 doi:10.20944/preprints202204.0078.v2

Review

Big Data in Laboratory Medicine – FAIR Quality for AI?

Tobias Ueli Blatter[1], Harald Witte[1], Christos Theodoros Nakas[1] and Alexander Benedikt Leichtle[1,2]

[1] Department of Clinical Chemistry, Inselspital – University Hospital Bern, Bern Switzerland
[2] Center of Artificial Intelligence in Medicine (CAIM), University of Bern, Switzerland

Abstract: Laboratory medicine is a digital science. Every large hospital produces a wealth of data each day - from simple numerical results from e.g. sodium measurements to highly complex output of "-omics" analyses, as well as quality control results and meta-data. Processing, connecting, storing, and ordering extensive parts of these individual data requires Big Data techniques. Whereas novel technologies such as artificial intelligence and machine learning have exciting application for the augmentation of laboratory medicine, the Big Data concept remains fundamental for any sophisticated data analysis in large databases. To make laboratory medicine data optimally usable for clinical and research purposes, they need to be FAIR: findable, accessible, interoperable, and reusable. This can be achieved for example by automated recording, connection of devices, efficient ETL (Extract, Transform, Load) processes, careful data governance, and modern data security solutions. Enriched with clinical data, laboratory medicine data allow a gain in pathophysiological insights, can improve patient care, or they can be used to develop reference intervals for diagnostic purposes. Nevertheless, Big Data in laboratory medicine do not come without challenges: The growing number of analyses and data derived from them is a demanding task to be taken care of. Laboratory medicine experts are and will be needed to drive this development, take an active role in the ongoing digitalization, and provide guidance for their clinical colleagues engaging with the laboratory data in research.

Keywords: digitalization; clinical chemistry; artificial intelligence; interoperability; FAIRification

1. Introduction

Laboratory medicine has always been one of the medical disciplines with the highest degree of digitalisation. Since its emergence, automation, electronic transmission of results and electronic reporting have become increasingly prevalent[1]. In addition, medical laboratories maintain extensive databases, not only with test results, but also with results from quality controls. Furthermore, they are usually equipped with elaborate quality management systems. It is therefore not surprising that laboratory medicine represents a paradigm discipline for the digitalisation of medicine. In contrast, the latest developments in the data science field, such as artificial intelligence (AI) and machine learning (ML), have not yet found their way into laboratory medicine across the board. Nevertheless, the time is now. Three key ingredients for augmenting laboratory medicine have become available to researchers on a wider scale: Learning and training algorithms, necessary computational power to run said algorithms, and high-volume data[2]. These latest and future developments of AI and ML in laboratory medicine, however, do not constitute the main focus of this experience-based opinion article, several recently published reviews can offer an excellent overview[1–5]. We will instead highlight the principles required for high quality, clinical, "big" data. Without solid data as a foundation, even the most refined algorithms will fail to draw reliable conclusions: "*Ex falso sequitur quodlibet*", or put more coarsely, "*garbage in, garbage out*". The manifold requirements and pitfalls for Big Data analysis in laboratory medicine shall be reviewed below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 July 2022 doi:10.20944/preprints202204.0078.v2

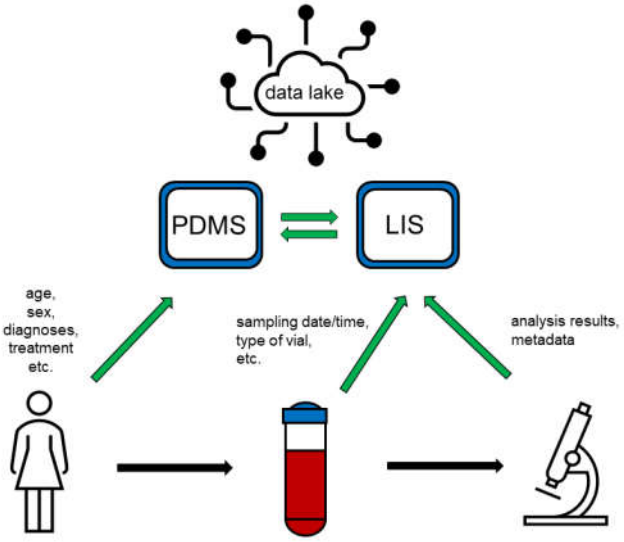

Figure 1. Patients´ data is entered into the patient data management system (PDMS) predominantly manually, while information about samples collected as well as about analyses conducted is entered into the laboratory information system (LIS) either manually or automatically. PDMS and LIS are connected and exchange parts of their stored data. Both systems feed a "data lake" comprising various types of data, which can be provided to researchers for Big Data applications.

2. Definitions surrounding Big Data

Since the terminology "Big Data" has gained traction in the late 1990's, driven by need to address the increasing data collection both in the private and public sector, there has never been a generally agreed upon understanding of the term[6]. Initial depictions of Big Data were observing the phenomena and stating the emergence of a new discipline without a cohesive clarifications on what the Big Data term encompasses[7]. A more formal definition, proposed recently, suggests three main V's (Volume, Velocity and Variety) as key dimensions of Big Data with the added requirement for specially designed technologies and analytics to translate data into value[8]. Subsequently, different scientific disciplines have attributed and highlighted other dimensionalities to Big Data (e.g. Value or Veracity[9,10]) while neglecting previously mentioned dimensionalities, making a holistic semantic definition rather difficult[11,12]. Ultimately, the strength of the "Big Data" conceptualization lays in its analytics, where Big Data is translated into clinical merit.

Big Data analysis is the assessment of massive amounts of information from multiple electronic sources in unison, by sophisticated analytic tools to reveal otherwise unrecognized patterns[13]. If we consider "standard" laboratory analyses, e.g. clinical chemistry, haematology or hemostaseology, the lion´s share of analysis results consists of numerical results, possibly enriched with reference ranges. Notably, in terms of size - not comprehensibility from the point of view of a human observer though - all laboratory results of a medium-sized university laboratory fit on a standard hard disk. The situation is different with "-omics" data, which, depending on the technology, can comprise of several hundred megabytes to several gigabytes, be it NGS (Next Generation Sequencing) data or proteome or metabolome data [14]. A distinction must also be made between the usually very extensive, often proprietary raw data, and pre-processed data, which are often available in tabular form and correspond to standard multiplex laboratory analyses in terms of volume. Other fields with extensive data volumes are diagnostic diagrams, whose information content may be limited, but who require large storage capacities, when saved in the form of graphs; and diagnostic image data, e.g. from microscopy or MRI (magnetic resonance imaging). Another Big Data resource not to be underestimated is also non-patient-related data, i.e. calibration and quality control data, which are often stored and administrated in specialised databases.

 doi:10.20944/preprints202204.0078.v2

Laboratory data are best suited to the Big Data concept if they are enriched with clinical data from the hospital's various IT systems.

