboratory (EMBL) in Heidelberg. From the very beginning, a cloud setup was developed that brings together all locations on a single cloud platform. All cloud sites are integrated into ELIXIR's authentication and authorization infrastructure (AAI) system, providing cloud users with easy access to the de.NBI Cloud via the web-based de.NBI Cloud portal. This way, users can access the portal and all de.NBI Cloud services via their own university user account. The de.NBI Cloud portal (https://cloud.denbi.de) is the central access point for de.NBI Cloud users. Via the portal, users can register for the cloud, apply for new and manage existing projects.

All de.NBI Cloud sites use the OpenStack software architecture for cloud computing. This controls large pools of computational, storage and networking resources throughout a data center, which is managed via a dashboard or the OpenStack API. OpenStack works with popular enterprise and open source technologies. This makes it ideal for the heterogeneous infrastructure of cloud locations and a critical building block for ensuring interoperability.

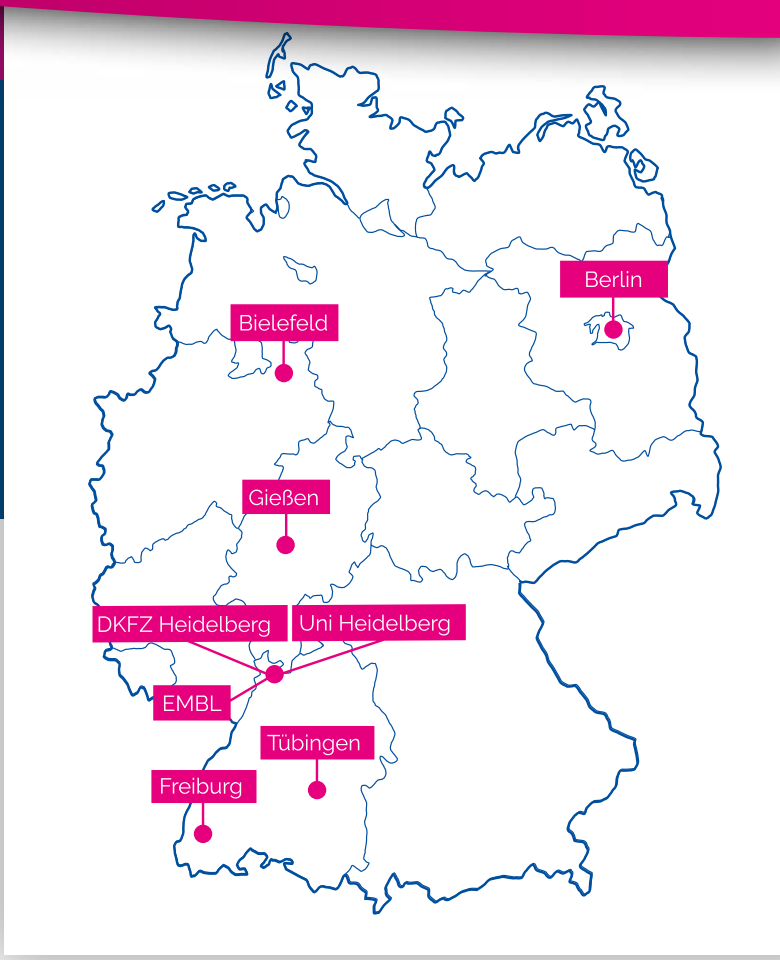


Figure 1: Distribution of the individual de.NBI Cloud locations within Germany (Source: https://cloud.denbi.de/cloud-federation).

The de.NBI Cloud is the largest German academic cloud used for non-commercial scientific purposes. More than 56,000 computing cores, 100 petabytes of file and object storage and 720 terabytes of RAM are currently available to the life sciences community. In order to be able to cover future requirements, the integration of new hardware is constantly being planned and implemented.

Who uses the de.NBI Cloud?

There are currently a total of over 1,000 cloud developers registered across 350 cloud projects (see Figure 2). These projects are services that have been integrated into the cloud by a handful of developers, but are often used by a large number of end users. One example is the Galaxy project at the Freiburg site, which supports a total of 15,000 users worldwide. Other examples include the SILVAngs service for ribosomal RNA analysis, which records about 200 users per month, and the eggNOG-mapper for DNA sequence annotation projects, which receives more than 4,500 requests per month from over 600 users.

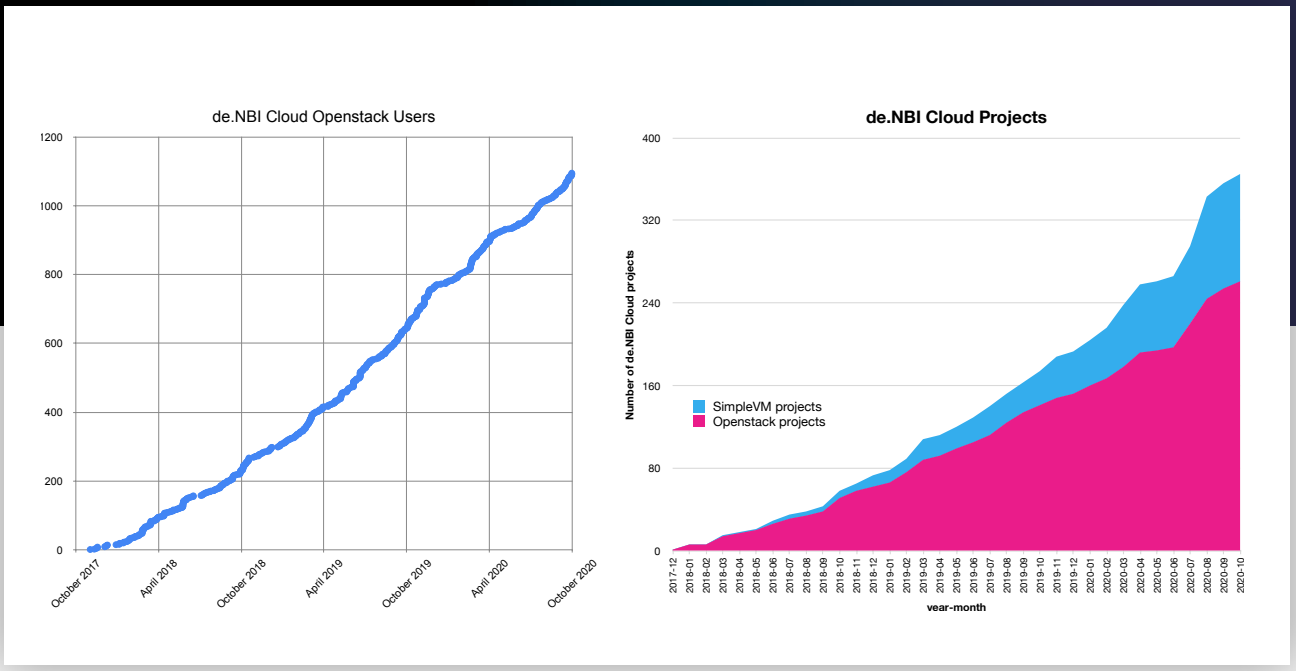


Figure 2: Development of user numbers and projects of the de.NBI Cloud (Source: Peter Belmann, https://cloud.denbi.de/cloud-numbers).

The de.NBI Cloud is used by scientists with varying degrees of background knowledge in cloud computing. It also addresses different user roles that reflect the respective levels of IT experience. Users with little IT experience are more likely to access individual virtual machines (VMs) via the centralized de.NBI Cloud portal using so-called SimpleVM projects. This allows specialized work flows to be made available in a particularly user-friendly manner. Performing high-throughput analyses with several hundred computing cores requires experienced bioinformaticians with direct access to the cloud infrastructure via the OpenStack API. And of course, many of the de.NBI service centers use the cloud infrastructure as a compute backend for their services. Cloud users are trained in a variety of tutorials, workshops, summer schools, hackathons and individual technical meetings that are tailored to various levels of background knowledge. In addition to providing technical assistance, the de.NBI locations bring their own particular scientific expertise to a variety of cloud projects. The various locations focus on topics such as human genomics, microbial (meta) genomics, plant genomics, proteomics, systems biology and integrative bioinformatics.

A special focus for the future: Data protection and IT security

Comprehensive IT security concepts are implemented at the de.NBI Cloud sites to ensure a high level of data protection and security. One of the central requirements for the de.NBI Cloud infrastructure is a solid security architecture. In the future, the implementation of these security concepts and data protection regulations should allow for personal data to be processed in the de.NBI Cloud. In order to implement these requirements in a sustainable and verifiable manner and to ensure the highest possible level of IT security, the de.NBI Cloud sites are planning to have their respective cloud systems assessed and to obtain either certification in accordance with IS0 27001 and 2/27017 or a certification seal in accordance with the Cloud Computing Compliance Controls Catalogue (C5) of the German Federal Office for Information Security (BSI). This would provide all de.NBI Cloud users with proof of a trustworthy cloud infrastructure with the necessary measures in place to ensure data integrity and confidentiality.

