

Mini Review

Tim Hulsen, David Friedecký, Harald Renz, Els Melis, Pieter Vermeersch* and Pilar Fernandez-Calle*

From big data to better patient outcomes

<https://doi.org/10.1515/cclm-2022-1096>

Received October 30, 2022; accepted December 12, 2022;

published online December 22, 2022

Abstract: Among medical specialties, laboratory medicine is the largest producer of structured data and must play a crucial role for the efficient and safe implementation of big data and artificial intelligence in healthcare. The area of personalized therapies and precision medicine has now arrived, with huge data sets not only used for experimental and research approaches, but also in the “real world”. Analysis of real world data requires development of legal, procedural and technical infrastructure. The integration of all clinical data sets for any given patient is important and necessary in order to develop a patient-centered treatment approach. Data-driven research comes with its own challenges and solutions. The Findability, Accessibility, Interoperability, and Reusability (FAIR) Guiding Principles provide guidelines to make data findable, accessible, interoperable and reusable to the research community. Federated learning, standards and ontologies are useful to improve robustness of artificial intelligence algorithms working on big data and to increase trust

in these algorithms. When dealing with big data, the univariate statistical approach changes to multivariate statistical methods significantly shifting the potential of big data. Combining multiple omics gives previously unsuspected information and provides understanding of scientific questions, an approach which is also called the systems biology approach. Big data and artificial intelligence also offer opportunities for laboratories and the *In Vitro* Diagnostic industry to optimize the productivity of the laboratory, the quality of laboratory results and ultimately patient outcomes, through tools such as predictive maintenance and “moving average” based on the aggregate of patient results.

Keywords: artificial intelligence; big data; data science; patient outcomes; personalized healthcare; precision medicine.

Introduction

In Medicine, much information is needed for adequate patient care. This information can be configured in unstructured (e.g., physician notes, data from medical imaging devices, echography, robotic surgery, patient monitor in intensive care units, wearable devices, biosensors) and structured data (e.g., laboratory result database). These structured data are based on a programming language called Structured Query Language (SQL). The structure of data is essential to fit well-defined fields in relational databases or spreadsheets, allowing quick searches and efficient management of the data.

The Institute of Medicine in the USA, in a report called “Crossing the Quality Chasm”, identified six aims for improving healthcare quality. These aims are: safe, effective, patient-centered, timely, efficient and equitable healthcare. Data science (DS) resources can be used to improve these six dimensions, helping physicians to deliver high-quality care in real-time [1]. The European Commission underlined the importance of data for the economy and society as a reason for the development of “a European strategy for data” in 2020 and that benefits will include “healthier lives and better health-care” [2]. Improving structured data collection and the use of health data standards will help to improve health care.

***Corresponding authors: Pieter Vermeersch**, Clinical Department of Laboratory Medicine, University Hospitals Leuven, Leuven, Belgium; Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium; and European Federation of Clinical Chemistry and Laboratory Medicine (EFLM), Milan, Italy, E-mail: pieter.vermeersch@uzleuven.be. <https://orcid.org/0000-0001-7076-061X>; and **Pilar Fernandez-Calle**, European Federation of Clinical Chemistry and Laboratory Medicine (EFLM), Milan, Italy; and Department of Laboratory Medicine, Hospital Universitario La Paz, Madrid, Spain, E-mail: pfernandez.hulp@gmail.com

Tim Hulsen, Department of Hospital Services & Informatics, Philips Research, Eindhoven, The Netherlands. <https://orcid.org/0000-0002-0208-8443>

David Friedecký, Department of Clinical Biochemistry, Laboratory for Inherited Metabolic Disorders, University Hospital Olomouc and Faculty of Medicine and Dentistry, Palacký University in Olomouc, Olomouc, Czech Republic

Harald Renz, Institute of Laboratory Medicine, member of the German Center for Lung Research (DZL), and the Universities of Giessen and Marburg Lung Center (UGMLC), Philipps University Marburg, Marburg, Germany; and Department of Clinical Immunology and Allergy, Laboratory of Immunopathology, I.M. Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia

Els Melis, Ortho Clinical Diagnostics, Zaventem, Belgium

Among the medical specialties, laboratory medicine is the largest producer of structured data and therefore can be clearly identified as one of the main targets to apply data science and specifically big data (BD). Data included in the laboratory report must be clear and easily understandable, not only for the healthcare providers but also for its final recipient, the patient. Ideally, the structure of laboratory data is standardized allowing access from outside of the health institution by healthcare professionals and the patient.

