NOVA ACTA LEOPOLDINA

Abhandlungen der Deutschen Akademie der Naturforscher Leopoldina

Herausgegeben von Diethard Tautz,
Director Ephemeridum der Akademie

NAL-conference NUMMER 423

Mission – Innovation

Telematics, eHealth and High-Definition Medicine in Patient-Centered Acute Medicine

Leopoldina Symposium

Berlin

February 28th and 29th, 2020

Editors:

Claudia Spies (Berlin)
Member of the Leopoldina

Gunnar Lachmann (Berlin)

Maria Heinrich (Berlin)

**Deutsche Akademie der Naturforscher Leopoldina –
Nationale Akademie der Wissenschaften, Halle (Saale) 2021
Wissenschaftliche Verlagsgesellschaft Stuttgart**

Redaktion: Dr. Renko Geffarth unter Mitarbeit von Robert Blumenau

Die Publikationsreihe Nova Acta Leopoldina erscheint bei der Wissenschaftlichen Verlagsgesellschaft Stuttgart, Birkenwaldstraße 44, 70191 Stuttgart, Bundesrepublik Deutschland.

Die Publikationsreihe wird gefördert durch das Bundesministerium für Bildung und Forschung sowie das Ministerium für Wirtschaft, Wissenschaft und Digitalisierung des Landes Sachsen-Anhalt.

Einbandbild: Franziska Joseph

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://portal.dnb.de abrufbar.

Die Abkürzung ML hinter dem Namen der Autoren steht für Mitglied der Deutschen Akademie der Naturforscher Leopoldina – Nationale Akademie der Wissenschaften.

© 2021 Deutsche Akademie der Naturforscher Leopoldina e. V. – Nationale Akademie der Wissenschaften
Postadresse: Jägerberg 1, 06108 Halle (Saale), Postfachadresse: 110543, 06019 Halle (Saale)
Hausadresse der Redaktion: Emil-Abderhalden-Straße 37, 06108 Halle (Saale)
Tel.: +49 345 47239134, Fax: +49 345 47239139
Herausgeber: Prof. Dr. Diethard Tautz, *Director Ephemeridum* der Deutschen Akademie der Naturforscher Leopoldina – Nationale Akademie der Wissenschaften
Printed in Germany 2021
Gesamtherstellung: Druck-Zuck GmbH Halle (Saale)
doi:10.26164/leopoldina_10_00361
ISSN (Print): 0369-5034
ISSN (Online): 2748-7431
ISBN: 978-3-8047-4204-8
Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung - Nicht kommerziell - Keine Bearbeitungen 4.0 International CC-BY-NC-ND 4.0 Lizenz.
Gedruckt auf chlorfrei gebleichtem Papier.

Contents

Teleneonatology

Social Policy Dimensions

Consolidation of Innovation Funds – Dream or Reality?

High Definition Medicine in Patient-Centered Acute Medicine

Connected Healthcare – How Does it Benefit the Patient?

Individualized Medicine – How Could it Work?

Patient-Centered Risk Reduction – Which Patient Does it Help?

Perspectives from the Industry

Scientific Symposium Abstract Winner

NAL-conference Nr. *423*, 77–82 (2021) doi:10.26164/leopoldina_10_00378

Knowledge Management for Precision Medicine

Ulf Leser (Berlin)

Summary

Medical decision-making is based on numerous types of knowledge about patients, diseases, pharmacology and the effects of interventions. Due in particular to advances in the molecular life sciences, the basis of this knowledge is growing faster than ever. At the same time, the extremely high dimensionality of molecular data makes the derivation of robust findings from limited sample numbers difficult, leading to a flood of publications with weak empirical evidence. This, in turn, challenges clinicians in their need to quickly find the most relevant facts for a given case. Modern computer-based technologies for extracting, aggregating and organizing knowledge, ranging from ontologies for structured knowledge codification to text mining algorithms for automatic extraction of findings and machine-learning based medical decision support, are expected to help cope with this challenge. Using precision oncology as an example for personalized medicine, this paper discusses recent advances in the (semi-) automatic extraction of clinically relevant findings from textual sources, their organization in knowledge bases, and their usage in medical decision support systems. We also describe requirements to accelerate the development of such technologies in Germany, especially when it comes to highly important medical documentation.

