Kalpa Publications in Computing

Volume 6, 2018, Pages 43–47

WDAM-2017. Workshop on Data Analysis in Medicine

New Reality in Clinical Informatics and Explanation-oriented Methods of Data Analysis

A.A. Neznanov

National Research University "Higher School of Economics", Moscow, Russia
ANeznanov@hse.ru

Abstract. Clinical informatics has been undergoing radical transformation. What are the causes and the drivers of this transformation? Which task can be solved well, and which cannot? How we should implement data analysis in clinical informatics projects in new reality? What is an importance of interpretability (comprehensibility) and explanation of data analysis methods in clinical informatics? At the workshop, we will try to answer some of such questions and setup a framework for later discussion.

Keywords: Clinical Informatics, Clinical Decision Support System, Ontology, Data Analysis, Text Mining, Formal Concept Analysis, Software.

1 Introduction

Clinical informatics (CI) [1] as a part of medical informatics has a deep development and a high impact on healthcare during the last years. Evidence-based medicine establishes one of the methodological foundations in modern CI. Evidence-based medicine (EBM) is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients [2]. CI not only implements EBM principles with high effectiveness and automatizes many process in clinical practice but also improves patient safety [3]. Another foundation is a data gathering and analysis methodology that now is known as Data Science. Data Science is an interdisciplinary field that combines machine learning, statistics, advanced analysis, and programming. It is a new form of art that draws out hidden insights and puts data to work in the cognitive era [4].

In the report, we will focus on workflows of data analysis tasks in clinical informatics. Also, we consider modern achievements in adjacent domains that could influence goals, structure, methods and tools of such workflows: text and graph mining methods, image processing, modern databases and Big Data, interoperability of medical information systems and clinical decision support systems, etc.

O.S. Pianykh, A. Neznanov, S. Kuznetsov, J. Baixeries and S. Boytcheva (eds.), WDAM-2017 (Kalpa Publications in Computing, vol. 6), pp. 43–47

2 Breakthrough in Clinical Informatics: New Reality

A breakthrough in CI at 2013 (approximately) consists in several achievements in different fields with high synergy.

1. Real **open data** in medical informatics. See "Project Data Sphere" (http://www.projectdatasphere.org), "NCI's Genomic Data Commons" (http://gdc.cancer.gov), "BMJ Open Data" (http://www.bmj.com/open-data), etc.
2. Integration of **main medical ontologies**. See "UMLS" – the project for linking health information (http://www.nlm.nih.gov/research/umls); "SNOMED CT" – the most comprehensive, multilingual clinical healthcare terminology (http://www.snomed.org/snomed-ct); "MeSH" (Medical Subject Headings) – the thesaurus used for indexing articles for PubMed (http://www.nlm.nih.gov/mesh); "LOINC" – the international standard for identifying health measurements, observations, and documents (http://loinc.org/), "ICD-10-CM" – 10th revision of the International Statistical Classification of Diseases and Related Health Problems, Clinical Modification (http://www.cdc.gov/nchs/icd/icd10cm.htm); "RxNorm" – normalized names for clinical drugs and links its names to many of the drug vocabularies (http://www.nlm.nih.gov/research/umls/rxnorm); etc. It is interesting that previous breakthrough had took place near 1998 – almost 20 ears ago.
3. Upsurge of interest in **clinical decision support systems** (CDSS) because of developments in data analysis and machine learning.
4. **Open API** to access health data including API for mobile devices. For example, see Healthcare API catalog at ProgrammebleWeb (http://www.programmableweb.com/category/healthcare).
5. Wide discussion of **legislation changes**. Even in Russia federal government at last approved so called "Bill on Telemedicine" [5] and the corresponding federal law will come into force from 2018.
6. Close attention to **data quality** in general [6] and randomized controlled trials (RCT) results quality in particular [7]. See reproducibility, provenance, and data citation discussions and efforts [8,9].
7. First practical and scalable achievements in **personalized medicine** [10].

Finalizing this short list of the achievements we can postulate accomplished revolution which change constraints and abilities in CI.

3 Data Analysis in Clinical Informatics

We suggest following high-level systematization of medical data from the point of view of an analyst.

1. **Raw data**, gathered by any healthcare institution in any case (frequently after a time and/or with hindsight). There is also a truism about 80% of unstructured data (see classical explanation by Seth Grimes [11]).
2. **Formalized online data**, gathered in real-time in EHR systems, certified automated lab systems and other medical information systems on the laboratory/clinical center level.
3. **Demographic data**, accumulated in citizen-oriented population registers on the regional/national-wide level.
4. **Epidemiological data**, for linking infectious diseases with geoinformation systems.
5. **RCT data**, accumulated in special clinical registers while conducting *randomized controlled trials* (RCT).
6. **Data of meta-analyses, systematic review, and ontologies**, as a highest level of formalized clinical knowledge.

