BIOS POLICY BRIEF

From big data to big impact– challenges of clinical registries

08/2020

Evidence and Analysis

Big data analysis offers an innovative solution to clinical research (Cacciola et al., 2019). However, at present and from a technical view, medical-scientific research is "not yet" dealing with "Big Data", which is measured in Petabyte (1015 byte). For example, the video platform YouTube handles 27 Petabyte each month. In medical care, the term "Big Data" is only loosely used to simply indicate "lots of" digital data, but the current production of data could soon reach those numbers (Tavazzi 2019).

Big Data is characterised by 3 "V's" – greater volume due to data-intense modalities; archives which are rapidly growing and processed at greater velocity; and data deriving from number and variety of sources. The fourth "V" of Big Data, recently introduced, veracity, dealing with differing interpretation and quality of data, is of fundamental importance as accuracy and trustfulness of decisions drawn from data is a must in medicine (Cacciola et al., 2019).

Big data, or better "huge" datasets, in medicine derive from electronical medical records, clinical registries, administrative datasets, genomic databases and wearable sensors amongst others, which can now be previously indiscernible using conventional analytic methods (Scott 2019).

More recently, in several countries, clinical (quality) registries have been activated for specific conditions, addressing both hospital-based and outpatients' practices. Clinical quality registries are an important component of healthcare surveillance and improvement and they also have a valuable mechanism for research (Aitken 2019).

Registers should also help to ensure traceability (especially in the event of damage). The mandatory creation and maintenance of such a register is contained in the Medical Devices Act (§73).

Thus, registries summarize the results of the management of a certain disease, managed in a specific way, the monitoring and quality of caring, and of tools, implants, or techniques used to treat a patient. Clinical registry databases include thousands of patients, and thus do not necessarily represent "Big Data". But the role of clinical databases as "observational medicine" has become the core of health systems, and observational scientific research is its guiding tool (Tavazzi 2019).

However, registries are a collection of data, and thus, are only as good as the fidelity of the data that are collected. This is valid for all (medical) data bases. While the paper file

Co-funded by the Erasmus+ Programme of the European Union

has now given way to various digital platforms, the fundamental constant for centuries has been the physician as collector, cultivator and keeper of patient data (George et al. 2019). Data is collected as "by-product" of routine transactions and very often are not part of a research protocol (Pass 2010). To be useful, data must be authentic and of good quality.

Challenges

Big data, its analyses, and thus the use of clinical registries, has limitations. If these limitations are not addressed, they can place patients at risk – from care derived from erroneous algorithms.

Hence, the enemy of each database is error. Clinical registry data is very often collected as a "by-product" of routine transactions. They may be handwritten, messy, do not include standardized variables, miss data, or include multiple data. Furthermore, databases contain structured as well as unstructured data, derive from different modalities and arrive at varying intervals in real time. Up to 70% of medical data come from unstructured, handwritten notes of a physician or nurse, which are inaccessible to analysis or must be transferred to data bases manually - which is a further source of error. Computational or analytic tools may be unable of processing handwritten data, and with confounding due to a lack of nuanced data. Furthermore, metadata are rarely available and this often makes the use or reuse of medical data from large databases impossible.

The more data is manipulated or transferred, the greater the chance for error. Using conventional analytical or statistical methods, trends and associations may be indiscernible. Previously unrecognised associations generated by Big Data will inevitably be spurious and misleading (e.g. false positives). Due to the enormous sample sizes of data bases, highly significant p-values for certain effects sizes are often generated, which should be cautiously interpreted and reasonably contested (Scott 2019).

In clinical registries, data is constrained by confounders and bias. Preoperative characteristics may differ between groups, although statistical techniques are able to control these differences. Nevertheless, unmeasured or unknown factors can influence the results.

Additionally, there exists the risk to create "calculated" conditions or computable phenotypes which are not applicable in clinical practice due to e.g. erroneous algorithms. Diseases may be fragmented into phenotypes that are clinically not relevant, or unrelated states are aggregated according to some phenotypic similarities (Tavazzi2019).

