Clinical Biochemistry 69 (2019) 1–7

Contents lists available at ScienceDirect

Clinical Biochemistry

journal homepage: www.elsevier.com/locate/clinbiochem

Review

Data science, artificial intelligence, and machine learning: Opportunities for laboratory medicine and the value of positive regulation

Damien Gruson[a,b,*], Thibault Helleputte[c], Patrick Rousseau[a], David Gruson[d]

[a] Department of Laboratory Medicine, Cliniques Universitaires St-Luc and Université Catholique de Louvain, Brussels, Belgium
[b] Pôle de recherche en Endocrinologie, Diabète et Nutrition, Institut de Recherche Expérimentale et Clinique, Cliniques Universitaires St-Luc and Université Catholique de Louvain, Brussels, Belgium
[c] Data-Driven Precision Medicine, DNAlytics, Louvain-la-Neuve, Belgium
[d] Genetics Regulation for Paris Descartes-University, Paris, France

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Deep learning
Big data
Health
Ethics
Biomarker

ABSTRACT

Artificial intelligence (AI) and data science are rapidly developing in healthcare, as is their translation into laboratory medicine. Our review article presents an overview of the data science domain while discussing the reasons for its emergence. We also present several perspectives of its applications in clinical laboratories, along with potential ethical challenges related to AI and data science.

1. Introduction

The potential of health data to stimulate the development of precision medicine is enormous [1–3]. Data science (DS) has begun to provide a new set of tools that will leverage enhanced laboratory medicine and reinforce its value in a continuously transforming healthcare ecosystem (Fig. 1). DS is a human-centered activity dedicated to the principled extraction of knowledge from complex data with the aim of generating further insights [2]. It is a general field encompassing artificial intelligence (AI), data capture and management, advances in databases, and computing infrastructures (*e.g.* local, distributed, or graphical processing units [GPUs] and the combination of conventional processors).

AI is a field of computing science that can be encompassed by DS. It is devoted to mimicking the human thought processes and behaviors used to make decisions or take actions [3]. AI employs quite different mathematical and algorithmic approaches, from operational research to constrained programming, and is therefore at the crossroads of neurocomputing, statistical inference, pattern recognition, data mining, knowledge discovery, and machine learning (ML) [4].

As a subfield of AI, ML is built upon statistical and optimization concepts. It can be described as the development of computer programs that learn from experience with respect to task and performance measures [5]. Both DS and AI have already been shown to be beneficial with regards to weather forecasting, face recognition, natural language processing, collaborative recommendations, improvement in industrial processes, and analyses of financial transactions [6,7].

As a specific feature, ML programs are able to adjust themselves when exposed to new datasets, *i.e.* “learn” without being explicitly programmed. Typically, they are designed to find patterns, trends, and associations; to discover inefficiencies; to learn and become better; to execute plans; to predict future outcomes based on historical trends; and to inform fact-based decisions [4]. ML is generally categorized into two types: supervised and unsupervised [8]. Supervised learning adapts a model to reproduce the known output from a training set, whereas in unsupervised learning, there are no outcome variables to predict. Several studies have compared the performances of these two learning techniques, notably in the process of gene selection or for identification of prognostic factors in patients with localized retroperitoneal sarcoma [9,10].

These techniques differ from automated procedures, such as computerized chess games where actual chess rules are encoded into the program. ML frameworks require (i) a definition of the task to accomplish (*e.g.* predict the amount of rain tomorrow), (ii) a performance metric to evaluate how “good” the model resulting from the ML setting will be (*e.g.* the square of the difference between the rain forecast in mm

* Corresponding author at: Pôle de recherche en Endocrinologie, Diabète et Nutrition (EDIN), Université Catholique de Louvain, Tour Claude Bernard, 54 Avenue Hippocrate, B-1200 Brussels, Belgium.

E-mail address: damien.gruson@uclouvain.be (D. Gruson).

https://doi.org/10.1016/j.clinbiochem.2019.04.013
Received 2 November 2018; Received in revised form 17 April 2019; Accepted 19 April 2019
Available online 22 April 2019

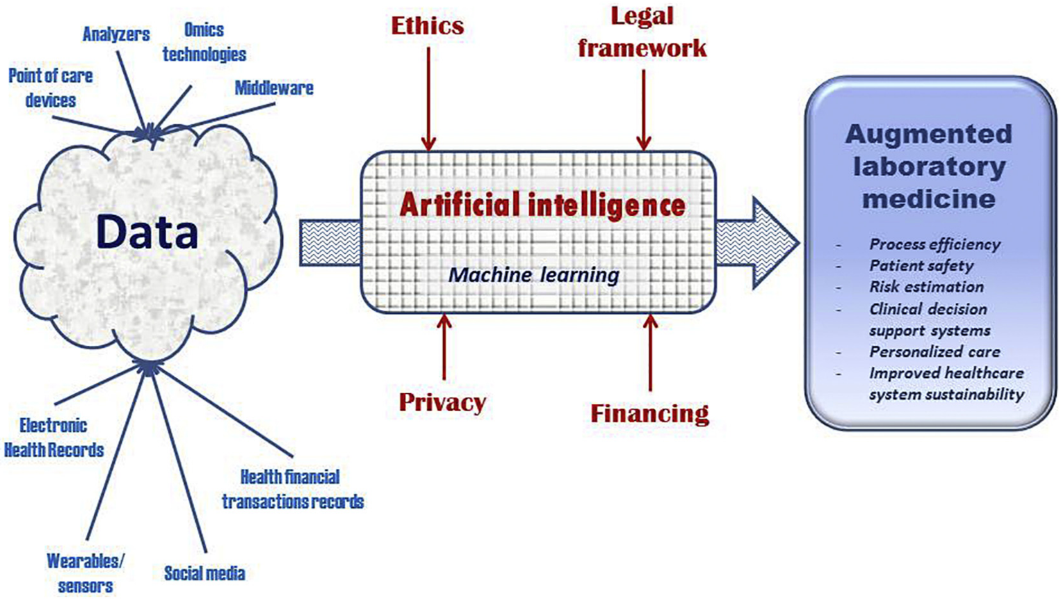

Fig. 1. From data to augmented laboratory medicine, the potential and challenges of artificial intelligence.