Because of a continuously increasing accessibility to high volumes of data (through electronic medical records, laboratory information systems), expectations about results derived from BD are increasing. There are multiple examples of evidence and benefits of BD use, such as for the management of non-communicable chronic diseases (I.e. cancer or cardiovascular disorders). BD clearly has the potential to support decision making by improving both diagnostic and prognostic performances [3]. Moreover, BD and artificial intelligence (AI) can facilitate early risk prediction, prevention and tailored intervention, and can therefore be seen as clear enablers for personalized medicine. The European Commission in 2020 stated that:

Personalized medicine will better respond to the patients' needs by enabling doctors to take data-enabled decisions. This will make it possible to tailor the right therapeutic strategy to the needs of the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention [2].

BD and AI hold great promise to augment the ability of clinical decision support (CDS) systems to improve patient safety, quality of care, and patient outcome. A major problem with current knowledge-based CDS systems is the large number of inadequate and clinically inconsequential alerts that hamper their effectiveness due to alert fatigue and distraction. It is estimated that 90–95% of alerts are overridden, with sometimes errors as an unintended consequence [4]. A recent randomized trial reported that interruptive CDS alerts at the time of ordering were not associated with improved adherence to 5 Choosing Wisely guidelines of the American Board of Internal Medicine Foundation [5]. BD and AI could significantly reduce the number of inconsequential alerts and improve the quality of the rules and algorithms that are used through self-learning.

Taken together, BD and AI, applied to healthcare data can revolutionize the diagnosis and treatment of patients, as well as disease prevention and control, significantly boosting patient safety and quality of care and opening the way for personalized medicine [6]. In addition to the clinical impact, BD and AI can also help to identify areas of

waste, improving processes and allowing a more rational laboratory test ordering, improving laboratory efficiency and sustainability [7].

Specialists in laboratory medicine must play a crucial role in the efficient and safe implementation of BD and AI in healthcare as well as in the training of the workforce and education of patients. The European Federation of Clinical Chemistry and Laboratory Medicine identified “Big data and how to utilize it to improve service, quality and patient outcomes” as one of the strategic challenges to be addressed during the third ELM strategic conference held virtually 25–27 May 2022. This article summarizes the important topics discussed during the lectures of the session “Big data and how to utilize it to improve service, quality and patient outcomes. Training the next generation to collect/analyze and use lab data in a more efficient manner with more focus on post-analytics than analytics.”

Big data in medicine – big data in laboratory medicine

Modern healthcare has now arrived in the area of personalized therapies and precision medicine. The implications of this are best illustrated in the field of non-communicable diseases (NCDs). NCDs comprise all major non-infectious diseases that affect major organs, including the heart and blood vessels (e.g., myocardial infarction, stroke, and chronic heart failure), pulmonary disease (e.g., COPD and asthma), kidneys (chronic kidney disease), the liver, the immune system (autoimmune diseases and allergies), the brain and others. Common to most of the NCDs is the development of the clinical condition based on complex gene $\times$ environment interactions [8]. Many environmental factors contribute to disease development either as risk factors or as protective exposures. Prime examples for risk factors are mode of delivery, particularly elective cesarean sections, industrialization, urbanization, western diet, loss of biodiversity, air pollution, smoking, overuse of antibiotics and others. On the other hand, vaginal deliveries, extended breastfeeding, exposure to a biodiverse environment, older siblings in the family and other factors have been identified as protective exposures. This emphasizes the complexity of environmental exposures taken into the equation. This complexity has been referred to as the “exposome” [9].

