doi:10.20944/preprints202204.0078.v1

*Review*

# Big Data in Laboratory Medicine – Ready for AI?

**Tobias Blatter[1], Harald Witte[1], Upal Nath[1], Filippo Franchini[1] and Alexander 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
*Corresponding author: alexander.leichtle@insel.ch

**Abstract:** Laboratory medicine is a digital science. Every large hospital produces a variety 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. Though overshadowed in recent years by the term "artificial intelligence", the big data concept remains fundamental for any sophisticated data analysis. 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 processes, careful data governance, and modern data security solutions. The possibilities for research are endless: Enriched with clinical data they serve projects to gain pathophysiological insights, improve patient care, or they can be used to develop reference intervals. Nevertheless, big data in laboratory medicine does 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 have 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 - on the other hand, the latest developments in the fields of Big Data and Artificial Intelligence have not yet found their way into laboratory medicine across the board. And although the term "big data" is now démodé and being pushed out of the spotlight by the artificial intelligence that builds on it, data is still the basis of every data science debate in the field. The manifold reasons for this shall be reviewed below.

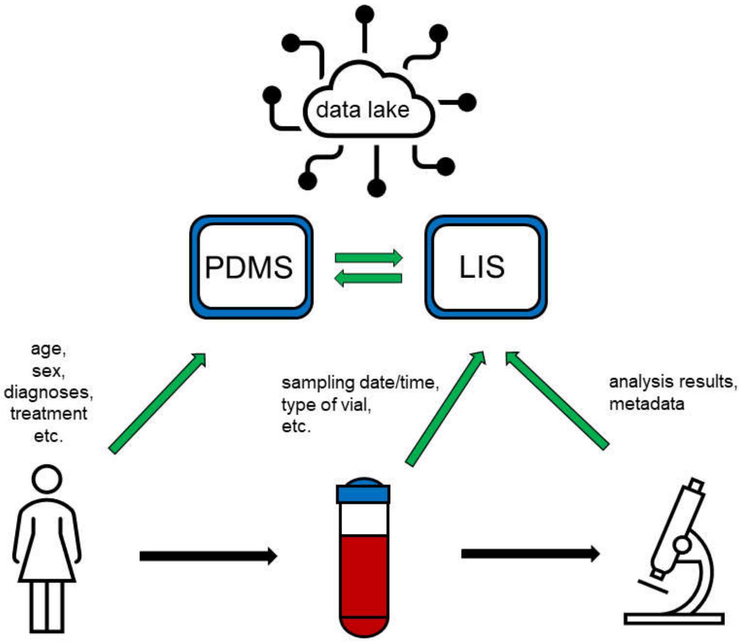

**Figure 1.** Patients´ data is entered into the patient data management system (PDMS) predominantly manually, while information about samples collected as well as about analyses conducted is entered into the laboratory information system (LIS) either manually or automatically. PDMS and LIS are connected 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. Definition of Big Data

According to Rajiv Kumar, "Big data" is the assessment of massive amounts of information from multiple electronic sources in unison, by sophisticated analytic tools to reveal otherwise unrecognized patterns[2]. If we consider "standard" laboratory analyses, e.g. clinical chemistry, hematology or hemostaseology, the lion´s share of analysis results consists of numerical results, possibly enriched with reference ranges. The resulting data volumes can by no means be considered "big" data - 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 several hundred megabytes to several gigabytes, be it NGS data or proteome or metabolome data[3]. A distinction must also be made between the usually very extensive, often proprietary raw data, and preprocessed 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. Another big data resource not to be underestimated is also non-patient-related data, such as calibration and quality control data, which are often stored and administrated in specialised databases.

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

## 3. Transforming Laboratory Medicine into Big Data Science

### *3.1. Requirements*

Even though laboratory medicine databases contain vast amounts of unused 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 more often than not mere data dumps.

So, what are the prerequisites for optimally usable laboratory medical data[4]? Central attributes data needs to have to be optimally suited for research use are summarized by the key word "FAIR": Finable, Accessible, Interoperable, Reusable[5]. (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 sound easier than it is. 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.