7. **Metaontologies and thesauruses**, as sources of clinical entity's description and semantic links between them.

Now we have an ability to analyze effectively all kinds of data due to the last technological revolution in Data Science. Firstly, the comprehensible implementation of basic methods and workflows have been developed and extensively tested for numerical, textual, and multimedia clinical data. Secondly, "AI as a Service" approach has received recognition. The most significant examples are Microsoft Cognitive Services (http://azure.microsoft.com/en-us/services/cognitive-services) and IBM Watson Services (http://www.ibm.com/watson/products-services). Thirdly, new approaches for clinical information systems interoperability have been adopted by main vendors. Now we can have direct links between data producers like laboratory equipment, data warehouses like PACS, and analytics components.

Consequently, new expectations and requirement are discussed during design and implementation of modern CDSSs as part of integrated clinical information systems. New generation of CDSSs becomes more intelligent and more interactive. Most of CDSSs also become more specialized, but they can be integrated through common ontologies.

3.1 Well-explained methods of data analysis

The choice of which data analysis methods to use for some problem became an important question. As an example of real world problems, we can mention recommendations for the users of the Microsoft Machine Learning Studio [12] and repository with walkthroughs, templates and documentation [13].

Form the clinician's point of view, interpretability, durability and simplicity of the explanation of results are significant features of a data analysis method. In one recent article in MIT Technology Review [14] you can see provocative subtitle: "No one really knows how the most advanced algorithms do what they do. That could be a problem.". The illustrative discussion about classification tasks can be found in position paper by Alex A. Freitas [15] and a good survey on interpreting modern machine learning methods can be found in [16]. Some examples of clinical applications of different methods, form simple ANOVA [17] up to recurrent neural network [18], can also be mentioned.

3.2 Latest projects, tools and data workflows

In International laboratory for intelligent systems and structural analysis at National Research University Higher School of Economics we have developed several explanation-oriented methods and have implemented several software tools for data analysis with clinical informatics needs in mind. Most of our original methods are based on Formal Concept Analysis [19,20,21] for knowledge representation, modern statistics and machine learning.

For example, N. Korepanova and clinicians from Dmitry Rogachev National Research Center of pediatric hematology, oncology and immunology [22] proposed the method of finding subgroups of patients with significant difference in efficiency between treatment strategies based on pattern structures [23]. The approach is not biased by local optimization heuristics and allows one to avoid binarization of attributes or using similarity measures on patients, which can result in artifacts [24]. This approach also builds a good foundation for solving adjacent tasks, for example, to develop recommender system for patient's actions.

For our software tools we use different platforms and software tools:

1. Our original systems like FCART (formal concept analysis research toolbox) [25].
2. Various open source tools like programming environments Jupiter Notebook and R Studio, machine learning libraries XGBoost and CatBoost.

3. Proprietary tools for dealing with big data like Microsoft Azure Services: Microsoft notebooks, Microsoft Machine Learning Studio, Azure Storage, R Server, Power BI.

Comprehensibility of our applications depends not only on chosen analysis methods but also on data workflow properties. We support dataset immutability principle, rich metadata management, and intermediate results visualization. Also, we support integration of our tools with familiar clinician's everyday tools like Microsoft Excel.

One of our next goals is to support the additional kinds of data sources. Illustrative example is CPACS – cardiology picture archiving and communication system – with a zoo of formats: SCP-ECG, HL7 aECG, GE Muse, PhilipsXML, EDF+ etc. [26]. Another goal is to support Arden syntax of HL7 framework [27] for the next level of interoperability with clinical information systems.

4 Conclusion

In the report, we described the latest progress in clinical informatics and postulated revolutionary changes in last 4 years.

We tried to cover the field and to show the real examples of discussions and projects, including the examples of research projects in International laboratory for intelligence systems and structural analysis at NRU HSE.

Acknowledgements

The report was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'.