3. Transforming Laboratory Medicine into Big Data Science

Requirements

Even though laboratory medicine databases constitute a rich source of data, frequently these are ill-suited for the application of data science techniques. Created to fit regulatory requirements instead of research purposes, most databases store data inefficiently and only for the minimally required retention period. Providing insufficient data quality for most research questions, databases are transformed into mere data dumps. So, what are the prerequisites for optimally usable laboratory medical data[15]? Central attributes data needs to have to be optimally suited for research use are summarized by the key word "FAIR": Finable, Accessible, Interoperable, Reusable[16]. (cf. *Table 1*)

Findable data must be stored in a way that enables easy retrieval. For "standard" examinations, this is usually realised though a patient identifier (PID) and date, so individual results can be assigned to the respective patients and collection times. Depending on the organisation of the laboratory, this is easier said than done. Potential pitfalls are for example, that the same PIDs might be assigned to different patients in different branch laboratories, or that analyses conducted for unidentified emergency patients, cannot be attributed to the correct person when their identity has been clarified. Additionally, results of different patients might be combined under a "collective" PID for research purposes. Also, data can be confusing when samples are registered with the planned collection date instead of the actual collection data resulting in analysis time points prior to collection. Equipment for special examinations poses particular challenges to findability, as they are frequently not connected to the laboratory information system (LIS). Here, the patient ID may be entered manually into the evaluation files in a way that does not conform to the standard, which can lead to confusion and incomplete entries. An example of this are "-omics" analyses: Analytical devices routinely produce and output files too large for transfer and storage in the central LIS. Therefore, they need to be linked, preferably in a searchable manner to enable offline findability. Likewise, findability has to be addressed in the sharing of machine-actionable (meta-)data online. Good meta-data makes data findable. In web 1.0/2.0 approaches, this was addressed by the Linked Data Principles, a set of best practices when publishing structured data to the web[17]. These principles were however proposed before the emergence of FAIR, meaning that little emphasis was put on standardisation and a variety of inherently different schemas were proposed[18]. One of the most recent effort for making semantic artefacts FAIR has been launched by the FAIRsFAIR project, where the authors list recommendations for findable (meta-)data, highlighting the need for GUPRIs (Globally Unique, Persistent and Resolvable Identifiers), highly enriched and searchable (meta-)data descriptions and, especially relevant for clinical laboratory sciences, the need to publish data and metadata separately[19]. Findability remains one of the most important aspects of the FAIRification of Big Data analysis, as a lack of appropriate metadata standards affects the availability of research data in the long term. A recent study observed decreased findability of UK health datasets over time[20], a trend also observed in a greater context of data-driven science both in terms of the findability of datasets and the reachability of the responsible authors[21].

The **accessibility** of laboratory data can also be a challenge. LISs usually do not have freely accessible query functionalities because of regulatory requirements. Therefore, LISs that are not connected to central clinical data warehouses, must be accessed through the laboratory IT personnel. This often leads to an enormous amount of additional work, since laboratory data are highly attractive for a variety of research projects[22]. For use in clinical data warehouses, the LISs must be electronically connected, and the data prepared via ETL processes (Extract, Transform, Load). This requires the use of universal web standards including HTTP (Hypertext Transfer Protocol), standardized data exchange formats (e.g. FHIR[23] and the semantic-based Resource Description Framework (RDF)[24,25])

 doi:10.20944/preprints202204.0078.v2

and tools which allow querying respective data (e.g. SPARQL[26]). Additionally, data models in particular OMOP[27,28] or i2b2[29] are in common use. In true FAIR fashion, LISs must present standard API (Application Programming Interface) with secure access protocols (SSL) for data management and retrieval[19]. Generally, not the entire content of the databases is transferred, but a limited subset of data (e.g. data records that can be clearly assigned to patients) is identified and transmitted. A special challenge in this context is posed by legacy systems that are solely operated in read-only mode and where the effort for the technical connection must be weighed against the benefit of the further use of the data contained. Moreover, as the available data for researchers grows there need to be mechanism in place to enable privacy protection with the use of de-identification or anonymization algorithms. While textbook method, for instance k-anonymity[30] or l-diversity[31], are often cited they do not come without their limitations[32–34]. In this context, the question arises as to who is allowed to access the laboratory data and under what conditions. For example, data relating to infection serologies or staff medical service is particularly sensitive and requires careful data governance[35]. Another important aspect is the question of patient consent for research projects access needs to be restricted according to regulatory requirements[36]. The use of patient data in research in Switzerland is governed by the Federal Act on Data Protection (FADP 1992, art. 3c) and the Human Research Act (HRA RS 810.30). Notably, the governance of Big Data is not different from "regular" research data: A request on the disposal and use of sensitive data must be submitted to a cantonal REC (Research Ethics Committee). Big Data research proposes novel ethical concerns[37], mostly surrounding the notions of privacy (hindrance of individual reidentification) and consent (possibility to later revoke consent), where traditional ethics oversight practice is often unaware of the direct societal impact of their decisions[38]. A recent study in Switzerland showed, that members of the seven Swiss RECs had broadly differing views in regards to the opportunities and challenges of Big Data citing insufficient expertise in big data analytics or computer science to adequately judge the use of Big Data in clinical research[39]. This can become especially cumbersome for researchers when data from different institutions are merged - in this case, modern systems that work with secure multiparty computing and homomorphic encryption, such as the MedCo system, can be a promising approach[40]. Wirth et al. offer a great overview regarding privacy-preserving data sharing infrastructures for medical research[41].