Zusammenfassung

Medizinische Entscheidungen beruhen auf einer Vielfalt an Wissen über Patienten, Krankheiten, Pharmakologie und den Auswirkungen von Behandlungen. Insbesondere durch die Fortschritte in den molekularen Biowissenschaften wächst die Wissensbasis schneller denn je. Gleichzeitig erschwert die extrem hohe Dimensionalität molekularer Daten die Ableitung robuster Erkenntnisse aus begrenzten Fallzahlen, was zu einer Fülle an Publikationen mit schwacher empirischer Evidenz führt. Dies wiederum fordert Mediziner heraus, schnell die relevantesten Fakten für einen bestimmten Patientenfall zu finden. Moderne computergestützte Technologien zur Gewinnung, Zusammenlegung und Organisation von Wissen sollen helfen, diese Herausforderung zu bewältigen. Diese Technologien reichen von Ontologien für eine strukturierte Wissenskodifizierung über Text-Mining-Algorithmen zur automatischen Gewinnung von Ergebnissen bis hin zu maschinellem Lernen für die Unterstützung von medizinischer Entscheidungsfindung. Anhand der Präzisionsonkologie als Beispiel der personalisierten Medizin zeigt dieser Beitrag die jüngsten Fortschritte bei der (halb-)automatischen Gewinnung klinisch relevanter Ergebnisse aus Textquellen, die Organisation dieser Ergebnisse in Wissensdatenbanken und ihre Verwendung als Unterstützungssysteme bei der medizinische Entscheidungsfindung. Darüber hinaus stellen wir die Voraussetzungen für eine beschleunigte Entwicklung solcher Technologien in Deutschland dar, insbesondere wenn es sich dabei um äußerst wichtige medizinische Unterlagen handelt.

A video of the presentation can be viewed online:

Introduction

Medicine is and has always been a knowledge-intensive science. When judging upon a patient's case, clinicians need to consider an intimidating amount of knowledge, including the detailed characterization of the patient, her history and family background, knowledge on disease etiology and phenotypes, the impact of possible (and approved) interventions, issues regarding the patient's quality of life and suitability of measures etc. The recent advances in the molecular life sciences have contributed another level of detail, making available information on thousands of genes, their function and interplay, and known implications of their inherited or acquired mutations on disease progression. Keeping pace with this enormous flood of information is a serious challenge for current medicine (Hey and Kesselheim 2016).

Molecular technologies produce measurements of the state of different molecules in a sample, such as the genome or transcriptome of a biopsy of a patient. Large studies, both preclinical and clinical, are required to condense such data into context-specific information, such as the overexpression of a certain gene under certain circumstances, and, eventually, into clinically relevant knowledge, such as the suitability of a treatment in a disease state as measured by a certain gene signature. The primary way of communicating such knowledge is natural language text – in the form of scientific publications, textbooks, clinical guidelines, case reports, medical documentation, discharge summaries etc. Given a clinical decision that needs to go beyond standard procedures, clinicians must search the most relevant texts, extract the relevant facts by reading, and assess their importance for the given case using personal experience and by comparing to results reported in other texts.

However, as the amount of potentially relevant published text is growing at an exponential scale[1], performing this step in a comprehensive manner becomes more and more difficult given the tense time constraints found in practice. The question arises how computer technology can help practitioners to find relevant information faster.

Extracting and Organizing Knowledge from Scientific Articles

Research in the (semi-)automatic extraction of facts from natural language text, summarized under the term "text mining", has seen on enormous growth over the last decades. In medicine, text mining can, for instance, be used to find relevant articles faster (Ševa et al. 2019), to automatically prepare digests of large text collections (Wei et al. 2019), or to support the creation of expert-curated knowledge bases (Lever et al. 2019). In the following, we describe applications of text mining in the field of precision oncology, before we give a short account on the current state of the art.

An important branch of current oncological research is concerned with the impact of somatic variations on disease formation, progression, therapy, and outcome (Gagan and Van Allen 2015). A famous example with approved therapeutic implications are EGFR mutations in colon and lung cancer (Seshacharyulu et al. 2012). A range of similar but not yet broadly confirmed relationships between individual mutations and certain drugs are currently studied in clinical practice (Lamping et al. 2020). The basis of these findings are numerous international trials investigating the associations between genomic features and

1 See, for instance, PubMed growth statistics at https://www.nlm.nih.gov/bsd/stats/cit_added.html

diseases. Therapeutic decisions are increasingly dependent on access to comprehensive knowledge bases that extract and organize large collections of such associations in the form of Variant Information Systems (VIS, Starlinger et al. 2018), a process illustrated in Fig. 1.

