Editorial

Hans-Ulrich Prokosch

Data integration in life sciences

https://doi.org/10.1515/itit-2017-0010

Big Data Analytics is one of the big buzzwords in the life sciences. Particularly in medicine expectations are high, considering future Learning Health Systems and Precision Medicine approaches. Nevertheless it is often undervalued, that data heterogeneity (variability) is one of the major aspects of big data and that integration of various types of data from extremely heterogeneous original data sources is therefore also one of the major obstacles towards efficient big data analytics. Research on data integration has already brought up a large spectrum of supporting tools that are helpful for a better understanding of heterogeneous sources, ease extraction and transformation tasks, and enable the specification of processing pipelines for ad hoc aggregation of large volumes of data. However, it has also become clear that the ambitious goals in healthcare require both, interdisciplinary cooperation among researchers, and different healthcare institutions to cooperate and share their data across institutional borders.

In the U.S. this barrier has been identified already some years ago and, just to name some exemplary research initiatives, the eMERGE network [1], the PCORNET grant program [2], as well as the recent precision medicine initiative [3] were launched. In all such research initiatives data integration at large scale across many clinical care and research sites, as well as especially the integration of clinical data and molecular/omics data are major research areas.

In Germany the Federal Ministry of Education and Research has identified this challenge for the German health system as well. In the recent medical informatics funding scheme it mentions that "researchers and doctors are facing new scientific, technical and organizational challenges posed by these [medical data – X-rays, medical reports, blood parameters – but also genome data generated by high-throughput technologies] enormous volumes of data." And it further states that "it is not only the volume but also the heterogeneity of the medical data which is so challenging. Widely divergent types of data must often be brought together: whereas the growth of a tumour becomes visible on an X-ray, we may only understand the cause of its growth once we are able to analyze the DNA of its cells" [4]. Thus, the German Ministry of Education and Research asks for the establishment – in an initial step – of data integration centres at selected university hospitals and further on to ensure the technical and organizational conditions which are necessary for a multi-site data exchange between health care and clinical and biomedical research [4].

In this context, the current special issue on "Data Integration in Life Sciences" addresses a very timely and challenging research area with lots of focus around the world. Exemplary four manuscripts have been selected, illustrating various issues, opportunities and experiences of data integration projects in the life sciences.

The special issue starts with Ying-Chi Lin, et al. and an article on "Integration and visualization of spatial data in LIFE". LIFE is a large epidemiological study managed by the LIFE Research Center for Civilization Diseases at Leipzig University. When integrating and analyzing huge amounts of heterogeneous data meaningful insights can often already be gained by visualizing medical data on geographical maps. Therefore, within the LIFE project the authors have developed an interactive web application (LIFE SDVS) allowing an effective data visualization by adding geographical facets to the data integration and analysis workflow.

The second manuscript "How to improve Information Extraction from German Medical Records" by Johannes Starlinger, et al. tackles a second challenge. This is the fact, that vast amounts of medical information are still recorded as unstructured narrative text, e.g. in radiology or pathology findings or in physician discharge letters. Nevertheless, all such data contain valuable information on the course of a patient, which is typically not found anywhere as structured, coded information in the patient's medical record. Therefore the Berlin research group from Humboldt University has initiated a project to make this information accessible by applying semantical analysis techniques and information extraction algorithms. Even though for English texts much research has already been published on natural language processing (NLP) and a comprehensive body of work and tools does al-

Hans-Ulrich Prokosch: Friedrich-Alexander-Universität Erlangen-Nürnberg, Chair of Medical Informatics, Wetterkreuz 13, 91058 Erlangen, Germany, e-mail: hans-ulrich.prokosch@fau.de

ready exist, for German texts analogue research results are still sparse, tools are mostly proprietary and can rarely be reused in other projects. In their article the authors first describe the challenges of information extraction, with a particular focus on German medical documents, secondly address the problems of missing German language resources and privacy implications and finally identify steps to overcome the current hurdles and fuel research in semantic integration of textual clinical data.

