Kalpa Publications in Computing

Volume 6, 2018, Pages 61–66

WDAM-2017. Workshop on Data Analysis in Medicine

Big Data and machine learning in medicine: the main problems

Shchelykalina S.P., Kiselev K.V., Zarubina T.V.

Pirogov Russian National Research Medical University(RNRMU), Moscow, Russia
svetlanath@gmail.com

Abstract. Big data and deep learning technologies play an important role in the modern scientific world. The tendency to work with huge data sets is now conquering the medical area. In this article, based on the experience of the Department of medical cybernetics and informatics of the RNRMU Medical and biological Faculty, we explain the main issues that researchers deal with in collection and processing of medical data. We explain that problems may relate to data sources issues, semantic interoperability, data relevance, multidimensionality, completeness, and comparability. Modern digital health records and their services like EHR nowadays cannot provide necessary "Big Data" information. The healthcare system makes it impossible to collect relevant big data sets in a short period. Further issues are certain irresponsibility of doctors and patients; their truthfulness about facts happened in reality and the difference between these facts and what is written in a medical record. This often leads to incorrect and incomplete data sets in medical information systems. We conclude by stating that "Big Data" in medicine today cannot be "Big" as in other scientific areas. Researchers should try to collect relevant, truthful, and complete information in observable amount and time and perform their studies.

Keywords: big data, deep learning, data analysis, medicine, medical data, medical information systems, experiment design problems.

1 Introduction

The Department of medical cybernetics and informatics of the RNRMU Medical and biological Faculty has been dealing with problems of computational diagnostics and prediction for the needs of the Russian healthcare service for more than 40 years. This time allowed gaining huge experience in solving diverse problems in functional diagnostics, cardiology, intensive care, surgery, neurology, pediatrics, preventive medicine, healthcare service management, and other branches of medicine.

Today, due to the wide spread of computer devices and their usage in the healthcare system, interest towards statistical analysis of big medical data has dramatically increased. For example, the number of publications within the last 5 years increased more than 17 times – from 29 in 2012 to 509 in 2016 (see Fig. 1). In Russia, this tendency is supported by the governmental focus on informatization of the

O.S. Pianykh, A. Neznanov, S. Kuznetsov, J. Baixeries and S. Boytcheva (eds.), WDAM-2017 (Kalpa Publications in Computing, vol. 6), pp. 61–66

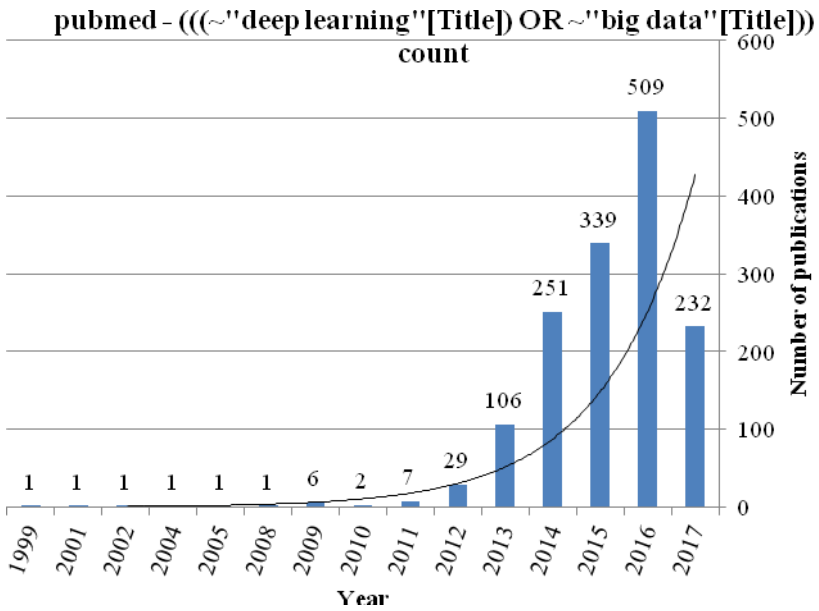

Fig. 1. Number of recent medical publications using Big Data methods and Deep Learning (by PubMed, 25 May 2017)

healthcare system. The recent statement of Minister of Health about plans to use Big Data in medicine in Russia from 2020 is indicative of the strength of this tendency [1].