Many of the NCDs are characterized by a chronic inflammatory response. A prime example for this are chronic lung diseases such as asthma [10]. Over the last

few years it has been clearly shown that type and pattern of chronic inflammation is highly diverse between patients (inter-individual heterogeneity), but also within patients (intra-individual heterogeneity) [11]. About half of the asthmatic patients suffer from a so-called type 2 inflammatory condition which is characterized by the presence of Th2-T-cells producing a unique panel of cytokines including IL-4, IL-5 and IL-13. These cytokines orchestrate the downstream inflammatory response of innate immune cells including eosinophils, basophils and mast cells and control the production of IgE antibodies by B-cells. This response is largely associated with allergic reactions. On the other side, at least as many asthmatic patients exhibit a differential inflammatory response which is termed “non-type 2”. Non-type 2 inflammation is in itself also heterogeneous, some patients show a predominant T-helper 1 immune response, others Th17 T-cell activation and others mixed or even different adaptive immune panel.

To identify the exact nature of inflammation in these patients has been of importance since the advent of biologicals in clinical use [11]. Many of these biologicals interfere on the level of type 2 inflammation with monoclonal antibodies directed against IL-5, IL-5 receptor, IL-4 receptor α (simultaneously targeting IL-4 and IL-13). Therefore it is critical to identify, characterize and use biomarkers which are precisely defining the type of inflammation of these patients in order to treat these patients in an optimal fashion.

One experimental approach to identify novel biomarkers is via mapping the inflammatory response, particularly on a single cell level (Mission: To create comprehensive reference maps of all human cells – the fundamental units of life – as a basis for both understanding human health and diagnosing, monitoring, and treating disease. Available at: www.humancellatlas.org [12]). This includes an unbiased approach on each level of gene regulation such as the transcriptome, epigenome, and metabolome. This effort creates huge data sets which need to be mapped and modeled, together with the exposome, the medication and other levels of regulation.

A further level of complexity is delivered by the microbiome. It has clearly been shown that the microbiome not only differs between diseased and healthy individuals, but that there exists also a cause-effector relationship between microbial dysregulation (dysbiosis) and development of inflammation [13].

One of the reasons why the mode of delivery is an important contributor to disease development is via microbial exposures, which fundamentally differ between babies delivered vaginally and babies delivered via caesarean section. Also, the overuse of antibiotics shifts

the microbiome towards an unhealthy pattern. The microbiome also differs between various anatomical sites and many of the microbes cannot be cultured, allowing analysis only via genomic and metabolomic analysis [13].

Another clinical use from the big data approach would be to derive biological variation (BV) estimates. BV is used for many applications in laboratory medicine, i.e. to define analytical performance specifications, to calculate the Reference Change Value for the interpretation of serial results of a biomarker in the monitoring or follow up of patients and recently it has been used to establish the personalized reference intervals for every individual to discriminate a significant change in their health status leading to the personalized Medicine concept. The use of data available from Laboratory Information Systems allows to obtain the BV estimates [14–16] by applying data mining statistical strategies avoiding the disadvantages of classical direct methods and working with a large number of subjects. Therefore, it allows us to explore seasonal variations and also different groups of ages or pathologies.

This exemplifies that in modern medicine, we are dealing with huge data sets not only for experimental and research approaches, but also in the “real world”. If the analysis of BD in research settings such as cohorts in clinical trials is already a big challenge, dealing with real world data is an even bigger challenge. Analysis of real world data requires development of legal, procedural and technical infrastructure. Another challenge is provided by the structure of data such as quality, interoperability, correctness, the issue of missing data and completeness of data sets [17–19]. However, the integration of all clinical data sets for any given patient is important and necessary in order to develop a patient-centered (in contrast to a disease-centered) treatment approach, which integrates individual data obtained from hospital visits, outpatient clinics and the daily life via bio wearables. Together with the development and availability of novel therapies, this real world BD approach will play an essential role to accelerate patient care to a new level.

Use of big data

When the term ‘big data’ was introduced, there were mixed responses: some researchers believed that upfront hypotheses were no longer necessary, while other researchers argued that new approaches were an irrelevant distraction from established methods [20]. It is now clear that neither were accurate: there is a place for both the ‘old’ hypothesis-driven research and the ‘new’ data-driven research, and they can even complement each other.

A hypothesis being tested generates (big) data, and these data can be used to generate new hypotheses. Hypothesis-driven research and data-driven research can also be carried out in parallel, as shown in figure 7 of McCue and McCoy [21]. In that case, only a general hypothesis is formulated upfront, after which the two research methods each follow their own track. Only at the end do they come together to accept, reject, revise or expand the hypothesis.