Furthermore, prediction rules drawn from Big Data-bases are more discriminatory, and may remain poorly calibrated when quantifying the risk in individual patients (Scott 2019).

None of these limitations are new to researchers but they are greatly magnified by Big Data. Indeed, high precision and statistical significance may be reported around wrong answers and misinterpreted as solid evidence (Scott 2019).

Large amounts of data, including those from clinical registries, coupled with machine learning/computing methods have the potential to effectively eliminate uncertainties in clinical medicine. Diagnoses, recommendations and predictions could provide "perfect" accuracy using computational methods - clinical judgment and the role of clinicians could be trivialized and reduced to a complex calculation performed by an "intelligent" computer (Scott 2019).

The future looks optimistic, but the expectations on doctors must be realistic and these aforementioned potential risks must be outlined. It is likely that the initial

Co-funded by the Erasmus+ Programme of the European Union

effects of the system on clinical practice will be disruptive, and it requires necessary adjustment time rather than incentives and sanctions (Tavazzi 2019).

However, the application of machine learning to Big Data may not necessarily overcome limitations of poor data and confounding (Scott 2019).

Tentative Showcase/Solution

An example of a well-functioning Big Data database with clear guidelines and an integrated quality management system is the Trauma Register DGU. The Trauma Register DGU data base sets standards worldwide for the quality management of severely injured patients. With over 800 participating hospitals, and institutions from more than 20 nations are taking part. The data base contributes to increasing safety and quality in the care of seriously injured patients.

On the one hand, the data base serves the participating clinics as an instrument for quality assurance; on the other hand, it can be used as a basis for further scientific work. Since its foundation in 1993, results data of almost 400,000 treatment courses have been documented. The TraumaRegister DGU is not only used by hospitals as an instrument for external quality assurance but has also been a basis for clinical and care research for years.

The register is based on standardised documentation in a central database. The data is recorded from the four consecutive phases A) pre-clinic, B) shock room and then OR, C) intensive care unit and D) discharge. The standard questionnaire requires the input of about 100 parameters per case. It contains detailed information on demography, injury patterns, comorbidities, preclinical and clinical management, intensive care history and important laboratory findings including transfusion data. Furthermore, outcome data such as the patient's condition after discharge are documented.

All participating hospitals are obliged to document certain services within the framework of the so-called external quality assurance. Clear guidelines and masks for data input help to avoid errors, bias and to identify confounders. Using a web-based application, the patient data is recorded anonymously in a central database. In order to ensure the highest possible data quality, a large part of the parameters is checked for plausibility as soon as they are entered.

Not only steady data growth but also growing interest in scientific evaluations from the Trauma Register DGU made controlled access to the data necessary. For this reason, since 2008 all evaluations from the Trauma Register DGU have been coordinated by a Review Board. This ensures the quality of the resulting work and avoids data misuse, for example for non-scientific purposes.

The participating clinics document their serious injuries and receive a detailed quality report once a year. In addition to information on process quality, such as the time required to perform a full-body CT or first emergency intervention, this report also contains comparative statements on the quality of results, such as the survival rate. This should enable a comparison with the overall figures of other hospitals and thus support the assessment of the current situation with regard to the hospital's own quality efforts.

The data from the complete register can be used by the clinics authorised to evaluate the data to produce scientific publications. In order to be authorised to evaluate, a clinic must have entered data using the standard form for at least two years. Applications are approved for evaluation by the Review Board of the TraumaRegister DGU. For the publication of scientific evaluations from the TraumaRegister DGU,

Co-funded by the Erasmus+ Programme of the European Union

the Board of the DGU has also adopted a binding publication guideline.

Policy Recommendations

In the near future, we will have to answer the following question: How can we integrate, validate and make new data sources trustworthy? This will be particularly important when large registry data conflict with traditionally accepted evidence-based medicine and research. New algorithms or validation mechanisms will be required to combine the knowledge from these sources and inform clinical decision making (George et al. 2019). While increasing availability, scope and amount of data, the Big Data flood will be of limited use if there is no conceptual framework that grounds the question to be asked and guides data collection, curation and interpretation.