In this regard, the aim of this paper is to summarize experience in medical data collection and analysis in order to assess prospects for the Big Data usage in medicine.

2 Types of Problems

2.1 Problems of data sources and subject area

The main data source about the health of a patient is a medical record. Today, there are both paper and digital medical records. When using medical records as a data source for data set formation, the following problems may arise:

1. Paper medical records are not possible to digitalize in high quality. Doctors are notorious for their bad handwriting, thus only specialists are capable of reading medical records, however, even then not always. The digitalization process is further challenged by a considerable volume of graphic material like ECG curves, X-rays, CT and MRI images, etc. On top of that, with time, paper data sources lose their color, get torn, and partly or completely lose the original information. For example, many modern cardiographs and laboratory equipment used for analysis of biological substances and tissues, print results on thermal paper. Depending on storage conditions, this paper can lose its color until it is not readable at all within half a year or a year. In some cases, this may happen within a few months.

2. The logical solution seems to be the usage of digital medical records as a data source for medical Big Data research (Electronic Health Record (EHR), Electronic Medical Record (EMR), Personal Health Record (PHR)). However, upon closer view, it turns out that digital records from different developers substantially vary in structure and often collect information mostly in a form of a plain text. The current projects of national standards and GOST P 52636-2006 do not define the structure and content of clinical protocols for Medical Health Records (MHR) [2-3].

3. Moreover, EHR/MHR implementation is not complete [4]. According to the official website of Mayor of Moscow, last year only 32% of doctors in outpatient clinics used the EHR service of UMIAS (United Medical Information and Analytics Service of Moscow) [5]. The scope of tasks and quality of support of medical workers by MHR/EHR systems is also not always satisfactory. Moreover, implementation of even a good medical record cannot be fast and comprehensive due to financial constraints and resistance of employees, which is well known to all IT-professionals [6-7]. Another reason is the need to adapt the system of digital medical records for a specific healthcare facility. The quality of the content (completeness) of primary medical documentation obviously depends on the time available. For example, a therapist can only have 15 minutes for initial examination of each patient (including filling in of a digital medical record) [8]. Obviously, filling in of a medical record under such conditions will be patterned or with many gaps in "irrelevant" for this specific clinical case places (in the opinion of this specific doctor).

4. Semantics in medical language sometimes differs from everyday language and languages of other subject areas. For example, "ядрышко" (nucleolus) is not a diminutive form of "ядро" (nucleus). These are separate structures in a living cell, connected by a relation of "the whole entity-part". The clinical parameter "P-wave dispersion" (in Russian "дисперсия P зубца") on an ECG has nothing to do with normal distribution parameters and formulas of variance (in Russian "дисперсия" too) and standard deviation. For doctors, "P-wave dispersion" is the difference between the maximal and minimal value of the P-wave duration on one ECG record.

5. Another issue is almost complete absence of semantic interoperability among medical information systems. Many of required reference books have not been created yet. Existing ones are often suitable for formation of reports, but not for deep scientific research as they use very broad language. Creation of such reference books is complicated by complete or partial synonyms. For example, the term "omega-3 polyunsaturated fatty acids" has more than 20 synonyms.

2.2 Problems with relevance of collected data

Even if we could collect a considerable data set, these data are most likely to contain inaccuracies and defects. There can be various reasons for these defects:

1. Data from a data set do not often represent the population or its certain part. Not all sick people go to see a doctor.

2. Those patients who do visit their doctors do not always follow prescriptions and many of them do not notify doctors about their decisions to change their treatment plan. Therefore, the information contained in medical records does not always reflect the reality and is very often incomplete.

3. In some cases, physicians themselves intentionally put inaccurate data due to the patient's conditions. For example, they may record that the patient's condition is "worse" than it is in reality, in order to get better legal justification for additional diagnostics, consultations, patient reference to specialized departments or facilities to exclude presumptive serious diseases. Any doubt is interpreted in support of a more dangerous outcome.

4. Data sets can contain false data. These may be falsified data used to improve reports to show full completion of planned work thus protecting doctors from penalties or dismissal. Also, very often we can see randomly distorted data: data of "test" patients. In some cases, false data can take up about

45% of the whole data set [9]. Some types of false data can be fixed, but a part of them still remains in the data set, because a certain incorrect record, especially the one intentionally made, does not differ from correct records. Sometimes, it is possible to indirectly find intentionally made false data because of incompliance between expected and observed distributions and too frequent certain values of certain parameters or their associations.