and the actual number of mm of rain), (iii) a search space/family of models for potential models to be learned *(e.g.* a neural network or much simpler: a linear regression), and (iv) most importantly, a set of examples for which the predictions to be made are already known (*e.g.* actual mm of rain for a series of days coupled with candidate predictors like pressure, temperature, altitude, latitude, longitude, or sea proximity).

The actual model is not programmed by the human programmer but learned by the ML framework based on the specifications it has, *e.g.* find a linear model that will output the lowest square difference between rain prediction and actual rain based on a limited set of examples. Historically, playing chess *via* AI required expert knowledge (the rules of the game) and did not necessarily require data. Currently, however, AI and particularly ML do not require explicit rule descriptions or expert human knowledge to gain impressive playing capabilities provided a sufficient amount of data (number of played chess games) are available to the artificial learner.

In the field of healthcare, DS has demonstrated success in image analysis in radiology, pathology, and dermatology; genomics analysis for accelerating diagnosis speed; improving accuracy paralleling; and enhancing medical expertise [12,13]. The market size of DS in healthcare is expected to grow 10-fold by 2021. The "big data" market in healthcare was estimated to be $14.25 billion in 2017 and could exceed $68 billion by the end of 2025 [13].

The potential of scalability could cover a broad spectrum of applications that AI could perform, such as robot-assisted surgery, virtual nursing assistance, administrative workflow assistance, fraud detection, reduction of dosage errors, integration of data from connected instruments and sensors, identification of potential participants in clinical trials, aiding in the formation of preliminary diagnoses, automated image analysis, nutrition support to treatment, and cybersecurity [12–14]. DS and AI clearly have the potential to revolutionize laboratory medicine in the very near future.

Moreover, considering the growing quantity and diversity of data to which healthcare professionals are exposed (*e.g.* imaging, biology, genomics, proteomics, clinical observations, and personal and environmental records), DS will no doubt prove useful for holistic interpretation of this wealth of information. In this review, we present several applications of DS in laboratory medicine. We also review the ethical challenges inherent in the implementation of DS into healthcare.

2. Why now?

To work properly, ML approaches require three key ingredients: learning algorithms, computational power, and data. ML algorithms were first developed several decades ago due to the interest in neural networks in the 1980s. Since then, numerous variants of decision tree-based approaches and kernel methods (namely support vector machines) have been developed and thoroughly documented. The more recent focus on deep learning is merely a resurgence of neural networks specifically adapted to image processing.

The increasing clinical interest and use of neural networks can be highlighted by the detection of cervical cancer. Cervical cancer is a common gynaecologic malignancy. Intelligent and objective classification of cervical samples (PAP smears) based on the use of convolutional neural networks are able to support pathologists for their diagnoses. Kudva and co-workers [15] reported a diagnostic accuracy close to 100% for such automated systems. The added value of decision-support scoring systems based on artificial neural networks for the prediction of histological diagnoses in women with cytological abnormalities when compared with the use of tests alone has also been evidenced [16]. In another study, Wu and co-workers reported an improvement of 3.85% when using augmented images as input data for the model [17]. An improvement in the prediction of adverse events caused by radical hysterectomies in cervical cancer patients using neuronal network classifiers of gene expression programming has also been documented with a prediction accuracy of more than 70% [18].

From a scientific perspective, while active research is still ongoing, this field pertaining to neuronal network classifiers is clearly maturing and its results, *e.g.* typically predictive models for cancer diagnosis, can now easily be integrated along with more general DS solutions into graphical interfaces like desktop or mobile solutions.

Enormous computational power is needed to solve optimization problems such as the objective function of the learning tasks or performance solvers to optimize the objective function for these standard measurements, which are the essence of these learning algorithms. As local machines that potentially combine central processing units (CPUs) and GPUs (which are better than CPUs at accomplishing certain computational tasks) with a large amount of memory and immense cloud computing infrastructures (*e.g.* Amazon's AWS; Microsoft Azure; and IBM's, Google's, and OVH's cloud infrastructures).

Since the early 2000s, the possibilities for patient monitoring have become incredibly diverse (*e.g.* sensors for continuous measurement of glucose, use of the internet of things for tension or weight collection), resulting in a real deluge of data. There are numerous types of these data, including structured clinical and environmental data, as well as huge volumes of data from DNA (and many epigenetic modifications), RNA (in various forms, such as messenger, long non-coding, or double-stranded), proteins, and metabolome and microbiota analysis. Accessing these huge volumes of data is not always easy, however, due to a lack of standardization (different medical software products used; structured *vs.* unstructured data; differences in syntax, ontology, and terminology for describing similar observations), legal challenges (issues about privacy), financial limitations, and ethical concerns.