References

1. AMIA, Clinical Informatics (https://www.amia.org/applications-informatics/clinical-informatics)
2. Sackett D.L., Rosenberg W.M., Gray J.A., Haynes R.B., Richardson W.S. Evidence based medicine: what it is and what it isn't. BMJ, 1996, 312:71.
3. Kilbridge P.M., Classen D.C. The Informatics Opportunities at the Intersection of Patient Safety and Clinical Informatics. JAMIA, 15(4), 2008, pp. 397-407.
4. IBM Analytics – Enterprise Data Science (http://www.ibm.com/analytics/us/en/technology/data-science/)
5. Федеральный закон от 29 июля 2017 г. N 242-ФЗ "О внесении изменений в отдельные законодательные акты Российской Федерации по вопросам применения информационных технологий в сфере охраны здоровья" (http://rg.ru/2017/08/04/zdorovie-dok.html)
6. AHIMA. Data Standards, Data Quality, and Interoperability (2013 update) (http://library.ahima.org/doc?oid=107104#.Wb4_7M3TR3g)
7. ProvCaRe – A Provenance Enabled Framework (http://provcare.case.edu)
8. Task Group on Data Citation Standards and Practice. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal. 12, pp. CIDCR1–CIDCR7.
9. P. Buneman, S. Davidson, J. Frew, Why Data Citation Is a Computational Problem, Communications of the ACM, 59 (9), 2016, pp. 50-57.
10. Hays P. Advancing Healthcare Through Personalized Medicine. CRC Press, 2017.
11. Grimes S. Unstructured Data and the 80 Percent Rule, 2008 (http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/)

12. Microsoft. How to choose algorithms for Microsoft Azure Machine Learning. 2017 (http://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-choice)
13. Azure-MachineLearning-DataScience GitHub Repository (http://github.com/Azure/Azure-MachineLearning-DataScience)
14. Knight W. The Dark Secret at the Heart of AI. 2017 (http://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/)
15. Freitas A.A. Comprehensible classification models: a position paper. SIGKDD Explorations Newsletter 15(1), 2014, pp. 1-10.
16. Hall P., Phan W., Ambati S. "Ideas on interpreting machine learning", 2017 (http://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning)
17. Van Belle V.M.C.A., Van Calster B., Timmerman D., Bourne T., Bottomley C, Valentin L., Neven P., Van Huffel S., Suykens J.A.K., Boyd S. A Mathematical Model for Interpretable Clinical Decision Support with Applications in Gynecology. PLoS ONE, 7(3):e34312, 2012.
18. Sha Y., Wang M.D. Interpretable Predictions of Clinical Outcomes with An Attention-based Recurrent Neural Network. The 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics, 2017. (preprint)
19. Ganter, B., Wille R. Formal Concept Analysis: Mathematical Foundations, Springer, 1999.
20. Poelmans J., Kuznetsov S.O., Ignatov D.I., Dedene G. Formal Concept Analysis in knowledge processing: A survey on models and techniques, Expert Systems with Applications, 40 (16), 2013, pp. 6601-6623.
21. Poelmans J., Ignatov D. I., Kuznetsov S.O., Dedene G. Formal concept analysis in knowledge processing: A survey on applications, Expert Systems with Applications, 40 (16), 2013, pp. 6538-6560.
22. Karachunskiy A., Roumiantseva J., Lagoiko S., Bührer C., Tallen G., Aleinikova O., Bydanov O., Korepanova N, Bajdun L., Nasedkina T., von Stackelberg A., Novichkova G., Maschan A., Litvinov D., Myakova N., Ponomareva N., Kondratchik K., Fechina L., Streneva O., Judina N., Scharapova G., Shamardina A., Gerbek I., Shapochnik A., Rumjanzew A., Henze G. Efficacy and toxicity of dexamethasone vs methylprednisolone – long-term results in more than 1000 patients from the Russian randomized multicentric trial ALL-MB 2002, Leukemia, 2015, 29(9), pp. 1955-1958.
23. Korepanova N.V., Kuznetsov S.O. Pattern Structures for Treatment Optimization, 13th International Conference on Concept Lattices and Their Applications, 2016, pp. 217-228.
24. Kuznetsov, S.O. Scalable Knowledge Discovery in Complex Data with Pattern Structures, 5th International Conference Pattern Recognition and Machine Intelligence (PReMI'2013), 2013, pp. 30-39.
25. Neznanov A.A., Parinov A.A. Distributed Architecture of Data Analysis System based on Formal Concept Analysis Approach, Intelligent Distributed Computing IX, 2015, pp. 265-271.
26. Trigo J.D., Alesanco A., Martınez I., Garcıa J. A Review on Digital ECG Formats and the Relationships Between Them, IEEE Transactions on Information Technology in Biomedicine, 16(3), 2012, pp. 432-444.
27. Arden Syntax v2.9 (Health Level Seven Arden Syntax for Medical Logic Systems, Version 2.9), 2013 (http://www.hl7.org/implement/standards/product_brief.cfm?product_id=290)