Biomedical Text Mining (BTM) can speed up and improve the creation and maintenance of such VIS considerably. To this end, typical BTM workflows process very large text collections (e.g. all PubMed abstracts) to find putatively relevant findings, which are subsequently aggregated into oncological knowledge and presented to experts for inclusion in a VIS. The three main steps in such workflows are (1) Named Entity Recognition (Habibi et al. 2017) concerned with finding textual mentions of genomic variants, genes, cancer types, drugs etc., (2) Named Entity Normalization (Leaman and Lu 2016) which assigns discovered mentions to unique biological entities, such as standardized gene and drug names, and (3) Relationship Extraction (Weber et al. 2020) for identifying the semantic relationships between those entities, such as the association of a mutation to a drug or of a drug to a cancer type.

Recent progress in these fields has been significant, boosting extraction accuracies for many cases into the 80% – 95% range. Extraction at that quality applied over large text collections can bring an enormous help for clinical practice. However, it is also important to keep in mind that BTM is not yet at a state where human judgement of extracted findings would be superfluous. For instance, a BTM process as described above would not be able to differentiate between the values of a mutation-drug association resulting from a large clinical trial, one found in single case report, or one found by in-vitro experiments with a small number of cancer cell lines. Further research on such problems is necessary, for instance to cope with conflicting results from different studies or to adequately consider the strength of empirical evidences.

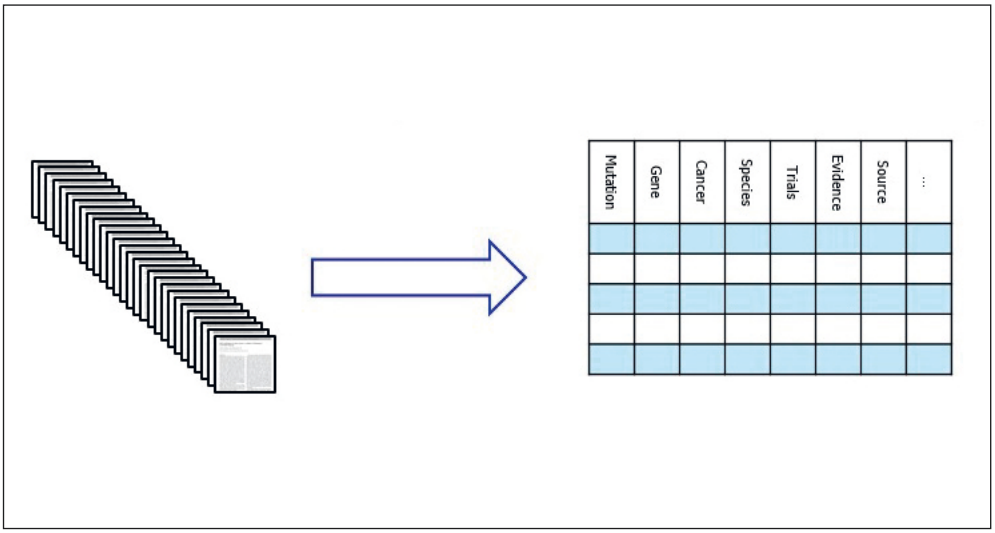

Fig. 1 Information on the role of mutations in specific types of cancer must be extracted from the literature and put into structured format for consideration in clinical decision support.

Supporting Medical Decision-Making

The most important application of BTM in clinical oncology is the support of therapeutic decisions in cases outside the guidelines, which often are taken by Molecular Tumor Boards (MTB). The preparation of individual patients for MTB meetings currently takes up to several hours, which are spent in searching publication databases, querying VIS, reading papers and database entries, judging upon the strength of the underlying evidences, and preparing an integrated report as basis for a rational decision. Several projects are ongoing to support this process by BTM to reduce preparation times and to improve the quality of decisions by making more relevant information available faster[2]. For instance, we recently performed a study comparing the content of five large curated variant-level knowledge bases for precision oncology, showing that using at least three of them significantly improves the recall of a search compared to using any single one (Pallarz et al. 2019). A similar analysis is described in (Perera-Bel et al. 2018), which also presents a software for generating integrated reports given the mutation signature of a patient. (Wagner et al. 2020) present a graphical user interface for intuitive browsing expert-curated information from multiple VIS.

Extracting Knowledge from Medical Reports in Germany

Access to comprehensive, integrated VIS can offer substantial support for clinical decision-making when the individual genome of a patient is to be taken into account. However, they inherently provide only a very narrow view on patients, i.e., their mutational signatures. Evidently, the utility of decision support systems could be improved considerably when also other patient information would be integrated, such as demographic information, family history, previous treatments, or clinical markers. Several clinics in Germany currently are building up clinical data warehouses (CDW) to provide such data fast and at high quality, but still very often it is available only in the form of medical reports[3]. It is thus a natural idea to develop BTM solutions also for medical text, as a complement to BTM on scientific text.