An important initiative in BD and ‘open science’ are the Findability, Accessibility, Interoperability, and Reusability (FAIR) Guiding Principles [22]. These provide guidelines to make data findable (F), accessible (A), interoperable (I) and reusable (R) to the research community. The FAIR Guiding Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data (e.g., using well-defined metadata), in addition to supporting its reuse by individuals. They help researchers to properly share their data with others in the spirit of ‘open science’ [23]. The FAIR principles have been recently extended towards research software as well [24], making sure that not only the data can be found and reused easily, but also the software and algorithms working with the data. Many of the principles could be directly applied to software, but some of them needed to be revised or extended to cope with some of the specific characteristics of software such as their executability, composite nature, and continuous evolution accompanied by frequent versioning.

The European GDPR law can be an issue when doing BD research on clinical data. However, there are technical solutions to help out, preserving privacy but still enabling researchers to do their job. For example, federated learning is a machine learning technique that trains an algorithm across multiple decentralized sources, without having to exchange the data. The AI algorithm travels to the data instead of the other way around. One example of a federated learning system using medical data is the Personal Health Train (PHT) [25, 26]. The PHT is designed to enable researchers to work with health data from various sources. It can give controlled access to data, while ensuring privacy protection and optimal engagement of individual patients and citizens [26]. The sensitive health data remain at the source and do not need to be transferred.

Data also needs to be structured and standardized well before it can be used by machine learning algorithms. A standard that is being used often is the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) standard [27]. It is free to use, defines a modern ReST interface and can easily be interpreted by machines. It also contains a human readable summary. Other widely used standards in the medical area include the Systematized Nomenclature Of MEDicine Clinical Terms (SNOMED CT) [28] and the

Logical Observation Identifiers Names and Codes (LOINC) [29]. SNOMED CT is a multilingual clinical healthcare terminology which supports the development of comprehensive high-quality clinical content in electronic health records (EHRs). LOINC is the international standard for identifying health measurements, observations, and documents. In genomics research, an often used standard is the gene ontology (GO) [30]. GO’s mission is to develop a comprehensive, cross-species computational model of biological systems, ranging from the molecular to the organism level. It is the world’s largest source of information on the functions of genes, and can be understood by both humans and machines.

The biggest disadvantage or risk with the use of AI in medicine, is that there is very limited trust in AI algorithms, because for many people (both clinicians and patients) these are like a ‘black box’ with limited explainability. Gunning et al. [31] state that “explainability is essential for users to effectively understand, trust, and manage powerful artificial intelligence applications” and introduce the term “Explainable AI” (XAI) for AI algorithms of which the behavior is made more intelligible to humans by providing explanations. However, just making AI explainable does not assure that an individual decision is correct, nor justify the acceptance of AI recommendations in clinical practice [32]. Therefore, Holzinger et al. [33] argues that in medicine we need to go beyond explainable AI, for which we need “causability” which encompasses measurements for the quality of explanations. Researchers need to be able to explain the AI algorithm using causability (increasing its transparency), and need to build trust by showing what AI can do (and with what accuracy). They should also show clinicians that AI can help them, not replace them [34]. Another way to increase trust is by improving the robustness of the algorithms. Robustness is the sensitivity of the decisions made by the AI models to the input dataset [35]. If the robustness is poor, it can lead to significant changes in the outcome of an AI model with small perturbations in the input data. Therefore AI models need to be proofed against such volatilities.

Handling big data

In laboratory medicine, univariate statistical methods have been used for many decades to provide exact evaluations of laboratory tests. Tools such as standard deviations, Z-scores [36] for data with normal distribution on the one hand, or quantiles for non-normal distribution on the other hand have become an essential part of most laboratory tests. One of the truly objective statistical tools is ROC analysis [37],

which provides information about the efficiency of tests when there is a sufficient number of samples. Despite the undeniable advances in data analysis, there are many areas where there is a potential for improvement – e.g., the handling of outliers, the definition of normality and its use in data transformation, and last but not least the problem of heteroscedasticity when applying statistics to tests with a wide range of values. In recent years, combinations of multiple biomarkers have been more widely used to provide better efficiency, e.g., using logistic or Cox regression [38].