The next big and perhaps most important aspect for Big Data in laboratory medicine is the necessary **semantic interoperability**. This means that the individual data items must be clearly assigned semantically, ideally by means of standardised coding along the lines of LOINC (Logical Observation Identifiers Names and Code). This represents an enormous challenge, which has been addressed in Switzerland, for example, by the L4CHLAB project[42]. It is not enough to identify laboratory analyses only by their trivial name (e.g. "potassium") - the necessary granularity is defined by the requirements of the research projects based on it. Thus, a creatinine measurement of any kind may be sufficient as a "safety lab measurement", but completely insufficient for a method comparison study or the establishment of reference intervals. It should be noted that currently there is no universal standard, as even LOINC does not specify e.g. device manufacturer and kit version, which need to be coded additionally. Unique identifiers for medical devices, e.g. from the GUDID[43] or EUDAMED database[44], or type identifiers, e.g. from medical device nomenclatures like GMDN[45] or EMDN[46], may enrich the LOINC system and increase its acceptance. Extensive preparatory work to address this issue has been done by the Swiss Personalized Health Network (SPHN), which established corresponding "concepts"[47]. Particular difficulties arise from historically grown LISs, which are often not structured according to the 1:1 principles of a LOINC nomenclature and prevent a clean assignment of laboratory analyses to unambiguous codes. This must be considered especially when replacing and updating LISs, so that the master data remains future-proof and interoperable[14]. The use of advanced data models like RDF is beneficial here as it allows a data scheme to evolve over time without the need to change the original data[25]. In the university environment, the latest test technology might be employed, using analyses

 doi:10.20944/preprints202204.0078.v2

which do not yet have a LOINC code assigned, making it necessary to deviate accordingly. For the consolidation of large amounts of data from different sources, a high semantic granularity, which is necessary for individual questions, can be problematic, equivalent analyses must be defined as such in order to enable comprehensive evaluations. Here, Minimum Information Checklists (MICs), stating the minimum requirements for quality and quantity to make data descriptions accurate and useful, could offer a needed standardisation to track data quality from various sources. It is essential that a core vocabulary features support for descriptions to be machine readable RDF[48], closely linking the commonly used semantics in laboratory medicine with machine-actionable descriptions. The use of semantic web technologies, such as RDF, in the laboratory environment could also help to establish the common use of Electronic Lab Notebooks (ELNs)[49]. Notably, the application of suitable data formats facilitates, but by itself does not guarantee actual interoperability of data sets from different data providers. Seemingly trivial details including spelling, cardinalities, datatypes, consistent use of GUPRIs or measurement units must be carefully assessed. In the context of RDF, the Shapes Constraint Language (SHACL) allows the testing and validating of data against a set of predefined requirements[50]. These conditions (SHACL rules) constitute a "shape graph" against which the actual data (as "data graph") is matched. The expression of complex constraints is facilitated by SHACL-extensions supporting SPARQL and JavaScript[51,52]. Despite the rise of user-friendly validation tools semantic standards alone are not a "silver bullet" against data mayhem. In fact, even with maximum semantic care, the competence of experts in laboratory medicine remains in high demand. Different automated approaches for resolving the semantic heterogeneity when mapping different ontologies have been launched but still require human oversight[53,54]. For many researchers who come from non-analytical subjects, the differences in the meaning of the analysis codes are not obvious at first glance. Considerable misinterpretations can occur, e.g. calculation of eGFR from urine creatinine. Here the laboratory holds responsibility since it has the necessary competence to avoid such errors.

The **reusability** of laboratory medical data depends to a large extent on the existence and level of detail of the associated metadata. This includes - as already mentioned - analysis-related data (mapped in the dimensions of LOINC), but also, batch numbers, quality management data and, if applicable, SPRECs (Sample PREanalytical Codes)[55] codes - in essence, everything that is or could be of importance for optimal replicability of the measurement results. It can be problematic that the metadata are stored in separate databases and cannot be provided automatically via the ETL processes, so that they can neither be exported nor viewed. Not only the (meta-)data needs to be reusable but also the algorithms and data processing scripts. With "FAIRly big", a functional framework for retracing and verifying the computational processing of large-scale data based on machine-actionable provenance records, high performance could be observed regarding data sharing, transparency and scalability, despite ignoring explicit metadata standards[56]. Reusability can also refer to the efficient use of statistical models that may arise using Machine Learning methodology. The latter may involve a feedback process where the model is validated and even further calibrated as information arrives through the expansion of the database with fresh data. Potential pitfalls impairing reusability may include legislative limitations imposed by national research acts or legal ambiguities in Data Transfer and Use Agreements (DTUA) of multicentre cohort studies involving several data providers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 July 2022 doi:10.20944/preprints202204.0078.v2

Table 1. Recommendations for the FAIRification of laboratory data.

Requirement	Implementation
Findability	- Assign PIDs meaningfully. • Each PID should uniquely identify a single patient, this needs to be consistent between branch laboratories with parallel systems. • Develop solutions for unknown emergency patients, that allows correct assignment of test results when personal data is identified later on. • Develop solutions for analyses conducted for research purposes. Avoid cumulative PIDs. - Record actual sampling time instead of planned sampling time - Connect all analytical devices to lab IT system to avoid manual entries. - Connect the lab IT system to the hospitals central IT system to enable searches by clinicians and researchers.
Accessibility	- Protect lab data adequately with • secure data storage solutions • careful data governance - Design ETL processes efficiently. - Consider general consent status of patients and allow access to data accordingly. - Employ modern technical solutions such as multiparty computing and homomorphic encryption for merging data from different sites.
Interoperability	- Code analyses in a standardized manner, e.g. with LOINC codes. - Additionally, code device manufacturer and kit version in a standardized way - Code newly developed analyses in a homogenous way, even if no standardized codes are available yet. - Enable consolidation of data from different labs.
Reusability	- Provide detailed meta-data to maximize reproducibility, including • LOINC codes • batch numbers • quality management data • SPREC codes
+	Offer your laboratory medicine expertise to clinicians and researchers, as no one knows the intricacies of your laboratory data better than you.
abbreviations: ETL: extract – transform – load, lab: laboratory, LOINC: Logical Observation Identifiers Names and Code, PID: patient identifier, SPREC: Standard Preanalytical Code	