These may constitute the last remaining obstacles to the deployment of DS approaches in healthcare. Some of these issues can be addressed through technical developments, for example, distributed learning now enables predictive models to be built in a centralized way without directly accessing the data [19,20]. When several hospitals or hospital departments agree to collaborate but cannot share their databases, the learning algorithm post requests for specific statistical data summaries within each data silo. These silos process the query and provide the summarized/computed values requested.

The central learning algorithm can build a strictly equivalent model or, depending on the approaches, an approximation of the model it would have built if all datasets had consolidated locally. This scheme fully complies with data privacy and addresses standardization issues to an extent given that each silo might implement the requests in a specific way. These approaches require an intense use of network communications when training the model, which is then reduced after the model is built.

There is a clear trend towards a higher integration of DS approaches into healthcare practices and the provision of improved decision support systems to assist physicians and laboratorians. Clinical laboratories will actively be engaged in this process as they represent a major source of healthcare data [12,21].

3. Off-the-shelf or tailor-made?

As cloud computing infrastructures have developed, numerous off-the-shelf ML frameworks have been proposed to the growing number of non-data scientists. These include Mahout from Apache [22], which was built on top of Hadoop distributed computing and uses the MapReduce paradigm, Amazon machine learning (AML) [23], and TensorFlow from Google [24]. There is great temptation to use these tools as black boxes, although ML requires a lot of fine-tuning and expertise to deliver relevant and robust results.

An example is setting the performance evaluation at the time of the experimental design with the aim of avoiding overly optimistic conclusions due to either over fitting or selection bias. Biases usually occur when the performance evaluation is determined using the same dataset employed to train the models. This was the case in a well-known study that focused on a breast cancer prognostic model based on a transcriptomic signature [25,26]. The authors validated the performance of their solution using a cohort that comprised one-third of the patients from the original cohort employed to identify the markers and build the predictive model. Selection bias has similarly been discussed in various publications [27,28].

Some mathematical aspects of ML should also be mentioned. Each model corresponds to an optimization problem and by default, the models in most ML libraries are designed to optimize the efficiency of a task (*e.g.* the ratio of diagnostic cases correctly predicted). Prediction tasks are rarely balanced in daily practice, however, and the incidence of a disease is often rather small compared to the whole population. Therefore, to optimize their accuracy, models tend to classify most, if not all, samples as being in the majority class, which ironically does not result in a good predictive model. For example, if arthritis affects 1% of the population, a model predicting that all individuals have no arthritis is totally irrelevant, despite its 99% accuracy.

Redesigning the optimization problems therefore requires specific mathematical and programming expertise. Several approaches have been proposed to improve the treatment of unbalanced data and the efficiency of algorithms, such as imputation, downsampling, and upsampling [29]. Imputation assumes that the actual data distribution is known, which is rarely the case, whereas downsampling reduces inference capabilities. Similarly, depending on the mathematical form of the models to be learned, upsampling does not necessarily represent a gain in inference capabilities.

Another temptation is to always apply the same model if the user has mastered it or finds it convenient for interpretation purposes. It has been shown, however, that there is no such thing as an ever-winning ML model type (this constitutes the “No Free Lunch” theorem) [30]. One must compare different model types for each case at hand. For relevant and robust results, it is better to rely on tailor-made approaches designed and implemented by specialists. In turn, these specialists may rely on convenient ML frameworks and adjust them to further improve productivity.

4. Application perspectives

4.1. Processes and care pathways

The complexity of laboratory processes and the challenges associated with their integration into care pathways are evident. DS is recognized for improving complex analytical tasks and flows in the laboratory domain. Therefore, DS could be used to examine data in real time and calculate and simulate the most efficient operational and clinical pathways. This avoids redundancies and provides the opportunity to adjust faster with minimal supervision.

Applying DS to the datasets of laboratory pathways and processes could enable supervisors to predict certain scenarios, adapt to them, and optimize task management. The translation of DS to laboratory operations and healthcare pathways offers the opportunity to mimic activities, redesign processes, and apply several process improvements based on next-generation technologies that assist laboratory staff members by removing repetitive, replicable, and routine tasks.

DS represents a shift towards the proactive management of processes and operations that are likely to reduce costs, increase the benefits of automated systems, optimize asset management, improve operational performances by identifying areas of downtime, enhance the use of reagents, improve the efficiency of staff resources, and/or reduce energy costs. DS is a crucial element in the application of intelligent process automation and smart workflows to laboratory medicine. It is key to the efficient integration of laboratory services, which provide optimized care pathways that benefit patients [31].

4.2. Laboratory test ordering and interpretation

With the current need to utilize healthcare resources more efficiently, the ordering of imaging and laboratory services requires improved control. DS can be used to assist physicians with imaging and laboratory test ordering based on the differential diagnoses provided. With proactive management of imaging and laboratory test orders provided by DS, the potential to improve patient safety, avoid unnecessary testing, and ensure the right tests are ordered for the right patient is huge.

In addition to improved control of laboratory test ordering, DS can trigger alerts when abnormal results occur. The development and validation of a dynamic, patient-tailored method designed to detect abnormal laboratory test results has previously been reported [32]. DS may also be used to facilitate the interpretation and integration of clinical data and laboratory tests in critical diseases, with the example of syndromic approaches in transmissible disorders requiring the analysis of complex datasets of *multi-omics* results [33,34]. The development of cognitive programs through DS will likely impact laboratory practices and test ordering and interpretation by applying natural language processing to integrate the up-to-date information, evidence, and guidelines coming from the rapidly expanding scientific literature [14].