Cloud locations

de.NBI Office – Cloud Governance (Peter Belmann)
Justus Liebig University Giessen (Prof. Dr. Alexander Goesmann)
University of Freiburg (Dr. Björn Grüning)
University of Tübingen (Dr. Jens Krüger)
Heidelberg University (Prof. Dr. Vincent Heuveline)
DKFZ (Dr. Ivo Buchhalter)
EMBL Heidelberg (Dr. Jan Korbel)
Bielefeld University (Prof. Dr. Alexander Sczyrba)
Charité – Universitätsmedizin Berlin (Harald Wagener)

Contact:

Prof. Dr. Alexander Sczyrba
Bielefeld University
Bielefeld Institute for
Bioinformatics Infrastructure
asczyrba@cebitec.uni-bielefeld.de

https://ekvv.uni-bielefeld.de/pers_publ/publ/PersonDetail.jsp?personId=17894

Christian Lawerenz
Steinbeis Transfer Center Health Data
christian.lawerenz@stw.de

www.steinbeis.de/en/su/2191

Photos: Sczyrba, Lawerenz

Figure 3: Schematic illustration for project application in the de.NBI Cloud (Source: Alexander Sczyrba and Susanne Konermann).

the data integration centre – a hub for health data from patient care and medical research

Data integration and its prerequisites[1]

by Björn Schreiweis, Danny Ammon, Martin Sedlmayr, Fady Albashiti, and Thomas Wendt

- In the Medical Informatics Initiative, almost all German university hospitals are establishing data integration centres (DICs).
- Health data from patient care and medical research is integrated into the DICs and made available for biomedical research projects.
- The use of interoperability standards plays a key role in accessing and processing the data.
- The DICs continue to support processes related to feasibility studies and applications for biomedical research projects as well as the review and implementation of such projects.

The data integration centre

Multisite use of data is still a hurdle for many research projects. Clinicians and medical researchers have long wanted reliable tools to help them answer questions through analysis of routine health data. That is why the Medical Informatics Initiative (MII) is setting up data integration centres (DICs) – in nearly all German university hospitals or the university (Schreiweis *et al.*, 2019). In the DICs, technical and organizational prerequisites are established so that medical data can be made available for optimal use in research and treatment in healthcare.

What can a data integration centre do?

How a data integration center can offer you support:

- You wonder how many suitable patients are likely to be available for a study next year
- You would like to know whether the use of the hospital's patient data is permitted for your project
- You require intensive care or microbiology data for your analyses
- You would like to merge and analyze data from your hospital with data from other institutions
- You would like to pseudonymize routine patient data for a research project in accordance with established procedures
- You would like to apply data from research results directly in patient care

In addition to research-motivated questions, DIC services can also be used for patient care, quality management, planning issues or other integration tasks. Instead of laboriously extracting data from different source systems and making it available via various channels, the DIC becomes the central point of contact for clinical data and services pertaining to this data.

[1] This is an updated version of the article "Das Datenintegrationszentrum – Ausgangspunkt für die datengetriebene medizinische Forschung und Versorgung" [The Data Integration Centre – Starting Point for Data-Driven Medical Research and Patient Care], which originally appeared in a special "Medical Informatics Initiative" issue (4/2019) of the magazine *mdi – Forum der Medizin, Dokumentation und Medizin-Informatik*. The magazine is published by BVMI e.V. and DVMD e.V. (www.bvmi.de/mdi).

Figure 1: Locations of the data integration centres of the Medical Informatics Initiative (Source: TMF e.V.).

What are data integration centres working on?

Technical/organizational structures and DIZ processes are currently being set up. These include:

- components through which data can be prepared and provided in a standardized form, including access management and pseudonymization,
- consent management,
- templates for user applications and contracts,
- data use and access committees (UACs).

Use cases have been defined in each MII consortium to demonstrate the capabilities of each of the DICs. Examples include diagnosis and care of acute respiratory failure and bloodstream infections, as well as a generic approach to phenotype identification (SMITH); development of classification/prediction models for patients with COPD and brain tumors (MIRACUM); visualization of local infection clusters and establishment of multi-institutional molecular tumor boards (HiGHmed); improvements in care for multiple sclerosis, Parkinson's disease and certain oncological diseases (DIFUTURE) (Hemmer *et al.*, 2019).

How can researchers access data?

Researchers first submit applications describing, among other things, the research topic and the data needed. At each requested location, a UAC evaluates applications according to certain organizational and legal criteria, but also with regard to the availability of the data. After a positive vote, the DIC concludes a user agreement that obligates researchers to handle the data carefully and to feed back the results, for example.

In addition to data exchange, decentralized analysis is enabled – also upon submission of an application. In times of "big data", the centralization of data is not always sensible, nor is it always possible under data protection laws. However, risks can be reduced if analyses are carried out locally at the sites and only (partial) results are returned centrally.

Applications, contracts and processes are harmonized at the national level. A central application and registration office will be established as a point of contact for researchers, for example, to forward applications to the DIC.

Interoperability

Interfaces between the DICs and users are consented in MII working groups. As a result, technical concepts pertaining to the DICs may differ to some extent, but they preserve interoperability on the path to a national data infrastructure. Among other things, this is represented by the jointly developed core data set (Ammon *et al.,* 2019). This defines how data must be prepared in terms of content and technology so that they can be used for further analyses.

Integration as a multi-topic task

The cooperation of the MII institutions is not without differences in terms of viewpoints, interests and obligations. The DICs collaborate on a common understanding of the opportunities and limits of using routine patient data for inter-institutional projects.

For each project, it must also be clarified in specific terms which data should be made available, in which formats and using which terminologies, and via which interfaces.

What is the current status?

The development of the DICs is funded for five years. During the first three years, the DICs have already been able to support projects by providing data. Very good progress was made, both technically and medically, particularly with regard to the use cases. Currently, due to the SARS-COV 2 pandemic, the topics of infection control (Smart Infection Control System, SmICS) and algorithmic surveillance of critically ill patients are extended to all DICs.

In addition, the MII shows that the DICs can also analyze data across consortia (see Demonstrator Study; Ganslandt *et al.,* 2019). Based on the demonstrator study, two continuing projects are conducted: CORD_MI (Collaboration on Rare Diseases) to improve documentation of rare diseases in clinical systems and POLAR_MI (Polypharmacy- Drug Side Effects - Risks) to minimize risks of polypharmacy.

Beyond the use cases, many DICs are active locally. User applications have been answered according to the individual capabilities of the DICs, even though routine structures are still being established in some cases.

The work to date has also revealed certain limitations. Not all clinical application systems used are capable of providing data in a suitable manner. This results in high expenses for processing. The harmonization of user applications and contracts, as well as broad consent, requires a good deal of coordination, also with partners outside the MII. Together with the data

About the Medical Informatics Initiative (MII)

The aim of the Medical Informatics Initiative (MII) is to enhance research opportunities and patient care through innovative IT solutions. These solutions are designed to enable data from healthcare contexts and from clinical and biomedical research to be used and exchanged across multiple entities and sites – transcending the boundaries of individual institutions and geographical locations. The German Federal Ministry of Education and Research (BMBF) is supporting the MII to the tune of around €160 million through 2022. All of Germany's university hospitals have joined forces with research institutions, businesses, health insurers and patient advocacy groups in four consortia – DIFUTURE, HiGHmed, MIRACUM and SMITH – to create a framework across more than 30 sites that harnesses research findings to the direct benefit of patients. Data protection and data security are given the highest priority.

A coordination office, which is located in Berlin and jointly run by the Technology, Methods and Infrastructure for Networked Medical Research (TMF), the German Association of Medical Faculties (MFT) and the German Association of Academic Medical Centers (VUD), is responsible for harmonizing developments at the national level.

Further information:
https://www.medizininformatik-initiative.de/en/ and https://medizininformatik-karte.de

protection authorities of the German states, ethics committees and biobanks, a nationally standardized consent form for data usage has been adopted (Medical Informatics Initiative, 2020).

Outlook

The development of the DICs is moving forward with great enthusiasm. The DICs not only create services for data-driven and networked medical research, but also bundle know-how about the lifecycle of data. They will soon also support studies in which artificial intelligence is used, along with quality management and monitoring processes.

In the future, DICs may also grow beyond university hospitals and link up with non-university service providers, as some already have.

The DICs are central points of contact when it comes to clinical data. Please ask your university hospital for more information!

Figure 2: In the DICs, the conditions are created to bring together and make available medical data in such a way that they can be optimally used for research (Photo: shutterstock / Panchenko Vladimir).

References:

Schreiweis, B., Ammon, D., Sedlmayr, M., Albashiti, F., Wendt, T. (2019). Das Datenintegrationszentrum – Ausgangspunkt für die datengetriebene medizinische Forschung und Versorgung. mdi 4 2019, 21, 106-110.