Risks

The use of laboratory medical data for Big Data analytics does not only have advantages, but it is also associated with a considerable number of risks: as all health data, laboratory values are worthy of special protection. As with all information compiled in large databases, there is an imminent risk of data leaks, especially if the data are accessible from the outside. Structured laboratory data can also be copied easily and quickly due to their small file size, so there is a considerable risk of unauthorized data duplication.

 doi:10.20944/preprints202204.0078.v2

Similarly, data governance must be ensured, which requires a comprehensive authorisation framework - this is easier to implement in closed LISs. Another essential aspect is data integrity, which must be ensured in particular through the ETL process pipelines and also for further processing. LISs, as medical products, usually fulfil the necessary standards, but with self-written transformation scripts this may be different and enforce a meticulous quality control. However, this has the advantage that non-data transfer-related errors can also be detected and deleted. In any case, certification of the IT processes is both sensible and costly. Post-analytics can also cause difficulties – the IT systems of the receivers (clinicians or researchers) must be able to handle the data formats supplied and must not alter or falsify their presentation. Another enormously problematic aspect is change tracking. In the LISs, laboratory tests are often identified by means of their internal analysis numbers - if changes occur here, e.g. due to the inclusion of new analyses, changes must be reported to the peripheral systems - preferably automatically and with confirmation of knowledge - otherwise serious analysis mix-ups can occur. Finally, when individual laboratory data are queried, the framework of the findings is no longer guaranteed - the analyses lose their context and thus their interpretability.

Chances

The introduction of "Big Data" technologies holds great potential for laboratory medicine, and some aspects will be specifically addressed here:

Setting up ETL processes inevitably leads to the detection of inadequacies in the structure and content of the laboratory's master data. Frequently, LISs have grown over years and - although continuously maintained - are not organised in a fundamentally consistent manner. Before one can begin with the extraction and processing of laboratory data, the data organisation, structure, and meta-information must already be disclosed in the source system. A thorough review of this data is recommended to be carried out in the mother database because tidying up is in any case necessary, and quite obviously better done in the source system than in subordinate databases. Another important aspect is the necessary introduction of clear semantics - this is a laborious process that initially represents a large workload but is subsequently relatively easy to maintain. Many laboratories are reluctant to take on this effort - here the diagnostics manufacturers are asked to supply the necessary codes (e.g.: extended LOINC codes, see above) for the analyses they offer, e.g. in tabular form, which makes bulk import considerably easier and a matter of a few days. For researchers in particular, it is also extremely helpful to have a data catalogue created in this context. Laboratory catalogues are often available electronically, but are often organized around request profiles, rather than individual analyses that are often of importance for research questions. The IT teams of the data warehouses will also be very grateful for appropriate documentation. This also offers the opportunity to make extensive metadata accessible and usable for interested researchers. Together with the introduction of semantics and data catalogues, transparent change tracking should be integrated so queries in the data warehouses can be adapted accordingly if, for example, analyses have changed, or new kits have been used. Change tracking is also clearly to be advocated from a good laboratory practice (GLP) point of view.

Another aspect of outstanding importance for laboratory medicine as a scientific subject is the visibility and documentability of the contribution of laboratory medicine to research projects. In the vast majority of clinical studies, laboratory data play an extremely important role, be it as outcome variables, as safety values, as quality and compliance indicators, or as covariates. With a transparent database and query structure, the use and publication impact of laboratory data can be shown more clearly and the position of the laboratory in the university environment as an essential collaboration and research partner can be strengthened. Other aspects include the improved use of patient data for research purposes - turning laboratory databases from graveyards of findings into fertile ground for research, an aspect that is certainly in the interest of patients in the context of improvement of treatment options. The improved indexability of laboratory data in large

 doi:10.20944/preprints202204.0078.v2

"data lakes" would also allow to link them to clinical data. Conversely, this also opens up completely new research possibilities for laboratory scientists, as the laboratory values no longer stand alone, but can be analysed in a clinical context. Last but not least, a cleanly curated database is an essential foundation for AI applications: It's like in most data science projects: 80% of the effort is data tidying, and 20% is the "fun part" of the analysis. Here the laboratories have to point out their very important, but little prestigious and extremely tedious role. - They are essential partners in the vast majority of research collaborations.

Fields of application

Big Data with its technological environment does not yet represent a translation into medical fields of application, but it should be regarded as a basis and facilitator for a large number of potential uses. Mainly applications come into consideration that already require a large amount of information to be processed and thus bring the human part of the evaluation pipeline to a processing limit. These include, of course, data-intensive "-omics" technologies including pattern recognition in specialised metabolic diagnostics and newborn screening, but also technical and medical validation and quality management. Further applications can be population-based evaluations such as the creation of reference value intervals. In the following, some of the potential fields of application are described.

An obvious field for Big Data technologies in laboratory medicine are "-omics" applications [57–59]: These have been developed for nucleic acid-based techniques as e.g. genomics[60,61], transcriptomics[62], and epigenomics[63], as well as for mass spectrometry-based methodologies such as proteomics[64,65] , metabolomics[66,67], lipidomics[68] and others. The particular challenges in this field include connecting the analysis systems to the corresponding data lakes - it is no longer possible to work with traditional database technologies and new approaches for example hadoop[69] become necessary. Even more than in the case of highly standardised routine procedures in classical laboratory medicine, metadata play an outstanding role in evaluability, comparability, and replicability. In addition, the raw data generated with these procedures are often formatted in a proprietary manner, and also of enormous size - comparable only with the data sets of the imaging disciplines. For retrieval, indexing and linking to the respective patient must be ensured; this can be achieved, for example, by linking tables of processed results instead of raw data output. The extent to which transformation and evaluation steps already make sense in the ETL process depends on the respective question, but following the FAIR principles, open file formats should be made available in addition to raw data, even if the transformation process is often accompanied by a loss of information (e.g. in mass spectrometry).