Developments in many scientific fields have recently allowed us to use new omics technologies and with them to generate BD. Here, we no longer make do with a univariate statistical approach, but multivariate statistical methods have begun to be used to advantage, significantly shifting the potential of BD. In particular, combining multiple omics (e.g., metabolomics, proteomics, transcriptomics, genomics and others) gives previously unsuspected information and provides understanding of scientific questions in a larger picture, which is also called the systems biology approach.

In the case of data handling in multivariate statistical analysis, we find similar problems, such as the requirement of normality and dealing with outliers. In addition, scaling and centering must be taken into account. In terms of statistical methods, both unsupervised and supervised statistical methods (e.g., principal component analysis, discriminant analysis and clustering) are being applied more frequently. Artificial intelligence approaches are also increasingly being utilized, but they are still a long way from being used in laboratory medicine. For example, we have published a promising method using DA for pancreatic cancer diagnosis based on plasma lipidome measurements that outperforms conventional biomarkers [39]. The method is now entering a screening validation study in the Czech Republic.

On the other hand, it is important to mention that BD from omics suffers from difficult long-term reproducibility due to the complex nature of the data (e.g., shift in sensitivity over long-term use of mass spectrometers). This and other issues present new challenges in harmonization across the laboratories. Therefore, in recent years, omics have been used more frequently for the discovery phase, and then selected diagnostic biomarkers of greatest importance are used for incremental development of laboratory tests. For example, the CERAM test for prediction of cardiovascular events based on selected plasma lipids (ceramides and phosphoglycerides) has been introduced in recent years and is successfully performed routinely at the Mayo Clinic [40]. Another very successful story is the application of the omics approach in neonatal screening for inherited metabolic disorders, where a comprehensive

online tool “CLIR – Collaborative Laboratory Integrated Reports” was developed by Piero Rinaldo to effectively screen newborns based on a comprehensive profile for amino acids and acylcarnitines [41]. This approach is a beautiful demonstration of how false positives or negatives can be significantly reduced through multivariate data analysis. In the near future, we are going to see intense developments just in the field of omics and BD and its use in disease prediction, diagnosis or treatment.

How big data can support partnerships between laboratories and the IVD industry

The BD era also offers opportunities for laboratories and the IVD Industry, to optimize the productivity of the laboratory and ultimately patient outcomes, through timely reporting of the test results.

Implementation of BD for the IVD industry is linked to the Internet of Things (IoT) and it is already in place today. Most core laboratory analyzers can be remotely connected to the manufacturer through a VPN connection and big amounts of data are continuously transmitted to the manufacturer’s servers, allowing analytics and learning [42].

IoT data transmitted by the analyzers can be split into two basic groups: analyzer subsystems performance and sample-analysis data (samples being either patients, QC or calibrators). Practical applications for analyzer subsystem data could be preventive maintenance and efficient transfer of info to users and field engineers, with the aim to maximize uptime. For privacy compliance (e.g., GDPR in Europe) any patient-related result must be anonymized, and data should be encrypted.

Analyzer subsystems data:

- Physical and electromechanics performance data is generated for many instrument components. As example, data items include temperatures, pressures, electrical signals, optical signals, among many others.
- Time and part usage counts are also recorded. Anomalies in any physical or electromechanics measurement can be compared to time or parts usage, so intelligence can be built about parts durability and its associated maintenance. This can be seen as real-life testing of components, helping to redefine maintenance and to run a continuous full-population quality control of parts and subsystems.
- At individual instrument level, this technology supports the predictive monitoring of each analyzer status.

Trends in the analyzer's subsystems status can be analyzed, failures can be anticipated before they occur, and pro-active maintenance planned to maximize the analyzer uptime.

- Instrument and assay performances can be analyzed real time and appropriate actions can be pushed to the right person (through mobile devices or the analyzer itself): to a lab technician if new calibration is required, to a field engineer if part requires replacement.
- IoT also enables a new model of support that anticipates user needs and makes the right document available when required. This is the paradigm to follow: Push the right information on the right devices to the right person at the right time.