Hemmer, B., Börries, M., Christoph, J., Marx, G., Maaßen, O., Schuppert, A., Scheithauer, S. (2019). Die klinischen Anwendungsbeispiele (Use Cases) der vier MII-Konsortien. mdi 4 2019, 21, 98-102.

Ammon, D., Bietenbeck, A., Boeker, M., Ganslandt, T., Heckmann, S., Heitmann, K., Sax, U., Schepers, J., Semler, S.C., Thun, S., Zautke, A. (2019). Der Kerndatensatz der Medizininformatik-Initiative – Interoperable Spezifikation am Beispiel der Laborbefunde mittels LOINC und FHIR. mdi 4 2019, 21, 113-117.

Ganslandt, T., Schaaf, J., Schepers, J., Storf, H., Balzer, F., Haferkamp, S., Lodahl, R., Prasser, F., Sax, U., Stenzhorn, H. and Prokosch, H.U. (2019). Experiences from the National Demonstrator Study within the German Medical Informatics Initiative. In: AMIA 2019 Annual Symposium, 16.–20. November 2019, Washington, D.C.

Medizininformatik-Initiative (2020). Medizininformatik-Initiative erhält grünes Licht für bundesweite Patienteneinwilligung. Pressemitteilung. 27.04.2020. https://www.medizininformatik-initiative.de/de/medizininformatik-initiative-erhaelt-gruenes-licht-fuer-bundesweite-patienteneinwilligung (letzter Zugriff: 17.08.2020)

Contact:

Prof. Dr. sc. hum. Björn Schreiweis
Professor for Medical Informatics
Institute for Medical Informatics and Statistics, Kiel University and University Hospital Schleswig-Holstein
bjoern.schreiweis@uksh.de

Danny Ammon
Data Integration Centre
Jena University Hospital
danny.ammon@med.uni-jena.de

Prof. Dr. rer. nat. Martin Sedlmayr
Chair of Medical Informatics
Technische Universität Dresden
martin.sedlmayr@tu-dresden.de

Dr. sc. hum. Fady Albashiti
CEO Medical Data Integration Center (MeDICLMU)
University Hospital of Munich (LMU)
fady.albashiti@med.uni-muenchen.de

Dr. rer. med. Thomas Wendt
Head of the Data Integration Centre
University of Leipzig Medical Center
thomas.wendt@medizin.uni-leipzig.de

Photos: Schreiweis, Ammon, Sedlmayr, Albashiti, Wendt

the ascent of mass spectrometry in systems medicine

Four research cores are searching for new biomarkers

by Jeroen Krijgsveld, Ursula Klingmüller, Carsten Müller-Tidow, Bernhard Küster, Daniel Teupser, Ulrich Keilholz, Frederick Klauschen, Markus Ralser, Matthias Selbach, Philipp Wild, and Stefan Tenzer

Within the framework of the German government's High-Tech strategy 2025 and the initiative "Research Cores for Mass Spectrometry in Systems Medicine" the Federal Ministry of Education and Research (BMBF) provides funding for the development of new analytical tools for health research. Four interdisciplinary research consortia ("research cores") have been launched across Germany to establish novel mass spectrometric methods to be used in clinical settings.

The holistic principles of systems medicine combined with mass spectrometry as a key technology to detect disease-relevant biomolecules (proteins, lipids and metabolites) offer considerable innovative potential. As disease progression is typically driven by multiple factors, it is essential to systematically characterize and quantify these factors during the course of disease. The generated knowledge can be used in various ways: On the one hand, it can contribute to a deeper mechanistic understanding of disease development. On the other hand, its implementation may improve numerous aspects of patient care, including (early) diagnosis, prognosis and therapy recommendations. Ultimately, this may result in the development of new, targeted and personalized therapeutic strategies.

Systems medicine approaches require a high level of interdisciplinary collaboration between clinical, biological, analytical, and computational scientists, to be able to generate, process and interpret large data sets, before implementation in the clinic. Mass spectrometric methods are still strongly underrepresented in medical diagnostics, and their potential is only beginning to be exploited. This is partly due to the lack of standardized procedures and previous technical limitations of mass spectrometers. To establish a mass spectrometry network for systems medicine in Germany, the BMBF is providing a total of €26.6 million over the next three years to four research cores in Berlin, Heidelberg, Mainz and Munich.

The research cores aim to build up and strengthen connections between the fields of mass spectrometry, medicine and informatics. Their specific objective is to foster close collaborations between scientists and clinicians to broaden the implementation of systems medicine in clinical routine (Figure 1). By mass spectrometric analyses of various clinical samples, the participating researchers aim to discover molecular signatures that may be used as novel biomarkers for diagnosis, prognosis, therapeutic decisions and treatment responses in the clinic. To reach this goal, best practice workflows shall be established and implemented across all research cores. In addition to the medical, logistical and technical infrastructures necessary for sample collection, extraction and analysis, the research cores will also optimize and standardize their newly developed methodological approaches with regard to sample throughput, robustness and reproducibility (Figure 2).

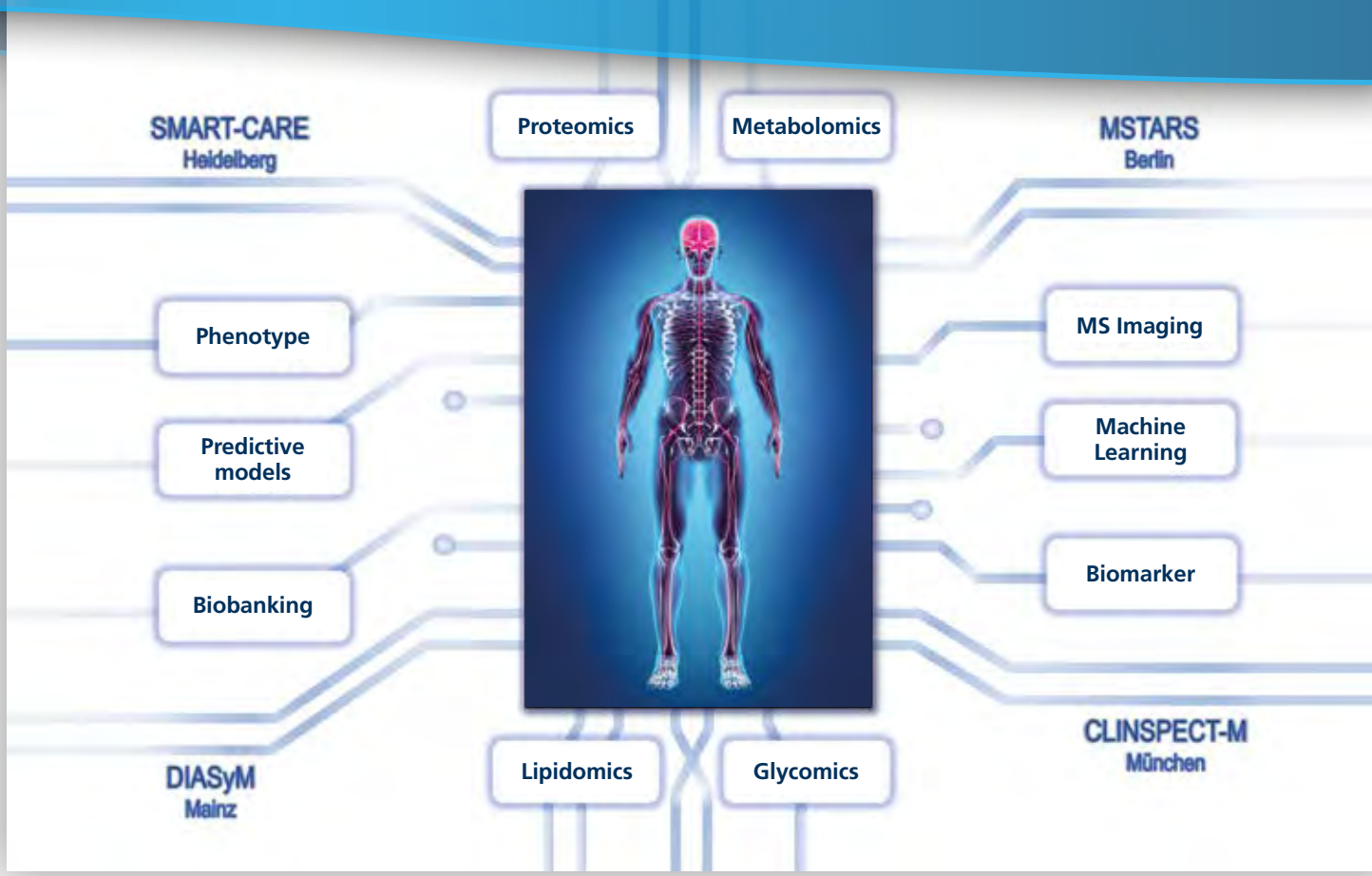


Figure 1: The four new research cores of the MSCoreSys initiative focus on the development of mass spectrometric methods to establish and improve systems medicine-based approaches to understand underlying pathophysiological (disease) mechanisms (Source: Philipp Wild and yodiyim – fotolia.com).