The accessibility of laboratory data can also be a challenge. LISs usually do not have freely accessible query functionalities because 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[6]. For use in clinical data warehouses, the LISs must be electronically connected and the data prepared via ETL processes (extract, transform, load). 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. 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. This data is particularly sensitive and required careful data governance[7]. Another important aspect is the question of patient consent for research projects access needs to be restricted according to regulatory requirements[8]. Access becomes complex 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[9].

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 such as LOINC. This represents an enormous challenge, which has been addressed in Switzerland, for example, by the L4CHLAB project (cf. https://www.famh.ch/qualitaet-sicherheit/l4chlab-dataset/). 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 (https://www.fda.gov/medical-devices/unique-device-identification-system-udi-system/global-unique-device-identification-database-gudid) or EUDAMED database(https://ec.europa.eu/tools/eudamed/), or type identifiers, e.g. from medical device nomenclatures like GMDN (https://www.gmdnagency.org) or EMDN (https://ec.europa.eu/health/system/files/2021-06/md_2021-12_en_0.pdf), may enrich the LOINC system and increase its acceptance. Extensive preparatory work to address this issue has been done by SPHN, which established corresponding "concepts" (https://sphn.ch/network/data-coordination-center/the-sphn-semantic-interoperability-framework/). 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 taken into account especially when replacing and updating LISs, so that the master data remains future-proof and interoperable[6]. In the university environment, the latest test technology might be employed, using analyses 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. Nevertheless, even with maximum semantic care, the competence of experts in laboratory medicine remains in demand. 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 such as those mapped in the dimensions of LOINC, but also, for example, batch numbers, quality management data and, if applicable, SPREC [10]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.

**Table 1.** Recommendations for the FAIRification of laboratory data.

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

### *3.2. Risks*

However, the use of laboratory medical data for Big Data analytics does not only have advantages, 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.

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. Last but not least, when individual laboratory data are queried, the framework of the findings is no longer guaranteed - the analyses lose their context and thus their interpretability.

### *3.3. 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 "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.

### *3.4. 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 such as 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.

Probably the most obvious field for Big Data technologies in laboratory medicine are "-Omics" methods[11–13]: These have been developed for nucleic acid-based techniques as e.g. genomics[14,15], transcriptomics[16], and epigenomics[17], as well as for mass spectrometry-based methodologies such as proteomics[18,19], metabolomics[20,21], lipidomics[22] 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 such as e.g. hadoop[23] 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. Newborn screening is a prime example of this[24], but complex metabolic diagnostics will also benefit from data that is machine learning-ready - there is still considerable potential for development[25].

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[2].

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[26] 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[27].

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[28,29] to the prediction of derailments in the blood glucose levels of diabetic patients[30] - 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[31]. 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 such as 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"[6] 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[32] - 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* 11, 1399 (2021).
2. 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* 61, 1433 1440 (2015).
3. Dash, S., Shakyawar, S. K., Sharma, M. & Kaushik, S. Big data in healthcare: management, analysis and future prospects. *J Big Data* 6, 54 (2019).
4. Cowie, M. R., Blomster, J. I., Curtis, L. H., Duclaux, S., Ford, I., Fritz, F., Goldman, S., Janmohamed, S., Kreuzer, J., Leenay, M., Michel, A., Ong, S., Pell, J. P., Southworth, M. R., Stough, W. G., Thoenes, M., Zannad, F. & Zalewski, A. Electronic health records to facilitate clinical research. *Clin Res Cardiol* 106, 1–9 (2017).
5. Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Santos, L. B. da S., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., Hoen, P. A. C. t, Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., Schaik, R. van, Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., Lei, J. van der, Mulligen, E. van, Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 9 (2016).
6. Dahlweid, F.-M., Kämpf, M. & Leichtle, A. Interoperability of laboratory data in Switzerland – a spotlight on Bern. *Laboratoriumsmedizin* 42, 251–258 (2018).
7. Cord, K. A. M. & Hemkens, L. G. Using electronic health records for clinical trials: Where do we stand and where can we go? *Cmaj* 191, E128–E133 (2019).
8. 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* 7, lsaa010 (2020).

9. Raisaro, J. L., Troncoso-Pastoriza, J. R., Pradervand, S., Cuendet, M., Misbach, M., Sa, J., Marino, F., Freundler, N., Rosat, N., Cavin, D., Leichtle, A., Fellay, J., Michielin, O. & Hubaux, J.-P. SPHN/PHRT - MedCo in Action: Empowering the Swiss Molecular Tumor Board with Privacy-Preserving and Real-Time Patient Discovery. *Stud Health Technol* 270, 1161–1162 (2020).