5. Attempts to perform analysis based on retrospective data also lead to false data sets. One of the most noticeable problems with retrospective data is the "time shift". During the data collection time, many important factors could have changed or did change in reality. These could be tracked parameters such as diagnostics and treatment methods, population structure, or hard-to-assess parameters such as ecology, living conditions, economics, as well as factors, the impact of which we do not even think of. Over time, notions of norm and pathology, health and disease considerably change. The second most frequent problem is so called "gaps" in retrospective and present data due to the absence of the diagnostic method, its low usage or unavailability, and due to information loss.

2.3 Problems of multidimensionality, completeness and comparability of data

When collecting data, it is essential to take into account, already at the planning stage, multidimensionality of medical data. Sources of multidimensionality are also diverse and can be explained by the biological nature of parameters, human factor or equipment-related issues, etc.

1. For the classical description of a specific condition, a comparatively high number of parameters are used, although limited. However, conditions can and almost always form different combinations. Parameters of a "classical case" are subject to complex mutual influences, often poorly understood by the strict medical science. Parameters themselves dynamically change under influence of physiological processes. As a norm, they substantially differ among various healthy people and even in the same person. Therefore, medical data must be fantastically multidimensional. Often, researchers cannot collect information about all valuable parameters also because they do not know themselves what the final set of parameters should be.

2. Many parameters, tracked by doctors, are under influence of circadian, weekly, seasonal, and other cyclic changes. Their impact depends on a certain individual and is not thoroughly studied, therefore it cannot be fully corrected mathematically [10]. At the same time, in some areas, this knowledge is already present in an explicit form and can be reflected in databases. For example, the female hormonal cycle is relatively well studied, including its impact on metabolism, endocrine, digestive, excretory systems, cognitive and emotional functions, etc.

3. Many parameters, tracked by doctors, are also quite subjective. This refers not only to just qualitative parameters, such as, for example, character and strength of pain, but also to quantitative parameters, such as the metric dimensions of chambers of heart during echocardiography. This factor leads to the need to take into account information about the subject – the doctor who performed the test.

4. Medical conclusions themselves are subjective, especially in analysis of CT, MRI, PET, ultrasound images, etc. Each specialist interprets the results differently in different moments of time: for example, interpretation of ECG by different specialists is reproduced by 80%.

5. Another aspect of medical data multidimensionality is the necessity to take into account characteristics of the equipment as additional parameters. Each device has its own settings and ways to evaluate values of parameters. In case of laboratory parameters, such as blood composition, for example, apart from characteristics of the equipment, reference ranges of a certain purchased kit of reagents should be taken into consideration, as they can vary, depending on the lot and manufacturer.

6. Physiological mobility of parameters further limits attempts to compare the results of tests for the same patient: studies must be performed within a relatively short period. Therefore, dates and time of tests, examinations, etc, are also valuable data parameters. The speed of providing diagnostic pro-

cedures in practical medicine is often too low to enable strict comparison of data of several methods. At the outpatient clinic level, it is almost impossible to organize synchronous performing of several surveys due to the queue system and, frankly speaking, due to irresponsibility of patients. For example, one has to wait for an echocardiography for about a month. Within this time some functional changes may appear and disappear. If we follow the condition that the time period between the digital ECG and echocardiography should not exceed one week, then we see the number of records, suitable for analysis, drop thrice: from 729 (both tests presented, continuous data collection in an outpatient clinic in Moscow within 5 years, 2006-2010) to 250.

7. An attempt to use qualitative assessment "norm or pathology" leads to, firstly, debates about what a "norm" is, and secondly, to the necessity to take into account characteristics of subpopulations, that is, to use standards, localized in space and time.

Multidimensionality generates the need to collect a huge number of research objects. At the same time, the collected material breaks down into smaller groups, often with a high percentage of "gaps". Rare diseases are very difficult to study this way. Problems appear even with common pathologies, especially in case of such a popular and perspective direction as prediction of pathological conditions. For example, when trying to study prognostic signs of acute myocardial infarction using the clinic database of 5 years with more than 50 thousand records of 20 thousand unique patients, we could actually include in our research only 266 records from 70 patients, with the most frequent form of myocardial infarction shown only 43 times (26 in men and 17 in women) [11].