Each of the four research cores focuses on different disease areas:

DIASyM in Mainz

In Mainz, the new research core DIASyM (Data-Independent Acquisition-Based Systems Medicine: Mass Spectrometry for High-Throughput Deep Phenotyping of Heart Failure Syndrome) is co-led by Stefan Tenzer, coordinator of the mass spectrometry technology platform and method development, and Philipp Wild, who is the systems medicine coordinator. DIASyM combines the expertise of researchers and physicians at the University Medical Center Mainz and Johannes Gutenberg-University Mainz. Heart failure affects roughly 15 million people in Europe and is a leading cause of hospitalization for patients over 65. The disease progresses continuously and leads to shortened life expectancy, since treatment options are fairly limited and only symptomatic. Thus, the disease places a significant burden on healthcare systems.

There has been little research so far on the underlying pathophysiological mechanisms of the disease. The research consortium in Mainz aims to establish optimized data-independent mass spectrometric methods for high-throughput phenotyping of the heart failure syndrome. DIASyM uses and analyzes mass spectrometric information from proteomics, lipidomics and metabolomics in combination with a variety of clinical parameters, employing state-of-the-art AI-based methods including deep and probabilistic machine learning. The researchers employ a systems medicine-based approach to provide a deeper understanding of the disease's underlying biological processes. In turn, those findings might serve as a starting point for developing targeted therapies that could counteract potential late adverse effects at an early disease stage, thus significantly improving patients' quality of life and life expectancy.

CLINSPECT-M in Munich

The Munich-based research core CLINSPECT-M (Clinical Mass Spectrometry Center Munich) is co-led by Bernhard Küster

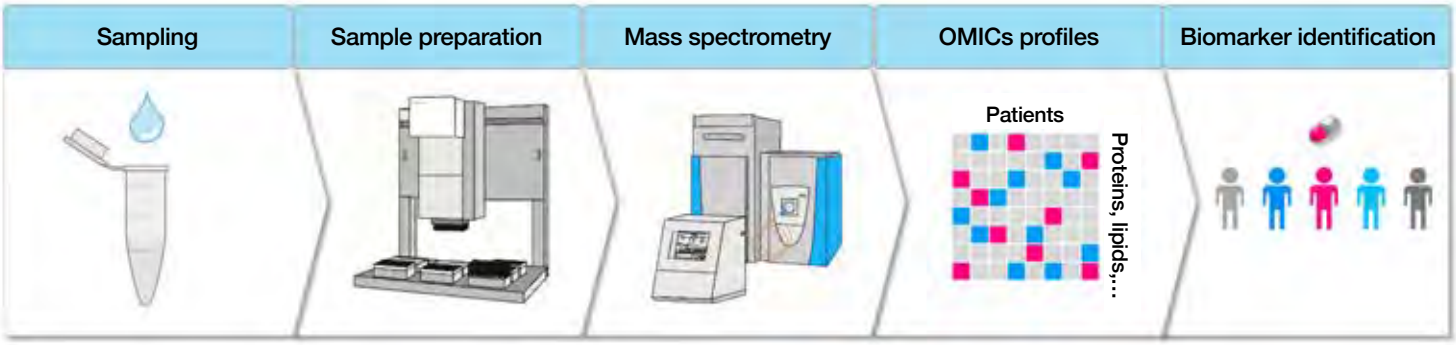


Figure 2: Schematic workflow sampling to (potential) identification of biomarkers (Source: Stephanie Heinzlmeir).

and Daniel Teupser and brings together the expertise of the Technical University of Munich, Ludwig-Maximilian-University Munich and their respective university hospitals as well as that of the Helmholtz Center Munich and the Max Planck Institute of Biochemistry. The interdisciplinary consortium is planning technological developments in the fields of proteomics, medicine and bioinformatics, with a biomedical research focus on diseases of the nervous system, in particular multiple sclerosis, Alzheimer's disease, stroke and cancer.

Despite their high incidence and social relevance, many aspects of these widespread diseases have so far not been understood. How does multiple sclerosis differ from other neuroinflammatory diseases? Why do some Alzheimer's patients respond better to drug therapy than others? How can patients be distinguished that suffered from ischemic strokes or from so-called stroke mimics? Can the proteomic profile of cancer patients help recommend more effective personalized therapies? These and other questions will be addressed by the CLINSPECT-M consortium using state-of-the-art mass spectrometric and bioinformatics approaches to provide new insights into pathological processes in the brain and possible treatment options. As the clinical applicability of proteomics for these diseases is at different stages of development, CLINSPECT-M will show examples of short-, medium- and long-term options for integrating the technology into daily clinical practice.

MSTARS in Berlin

The Berlin research core MSTARS (Multimodal Clinical Mass Spectrometry to Target Treatment Resistance) is co-led by Ulrich Keilholz (Charité Comprehensive Cancer Center), Frederick Klauschen (Charité, Institute for Pathology and Institute for Pathology, Ludwig-Maximilians-University Munich), Markus Ralser (Charité, Institute of Biochemistry) and Matthias Selbach (Max Delbrück Center for Molecular Medicine in the Helmholtz Association, MDC). Most chronic diseases exhibit complex pathophysiologies. In most cases, the underlying fundamental pathomechanisms are better understood on the molecular level than those mechanisms regulating healing and progression. For many diseases, the treatment response at the level of the individual patient represents a challenge in precision medicine.

The Berlin research core will leverage and combine complementary technologies of (phospho)-proteomics, metabolomics, glycomics and mass spectrometry imaging, and it will integrate these with clinical expertise in a multimodal approach supported by state-of-the-art computational research and machine learning methods. While the research concept with its mechanistic and signature-derived strategy is broadly applicable, MSTARS' medical focus will be on cancer and inflammatory diseases. The consortium has access to a unique and large collection of clinical samples and preclinical models, with head and neck squamous cell carcinoma representing the primary use case.

SMART-CARE in Heidelberg

SMART-CARE (Systems Medicine Approach to Stratification of Cancer Recurrence), the Heidelberg-based research core, is coordinated by Jeroen Krijgsveld (Heidelberg University Hospital, UKHD), Ursula Klingmüller (German Cancer Research Center, DKFZ) and Carsten Müller-Tidow (UKHD). Also involved are researchers and physicians from five different clinics at UKHD as well as from Heidelberg University, the European Molecular Biology Laboratory (EMBL) and the Center for Mass Spectrometry and Optical Spectroscopy (CeMOS) at Mannheim University of Applied Sciences. The aim of the research core is to gain a deeper molecular understanding of the progression of various individual cancers, and especially with regard to cancer relapse.

Although primary tumors can be treated in an increasing number of cases, cancer relapse remains the main cause of cancer-related deaths. The researchers in SMART-CARE therefore aim to identify molecular markers for cancer relapse in various clinical specimen, such as tissue, blood or cerebrospinal fluid. This will facilitate the detection of tumor recurrence at an early stage to then prompt treatment of the patient, or to monitor cancer progression and predict the success of the therapeutic approach. The focus of SMART-CARE is on blood cancers, lung cancers, brain tumors and sarcomas, each treated with innovative intervention strategies. The research core will generate systematic and quantitative protein and metabolic profiles of these four cancers, which will feed into artificial intelligence-based methods, machine learning and mathematical modeling. This will yield a rich treasure trove of data that can be used to develop a new generation of biomarker patterns. Ultimately, this will inform physicians to propose therapies for individual cancer patients.

Contact:

Current spokespersons for the MSCoreSys initiative:

Prof. Dr. Jeroen Krijgsveld
German Cancer Research Center (DKFZ) and
Heidelberg University Hospital (UKHD)
j.krijgsveld@dkfz-heidelberg.de

https://www.dkfz.de/en/proteomik-stammzellen-krebs/index.php

Prof. Dr. Stefan Tenzer
Institute for Immunology
University Medical Center of the
Johannes Gutenberg-University Mainz
tenzer@uni-mainz.de

http://www.tenzerlab.de

Homepage of the initiative:
www.mscoresys.de

Photos: Jutta Jung / DKFZ, Thomas Böhm / University Medical Center of the Johannes Gutenberg-University Mainz

modeling infectious respiratory diseases

Systems medicine modeling of treatment options for community-acquired pneumonia and COVID-19

by Peter Ahnert and Markus Scholz

Even before COVID-19, respiratory infections were among the most common infectious diseases worldwide, with lower respiratory tract infections ranking as one of the top five leading causes of death (Michaud, 2009). Community-acquired pneumonia (CAP) accounts for a significant proportion of these deaths. The clinical course of CAP varies greatly from person to person and may include rapid and sometimes unpredictable deterioration. Using systems medicine approaches, the CAPSyS consortium is investigating the mechanisms that impact the disease's progression. For the current COVID-19 pandemic, we would also like to use this approach to contribute to better prognostic and therapeutic options.