Also in other diagnostic fields where a large number of different analyses have to be medically validated synoptically, Big Data technologies offer a good basis for the development of pattern recognition and AI algorithms that not only help to automate workflows efficiently but can also recognise conspicuous patterns without fatigue and thus lead to a reduced false negative rate. New-born screening is a prime example of this[70], but complex metabolic diagnostics will also benefit from data that is machine learning-ready - there is still considerable potential for development[71]. For algorithms to be registered as “medical devices”, the hurdles to be taken are fairly high, including proper assessment of potential risks, detailed software design specifications, traceability, data security, etc. just to name a few obligations to be compliant with the new “Medical Device Regulation” (MDR) of the European Union[72]. Moreover, to be used in hospital settings, data collection requires strict quality management systems certified in accordance with ISO 13485[73]. Currently, European notified bodies or other authorities like the U.S. Food and Drug Administration (FDA) or the UK Medicines and Healthcare products Regulatory Agency (MHRA) have started to adapt guidelines for Good Machine Learning Practice (GMLP) for the development of AI and ML applications as medical devices or overhauled its existing regulations [74–76].We are now witnessing the clearance of the first AI-

 doi:10.20944/preprints202204.0078.v2

based algorithms for prediction and diagnostics for use with patients. The “IDx-DR”-algorithm which detects diabetic retinopathy from retinal images is an inspiring example[77]. It was the first medical device using artificial intelligence to be approved by the FDA in April 2018 and for use on the European market in April 2019.[78,79] Data from a multi-centre study with 900 patients enrolled at ten different sites was a cornerstone for the approval of the “IDx-DR”-algorithm – a masterpiece, unthinkable without proper “Big Data”-management[77].

Besides laboratory diagnostics itself, there are a large number of other fields of application for Big Data in laboratory medicine. For example, the field of quality management. Mark Cervinski notes that "modelling of Big Data allowed us to develop protocols to rapidly detect analytical shifts" - additionally administrative and process-oriented aspects, such as optimising turnaround time (TAT), can also benefit from Big Data[13].

Clinical decision support systems are more oriented towards clinical needs and are essentially based on laboratory data. This can be in the context of integrated devices[80] or more or less complex algorithms that enable the integration of multimodal information and allow clinicians to quickly and reliably make statements about the diagnostic value of constellations of findings. An example of this is the prediction of the growth of bacteria in urine culture based on urine flow cytometric data[81].

Perhaps the most exciting field of application for Big Data in laboratory medicine, however, is predictive and pre-emptive diagnostics. With the help of laboratory data, probabilities for a variety of patient-related events can be calculated and, in the best case, therapeutic countermeasures can be initiated so that the events do not occur in the first place. This can range from the prediction of in-house mortality in the sense of an alarm triage[82,83] to the prediction of derailments in the blood glucose levels of diabetic patients[84] - the possible applications are almost unlimited.

4. Conclusion and Outlook

Laboratory medicine has always been a data-driven discipline - more so than ever with the advent of multi-parametric and "-omics" technologies. On the other hand, the discipline has been largely fossilised by a way of working that has remained almost unchanged for decades and by specific requirements of clinicians and regulatory bodies for reporting findings[85]. This is especially true for routine clinical diagnostics, so opening up to "Big Data" represents a challenge that should not be underestimated. Yet this openness represents the basis for modern technologies in particular deep learning or artificial intelligence which can bring diverse advantages for diagnostics, but also for laboratory medicine as an academic and research-based medical discipline. Many steps that are required in the transformation of laboratory medicine data into "Big Data"[22] that can be used for research make sense anyway for lean, efficient, sustainable, and complete data management and can lead to a cleansing and "aggiornamento" of laboratory data. If laboratory medicine shies away from these developments, it will be degraded to a pure number generator in the foreseeable future or disappear completely as an academic subject in integrated diagnostic devices. On the other hand, the importance of comprehensive, quality-assured laboratory medical data and metadata for clinical research can hardly be underestimated. It is important to set standards in openness, willingness to collaborate, and FAIRification of medical data. After all, health data is the new blood[86] - which can also revitalise laboratory medicine not only in a figurative sense.

Acknowledgments: The authors would like to thank Aurélie Pahud de Mortanges, MD for reviewing the manuscript and her outstanding support.