4.3. Data mining, early diagnosis, and proactive disease monitoring

There is the possibility of real-time data mining, curation, and integration (patients' histories, clinical records, ongoing medications and treatments, results of imaging and laboratory tests, *etc.*) with DS tools. Such data integration could join-up fragmented care records, resulting in increased patient safety and fewer medical errors. DS could also speed up the process of identifying patient clusters by digging deep into electronic records [35].

The integration of data combined with medical expertise could shift the paradigm of diagnosis, patient risk estimation, and treatment selection. The potential of data mining in laboratory information systems is clear. The continuous expansion of these databases generates strong interest in the application of powerful deep-learning methods for analyzing these data and improving patient diagnosis and risk estimation [36]. Improvement in outcomes has already been achieved in cardiovascular medicine using ML [37,38]. DS technologies have been applied in cardiovascular medicine to explore novel genotypes and phenotypes in existing diseases, improve the quality of patient care, enable cost-effectiveness, and reduce hospital readmission and mortality rates.

Intelligent decision support has therefore been proven useful for diagnosing cardiovascular diseases and assisting clinical decisions [39]. The impact of applying DS to decision-making systems involving multidisciplinary team decisions to avoid potential mistakes in selected cases has been documented. Buzaev and co-workers [40] reported the use of a neural network model to assist the choice between coronary aortic bypass surgery and percutaneous coronary intervention. This selection system relies on a registry with significant factors, decisions, and results obtained through DS processing.

Enhanced risk estimation using DS tools is likewise a reality in cardiovascular medicine [38]. The Framingham Heart Study showed that integrated genetics and epigenetics can improve the prediction of coronary heart disease by means of methylation analysis and the mapping of risk factor signatures [41]. ML methods have likewise been applied in the Swedish Heart Failure Registry with the aim of improving the ability to risk-stratify heart failure patients [42]. In this study, cluster analysis identified four distinct phenotypes that significantly differed in both outcome and response to various therapies, thus optimizing personalized care for heart failure [42].

Other studies have shown that ML could be used to predict the risk of bleeding from esophageal varices in children with hepatic impairment [43]. The risk-scoring algorithm identified varices grade, fibrinogen level, and red spots, out of several dozen clinical and laboratory parameters, as potential determinants of bleeding risk in these patients. When appropriately combined with a mathematical model, the algorithm accurately predicted variceal bleeding in about 85% of the patients, enabling at-risk pediatric patients to be prioritized for urgent liver transplants. These results have been validated on an independent patient cohort, with an online application developed (http://hrs2c2.com).

Other examples can be found in the oncology domain. A recent exploratory study was performed to develop and assess a prediction model based on anthropometric data and routine blood analysis parameters, which can potentially be applied as a surrogate biomarker of breast cancer [44]. The study concluded that support vector machine models using glucose, resistin, age, and BMI as determinants could predict the presence of breast cancer in women. These findings provided promising evidence that such parameters can be powerful, inexpensive, and effective tools for identifying potential cancer biomarkers. Clearly, in order to apply optimal DS algorithms, each experiment or case requires some understanding of the problem in terms of medical impact, outcomes, and statistical features.

4.4. Personalized treatment and clinical trials

Molecular biomarkers and pharmacogenetics tools can predict drug efficacy and treatment responses in patients; these are crucial components for the advancement of precision medicine [33,45]. In this field, DS is relevant at several levels. First, the identification of genetic variants and molecular factors can be optimized by data mining the increasing amount of literature and databases [33]. Second, the development of DS-based algorithms can contribute to the translation of harvested evidence into the daily practice of physicians and pharmacists. Identifying and offering new treatment options is extremely challenging, but in this context DS could facilitate identification and utilization of molecular disease pathways, leading to new and specific therapeutic targets.

In a more prospective context, DS could improve the identification and enrollment of subjects into clinical trials [38]. The identification of patients matching the trials' eligibility criteria could be improved by using structured data, thereby automating the trial screening process, creating databases of clustered patients, and developing more customized trial materials. The integration of DS with laboratory records and biobanks may also contribute to identifying potential participants for trials.

5. Data privacy

There are several challenges associated with the translation of DS and AI into daily practices. The use of data means the collection and sharing of data with potentially sensitive information between healthcare professionals and researchers. This has generated major concerns about privacy [46,47]. Privacy was defined as a fundamental human right in the Universal Declaration of Human Rights at the 1948 United Nations General Assembly [48]. The question of privacy impacts the adoption of AI and ML tools, with the rate of adoption remaining relatively slow [49]. Our community and our patients should therefore be aware of the ways in which their privacy could be compromised and of the potential sources of privacy breaches.

Major progress is being made in terms of data protection, de-identification, and encryption, however [46,48]. To combat privacy challenges, laboratories, hospitals, scientific societies, and governmental agencies must provide guidance, education, and training to healthcare professionals, patients, and helpers alike [50]. Legal frameworks are evolving, allowing us to gain better control of these issues, *e.g.* the recent European General Data Protection Regulations [49].

6. Ethical considerations: towards the positive regulation of DS in healthcare

DS and AI trigger major ethical concerns and questions regarding the ability to disseminate their benefits in an equitable manner [1,51]. Joint actions are required to pave such a challenging route and encourage their adoption by the clinical community. Several regulatory and ethical frameworks have been proposed. For example, the Ethics Committee of the American College of Epidemiology warns of the implications of big data and computing, the fallacy of “secondary use” of data, and the duty of citizens to contribute to big data. This committee also emphasizes the need for balanced perspectives that allow safeguards to be provided for individuals without handcuffing research efforts to improve the health of the population [51].