CAP, an infectious lung disease with systemic effects

Community-acquired pneumonia (CAP) is a common infectious disease of the lower respiratory tract that can take a very severe course, especially in the elderly and in people with pre-existing conditions. The wide range of progressions includes rapid deteriorations that are often difficult to predict. The arsenal of therapeutic interventions is limited. In critical progression the function of the epithelial-endothelial barrier, which isolates the infectious process in the lungs from the bloodstream, is often impaired to such a degree that the infection becomes systemic. It is no longer limited to a specific location and spreads throughout the organism. This results in sepsis and associated damage to other organ systems, e.g., the liver and kidneys. Yet, even when the barrier remains intact, the systemic immune response may cause damage. Understanding the processes surrounding the barrier is a major CAPSyS research focus as there is hope for new therapeutic approaches in this area. As a result, in addition to examining disease progression in humans, we are conducting targeted mouse studies to investigate barrier function under different interventions during the course of the infection.

Pneumonia may also be caused by an infection of the respiratory tract with the novel pathogen SARS-CoV-2. However, the progression of this disease differs – sometimes significantly – from that of CAP with other causes, both in its manifestation and dynamics (Zhou *et al.*, 2020). In COVID-19, patients may feel relatively well despite severe lung damage. As the disease progresses, the lungs may suddenly fail. We are therefore working to apply our systems medicine research approach, developed within the framework of CAPSyS, to lung inflammation triggered by SARS-CoV-2 as quickly as possible.

As part of the Genetical Statistics and Systems Biology Group (Figure 1) at Leipzig University's Institute of Medical Informatics, Statistics and Epidemiology, we are responsible for the statistical and bioinformatics analysis of the extensive data derived from clinical trials and mouse experiments within CAPSyS. In addition, we are developing dynamic biomathematical models with which to define and predict different aspects of infection outbreaks. This article presents our research and some of our results in brief.

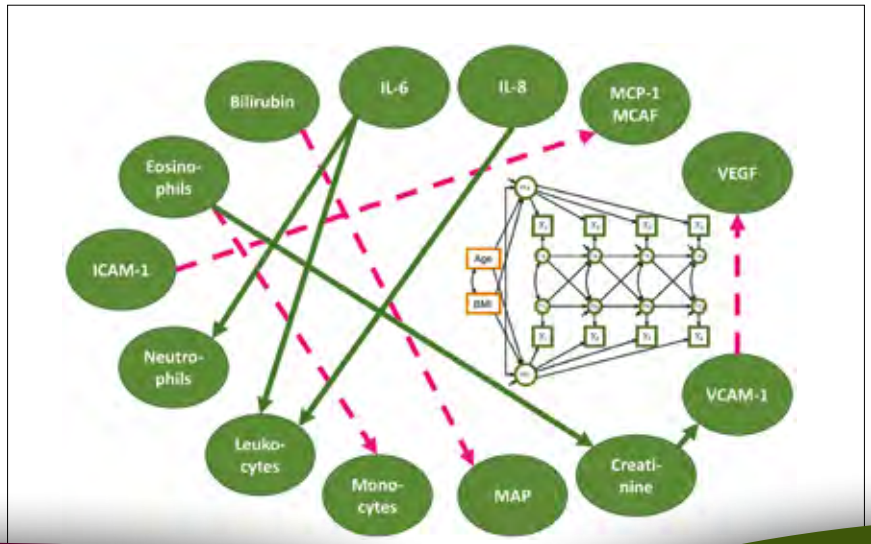
Figure 1: Genetical Statistics and Systems Biology Group at the Institute for Medical Informatics, Statisics and Epidemiology (IMISE) at Universität Leipzig (Photo: Medical Faculty of Universität Leipzig).

Molecular network analyses reveal causal relationships

Extensive molecular data from several omics levels (genome, transcriptome, metabolome, proteome, cytokines) are available from the PROGRESS study[1] on community-acquired pneumonia in hospitalized patients as well as other studies. Some of these data are available as time series. The analysis of these extensive data is a key focus of our work in CAPSyS. The approach is two-pronged: We are developing and validating diagnostic and prognostic biomarker signatures, and we aim to establish causal relationships between characteristics within and between omics levels or clinical parameters. The latter is accomplished through both, cross-sectional analyses (e.g., using Mendelian randomization methods or mediation analyses) and time series data analyses (e.g., using structural equation models). The successive application of these methods allows us to construct extensive causal networks (see example in Figure 2). The goal of these analyses is to establish relevant omics relationships that inform mechanistic modeling of disease processes.

[1] http://capnetz.de/html/progress/project?set-language-to=en

Figure 2: Results of causal network analyses based on time series data in PROGRESS



Inferred causal relations between cytokines, blood cells and clinical parameters based on the analysis of time series data. Green arrows: positive causal relationships, red dashed arrows: negative causal relationships. Only parameters (nodes) that were significantly associated with at least one other parameter are shown. Significant relationships were defined by raw $P < 0.001$ in the corresponding latent growth curve model with structured residuals (LCM-SR, see inset for simplified path diagram). (Source: Genetical Statistics and Systems Biology Group, IMISE, Universität Leipzig)

Biomathematical modeling of barrier function and therapy

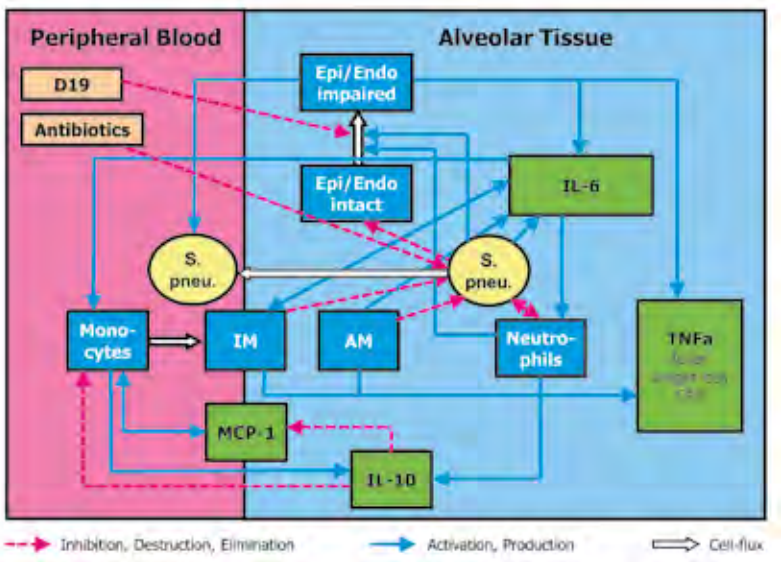
Based on relationships between molecular and cellular factors identified as essential, we have established a biomathematical model of pneumococcal infections in mice. Using ordinary differential equations, the model describes the evolution of the bacterial load in the lungs and circulation, the populations of important immune cells and the concentrations of essential cytokines and chemokines (proteins that regulate cell growth and migration, respectively), as well as corresponding interactions between these immune response components. Mechanisms of antibiotic action were also taken into account. The model was developed and parameterized using extensive animal time series data, particularly from the early phases of infection (Schirm *et al.,* 2016). Recently, this model was extended significantly, with particular emphasis on (1) describing in greater detail the function of the epithelial-endothelial barrier located between the alveoli (pulmonary alveoli) and blood circulation, and (2) investigating the effect of barrier-protective therapeutic strategies such as complement component 5a inactivation (C5a inhibitor, Figure 3). This allows us to simulate and predict the impact of combined therapies of antibiosis and barrier-protective factors. One of the insights gained was that an earliest possible antibiosis combined with a high dosage of the C5a inhibitor leads to a particularly favorable course of disease with significantly stabilized barrier function and correspondingly lower systemic inflammation.

Model of disease severity in humans

The model for immune response and therapy of pneumococcal pneumonia in mice shown in Figure 3 cannot be readily applied to humans. For one, essential information and data from the site of infection that are required for parameterization are lacking. Moreover, data from the early phase of infection in humans are unavailable. Information on when the infection occurred is also usually missing, so that even with close monitoring in the hospital, patients are observed starting at different times after the onset of infection, resulting in an asynchronous time series. To circumvent this problem, we first created a phenomenological model in humans that describes transitions between disease stages as random processes in time. Known as a continuous-time Markov chain, this model was developed and parameterized with data from the PROGRESS study and from patients with pneumogenic sepsis (i.e., sepsis caused by respiratory tract infection) in large multicenter clinical trials conducted by the SepNet study group.[2] Disease severity was operationalized using the Sequential Organ Failure Assessment (SOFA) score (Ahnert *et al.,* 2019, Figure 4). The model provides solid predictions in an independent data set regarding 28-day mortality of CAP patients. It will be expanded to include individual risk factors (Przybilla *et al.,* 2020).