Sample-analysis data:

- The aggregation of all the sample results produced over time by one or more analyzers, including analyzers in distant sites belonging to the same network, can be visualized in Dashboards. Dashboards turn data into visual presentation facilitating the data analysis.
 - Dashboards can present the workload of labs over time and its distribution across sites and analyzers. This supports the allocation of resources and the redistribution of workloads to maximize the capacity of laboratories.
 - Distribution of turnaround times by hour of the day or by test can support decisions to keep service commitments to medical departments like the emergency room.
 - During the COVID-19 pandemic, the register of samples turning positive for antigen or antibody testing allowed to establish a percentage of positivity or seroconversion in the tested population. Such public health contributions can be extended for instance by following the trends of reactive samples for the infectious diseases screened in blood donations.
 - Efficiency in the use of reagents can be tracked as well. Analyzers can split the test that have led to a reported result, the ones that have not produced a reportable result, the tests used for quality control and calibration and even the wasted tests (e.g., reagent expired on board). This efficiency checks can help to reorganize the testing of certain analytes by concentrating some tests in fewer analyzers or even deciding which tests have a demand scale which makes the outsourcing more favorable.
- For some analytes and some populations, the aggregate of patient results in a single instrument calculated as “moving average” can be used as an indicator of test

accuracy, becoming an additional quality check for the lab [43].

Conclusions

- An essential requirement for BD and AI in healthcare is the need to increase the number of information stored as structured data and better define and standardize them. It will allow BD to impact the diagnosis and treatment of patients improving patient safety and quality of care and opening the way for personalized medicine.
- Data-driven research is complementary to hypothesis-driven research. By making data and software FAIR, by respecting privacy laws and by using standards and ontologies, we can generate trust in AI algorithms working with BD so that clinicians (and patients) are confident in using them.
- BD requires new statistical perspectives on the multivariate structure of data that outperform the traditional univariate statistical methods used in laboratory medicine over the past decades.
- Analyzer subsystems performance data and sample-analysis data can help the IVD industry and clinical laboratories to improve productivity, quality and ultimately patient outcomes.

Research funding: None declared.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Authors state no conflict of interest.

Informed consent: Not applicable.

Ethical approval: Not applicable.

References

1. Broughman JR, Chen RC. Using big data for quality assessment in oncology. *J Comp Eff Res* 2016;5:309–19.
2. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: a European strategy for data. COM/2020/66 final; 2020. Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066>.
3. Kapoor A. Hands-on artificial intelligence for IoT: expert machine learning and deep learning techniques for developing smarter IoT systems. Birmingham, UK: Packt Publishing Ltd; 2019.
4. Powers EM, Shiffman RN, Melnick ER, Hickner A, Sharifi M. Efficacy and unintended consequences of hard-stop alerts in electronic health record systems: a systematic review. *J Am Med Inf Assoc* 2018;25: 1556–66.