Char and co-workers have recently addressed the ethical challenges associated with implementing ML in the healthcare domain [52]. They underlined several key points, such as the need to avoid bias (*e.g.* racial biases in the delivery of healthcare or tampering with the allocation of scarce resources, such as organs for transplantation for patients with neurodevelopmental delays or those with certain genetic profiles) and preventing any intent behind the design of ML systems that could create ethical strain, *e.g.* the car industry using certain algorithms to enable vehicles to pass emission tests [52]. They also point out the need for the active involvement of healthcare workforces in the construction of ML systems so that they are built to the most recent ethical standards that guide the healthcare sector [52].

Multidisciplinary sharing of information is mandatory for success when implementing ML in the healthcare domain. Accordingly, the EthiK IA brought together a team of researchers and professionals from the algorithmic and healthcare sectors to work on the regulation of AI and robotics [53].

Their operational objective was to ensure that France and the European Union were at the forefront of the development of several AI soft regulations in accordance with the requirements of an advanced democratic society. This approach aimed to unite the actors of research and management in the ethical, legal, and social regulations of the deployment of robotization and AI in the health and medico-social fields. It relied on the work already undertaken in the framework of the Sciences-Po Paris Healthcare Chair and the Paris-Descartes Healthcare Law Institute.

In the context of the review process of bioethics' laws, the principles of positive regulations for the deployment of innovation will have to be identified [53,54]. Five keys to the regulation of AI and robotics in health have already been identified and are designed to help authorities reduce the minimum requested level of formal legislation. The appropriate tool to nurture AI innovation in healthcare is obviously soft law.

6.1. Key 1: patient information and consent

The framework proposes that the patient must be informed prior to the use of any AI technology in the course of their care. The AI device should not replace the collection of the patient's consent. Specific modalities such as the use of a trusted person, prior aids for a range of care options, or enhanced protection for the vulnerable must, where appropriate, be arranged to ensure the effectiveness of the collection of this consent. There is currently no more obligation to request informed consent for the use of AI than there is for the use of any drug.

Specific AI applications will fall under the category of “healthcare product (treatment/*in vitro* diagnostic device/medical device)” and follow the same rules for market clearance. If an AI-based diagnostic tool is proven to correctly diagnose a certain disease and is satisfactorily validated through a clinical trial, how different will it be from any other IVD device? The recent European regulations on IVDs explicitly includes software applications in their scope, meaning the possibility of extension to DS-derived tools. Nevertheless, it might be useful to enhance the level of information and thus the autonomy of patients regarding these new developments.

6.2. Key 2: AI human warranty

The principle of human warranty of AI in healthcare must be respected. This warranty should be ensured by the regular verification (both targeted and random) of the management options proposed by the AI device and patients and health professionals should be able to request a second opinion from a human medical expert. This second opinion can be implemented *via* telemedicine devices if necessary.

6.3. Key 3: graduation of regulation according to the level of sensitivity of healthcare data

According to the principles of bioethical law, the regulation of the deployment of an AI device for the processing of large quantities of health data must be graduated according to the sensitivity of these data.

6.4. Key 4: accompaniment of the adaptation of healthcare professions

The implementation of DS-based devices for data processing or robotization in healthcare should not prevent the application of ethical principles. Instead, new principles applicable to the health professions using these devices should be implemented. The resources saved with the increase in efficiency achieved by the deployment of DS in healthcare could be used to finance the training (both initial and continuous) of professionals. The definition of competences and qualifications to acquire will also be mandatory.

6.5. Key 5: need for independent guidance

An independent guidance group must be implemented to examine the efforts made to promote the four keys outlined above. Such a concept has been widely spoken of in the public debate surrounding the practical applications of human warranty to AI. In a more specific domain, EthiK IA presented, in cooperation with the teams of the Hospitalo-University Institute Imagine, a prototype of the appropriate use of AI applied to genomic data [26]. The reasoning is fairly simple: its aim is to establish a “perimeter of sensibility” with increased safety rules, within which data processing (and in particular the crossings between genomic and given phenotypic data) can be significantly facilitated.

This prototype establishes the first stage of defining the segments of data and sensitive data processing as it proposes a framework in which the application of DS must be the object of strengthened protection. The objective of forward standardization is to devise a reference frame like the hazard classes (P1–P4) for biosafety implemented for laboratories. From this perspective, genetic data can be considered as a possible domain where a higher level of protection should be implemented due to the risk of healthcare data capture by a non-specialized AI.

7. Concluding remarks

The use of DS and AI is increasing and this will contribute to the transformation of laboratory medicine. These technologies carry the potential to address unmet clinical needs by enhancing personalized patient care and improving the efficiency of laboratory processes and care pathways. Such evolution is in its infancy, however, and the balance between usability and desirability remains a matter of concern. Multiple challenges are paving the way prior to a transfer of these concepts into clinical practices, such as recovery plans and validation.

As for other laboratory activities and medical devices, continuous operation and care continuity will have to be guaranteed. In the high-velocity data environment, efficient back-ups and the redundancy of servers in off-premises locations are critical for minimising potential downtime and restoring data, operating systems, applications, files, and folders. Teams will therefore have to define recovery point objectives and the maximum acceptable amount of data loss and conduct regular recovery tests. The need for extensive validation of clinical features and performance will also be a key element. Multidisciplinary teams that include physicians, laboratorians, data scientists, and healthcare professionals will have to be involved.