However, the situation in BTM on German medical text is much worse than for English (Starlinger et al. 2016). The main reason is that all state-of-the-art methods for BTM are based on machine learning and must be trained on large, language- and domain-specific annotated text collections, also called Gold Standard Corpora (GSC). Over the last two decades, the international research community has created dozens of such GSC for various tasks for English scientific texts, and large corpora also are freely available for English medical text (Johnson et al. 2016). However, currently not a single such GSC exists for German medical text (Hahn et al. 2018). This not only means that the best current BTM algorithms are not usable, but also that different methods cannot be compared to each other in a systematic manner.[4] Many English GSC have been created in the course of international BTM competitions, which boosted the performance of methods enormously. Such competitions are also emerging

2 See, for instance, the BMBF initiative id:SEM: https://www.gesundheitsforschung-bmbf.de/de/i-dsem-integrative-datensemantik-in-der-systemmedizin-3367.php

3 Note that any CDW faces the problem of integrating the Millions of treatments that were handled prior to building the system – which are available only in textual form.

4 Several companies offer solutions for German medical BTM, but the quality of their tools cannot be compared in an unbiased manner due to the lack of German GSC.

for other western languages, like French or Spanish, but none for German. This also hinders the development of other important applications of BTM on medial reports, such as systems for automatically quality controlling medical accounting systems, e.g. DRG codes, or quality assurance of treatments by controlling compliance to predefined clinical pathways.

There are multiple reasons for this situation (Starlinger et al. 2016). Undoubtedly, the most important one is data protection regulations. Medical reports, in principle, can be shared publically after complete anonymization. However, there exists no common understanding or clear guidelines what a ”complete anonymization” of a medical report implies, and different data protection authorities in different federal states and different organizations tend to communicate very different points-of-view on this subject (Dierks and Rossnagel 2020). This creates a situation in which none of the actors dares to share such data, which severely hinders research progress in German BTM.

Conclusions

Biomedical text mining and knowledge management has great potential to improve medical treatments. The potential is the larger, the more precise characterizations of patients and treatments are available and actually taken into account i.e., in precision medicine. While research in obtaining such information semi-automatically from English texts is well developed, our abilities to analyse German medical text are still very limited. German clinics currently trail far behind in their means to exploit the wealth of medical knowledge buried in their EHR systems, so to change this situation it is of utmost importance to define safe and reliable ways of exchanging carefully selected, carefully anonymized, and carefully annotated medical text corpora. The availability of such corpora would quickly lead to the free availability of high-quality BTM tools, especially because it would open the door for the large German communities of Computational Linguistics and Artificial Intelligence to work on such problems.

Acknowledgements

We acknowledge generous funding from DFG, especially RTG SOMAED, RU Beyond the Exome, and RTG CompCancer, and from BMBF, especially eMED Projects PERSONS, PREDICT, and Map-Tor-Net.

References

Dierks, C., und Rossnagel, A.: Sekundärnutzung von Sozial- und Gesundheitsdaten – Rechtliche Rahmenbedingungen. *Schriftenreihe der TMF Band 17* Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft 2020

Gagan, J., and Van Allen, E. M.: Next-generation sequencing to guide cancer therapy. Genome Med. *7*/1, 80. doi:10.1186/s13073-015-0203-x (2015)

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., and Leser, U.: Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics *33*/14, i37–i48. doi:10.1093/bioinformatics/btx228 (2017)

Hahn, U., Matthies, F., Lohr, C., and Löffler, M.: 3000PA–Towards a national reference corpus of German clinical language. Stud. Health Technol. Inform. *247*, 26–30. doi:10.3233/978-1-61499-852-5-26 (2018)

Hey, S. P., and Kesselheim, A. S.: Countering imprecision in precision medicine. Science *353/6298*, 448–449. doi:10.1126/science.aaf5101 (2016)

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehmann, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G.: MIMIC-III, a freely accessible critical care database. Sci. Data *3/160035*, 1–9. doi:10.1038/sdata.2016.35 (2016)