References

1. Cadamuro, J. Rise of the Machines: The Inevitable Evolution of Medicine and Medical Laboratories Intertwining with Artificial Intelligence—a Narrative Review. *Diagnostics* **2021**, *11*, 1399, doi:10.3390/diagnostics11081399.
2. Gruson, D.; Helleputte, T.; Rousseau, P.; Gruson, D. Data Science, Artificial Intelligence, and Machine Learning: Opportunities for Laboratory Medicine and the Value of Positive Regulation. *Clin. Biochem.* **2019**, *69*, 1–7, doi:10.1016/j.clinbiochem.2019.04.013.
3. Cabitza, F.; Banfi, G. Machine Learning in Laboratory Medicine: Waiting for the Flood? *Clin. Chem. Lab. Med.* **2018**, *56*, 516–524, doi:10.1515/cclm-2017-0287.
4. Ronzio, L.; Cabitza, F.; Barbaro, A.; Banfi, G. Has the Flood Entered the Basement? A Systematic Literature Review about Machine Learning in Laboratory Medicine. *Diagnostics* **2021**, *11*, 372, doi:10.3390/diagnostics11020372.
5. Mannello, F.; Plebani, M. Current Issues, Challenges, and Future Perspectives in Clinical Laboratory Medicine. *J. Clin. Med.* **2022**, *11*, 634, doi:10.3390/jcm11030634.
6. Hitzler, P.; Janowicz, K. Linked Data, Big Data, and the 4th Paradigm. *Semant. Web* **2013**, *4*, 233–235, doi:10.3233/SW-130117.
7. Diebold, F.X. On the Origin(s) and Development of the Term "Big Data." *SSRN Electron. J.* **2012**, doi:10.2139/ssrn.2152421.
8. De Mauro, A.; Greco, M.; Grimaldi, M. A Formal Definition of Big Data Based on Its Essential Features. *Libr. Rev.* **2016**, *65*, 122–135, doi:10.1108/LR-06-2015-0061.
9. Lukoianova, T.; Rubin, V.L. Veracity Roadmap: Is Big Data Objective, Truthful and Credible? *Adv. Classif. Res. Online* **2014**, *24*, 4, doi:10.7152/acro.v24i1.14671.
10. Reimer, A.P.; Madigan, E.A. Veracity in Big Data: How Good Is Good Enough. *Health Informatics J.* **2019**, *25*, 1290–1298, doi:10.1177/1460458217744369.
11. Kitchin, R. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*; SAGE Publications Ltd: 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, 2014; ISBN 9781446287484.
12. Kitchin, R.; McArdle, G. What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. *Big Data Soc.* **2016**, *3*, 205395171663113, doi:10.1177/2053951716631130.
13. Tolan, N. V.; Parnas, M.L.; Baudhuin, L.M.; Cervinski, M.A.; Chan, A.S.; Holmes, D.T.; Horowitz, G.; Klee, E.W.; Kumar, R.B.; Master, S.R. "Big Data" in Laboratory Medicine. *Clin. Chem.* **2015**, *61*, 1433–1440, doi:10.1373/clinchem.2015.248591.
14. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big Data in Healthcare: Management, Analysis and Future Prospects. *J. Big Data* **2019**, *6*, 54, doi:10.1186/s40537-019-0217-0.
15. Cowie, M.R.; Blomster, J.I.; Curtis, L.H.; Duclaux, S.; Ford, I.; Fritz, F.; Goldman, S.; Janmohamed, S.; Kreuzer, J.; Leenay, M.; et al. Electronic Health Records to Facilitate Clinical Research. *Clin. Res. Cardiol.* **2017**, *106*, 1–9, doi:10.1007/s00392-016-1025-6.
16. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, Ij.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018 9, doi:10.1038/sdata.2016.18.
17. Heath, T.; Bizer, C. Linked Data: Evolving the Web into a Global Data Space. *Synth. Lect. Semant. Web Theory Technol.* **2011**, *1*, 1–136, doi:10.2200/S00334ED1V01Y201102WBE001.
18. Euzenat, J.; Shvaiko, P. *Ontology Matching*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2013; ISBN 978-3-642-38720-3.
19. Hugo, W.; Le Franc, Y.; Coen, G.; Parland-von Essen, J.; Bonino, L. FAIR Semantics Recommendations - Second Iteration. **2020**.
20. Griffiths, E.; Joseph, R.M.; Tilston, G.; Thew, S.; Kapacee, Z.; Dixon, W.; Peek, N. Findability of UK Health Datasets Available for Research: A Mixed Methods Study. *BMJ Heal. care informatics* **2022**, *29*, doi:10.1136/bmjhci-2021-100325.
21. Vines, T.H.; Albert, A.Y.K.; Andrew, R.L.; Débarre, F.; Bock, D.G.; Franklin, M.T.; Gilbert, K.J.; Moore, J.-S.; Renaut, S.; Rennison, D.J. The Availability of Research Data Declines Rapidly with Article Age. *Curr. Biol.* **2014**, *24*, 94–97, doi:10.1016/j.cub.2013.11.014.
22. Dahlweid, F.M.; Kämpf, M.; Leichtle, A. Interoperability of Laboratory Data in Switzerland - A Spotlight on Bern. *J. Lab. Med.* **2018**, *42*, 251–258, doi:10.1515/labmed-2018-0072.
23. FHIR Management Group Website for HL7 FHIR. **2022**.
24. Brickley, D.; Guha, R.V. RDF Schema 1.1. - W3C. **2004**.
25. Boldi, P.; Vigna, S. The Webgraph Framework I. In Proceedings of the Proceedings of the 13th conference on World Wide Web - WWW '04; ACM Press: New York, New York, USA, 2004; p. 595.
26. Coyle, K. Semantic Web and Linked Data. *Libr. Technol. Rep.* **2012**, *48*, 10–14.
27. Hripcsak, G.; Duke, J.D.; Shah, N.H.; Reich, C.G.; Huser, V.; Schuemie, M.J.; Suchard, M.A.; Park, R.W.; Wong, I.C.K.; Rijnbeek, P.R.; et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud. Health Technol. Inform.* **2015**, *216*, 574–578.
28. Informatics, O.H.D.S. and The Book of OHDSI. **2021**.
29. tranSMART Foundation I2b2 Website. **2022**.
30. Sweeney, L. K-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertainty, Fuzziness Knowlege-Based Syst.* **2002**, *10*, 557–570, doi:10.1142/S0218488502001648.
31. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. L -Diversity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3, doi:10.1145/1217299.1217302.
32. Aggarwal, C.C. On K-Anonymity and the Curse of Dimensionality. In Proceedings of the VLDB; 2005; Vol. 5, pp. 901–909.
33. Li, N.; Li, T.; Venkatasubramanian, S. T-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering; IEEE, April 2007; pp. 106–115.