5. Ho VT, Aikens RC, Tso G, Heidenreich PA, Sharp C, Asch SM, et al. Interruptive electronic alerts for choosing wisely recommendations: a cluster randomized controlled trial. *J Am Med Inf Assoc* 2022;29: 1941–8.
6. Tan SSL, Gao G, Koch S. Big data and analytics in healthcare. *Methods Inf Med* 2015;54:546–7.
7. SAS. Big data: what it is and why it matters [online]. Available from: https://www.sas.com/en_au/insights/big-data/what-is-big-data.html [Accessed 2 Oct 2022].
8. Rappaport SM. Genetic factors are not the major causes of chronic diseases. *PLoS One* 2016;11:e0154387.
9. Renz H, Holt PG, Inouye M, Logan AC, Prescott SL, Sly PD. An exposome perspective: early-life events and immune development in a changing world. *J Allergy Clin Immunol* 2017;140:24–40.
10. von Hertzen L, Beutler B, Bienenstock J, Blaser M, Cani PD, Eriksson J, et al. Helsinki alert of biodiversity and health. *Ann Med* 2015;47: 218–25.
11. Holgate ST, Wenzel S, Postma DS, Weiss ST, Renz H, Sly PD. Asthma. *Nat Rev Dis Prim* 2015;1:15025.
12. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* 2016;17:175–88.
13. Renz H, Skevaki C. Early life microbial exposures and allergy risks: opportunities for prevention. *Nat Rev Immunol* 2021;21:177–91.
14. Loh TP, Ranieri E, Metz MP. Derivation of pediatric within-individual biological variation by indirect sampling method: an LMS approach. *Am J Clin Pathol* 2014;142:657–63.
15. Jones GRD. Estimates of within-subject biological variation derived from pathology databases: an approach to allow assessment of the effects of age, sex, time between sample collections, and analyte concentration on reference change values. *Clin Chem* 2019;65: 579–88.
16. Marqués-García F, Nieto-Librero A, González-García N, Galindo-Villardón P, Martínez-Sánchez LM, Tejedor-Ganduxé X, et al. Within-subject biological variation estimates using an indirect data mining strategy. Spanish multicenter pilot study (BiVaBiDa). *Clin Chem Lab Med* 2022;60:1804–12.
17. Bunyavanich S, Schadt EE. Systems biology of asthma and allergic diseases: a multiscale approach. *J Allergy Clin Immunol* 2015;135: 31–42.
18. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016;6:26094.
19. Woodhouse S, Moignard V, Göttgens B, Fisher J. Processing, visualising and reconstructing network models from single-cell data. *Immunol Cell Biol* 2016;94:256–65.
20. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, et al. From big data to precision medicine. *Front Med* 2019;6:34.
21. McCue ME, McCoy AM. The scope of big data in one medicine: unprecedented opportunities and challenges. *Front Vet Sci* 2017;4:194.
22. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018.
23. Hulsen T. Sharing is caring-data sharing initiatives in healthcare. *Int J Environ Res Publ Health* 2020;17:3046.
24. Lamprecht AL, Garcia L, Kuzak M, Martinez C, Arcila R, Martin Del Pico E, et al. Towards FAIR principles for research software. *Data Sci* 2020;3: 37–59.
25. Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, et al. Distributed learning on 20 000+ lung cancer patients – the Personal Health Train. *Radiother Oncol* 2020;144:189–200.
26. Health RI. Personal health train [online]. Available from: <https://www.health-ri.nl/initiatives/personal-health-train> [Accessed 21 Oct 2022].
27. Bender D, Sartipi K. HL7 FHIR: an Agile and RESTful approach to healthcare information exchange. In: *Proceedings of the 26th IEEE international symposium on computer-based medical systems*. Porto, Portugal: IEEE; 2013.
28. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. *Stud Health Technol Inf* 2006;121:279–90.
29. Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem* 1996;42:81–90.
30. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* 2019;47:D330–8.
31. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI-Explainable artificial intelligence. *Sci Robot* 2019;4:eaay7120.
32. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021;3:e745–50.
33. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019;9:e1312.
34. Hulsen T. Challenges and solutions for big data in personalized healthcare. In: Moustafa AA, editor. *Big data in psychiatry & neurology*. Amsterdam, The Netherlands: Elsevier; 2021.
35. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res* 2020;22:e15154.
36. Altman EI. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J Finance* 1968;23:589–609.
37. Robertson EA, Zweig MH. Use of receiver operating characteristic curves to evaluate the clinical performance of analytical systems. *Clin Chem* 1981;27:1569–74.
38. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol* 1972;34:187–202.
39. Wolrab D, Jirásko R, Cífková E, Höring M, Mei D, Chocholoušková M, et al. Lipidomic profiling of human serum enables detection of pancreatic cancer. *Nat Commun* 2022;13:124.
40. Mayo Clinic. CERAM: MI-heart ceramides, plasma [online]. Available from: <https://www.mayocliniclabs.com/test-catalog/Overview/606777> [Accessed 2 Oct 2022].
41. Mayo Clinic. CLIR – Collaborative Laboratory Integrated Reports [online]. Available from: <https://clir.mayo.edu/> [Accessed 2 Oct 2022].
42. Aris-Brosou S, Kim J, Li L, Liu H. Predicting the reasons of customer complaints: a first step toward anticipating quality issues of in Vitro diagnostics assays with machine learning. *JMIR Med Inform* 2018;6:e34.
43. Badrick T, Graham P. Can a combination of average of normals and “real time” external quality assurance replace internal quality control? *Clin Chem Lab Med* 2018;56:549–53.