Summarizing, the following major challenges of clinical registries or other Big Databases are envisioned: findability, accessibility, interoperability, reusability, and veracity. The first four principles are equally part of the FAIR principle for scientific data (Wilkinson2016).

FAIRness is a prerequisite for proper data management. Data needs to be findable, and thus need to be assigned by a persistent identifier. They should be enriched and described with meta-data to allow further use or linking to the same and different clinical specialities. Data need to be accessible at open, free, and universally implementable protocols, allowing an authentication or authorization, where necessary, to follow privacy policies. Data needs to use a formal, accessible, shared and applicable language or vocabularies to make them interoperable and thus should assure communication and interoperability among the components of same and different clinical specialities. Data must be prepared in such a way that it can be used over and over again and is therefore reusable in order to save costs and not burden patients with repeated research measures.

Finally, veracity, the fourth and most fundamental "V" of big data in medicine deals with the interpretation and quality of data (Cacciola et al. 2019). Advanced algorithms to interpret and meaningful aggregate free text data will need to be developed and used that move beyond methods used for binary and defined data typed (George et al. 2019) However, narrative skills of recognising, absorbing, interpreting and responding to stories of illness will remain integral to clinical reasoning – in recognition of the irreducible uncertainty of clinical medicine.

The US scientific community has already placed medical (quality) registries at the centre of quality-based medicine. They give basic principles which are recommended for the use of registries. First, best clinical practices should base on methodologically correct evidences and second, should measure fatal as well as non-fatal of outcomes including a systematic patient follow up. Third, validated techniques for data-quality control and in particular a standardization of nomenclature should be available starting from the event`s definition. Fourth, medical registries shall include clinical data and not only administrative data, which shall be used to improve quality of care and outcomes. Fifth, a feedback useful for clinicians should be derived from registries and sixth and finally, registries should always consider the complexity and frailty of the patients (Tavazzi 2019).

Whether Big Data outputs are readily interpretable to clinicians, are sufficiently sensitive to values, preferences, and risk perceptions of individuality, and can influence clinical decision making that leads to improved patient outcomes, need to be impartially assessed. Predictive algorithms generated from Big Data need to be tested prospectively before widespread adoption.

So there is no doubt that one need to train aspiring clinical doctors with programming and computer skills, and also ensure an advanced level of computational know-how

Co-funded by the Erasmus+ Programme of the European Union

to understand the concepts behind the theory and the computational software used to extract information from biomedical data sets. They will also develop innovative approaches that may lead to a paradigm shift in healthcare. This requires trifold time and effort and passion from both clinicians and engineers to face easily accessible and high-yield didactic seminars on machine learning, in-depth learning and the development of computational methods.

BioS Project

This Policy Brief was created by the BioS Consortium, managing the BioS Project "Digital Skills on Computational Biology". The project aims at advancing the digital skills of medical doctors through the design, development and delivery of new modular vocational curricula on Computational Biology & Bioinformatics. The BioS Course provides basic knowledge about methods and tools of bioinformatics as well as computer-aided statistics with a focus on the interpretation of biomedical data. By participating in the BioS course, health care professionals learn how to use genetic information services in their daily work with patients.

The course is web-based and free of charge.

Interested learners can register under: https://moos.bios-project.eu

BioS at a glance

Project Name: BioS: Digital Skills on Computational Biology

Consortium: Steinbeis University Berlin (SHB), Enios Applications Idiotiki Kefalaiouchiki Etaireia (e-NIOS), OLYMPIC TRAINING AND CONSULTING LTD (OT), Skybridge Partners, Bioinformatics Barcelona Association (BIB), University of Patras (UPAT), European Medical Association (EMA), European Recreation and Health Valley (EUREHVA), BG Klinikum Murnau gGmbH (BGU Murnau), FOR SRL, HiDucator Ltd, EPRALIMA_Vocational School of Alto Lima, C.I.P.R.L. (EPRALIMA), German Oncology Centre (GOC)

Duration: 01.01.2018 – 31.12.2020

Funding Source: EACEA, Erasmus+ / Sector Skills Alliances Programme

Website: https://www.bios-project.eu/