10. Lehmann, S., Guadagni, F., Moore, H., Ashton, G., Barnes, M., Benson, E., Clements, J., Koppandi, I., Coppola, D., Demiroglu, S. Y., DeSouza, Y., Wilde, A. D., Duker, J., Eliason, J., Glazer, B., Harding, K., Jeon, J. P., Kessler, J., Kokkat, T., Nanni, U., Shea, K., Skubitz, A., Somiari, S., Tybring, G., Gunter, E. & Science], F. B. [International S. for B. and E. R. (ISBER) W. G. on B. Standard Preanalytical Coding for Biospecimens: Review and Implementation of the Sample PREanalytical Code (SPREC). *Biopreserv Biobank* 10, 366–374 (2012).

11. 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. *Metabolis* 87, A1–A9 (2018).

12. Li, R., Li, L., Xu, Y. & Yang, J. Machine learning meets omics: applications and perspectives. *Brief Bioinform* 23, (2022).

13. Wang, Z. & He, Y. Precision omics data integration and analysis with interoperable ontologies and their application for COVID-19 research. *Brief Funct Genomics* 20, 235–248 (2021).

14. 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 Inform Assn* 29, 592–600 (2021).

15. Nydegger, U., Lung, T., Risch, L., Risch, M., Escobar, P. M. & Bodmer, T. Inflammation Thread Runs across Medical Laboratory Specialities. *Mediat Inflamm* 2016, 4121837 (2016).

16. 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* 15, S3 (2014).

17. Ehrlich, M. Risks and rewards of big-data in epigenomics research: an interview with Melanie Ehrlich. *Epigenomics-uk* 0, (2022).

18. Halder, A., Verma, A., Biswas, D. & Srivastava, S. Recent advances in mass-spectrometry based proteomics software, tools and databases. *Drug Discov Today Technologies* 39, 69–79 (2021).

19. 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. & Mann, M. A knowledge graph to interpret clinical proteomics data. *Nat Biotechnol* 1–11 (2022). doi:10.1038/s41587-021-01145-6

20. Tolani, P., Gupta, S., Yadav, K., Aggarwal, S. & Yadav, A. K. Proteomics and Systems Biology. *Adv Protein Chem Str* 127, 127–160 (2021).

21. 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* 12, 14 (2021).

22. 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* 22, 1531–1542 (2020).

23. Petrillo, U. F., Palini, F., Cattaneo, G. & Giancarlo, R. FASTA/Q data compressors for MapReduce-Hadoop genomics: space and time savings made easy. *Bmc Bioinformatics* 22, 144 (2021).

24. 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. *Frontiers Mol Biosci* 7, 115 (2020).

25. Marwaha, S., Knowles, J. W. & Ashley, E. A. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Med* 14, 23 (2022).

26. Mejía-Salazar, J. R., Cruz, K. R., Vásques, E. M. M. & Oliveira, O. N. de. Microfluidic Point-of-Care Devices: New Trends and Future Prospects for eHealth Diagnostics. *Sensors Basel Switz* 20, 1951 (2020).

27. 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* 13, e0193255 (2018).

28. 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* 24, 1–15 (2018).

29. Nakas, C. T., Schütz, N., Werners, M. & Leichtle, A. B. Accuracy and Calibration of Computational Approaches for Inpatient Mortality Predictive Modeling. *Plos One* 11, e0159046 (2016).

30. Witte, H., Nakas, C. T., Bally, L. & Leichtle, A. B. Machine-learning based prediction of hypo- and hyperglycemia from electronic health records (Preprint). doi:10.2196/preprints.36176

31. Cadamuro, J., Hillarp, A., Unger, A., Meyer, A. von, Bauçà, J. M., Plekhanova, O., Linko-Parvinen, A., Watine, J., Leichtle, A., Buchta, C., Haschke-Becher, E., Eisl, C., Winzer, J. & Kristoffersen, A. H. 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 Cl Lab Sci* 58, 329–353 (2021).

32. Perakslis, E. & Coravos, A. Is health-care data the new blood? *Lancet Digital Heal* 1, e8–e9 (2019).