Another key element of the integration of DS into laboratory practices will be the ability of international scientific societies to outline best practice guidance. Addressing the ethical challenges associated with AI and ML is also essential. Prospective and interdisciplinary validation of AI systems is therefore needed to ensure their reliability and their integration with an adequate body of positive soft-law regulations.

References

[1] M. Sonja, G. Ioana, Y. Miaoqing, K. Anna, Understanding value in health data ecosystems: A review of current evidence and ways forward, Rand Health Q. 7 (2) (2018 Jan) 3 [Internet]. [cited 2018 Dec 31]. Available from: http://www.ncbi.

nlm.nih.gov/pubmed/29416943.

[2] L.N. Sanchez-Pinto, Y. Luo, M.M. Churpek, Big data and data science in critical care, Chest 154 (5) (2018 Nov) 1239–1248 [Internet]. [cited 2019 Jan 1]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29752973.

[3] S.J. Russell, J. Stuart, P. Norvig, J. Canny, Artificial Intelligence: A Modern Approach, [Internet] Pearson Education, 2003, p. 1081 [cited 2018 Aug 26]. Available from: https://dl.acm.org/citation.cfm?id=773294.

[4] D.D. Miller, E.W. Brown, Artificial intelligence in medical practice: The question to the answer? Am. J. Med. 131 (2) (2018 Feb) 129–133 [Internet]. [cited 2018 Aug 26]. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002934317311178.

[5] T.M. Mitchell, M. Tom, Machine Learning, [Internet] McGraw-Hill, 1997, p. 414 [cited 2018 Aug 26]. Available from: http://www.cs.cmu.edu/~tom/mlbook.html.

[6] R. Liu, B. Yang, E. Zio, X. Chen, Artificial intelligence for fault diagnosis of rotating machinery: a review, Mech. Syst. Signal Process. 108 (2018 Aug) 33–47 [Internet]. [cited 2018 Aug 26]. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0888327018300748.

[7] C. Linnhoff-Popien, R. Schneider, M. Zaddach (Eds.), Digital Marketplaces Unleashed, Springer Berlin Heidelberg, Berlin, Heidelberg, 2018[Internet]. [cited 2018 Aug 26]. Available from: http://link.springer.com/10.1007/978-3-662-49275-8.

[8] P. Sajda, Machine learning for detection and diagnosis of disease, Annu. Rev. Biomed. Eng. 8 (1) (2006 Aug) 537–565 [Internet]. [cited 2019 Jan 4]. Available from: http://www.annualreviews.org/doi/10.1146/annurev.bioeng.8.061505.095802.

[9] J.C. Ang, A. Mirzal, H. Haron, Hamed HNA, Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection, IEEE/ACM Trans. Comput. Biol. Bioinforma. 13 (5) (2016 Sep 1) 971–989 [Internet]. [cited 2018 Dec 31]. Available from: http://ieeexplore.ieee.org/document/7264992/.

[10] R. De Sanctis, A. Viganò, A. Giuliani, A. Gronchi, A. De Paoli, P. Navarria, et al., Unsupervised versus supervised identification of prognostic factors in patients with localized retroperitoneal sarcoma: a data clustering and mahalanobis distance approach, Biomed. Res. Int. (Apr 23 2018) 2786163 [Internet]. [cited 2018 Dec 31]. Available from: https://www.hindawi.com/journals/bmri/2018/2786163/.

[12] W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, Health Inf. Sci. Syst. 2 (2014) 3 [Internet]. [cited 2018 Aug 26]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25825667.

[13] Global Big Data in Healthcare Market Research Report, Trends & Forecast, [Internet]. [cited 2018 Aug 26]. Available from: https://bisresearch.com/industry-report/global-big-data-in-healthcare-market-2025.html.

[14] B. Meskó, G. Hetényi, Z. Győrffy, Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Serv. Res. 18 (1) (2018 Jul 13) 545 [Internet]. [cited 2018 Aug 26]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30001717.

[15] V. Kudva, K. Prasad, S. Guruvare, Automation of detection of cervical cancer using convolutional neural networks, Crit. Rev. Biomed. Eng. 46 (2) (2018) 135–145 [Internet]. [cited 2019 Mar 25]. Available from: http://www.dl.begellhouse.com/journals/4b27cbfc562e21b8,6b7afd80398dbd73,6bd9a39e101b6402.html.

[16] M. Kyrgiou, A. Pouliakis, J.G. Panayiotides, N. Margari, P. Bountris, G. Valasoulis, et al., Personalised management of women with cervical abnormalities using a clinical decision support scoring system, Gynecol. Oncol. 141 (1) (2016 Apr) 29–35 [Internet]. [cited 2019 Mar 25]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27016226.

[17] M. Wu, C. Yan, H. Liu, Q. Liu, Y. Yin, Automatic classification of cervical cancer from cytological images by using convolutional neural network, Biosci. Rep. 38 (6) (2018 Dec 21) [Internet]. [cited 2019 Mar 25]. BSR20181769. Available from: http://www.bioscirep.org/content/38/6/BSR20181769.long.

[18] M. Kusy, B. Obrzut, J. Kluska, Application of gene expression programming and neural networks to predict adverse events of radical hysterectomy in cervical cancer patients, Med. Biol. Eng. Comput. 51 (12) (2013 Dec 18) 1357–1365 [Internet]. [cited 2019 Mar 25]. Available from: http://link.springer.com/10.1007/s11517-013-1108-8.