[2] https://www.sepsis-stiftung.eu/sepnet

Figure 3: Model of pathomechanisms and therapy of pneumococcal pneumonia in mice



Schematic representation of the model describing the cellular and molecular components and bacterial populations in murine alveolar tissue and peripheral blood upon infection with S. pneumonia (S. pneu.), considering the impact of antibiotics and barrier stabilizers (D19). Epi/Endo intact: unaffected epithelial and endothelial cells, Epi/Endo impaired: affected epithelial and endothelial cells during barrier failure, IM: inflammatory macrophages, AM: alveolar macrophages, CRP: C-reactive protein. The mathematical model based on this scheme employs ordinary differential equations. (Source: Genetical Statistics and Systems Biology Group, IMISE, Universität Leipzig)

COVID-19

While the research focus of CAPSyS is pneumococcal-associated pneumonia, there are some parallels to COVID-19. So, in collaboration with colleagues at Charité – Universitätsmedizin Berlin, we are trying to apply the systems medicine approaches developed in CAPSyS to this new disease. We are currently using the PROGRESS and CAPSyS infrastructures to develop a qualified database from hospitalized COVID-19 patients. Using this database, we aim to apply the Markov chain models described in the previous section to COVID 19 to arrive at a better understanding of disease progression.

Barrier function also plays a crucial role in COVID-19. Therefore, we are currently working to transfer our general pneumonia model to the COVID-19 disease pattern. As for general pneumonia, the COVID-19 model will be calibrated on the basis of data from animal experiments of the infection process in the lung and the function of the barrier. To support this, our partners at Charité will conduct extensive experiments with SARS-CoV-2, analogous to the experiments previously performed in CAPSyS with pneumococci.

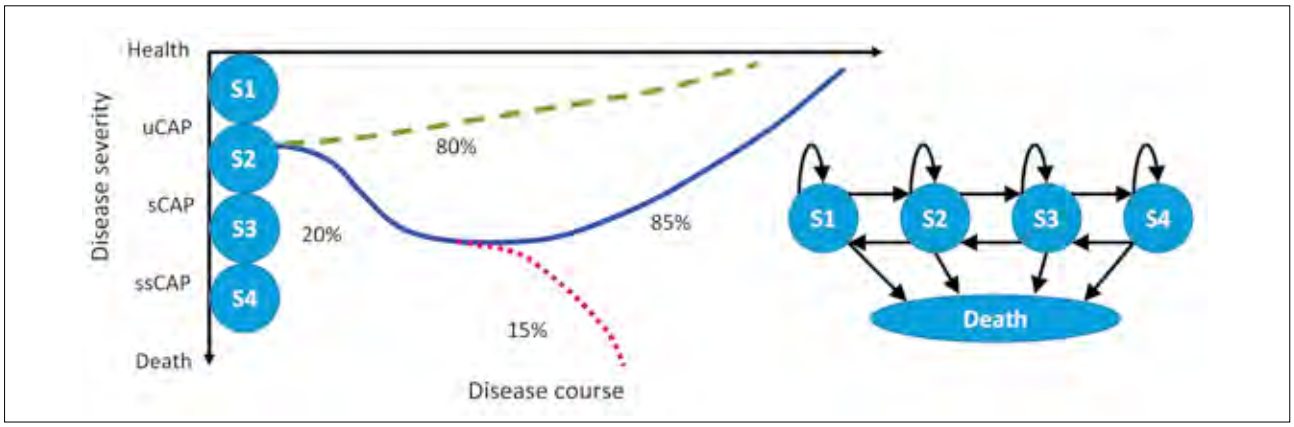
Finally, we are participating in the epidemiological modeling of the COVID-19 pandemic, again relying on models we built for the epidemiological description of pneumococcal pneumonia in other projects. Predictions on the development of COVID-19 deaths in Germany based on a first version of this model are already available to the public via a platform on which predictions of different modeling groups are visualized collectively (https://kitmetricslab.github.io/forecasthub/forecast).

With our research in the CAPSyS consortium, we hope to make a systems medicine-based contribution to the development of better treatment options as well as to a better understanding of the pathomechanisms of COVID-19 and its spread.

Long-term effects

Little is known at this point about the long-term effects of conventional pneumonia and pneumonia caused by SARS-CoV-2. In conventional pneumonia, epidemiological data suggest, among other things, an increase in atherosclerosis and associated events. The underlying pathomechanisms are largely unknown. Under the direction of Charité (Martin Witzenrath) and Leipzig University (Markus Scholz), the collaborative e:Med project SYMPATH (Systems Medicine of Pneumonia-Aggravated Atherosclerosis) was recently initiated to investigate these pathomechanisms following a systems medicine approach (https://www.sys-med.de/de/verbuende/sympath). As in CAPSyS, researchers are conducting parallel studies in mice and humans in order to probe and model the underlying molecular pathomechanisms. A transfer of the approach to COVID-19 is planned.

Figure 4: Markov model describing clinical courses of community-acquired pneumonia



Exemplary disease trajectories from mild community acquired pneumonia (uCAP) to severe CAP (sCAP) and severe CAP with septic shock and high lethality (ssCAP). For the Markov model (scheme on right) 4 disease states defined by the SOFA-Score were assumed (S1-S4). For simplicity, these states are transversed sequentially. Death is an absorbing state and can be reached directly from each disease state. (Source: Genetical Statistics and Systems Biology Group, IMISE, Universität Leipzig)

Project profile:

The consortium "CAPSyS – Medical Systems Biology of Pulmonary Barrier Failure in Community-Acquired Pneumonia" was launched in 2014 as part of the research and funding concept "e:Med – Measures Establishing Systems Medicine" of the German Federal Ministry of Education and Research (BMBF). Its aim is to use systems biology approaches to analyze the progression of pneumococcal-associated pneumonia, with a special focus on how to describe and predict progression on the basis of clinical, cellular and molecular features. The consortium is comprised of seven institutions: Leipzig University (Institute for Medical Informatics, Statistics and Epidemiology, IMISE), Charité – Universitätsmedizin Berlin (Department of Infectious Diseases and Respiratory Medicine), Universitätsklinikum Erlangen (Laboratory of Systems Tumor Immunology), Justus Liebig University Giessen (Institute of Medical Microbiology), University of Greifswald (Department of Functional Genomics), Jena University Hospital (Institute of Clinical Chemistry and Laboratory Diagnostics) and the Philipps University of Marburg (Institute for Lung Research). To date, the consortium's research has been published in 57 papers in international scientific journals.

CAPSyS
Systems Medicine of Community Acquired Pneumonia

References:

Ahnert, P., Creutz, P., Horn, K., Schwarzenberger, F., Kiehntopf, M., Hossain, H., Bauer, M., Brunkhorst, F.M., Reinhart, K., Völker, U., *et al.* (2019). Sequential organ failure assessment

score is an excellent operationalization of disease severity of adult patients with hospitalized community acquired pneumonia - results from the prospective observational PROGRESS study. Critical care (London, England) 23, 110.
Michaud, C.M. (2009). Global Burden of Infectious Diseases. Encyclopedia of Microbiology, 444-454.
Przybilla, J., Ahnert, P., Bogatsch, H., Bloos, F., Brunkhorst, F.M., SepNet, C.C.T.G., Progress, S.G., Bauer, M., Loeffler, M., Witzenrath, M., *et al.* (2020). Markov State Modelling of Disease Courses and Mortality Risks of Patients with Community-Acquired Pneumonia. Journal of clinical medicine 9.
Schirm, S., Ahnert, P., Wienhold, S., Mueller-Redetzky, H., Nouailles-Kursar, G., Loeffler, M., Witzenrath, M., and Scholz, M. (2016). A Biomathematical Model of Pneumococcal Lung Infection and Antibiotic Treatment in Mice. PloS one 11, e0156047.
Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., *et al.* (2020). Clinical course and risk factors for mortality of adult in patients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet.

Contact:
Prof. Dr. Markus Scholz
Genetical Statistics and Systems Biology Group
Institute for Medical Informatics, Statistics and Epidemiology (IMISE)
Medical Faculty
Universität Leipzig
markus.scholz@imise.uni-leipzig.de

https://www.genstat.imise.uni-leipzig.de
https://www.capsys.imise.uni-leipzig.de

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

The German Network for Bioinformatics Infrastructure offers training events for life scientists

Online Training

Basis bioinformatics training for biologists
Analysis, Vsualization and Integration of Multi-Level Omics Data

Introduction to Oxford Nanopore sequence data analysis
Nanopore Workshop - Best Practice and SARS-CoV-2 Applications

Introduction to metagenome data analysis
Metagenomics Bioinformatics

Introduction to RNA-seq data analysis
RNAseq and High Throughput Omics Analysis for Plants

Advanced methods for differential analysis of proteomics data
Advanced Analysis of Auantitative Proteomics Data Using R

Primary means to store biological data
Ontologies - Statistics, Biases, Tools, Networks and Interpretation

Analyzing scientific data using machine learning algorithms
Machine Learning using Galaxy

Introduction to the de.NBI Cloud
de.NBI Cloud User Meeting

And many more at:
www.denbi.de/training

linkedin.com/company/de-nbi
contact@denbi.de
@denbiOffice

SPONSORED BY THE
Federal Ministry of Education and Research
Projektträger Jülich Forschungszentrum Jülich

From data to knowledge – standards for personalized medicine

EU-STANDS4PM – a pan-European Expert Forum aimed at developing standards for data integration and data-driven in silico models for personalized medicine

by Dr. Marc Kirschner, Project Management Jülich (PtJ)

The systematic analysis and interpretation of large, complex data sets (big data) from various areas of application has the potential to significantly improve the diagnosis and treatment of diseases. This includes, for example, data from the life sciences, the health sector and clinical research. Particularly in personalized medicine, data-driven methods using computer-based models can make a significant contribution to the early detection and prevention of diseases, as well as the prediction of therapeutic success.