The so-called "omics" data, spanning a wide range of research from metabolomics, transcriptomics, proteomics, to genomics play an important role in translational research. In our third article Björn Sommer and Falk Schreiber have focused on a small subset of such research, the modelling and analysis of metabolic processes (a special field of research in computational systems biology and medicine) by capturing also information on subcellular localization of proteins, enzymes and transporters involved in the metabolic processes. In the recent years, the authors have developed two tools (CELL microcosmos 4 PathwayIntegration: CmPI and VANTED) and present those in the manuscript on "Integration and Virtual Reality Exploration of Biomedical Data with CmPI and VANTED". While CmPI is used to analyze, visualize and explore potential subcellular localizations in a cell environment, e.g. with 2D and 3D visualization of directed networks showing the enzyme-compound relationships, VANTED allows to analyze and visualize biological networks, to integrate data into these networks and to simulate their dynamics. In the article the authors especially focus on presenting future perspectives of these approaches by integrating them in order to provide complex modelling, visualization and exploration of metabolic networks in combination with omics data. Furthermore, they illustrate the potential of immersive analytics of the spatially distributed networks used in different Virtual Reality-based technologies.

Last but not least, Benjamin Baum, et al. in their opinion paper "Data Provenance Challenges in Biomedical Research", provide an overview of challenges concerning data provenance in biomedical research. The authors reflect current literature and depict some examples of existing implicit or explicit provenance aspects in some standard data types in translational research. In data integration projects, the comparability of data from different sources is of utmost importance. Towards this end data provenance should provide a recall about the origin of the data, transformation process steps, support replication and presentation of the data. Even though usable concepts for the documentation of data provenance can be found in other fields as early as 2005, the penetration rate in biomedical projects and in the biomedical literature is still quite low. Thus, the authors' plea is, that awareness for the necessity of basic data provenance in biomedical research has to be raised and respective concepts need to be defined and standardized in the biomedical informatics community.

References

1. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, Sanderson SC, Kannry J, Zinberg R, Basford MA, Brilliant M, Carey DJ, Chisholm RL, Chute CG, Connolly JJ, Crosslin D, Denny JC, Gallego CJ, Haines JL, Hakonarson H, Harley J, Jarvik GP, Kohane I, Kullo IJ, Larson EB, McCarty C, Ritchie MD, Roden DM, Smith ME, Böttinger EP, Williams MS; eMERGE Network: The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med. 2013 Oct;15(10):761–71.
2. PCORnet PPRN Consortium, Daugherty SE, Wahba S, Fleurence R: Patient-powered research networks: building capacity for conducting patient-centered clinical outcomes research. J Am Med Inform Assoc. 2014 Jul–Aug;21(4):583–6.
3. https://allofus.nih.gov/.
4. BMBF – Federal Ministry of Education and Research. Medical Informatics Funding Scheme – Networking data – improving healthcare. October 2015. http://www.gesundheitsforschung-bmbf.de/_media/Medizininformatik_englisch_barrierefrei.pdf.

Bionotes

Prof. Dr. Hans-Ulrich Prokosch
Friedrich-Alexander-Universität Erlangen-Nürnberg, Chair of Medical Informatics, Wetterkreuz 13, 91058 Erlangen, Germany
hans-ulrich.prokosch@fau.de

Hans-Ulrich Prokosch received a degree in Mathematics and his Ph.D. in Medical Informatics from the Justus-Liebig University of Gießen. He started his scientific research at the Department of Medical Informatics (Gießen University) and gained international experience during a two year research visit at the University of Utah (Department of Medical Informatics, Salt Lake City). After his habilitation in Medical Informatics he initiated the Clinical Information Systems Group in the Department of Medical Informatics and Biometrics at Münster University. Since 2003 he holds the Chair of Medical Informatics within the Department of Medical Informatics, Biometrics and Epidemiology at the University of Erlangen-Nürnberg. He is in parallel Chief Information Officer of Erlangen University Hospital and thus responsible for the hospital's strategic development of its IT architecture, solutions and management. He is a member of the German Association of Medical Informatics, Biometrics and Epidemiology (GMDS E.V.) and the German Informatics Society (GI e.V.). He is in the editorial board of numerous international scientific journals in biomedical informatics and member of the American College of Medical Informatics.