[19] D. Peteiro-Barral, B. Guijarro-Berdiñas, A survey of methods for distributed machine learning, Prog. Artif. Intell. 2 (1) (2013 Mar 15) 1–11 [Internet]. [cited 2018 Aug 26]. Available from: http://link.springer.com/10.1007/s13748-012-0035-5.

[20] D. Caragea, A. Silvescu, V.A. Honavar, Framework for learning from distributed data using sufficient statistics and its application to learning decision trees, Int. J. Hybrid Intell. Syst. 1 (1–2) (2004 Apr 1) 80–89 [Internet]. [cited 2018 Aug 26]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20351798.

[21] B.H. Shirts, B.R. Jackson, G.S. Baird, J.M. Baron, B. Clements, R. Grisson, et al., Clinical laboratory analytics: challenges and promise for an emerging discipline, J. Pathol. Inform. 6 (2015) 9 [Internet]. [cited 2018 Aug 26]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25774320.

[22] Apache Mahout, [Internet]. [cited 2018 Nov 2]. Available from: https://mahout.apache.org/.

[23] Amazon Machine Learning – Analyses prédictives Avec AWS, [Internet]. [cited 2018 Nov 2]. Available from: https://aws.amazon.com/fr/aml/.

[24] TensorFlow White Papers | TensorFlow, [Internet]. [cited 2018 Nov 2]. Available from: https://www.tensorflow.org/about/bib.

[25] M.J. van de Vijver, Y.D. He, L.J. van 't Veer, H. Dai, Hart AAM, D.W. Voskuil, et al., A Gene-Expression Signature as a Predictor of Survival in Breast Cancer, N. Engl. J. Med. 347 (25) (2002 Dec 19) 1999–2009 [Internet]. [cited 2018 Nov 2]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12490681.

[26] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, Hart AAM, M. Mao, et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (6871) (2002 Jan 31) 530–536 [Internet]. [cited 2018 Nov 2]. Available from: http://www.nature.com/articles/415530a.

[27] Feature Extraction: Foundations and Applications - Google Books, [Internet]. [cited 2018 Nov 2]. Available from: https://books.google.be/books?id=FOTzBwAAQBAJ&pg=PR3&lpg=PR3&dq=Isabelle+Guyon,+fuzziness&source=bl&ots=5Tk4L68urZ&sig=M4zOuJG76uAnj8gxKw3r0Dgs0mw&hl=en&sa=X&ved=2ahUKEwiQit2ZybXeAhUK6RoKHRuZCUgQ6AEwCnoECAAQAQ#v=onepage&q=Isabelle Guyon%2C fuzziness&f.

[28] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics 26 (3) (2010 Feb 1) 392–398 [Internet]. [cited 2018 Nov 2]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19942583.

[29] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, Francisco Herrera, Learning from Imbalanced Data Sets, Google Books, 2018 [Internet]. [cited 2019 Jan 4]. Available from: https://books.google.be/books?id=8Fp0DwAAQBAJ&pg=PA117&lpg=PA117&dq=Almogahed+BA,+Kakadiaris+IA.+Neater:+filtering+of+over-sampled+data+using+noncooperative&source=bl&ots=3vQM8wsG2v&sig=_v0FwrL6pKBUUUFhEaPcfUMrZ0g&hl=en&sa=X&ved=2ahUKEwjqt5jawdTfAhVOLFAKH.

[30] The Journal of the Acoustical Society of America - Acoustical Society of America, Google Books, 2005 [Internet]. [cited 2018 Nov 2]. Available from: https://books.google.be/books?id=3MGGAAAAIAAJ&q=The+Lack+of+A+Priori+Distinctions+Between+Learning+wolpert&dq=The+Lack+of+A+Priori+Distinctions+Between+Learning+wolpert&hl=en&sa=X&ved=0ahUKEwiB78T-ybXeAhUI3KQKHRXIABw4ChDoAQgqMAE.

[31] L.D. Fiore, P.W. Lavori, Integrating randomized comparative effectiveness research with patient care, Drazen JM, Harrington DP, McMurray JJV, Ware JH, Woodcock J, editors. N. Engl. J. Med. 374 (22) (2016 Jun 2) 2152–2158 [Internet]. [cited 2018 Aug 26]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27248620.

[32] P. Fraccaro, B. Brown, M. Prosperi, M. Sperrin, I. Buchan, N. Peek, Development and preliminary validation of a dynamic, patient-tailored method to detect abnormal laboratory test results, Stud Health Technol Inform, 216 2015, pp. 701–705 [Internet]. [cited 2018 Aug 26]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26262142.

[33] K. Lee, B. Kim, Y. Choi, S. Kim, W. Shin, S. Lee, et al., Deep learning of mutation-gene-drug relations from the literature, BMC Bioinforma. 19 (1) (2018 Jan 25) 21 [Internet]. [cited 2018 Aug 26]. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2029-1.

[34] B.R. Lauwerys, D. Hernández-Lobato, P. Gramme, J. Ducreux, A. Dessy, I. Focant, et al., Heterogeneity of synovial molecular patterns in patients with arthritis, PLoS One 10 (4) (2015) [Internet]. [cited 2018 Aug 26]. e0122104. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25927832.

[35] M. Henglin, G. Stein, P.V. Hushcha, J. Snoek, A.B. Wiltschko, S. Cheng, Machine learning approaches in cardiovascular imaging, Circ. Cardiovasc. Imaging 10 (10) (2017 Oct 27) [Internet]. [cited 2018 Aug 26]. e005614. Available from: http://circimaging.ahajournals.org/lookup/doi/10.1161/CIRCIMAGING.117.005614.