Therefore, medical data sets are rather "thick" than "big": the number of parameters very often exceeds or is equal to the number of observations. Additional challenges and restrictions for data collection are caused by the structure of the healthcare system, legal aspects, personal data security, medical privacy, etc.

3 Conclusions

Taking into account all the above, it is clear that data sets for Big Data analysis can only be collected with the usage of super secure electronic health record, strict equipment standardization, and rapprochement of "medical schools". In practice, nowadays, this means that for a researcher it is better to use data of a few years from similar clinics of the same "medical school" that utilize mostly the same equipment and medical information systems working with the same subpopulations. However, in this case, "Big Data" will mean only several tens of thousands of records.

For example, a study was conducted in one of Moscow's oncological hospitals, where the EHR/MHR system had been operating for a long time [12]. In this hospital for 3 years (2011-2014), 4,884 patients underwent 30,024 courses of chemotherapy. After the exclusion of courses with incomplete and incorrectly filled in plans and protocols of chemotherapy and patients with less than two blood tests before and after the course of chemotherapy, the database for the analysis significantly decreased: there remained only 3,193 patients and 15,504 courses of chemotherapy.

References

1. TASS, http://tass.ru/obschestvo/4113127, last access 27/05/2017.
2. Consultant plus, http://www.consultant.ru/cons/cgi/online.cgi?req=doc;base=OTN;n=9320#0, last access 27/05/2017.
3. EGISZ, http://portal.egisz.rosminzdrav.ru/files/%D0%9F%D1%80%D0%BE%D0%B5%D0%BA%D1%82%20%D

0%93%D0%9E%D0%A1%D0%A2%20%D0%AD%D0%BB%D0%B5%D0%BA%D1%82%D1%80%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F%20%D0%BC%D0%B5%D0%B4%D0%B8%D1%86%D0%B8%D0%BD%D1%81%D0%BA%D0%B0%D1%8F%20%D0%BA%D0%B0%D1%80%D1%82%D0%B0.%20%D0%A2%D0%B5%D1%80%D0%BC%D0%B8%D0%BD%D1%8B%20%D0%B8%20%D0%BE%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D1%8F.docx, last access 27/05/2017.

4. RBC, http://www.cnews.ru/reviews/publichealth2016/articles/sozdanie_rossijskogo_ehealth_tormozitsya_otsutstviem_deneg_i, last access 27/05/2017.
5. Moscow mayor site, https://www.mos.ru/mayor/themes/18299/3437050/, last access 27/05/2017.
6. Paschemno D, Implementation of changes problems during software development. Izvestiya Tulskogo gosudarstvennogo universiteta. Economicheskie i uridicheskie nauki 4-1, 303-314 (2014).
7. Estratova E, Lastochkin P: Specificity of medical information systems implementation in healthcare facilities. Medicina i obrazovanie s sibiry, 6 (2014).
8. Russia Ministry of Health, https://www.rosminzdrav.ru/documents/9082-prikaz-ministerstva-zdravoohraneniya-rossiyskoy-federatsii-ot-2-iyunya-2015-g-290n-ob-utverzhdenii-tipovyh-otraslevyh-norm-vremeni-na-vypolnenie-rabot-svyazannyh-s-posescheniem-odnim-patsientom-vracha-pediatra-uchastkovogo-vracha-terapevta-uchastkovogo-vracha-obschey-praktiki-semeynogo-vracha-vracha-nevrologa-vracha-otorinolaringologa-vracha-oftalmologa-i-vracha-akushera-ginekologa, last access 27/05/2017.
9. Rudnev SG et al.: Health Centres: technology to process mass data on preventive screening. Social aspects of population health 6(46), 1–19 (2015).
10. Soroko S, Long term variation of cosmic rhythms influence on biochemical parameters in humans. Phisiologiya cheloveka 30(1), 82-94 (2004).
11. Chernykh SP et al.: Estimation of informative value of electrocardiogram parameters for diagnosis of Q-forming myocardial infarction. Functional diagnostics 3, 46–47 (2010).
12. Penzin OV et al.: Prediction of severe myelotoxic complications of cancer chemotherapy based on clinical and laboratory data. Journal of new medical technologies 3(30), 67–75, (2016).