Yet the process of generating new medical knowledge currently falls far short of its potential. This is largely due to the heterogeneity of big data and the lack of widely accepted standards for data collection, harmonization and integration. Working with personal and patient-related data also brings with it strict ethical and legal requirements regarding the fundamental right of patients to information and data protection.

In the future, ethically and legally compliant standardization guidelines will therefore represent a central component of personalized medicine in the field of computer-based modeling. These guidelines are a basic prerequisite for customized treatment strategies and for targeted early detection and preventive measures.

This is why the European Expert Forum EU-STANDS4PM was established: to accelerate the further development of such standardization and to make a sustainable contribution to the establishment of transnational guidelines for data-driven modeling methods in personalized medicine.

EU-STANDS4PM is a Coordination and Support Action (CSA) that has been funded by the European Commission's Horizon 2020 program since January 2019 and aims to further advance the use of big data in health for personalized medicine. The EU-STANDS4PM consortium comprises 16 partners from eight European countries, and brings together cross-disciplinary expertise from scientific organizations, industry, European ESFRI infrastructure projects, legal and ethics experts, and standardization organizations.

Core activities during the three-year duration of the EU-STANDS4PM project include the harmonization of data integration strategies in medical research and practice in Europe and the development of flexible standardization guidelines for European research networks, taking into account the need for a transparent ethical and legal framework. The project also aims to strengthen data-driven predictive modeling strategies in personalized medicine in a targeted manner, and to harmonize the handling of data within collaborative research projects. To help achieve this, a harmonized Data Access Agreement was recently developed, which simplifies access to data stored centrally in the European Genome-phenome Archive (EGA).

EU-STANDS4PM is an open network and seeks input from all relevant stakeholders that have an interest in advancing predictive in silico methodologies in personalized medicine through broadly applicable standards for data integration and harmonization.

Further information:
www.eu-stands4pm.eu

MTZ®-Award for Medical Systems Biology 2020

This year's winners of the MTZ® Award for Medical Systems Biology have been announced. **Dr. Fabian Fröhlich** (Technical University of Munich), **Dr. Carolin Loos** (Technical University of Munich) and **Dr. Martin Scharm** (University of Rostock) won over the national review panel and the board of the MTZ® Foundation with their excellent doctoral theses to emerge victorious. The MTZ® Foundation grants the award to young scientists whose doctoral theses have achieved groundbreaking and outstanding results in the field of medically oriented systems biology. The award is intended to draw special attention to promising young scientists and help them gain public recognition. The MTZ® Foundation works together with the Federal Ministry of Education and Research (BMBF) and Project Management Jülich (PtJ).

MTZ® stiftung

– for a better future –

The awarding of the prize, consisting of a certificate and prize money, took place for the seventh time. Due to the COVID-19 pandemic it was solemnly presented in a virtual ceremony.

Further information:

www.mtzstiftung.de

www.mtzstiftung.de/stiftung/mtzfoundation/the-work

Source: Project Management Jülich (PtJ)

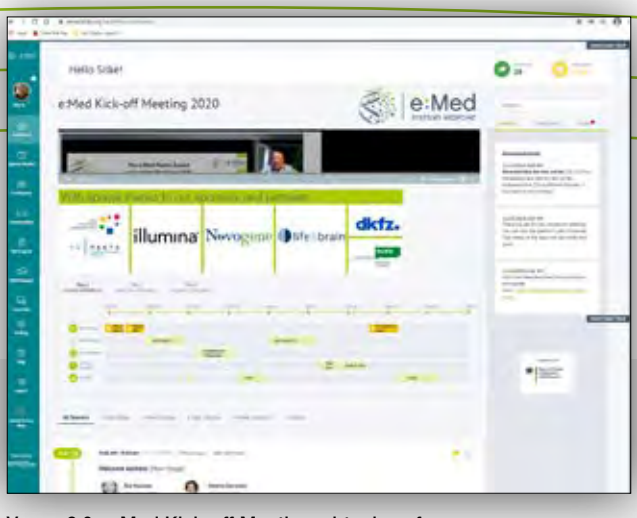
events

Despite the Pandemic:

e:Med Kick-off Meeting 2020

24. November - 25. November, vitual meeting

by Asli Kayserili, e:Med Office



Venue 2.0: e:Med Kick-off Meeting, virtual conference space

While many scientists were rescheduling their experiments to continue after the lockdown of March 2020, the e:Med Office announced that the Kick-off Meeting had to be postponed until later in the year. That was just a week before the event.

The pandemic was untimely for the e:Med community. The second term of BMBF funding had just started. The new groups were looking forward to presenting their research to each other. The meeting was an opportunity to network. Hence, cancelling the event was not an option. The organization committee quickly adapted the program to the digital format. The virtual meeting took place on November 24 – 25 and consisted of 3 outstanding keynotes, 40 scientific talks, 29 of which were new projects, 50 poster presentations, 19 flash talks, 15.5 hours of streaming, and attracted 270 participants.

The keynotes reflected main themes of the field. **Dagmar Kulms** presented her research aimed at identifying molecular alterations that confer therapy resistance in melanoma. How modelling RNA dynamics can help us gain insight into development was described by **Fabian Theis**. **Joachim Schulze** discussed the technological requirements to make systems medicine a reality for patients. **Peter Krawitz** demonstrated that the AI-driven medical image analysis supports the identification of genetic variants. Attendees cherished the diversity of topics.

Besides outstanding talks, there were also lively community activities. The e:Med Office introduced the 'Project Groups' on overlapping topics that boost networking within the community. **Maria Fedorova** and **Christoph Schickhardt** promoted MS-based Omics and Data Security and Ethics, respectively. Other highlights of the program included poster sessions and flash talks that preceded these. Best posters were contributed by young researchers from SysMedSUDs, SYS-Stomach, and coNfirm. Our sponsors HMG Systems Engineering, Illumina, Novogene, and Life&Brain demonstrated how the latest innovations could further progress in systems medicine.

The virtual meeting took place on a conference platform which kept the community together. For convenience, a central dashboard made all key elements of the meeting accessible on the platform (live video, chat, program, abstracts, short CVs, posters, abstract book, other info). Participants could create their profile and benefit from customized networking. Alternatively, they could meet one another randomly, making the platform very close to an on-site conference.

Our survey results show that many liked the interactive and responsive platform. Technical support during streaming was valuable to our speakers and chairs. Participants loved the high quality of the talks, representation of various fields and the chance to contact everyone. We also experienced limitations of the format. For example, asking questions per chat couldn't substitute a real-life exchange. Nevertheless, it was a beneficial experience.

As organizers, we are ready to implement the experience at the next event. Although COVID-19 forced us to meet virtually, it also helped us realize the value of online conferences. It also has advantages over meeting on-site: it requires no travel, allows asynchronous participation, is cost-effective and it is more environmentally friendly. Nevertheless, a face-to-face meeting offers completely different possibilities.

We are already looking forward to the next event and to bringing the e:Med Systems Medicine community together. In 2021, we will meet again in a digital medium. Until conditions truly allow for an on-site meeting, the digital format can make an important contribution to scientific success.

Contact:

Dr. Asli Kayserili
Scientific Manager e:Med Management Office
German Cancer Research Center (DKFZ) Heidelberg
asli.kayserili@dkfz.de

imprint

gesundhyte.de

The magazine for Digital Health in Germany – International Edition Issue 13, July 2021

gesundhyte.de is an annually published magazine with information from German research in the field of digital medicine.

ISSN 2748-3533

This issue is an adapted version of the already published German magazine.

Publisher:

gesundhyte.de is published by the Berlin Institute of Health (BIH) at Charité and the Project Management Jülich (PtJ).

Editors:

Editor-in-Chief: Prof. Dr. Roland Eils (BIH at Charité)

Editorial Coordination: Franziska Müller (BIH at Charité)

Editorial Team:

Dr. Silke Argo (e:Med), Dr. Stefanie Gehring (DLR-PT), Prof. Dr. Lars Kaderali (University of Greifswald), Dr. Marco Leuer (DLR-PT), Dr. Yvonne Pfeiffenschneider (PtJ), Dr. Matthieu-P. Schapranow (Hasso-Plattner-Institute), Dr. Thomas Schmidt (Plattform Lernende Systeme), Dr. Stefanie Seltmann (BIH at Charité), and Dr. Gesa Terstiege (PtJ).