Lamping, M., Benary, M., Leyvraz, S., Messerschmidt, C., Blanc, E., Kessler, T., Schütte, M., Lenze, D., Jöhrens, K., Burock, S., Klinghammer, K., Ochsenreither, S., Sers, C., Schäfer, R., Tinhofer, I., Beule, D., Klauschen, F., Yaspo, M.-L., Keilholz, U., and Rieke, D. T.: Support of a molecular tumour board by an evidence-based decision management system for precision oncology. Eur. J. Cancer *127*, 41–51. doi:10.1016/j.ejca.2019.12.017 (2020)

Leaman, R., and Lu, Z.: TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics *32*/18, 2839-2846. doi:10.1093/bioinformatics/btw343 (2016)

Lever, J., Jones, M. R., Danos, A. M., Krysiak, K., Bonakdar, M., Grewal, J. K., Culibrk, L. Griffith, O. L., Griffith, M., and Jones, S. J. M.: Text-mining clinically relevant cancer biomarkers for curation into the CIViC database. Genome Med. *11*/78. doi:10.1186/s13073-019-0686-y (2019)

Pallarz, S., Benary, M., Lamping, M., Rieke, D., Starlinger, J., Sers, C., Wiegandt, D. L., Seibert, M., Ševa, J., Schäfer, R., Keilholz, U., and Leser, U.: Comparative analysis of public knowledge bases for precision oncology. JCO Precis. Oncol. *3*, eCollection 2019. doi:10.1200/PO.18.00371 (2019)

Perera-Bel, J., Hutter, B., Heining, C., Bleckmann, A., Fröhlich, M., Fröhling, S., Glimm, H., Brors, B., and Beissbarth, T.: From somatic variants towards precision oncology: Evidence-driven reporting of treatment options in molecular tumor boards. Genome Med. *10*/1, 18. doi:10.1186/s13073-018-0529-2 (2018)

Seshacharyulu, P., Ponnusamy, M. P., Haridas, D., Jain, M., Ganti, A. K., and Batra, S. K.: Targeting the EGFR signaling pathway in cancer therapy. Expert Opin. Ther. Targets *16*/1, 15–31. doi:10.1517/14728222.2011.648617 (2012)

Ševa, J., Wiegandt, D. L., Götze, J., and Lamping, M.: VIST – A Variant-Information Search Tool for precision oncology. BMC Bioinformatics *20*/429. doi:10.1186/s12859-019-2958-3 (2019)

Starlinger, J., Kittner, M., Blankenstein, O., und Leser, U.: How to improve information extraction from German medical records. it - Information Technology *59*/4, 171–180. doi:10.1515/itit-2016-0027 (2016)

Starlinger, J., Pallarz, S., Ševa, J., Rieke, D., Sers, C., Keilholz, U., and Leser, U.: Variant Information Systems for Precision Oncology. BMC Med. Inform. Decis. Mak. *18*/1, 107. doi:10.1186/s12911-018-0665-z (2018)

Wagner, A. H., Walsh, B., Mayfield, G., Tamborero, D., Sonkin, D., Krysiak, K., Deu-Pons, J., Duren, R. P., Gao, J., McMurry, J., Patterson, S., del Vecchio Fitz, C., Pitel, B. A., Sezerman, O. U., Ellrott, K., Warner, J. L., Rieke, D. T., Aittokallio, T., Cerami, E., Ritter, D. I., Schriml, L. M., Freimuth, R. R., Haendel, M., Raca, G., Madhavan, S., Baudis, M., Beckmann, J. S., Dienstmann, R., Chakravarty, D., Li, X. S., Mockus, S., Elemento, O., Schultz, N., Lopez-Bigas, N., Lawler, M., Goecks, J., Griffith, M., Obi, L., Griffith, O. L., Margolin, A. A., and *Variant Interpretation for Cancer Consortium*: A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. Nat. Genet. *52*/4, 448–457. doi:10.1038/s41588-020-0603-8 (2020)

Weber, L., Thobe, K., Lozano, O. A. M., Wolf, J., and Leser, U.: PEDL: extracting protein-protein associations using deep language models and distant supervision. Bioinformatics *36*/Supplement 1, i490–i498. doi:10.1093/bioinformatics/btaa430 (2020)

Wei, C.H., Allot, A., Leaman, R., and Lu, Z.: PubTator central: automated concept annotation for biomedical full text articles. Nucleic. Acids Res. *47*/W1, W587–W593. doi:10.1093/nar/gkz389 (2019)

Prof. Dr. Ulf Leser
Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik
Wissnesmanagement in der Bioinformatik
Rudower Chausee 25, Raum 4.401
D-12489 Berlin
Germany
Tel.: +49 30 2093-3902
Fax: +49 30 2093-5484
E-Mail: leser@informatik.hu-berlin.de