34. Yin, C.; Zhang, S.; Xi, J.; Wang, J. An Improved Anonymity Model for Big Data Security Based on Clustering Algorithm. *Concurr. Comput. Pract. Exp.* **2017**, *29*, e3902, doi:10.1002/cpe.3902.
35. McCord, K.A.; Hemkens, L.G. Using Electronic Health Records for Clinical Trials: Where Do We Stand and Where Can We Go? *Cmaj* **2019**, *191*, E128–E133, doi:10.1503/cmaj.180841.
36. Scheibner, J.; Ienca, M.; Kechagia, S.; Troncoso-Pastoriza, J.R.; Raisaro, J.L.; Hubaux, J.P.; Fellay, J.; Vayena, E. Data Protection and Ethics Requirements for Multisite Research with Health Data: A Comparative Examination of Legislative Governance Frameworks and the Role of Data Protection Technologies. *J. Law Biosci.* **2020**, *7*, lsaa010, doi:10.1093/jlb/lsaa010.
37. Price, W.N.; Cohen, I.G. Privacy in the Age of Medical Big Data. *Nat. Med.* **2019**, *25*, 37–43, doi:10.1038/s41591-018-0272-7.
38. Samuel, G.; Chubb, J.; Derrick, G. Boundaries Between Research Ethics and Ethical Research Use in Artificial Intelligence Health Research. *J. Empir. Res. Hum. Res. Ethics* **2021**, *16*, 325–337, doi:10.1177/15562646211002744.
39. Ferretti, A.; Ienca, M.; Velarde, M.R.; Hurst, S.; Vayena, E. The Challenges of Big Data for Research Ethics Committees: A Qualitative Swiss Study. *J. Empir. Res. Hum. Res. Ethics* **2022**, *17*, 129–143, doi:10.1177/15562646211053538.
40. Raisaro, J.L.; Troncoso-Pastoriza, J.R.; Misbach, M.; Sousa, J.S.; Pradervand, S.; Missiaglia, E.; Michielin, O.; Ford, B.; Hubaux, J.P. MEDCO: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2019**, *16*, 1328–1341, doi:10.1109/TCBB.2018.2854776.
41. Wirth, F.N.; Meurers, T.; Johns, M.; Prasser, F. Privacy-Preserving Data Sharing Infrastructures for Medical Research: Systematization and Comparison. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 242, doi:10.1186/s12911-021-01602-x.
42. Medical Laboratories of Switzerland L4CHLAB Project. **2021**.
43. FDA Global Unique Device Identification Database Submission.
44. IDABC IDABC - EUDAMED: European Database on Medical Devices Available online: http://ec.europa.eu/idabc/en/document/2256/5637.html.
45. GMDN Agency GMDN Agency. **2021**.
46. Commission, E.; Emdn, T.; Commission, E. European Medical Device Nomenclature (EMDN) Available online: https://ec.europa.eu/health/system/files/2021-06/md_2021-12_en_0.pdf.
47. SPHN The SPHN Semantic Interoperability Framework Available online: https://sphn.ch/network/data-coordination-center/the-sphn-semantic-interoperability-framework/ (Accessed 13.07.2022).
48. Gamble, M.; Goble, C.; Klyne, G.; Zhao, J. MIM: A Minimum Information Model Vocabulary and Framework for Scientific Linked Data. In Proceedings of the 2012 IEEE 8th International Conference on E-Science; IEEE, October 2012; pp. 1–8.
49. Hughes, G.; Mills, H.; De Roure, D.; Frey, J.G.; Moreau, L.; Schraefel, m. c.; Smith, G.; Zaluska, E. The Semantic Smart Laboratory: A System for Supporting the Chemical EScientist. *Org. Biomol. Chem.* **2004**, *2*, 3284, doi:10.1039/b410075a.
50. Knublauch, H.; Kontokostas, D. Shapes Constraint Language (SHACL) Website - W3C. **2017**.
51. Knublauch, H.; Allemang, D.; Steyskal, S. SHACL Advanced Features - W3C. **2017**.
52. Knublauch, H.; Maria, P. SHACL JavaScript Extensions - W3C. **2017**.
53. Bilke, A.; Naumann, F. Schema Matching Using Duplicates. In Proceedings of the Proceedings - International Conference on Data Engineering; 2005; pp. 69–80.
54. Nikolov, A.; Motta, E. Capturing Emerging Relations between Schema Ontologies on the Web of Data. *CEUR Workshop Proc.* **2011**, *665*.
55. Lehmann, S.; Guadagni, F.; Moore, H.; Ashton, G.; Barnes, M.; Benson, E.; Clements, J.; Koppandi, I.; Coppola, D.; Demiroglu, S.Y.; et al. Standard Preanalytical Coding for Biospecimens: Review and Implementation of the Sample PREanalytical Code (SPREC). *Biopreserv. Biobank.* **2012**, *10*, 366–374, doi:10.1089/bio.2012.0012.
56. Wagner, A.S.; Waite, L.K.; Wierzba, M.; Hoffstaedter, F.; Waite, A.Q.; Poldrack, B.; Eickhoff, S.B.; Hanke, M. FAIRly Big: A Framework for Computationally Reproducible Processing of Large-Scale Data. *Sci. data* **2022**, *9*, 80, doi:10.1038/s41597-022-01163-2.
57. Perakakis, N.; Yazdani, A.; Karniadakis, G.E.; Mantzoros, C. Omics, Big Data and Machine Learning as Tools to Propel Understanding of Biological Mechanisms and to Discover Novel Diagnostics and Therapeutics. *Metabolism.* **2018**, *87*, A1–A9, doi:10.1016/j.metabol.2018.08.002.
58. Li, R.; Li, L.; Xu, Y.; Yang, J. Machine Learning Meets Omics: Applications and Perspectives. *Brief. Bioinform.* **2022**, *23*, doi:10.1093/bib/bbab460.
59. Wang, Z.; He, Y. Precision Omics Data Integration and Analysis with Interoperable Ontologies and Their Application for COVID-19 Research. *Brief. Funct. Genomics* **2021**, *20*, 235–248, doi:10.1093/bfgp/elab029.
60. Kahn, M.G.; Mui, J.Y.; Ames, M.J.; Yamsani, A.K.; Pozdeyev, N.; Rafaels, N.; Brooks, I.M. Migrating a Research Data Warehouse to a Public Cloud: Challenges and Opportunities. *J. Am. Med. Informatics Assoc.* **2022**, *29*, 592–600, doi:10.1093/jamia/ocab278.
61. Nydegger, U.; Lung, T.; Risch, L.; Risch, M.; Medina Escobar, P.; Bodmer, T. Inflammation Thread Runs across Medical Laboratory Specialities. *Mediators Inflamm.* **2016**, *2016*, 4121837, doi:10.1155/2016/4121837.
62. Wang, S.; Pandis, I.; Wu, C.; He, S.; Johnson, D.; Emam, I.; Guitton, F.; Guo, Y. High Dimensional Biological Data Retrieval Optimization with NoSQL Technology. *BMC Genomics* **2014**, *15*, S3, doi:10.1186/1471-2164-15-S8-S3.
63. Ehrlich, M. Risks and Rewards of Big-Data in Epigenomics Research: An Interview with Melanie Ehrlich. *Epigenomics* **2022**, *14*, 351–358, doi:10.2217/epi-2022-0056.
64. Halder, A.; Verma, A.; Biswas, D.; Srivastava, S. Recent Advances in Mass-Spectrometry Based Proteomics Software, Tools and Databases. *Drug Discov. Today Technol.* **2021**, *39*, 69–79, doi:10.1016/j.ddtec.2021.06.007.