[36] https://open.epic.com/.

[37] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, T. Kitai, Artificial intelligence in precision cardiovascular medicine, J. Am. Coll. Cardiol. 69 (21) (2017 May 30) 2657–2664 [Internet]. [cited 2018 Aug 26]. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0735109717368456.

[38] K. Shameer, K.W. Johnson, B.S. Glicksberg, J.T. Dudley, P.P. Sengupta, Machine learning in cardiovascular medicine: are we there yet? Heart 104 (14) (2018 Jul) 1156–1164 [Internet]. [Cited 2018 Aug 26]. Available from http://www.ncbi.nlm.nih.gov/pubmed/29352006.

[39] A. Dudchenko, G. Kopanitsa, Decision support systems in cardiology: a systematic review, Stud Health Technol Inform, 237 2017, pp. 209–214 [Internet]. [cited 2018 Aug 26]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28479570.

[40] I.V. Buzaev, V.V. Plechev, I.E. Nikolaeva, R.M. Galimova, Artificial intelligence: Neural network model as the multidisciplinary team member in clinical decision support to avoid medical mistakes, Chronic Dis. Transl. Med. 2 (3) (2016 Sep) 166–172 [Internet]. [cited 2018 Aug 26]. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2095882X16300238.

[41] M.V. Dogan, I.M. Grumbach, J.J. Michaelson, R.A. Philibert, Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham Heart study, Zeller T, editor, PLoS One 13 (1) (2018 Jan 2) [Internet]. [cited 2018 Aug 26]. e0190549. Available from: http://dx.plos.org/10.1371/journal.pone.0190549.

[42] I. Ahmed, N.S. Ahmad, S. Ali, S. Ali, A. George, H. Saleem Danish, et al., Medication adherence apps: review and content analysis, JMIR mHealth uHealth 6 (3) (2018 Mar 16) e62 [Internet]. [cited 2018 Sep 4]. Available from: http://mhealth.jmir.org/2018/3/e62/.

[43] C. Wanty, T. Helleputte, F. Smets, E.M. Sokal, X. Stephenne, Assessment of risk of bleeding from esophageal varices during management of biliary atresia in children, J. Pediatr. Gastroenterol. Nutr. 56 (5) (2013 May) 537–543 [Internet]. [cited 2018 Nov 2]. Available from: https://insights.ovid.com/crossref?an=00005176-201305000-00015.

[44] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, et al., Using Resistin, glucose, age and BMI to predict the presence of breast cancer, BMC Cancer 18 (1) (2018 Dec 4) 29 [Internet]. [cited 2018 Aug 26]. Available from: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-017-3877-1.

[45] Guebila M. Ben, I. Thiele, Model-based dietary optimization for late-stage, levodopa-treated, Parkinson's disease patients, NPJ Syst. Biol. Appl. 2 (2016) 16013 [Internet]. [cited 2018 Aug 26]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28725472.

[46] C.-A. Azencott, Machine learning and genomics: precision medicine versus patient privacy, Philos. Trans. A Math. Phys. Eng. Sci. 376 (2128) (2018 Sep 13) [Internet].

[cited 2018 Dec 31]. 20170350. Available from: http://rsta.royalsocietypublishing.org/lookup/doi/10.1098/rsta.2017.0350.

[47] Y. Ming, T. Zhang, Efficient privacy-preserving access control scheme in electronic health records system, Sensors (Basel) 18 (10) (2018 Oct 18) 3520 [Internet]. [cited 2018 Dec 31]. Available from: http://www.mdpi.com/1424-8220/18/10/3520.

[48] M. Kayaalp, Patient privacy in the era of big data, Balkan Med. J. 35 (1) (2018 Jan 20) 8–17 [Internet]. [cited 2018 Dec 31]. Available from: http://balkanmedicaljournal.org/pdf.php?&id = 1788.

[49] Y. Flaumenhaft, O. Ben-Assuli, Personal health records, global policy and regulation review, Health Policy 122 (8) (2018 Aug) 815–826 [Internet]. [cited 2018 Dec 31]. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168851018301325.

[50] B. John, Are you ready for general data protection regulation? BMJ 360 (2018 Mar 2) k941 [Internet]. [cited 2018 Dec 31]. Available from: http://www.bmj.com/lookup/doi/10.1136/bmj.k941.

[51] J. Salerno, B.M. Knoppers, L.M. Lee, W.M. Hlaing, K.W. Goodman, Ethics, big data and computing in epidemiology and public health, Ann. Epidemiol. 27 (5) (2017 May) 297–301 [Internet]. [cited 2018 Dec 31]. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1047279717300017.

[52] D.S. Char, N.H. Shah, D. Magnus, Implementing machine learning in health care – addressing ethical challenges, N. Engl. J. Med. 378 (11) (2018 Mar 15) 981–983 [Internet]. [cited 2018 Dec 31]. Available from: http://www.nejm.org/doi/10.1056/NEJMp1714229.

[53] Intelligence artificielle - Ethik IA prône la régulation, [Internet]. [cited 2018 Aug 26]. Available from: https://www.lequotidiendumedecin.fr/actualites/article/2018/01/25/ethik-ia-prone-la-regulation_854511.

[54] D. Gruson, Artificial intelligence in health: 10 key messages, Rev Prat 68 (10) (2018 Dec) 1152 [Internet]. [cited 2019 Apr 1]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30869232.