Address:

Editorial office gesundhyte.de
c/o BIH at Charité
Center for Digital Health
Kapelle-Ufer-2, 10117 Berlin, Germany

Design and layout:

Kai Ludwig, LANGEundPFLANZ Werbeagentur GmbH, Speyer (www.LPsp.de)

Translation:

Todd Brown, Berlin, Deutschland

Printed by:

Ottweiler Druckerei und Verlag GmbH, Ottweiler; Germany

FSC Mix

Subscriptions:

The magazine is funded by the German Federal Ministry of Education and Research (BMBF). It is published as part of the public relations work of the initiatives listed as "Publisher". It is supplied free of charge and must not be sold.

For subscription please visit www.gesundhyte.de or contact:

Editorial office gesundhyte.de
c/o BIH at Charité
Center for Digital Health
Kapelle-Ufer-2; D-10117 Berlin
franziska.mueller@bih-charite.de

systembiologie.de becomes gesundhyte.de!

apops - Fotolia

Please visit our homepage if you would like to find out more about the topics of our magazine.

What you can expect:

- Exciting stories from everyday research – find out more about ongoing projects
- Interviews and profiles – get to know the faces behind the research
- Extensive overview of events – never miss another important date
- Information about current funding – stay up to date
- Get involved – suggest your own topic

We look forward to your visit on our homepage!

about us

Presenting the gesundhyte.de editorial team

gesundhyte.de aims to communicate the successes of the German research accessible to a broad audience in a descriptive way. The magazine is created once a year in German and English by a multidisciplinary editorial team from various German research institutions: Berlin Institute of Health at Charité, Hasso Plattner Institute Potsdam, University of Greifswald, Project Management Jülich, DLR Project Management and representatives of the initiatives: Lernende Systeme – the Platform for Artificial Intelligence in Germany and e:Med Systems Medicine. The magazine is financed by the Berlin Institute of Health (BIH) at Charité and the German Federal Ministry of Education and Research (BMBF).

The editorial team of gesundhyte.de:

Front row, left to right: Prof. Dr. Roland Eils (BIH at Charité), Franziska Müller (BIH at Charité), Dr. Silke Argo (e:Med), Dr. Stefanie Gehring (DLR-PT), **second row, left to right:** Prof. Dr. Lars Kaderali (University of Greifswald), Dr. Marco Leuer (DLR-PT), Dr. Yvonne Pfeiffenschneider (PtJ), Dr. Matthieu-P. Schapranow (Hasso Plattner Institute), **third row, left to right:** Dr. Thomas Schmidt (Plattform Lernende Systeme), Dr. Stefanie Seltmann (BIH at Charité), Dr. Gesa Terstiege (PtJ), and Kai Ludwig (LANGEundPFLANZ)

contact

Berlin Institute of Health at Charité
Editor in Chief: Prof. Dr. Roland Eils
Editorial coordination and subscription services:
Franziska Müller
c/o BIH at Charité
Center for Digital Health
Kapelle-Ufer-2; D-10117 Berlin, Germany
E-Mail: franziska.mueller@bih-charite.de
www.hidih.org

Communication and Marketing, Press Office:
Dr. Stefanie Seltmann
Anna-Louisa-Karsch-Str. 2; D-10178 Berlin, Germany
E-Mail: stefanie.seltmann@bih-charite.de
www.bihealth.org

Lernende Systeme – Germany's Platform for Artificial Intelligence
Contact persons:
Dr. Thomas Schmidt and Dr. Ursula Ohliger
Managing Office: c/o acatech
Pariser Platz 4a; D-10117 Berlin, Germany
E-Mail: t.schmidt@acatech.de, ohliger@acatech.de
www.plattform-lernende-systeme.de

Project Management Jülich
Forschungszentrum Jülich GmbH
Life- and Health Sciences (LGF)
Contact persons:
Dr. Yvonne Pfeiffenschneider and Dr. Gesa Terstiege
D-52425 Jülich, Germany
E-Mail: y.pfeiffenschneider@fz-juelich.de, g.terstiege@fz-juelich.de
www.ptj.de

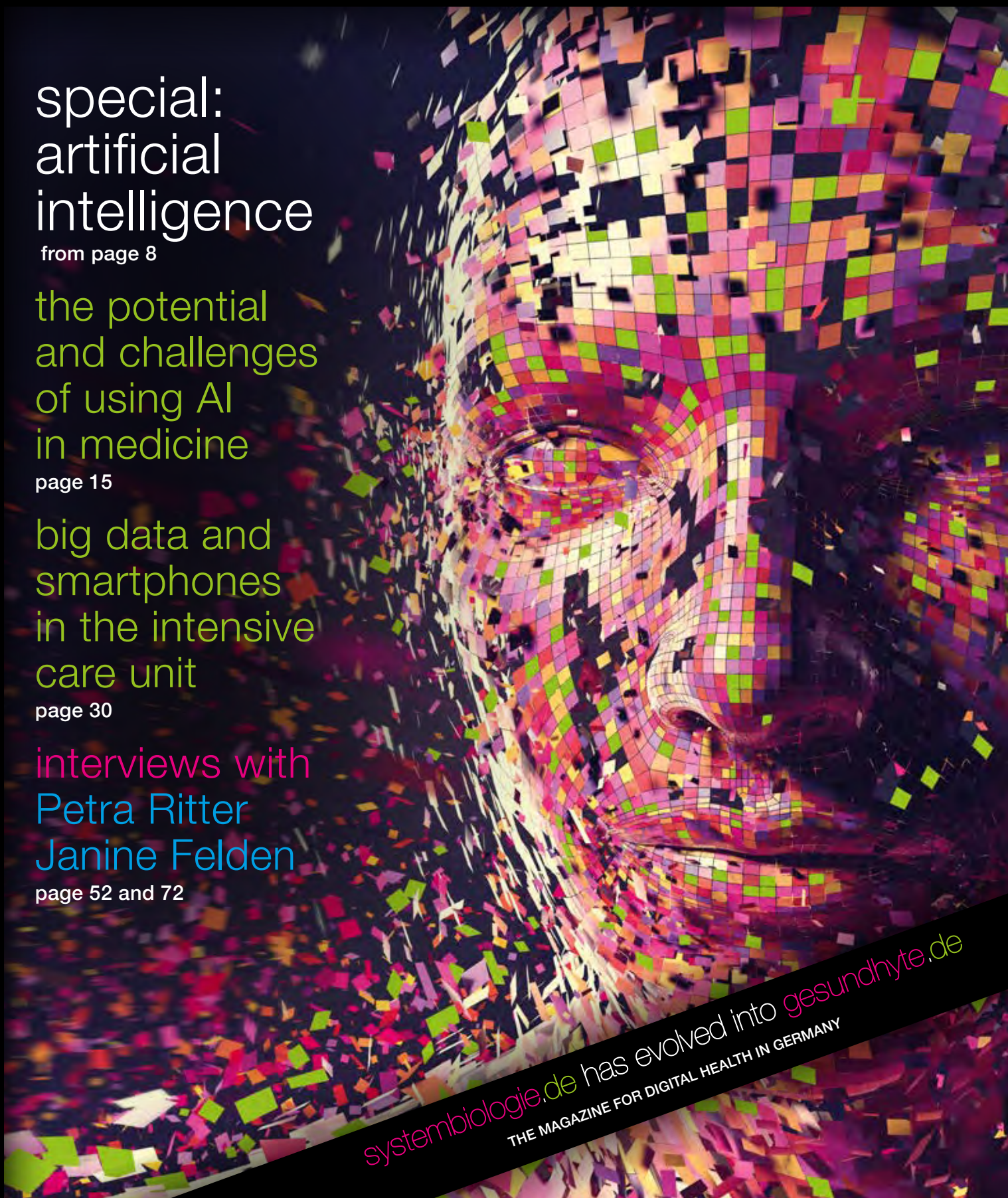
German Aerospace Center (DLR)
Project Management Agency
Department Health (OE20)
Contact persons:
Dr. Marco Leuer and Dr. Stefanie Gehring
Heinrich-Konen-Str. 1; D-53227 Bonn, Germany
E-Mail: marco.leuer@dlr.de, stefanie.gehring@dlr.de
www.dlr-pt.de

e:Med Management Office
Managing Director:
Dr. Silke Argo
c/o Deutsches Krebsforschungszentrum (DKFZ) – V025
Im Neuenheimer Feld 581; D-69120 Heidelberg, Germany
E-Mail: s.argo@dkfz.de
www.sys-med.de

systembiologie.de

THE MAGAZINE FOR SYSTEMS BIOLOGY RESEARCH IN GERMANY

ISSUE 13 JULY 2021



special: artificial intelligence

from page 8

the potential and challenges of using AI in medicine

page 15

big data and smartphones in the intensive care unit

page 30

interviews with Petra Ritter Janine Felden

page 52 and 72

systembiologie.de has evolved into gesundhyte.de

THE MAGAZINE FOR DIGITAL HEALTH IN GERMANY