65. Santos, A.; Colaço, A.R.; Nielsen, A.B.; Niu, L.; Strauss, M.; Geyer, P.E.; Coscia, F.; Albrechtsen, N.J.W.; Mundt, F.; Jensen, L.J.; et al. A Knowledge Graph to Interpret Clinical Proteomics Data. *Nat. Biotechnol.* **2022**, 1–11, doi:10.1038/s41587-021-01145-6.
66. Tolani, P.; Gupta, S.; Yadav, K.; Aggarwal, S.; Yadav, A.K. Big Data, Integrative Omics and Network Biology. In *Advances in Protein Chemistry and Structural Biology*; 2021; Vol. 127, pp. 127–160 ISBN 9780323853194.
67. Passi, A.; Tibocha-Bonilla, J.D.; Kumar, M.; Tec-Campos, D.; Zengler, K.; Zuniga, C. Genome-Scale Metabolic Modeling Enables in-Depth Understanding of Big Data. *Metabolites* **2022**, *12*, 14, doi:10.3390/metabo12010014.
68. Sen, P.; Lamichhane, S.; Mathema, V.B.; McGlinchey, A.; Dickens, A.M.; Khoomrung, S.; Orešič, M. Deep Learning Meets Metabolomics: A Methodological Perspective. *Brief. Bioinform.* **2021**, *22*, 1531–1542, doi:10.1093/bib/bbaa204.
69. Ferraro Petrillo, U.; Palini, F.; Cattaneo, G.; Giancarlo, R. FASTA/Q Data Compressors for MapReduce-Hadoop Genomics: Space and Time Savings Made Easy. *BMC Bioinformatics* **2021**, *22*, 144, doi:10.1186/s12859-021-04063-1.
70. Zhu, Z.; Gu, J.; Genchev, G.Z.; Cai, X.; Wang, Y.; Guo, J.; Tian, G.; Lu, H. Improving the Diagnosis of Phenylketonuria by Using a Machine Learning–Based Screening Model of Neonatal MRM Data. *Front. Mol. Biosci.* **2020**, *7*, 115, doi:10.3389/fmolb.2020.00115.
71. Marwaha, S.; Knowles, J.W.; Ashley, E.A. A Guide for the Diagnosis of Rare and Undiagnosed Disease: Beyond the Exome. *Genome Med.* **2022**, *14*, 23, doi:10.1186/s13073-022-01026-w.
72. The European Parliament and Council Regulation on Medical Devices Available online: http://data.europa.eu/eli/reg/2017/745/2020-04-24.
73. Goodall, A.; Bos, G. ISO 13485:2003 Medical Devices — Quality Management Systems — Requirements for Regulatory Purposes Available online: https://www.iso.org/standard/59752.html.
74. European Commission Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.
75. U.S. Food and Drug Administration Good Machine Learning Practice for Medical Device Development: Guiding Principles Available online: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.
76. The Medicines and Healthcare products Regulatory Agency (MHRA) Transforming the Regulation of Software and Artificial Intelligence as a Medical Device Available online: https://www.gov.uk/government/news/transforming-the-regulation-of-software-and-artificial-intelligence-as-a-medical-device.
77. Abràmoff, M.D.; Lavin, P.T.; Birch, M.; Shah, N.; Folk, J.C. Pivotal Trial of an Autonomous AI-Based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices. *npj Digit. Med.* **2018**, *1*, 39, doi:10.1038/s41746-018-0040-6.
78. FDA Permits Marketing of Artificial Intelligence-Based Device to Detect Certain Diabetes-Related Eye Problems Available online: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye.
79. Polish Center for Testing and Certification EC Certificate No. 1434-MDD-228/2019 Available online: https://uploads-ssl.webflow.com/5c118f855cb29ab026a90802/5dc09f28b316f423d17ce52b_Certyfikaty PCBC.pdf.
80. Mejía-Salazar, J.R.; Cruz, K.R.; Vásques, E.M.M.; de Oliveira, O.N. Microfluidic Point-of-Care Devices: New Trends and Future Prospects for Ehealth Diagnostics. *Sensors (Switzerland)* **2020**, *20*, 1951, doi:10.3390/s20071951.
81. Müller, M.; Seidenberg, R.; Schuh, S.K.; Exadaktylos, A.K.; Schechter, C.B.; Leichtle, A.B.; Hautz, W.E. The Development and Validation of Different Decision-Making Tools to Predict Urine Culture Growth out of Urine Flow Cytometry Parameter. *PLoS One* **2018**, *13*, e0193255, doi:10.1371/journal.pone.0193255.
82. Schütz, N.; Leichtle, A.B.; Riesen, K. A Comparative Study of Pattern Recognition Algorithms for Predicting the Inpatient Mortality Risk Using Routine Laboratory Measurements. *Artif. Intell. Rev.* **2019**, *52*, 2559–2573, doi:10.1007/s10462-018-9625-3.
83. Nakas, C.T.; Schütz, N.; Werners, M.; Leichtle, A.B.L. Accuracy and Calibration of Computational Approaches for Inpatient Mortality Predictive Modeling. *PLoS One* **2016**, *11*, e0159046, doi:10.1371/journal.pone.0159046.
84. Witte, H.; Nakas, C.T.; Bally, L.; Leichtle, A.B. Machine-Learning Prediction of Hypo- and Hyperglycemia from Electronic Health Records: Algorithm Development and Validation. *JMIR Form. Res.* **2022**, doi:10.2196/36176.
85. Cadamuro, J.; Hillarp, A.; Unger, A.; von Meyer, A.; Bauçà, J.M.; Plekhanova, O.; Linko-Parvinen, A.; Watine, J.; Leichtle, A.; Buchta, C.; et al. Presentation and Formatting of Laboratory Results: A Narrative Review on Behalf of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group "Postanalytical Phase" (WG-POST). *Crit. Rev. Clin. Lab. Sci.* **2021**, *58*, 329–353, doi:10.1080/10408363.2020.1867051.
86. Perakslis, E.; Coravos, A. Is Health-Care Data the New Blood? *Lancet Digit. Heal.* **2019**, *1*, e8–e9, doi:10.1016/s2589-7500(19)30001-9 PMID - 33323242.