



VirtualAB
2023



Big data for artificial intelligence applications in laboratory medicine: challenges and opportunities

Andrea Padoan

Department of Medicine (DIMED), University of Padova, Italy

Big data for artificial intelligence applications in laboratory medicine: challenges and opportunities

In recent years, clinical laboratories have experimented a huge improvements in technological tools and instrumentation. Laboratory information systems (LIS) have rapidly evolved from simple software to sophisticated tools able to retrieve and exchange information with several instrumental middleware, other laboratories and the hospital database. Overall, the increase in capabilities of LIS, in addition to recent updates of several technologies, including "-omics" have determined an increase in the flow of laboratory data in clinical laboratories. In addition to demographic details, relevant medical history or diagnosis and test results, other pieces of information are usually documented in the Laboratory Information System (LIS). These additional details mainly encompass the test name, timing of blood withdrawal, any changes made to records tracked through an audit trail, and the technical or medical validations, with the respective wards for inpatients' requests, and general practitioners for outpatients' records. Further, some LIS might include data from the quality system of the lab, not only limited to external and internal quality controls but also as additional resources about the entire process of verification and validation of analytical methods. These data, which present the characteristics of big data, can represent a richness and can be used in the development of several laboratory tools, for improving the entire laboratory testing process.



History Timeline - Understanding the flow of "laboratory information"



ANCIENT ERA

Small and basic facilities equipped with relatively simple methods and relied heavily on manual techniques (e.g. microscopes). Manual testing with mainly qualitative assays



ARCAIC ERA

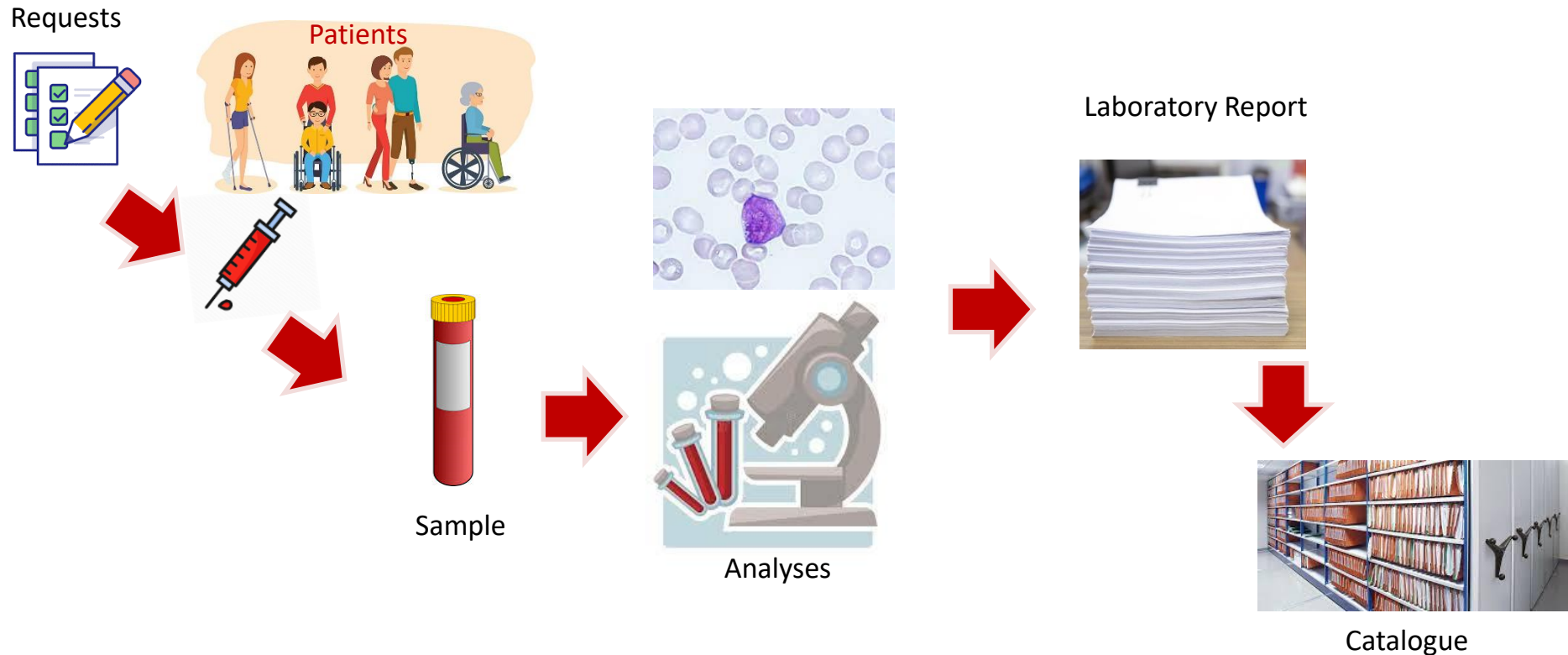
Primarily based on physical examination and qualitative evaluations



Adapted from: Plebani M. Exploring the iceberg of errors in laboratory medicine. Clin Chim Acta 2009;404(1):16–23.

Understanding the flow of "laboratory information"

The "ancient" era



History Timeline - Understanding the flow of "laboratory information"



ANCIENT ERA

Small and basic facilities equipped with relatively simple methods and relied heavily on manual techniques (e.g. microscopes). Manual testing with mainly qualitative assays



ARCAIC ERA

Primarily based on physical examination and qualitative evaluations



MIDDLE ERA

Specialized facilities, subdivided in areas, characterized by automation of main instruments, trained personnel, wide range of tests



Adapted from: Plebani M. Exploring the iceberg of errors in laboratory medicine. Clin Chim Acta 2009;404(1):16–23.

Understanding the flow of “laboratory” information

The “mid era”

Drug Information Journal, Vol. 28, pp. 397–402, 1994
Printed in the USA. All rights reserved.

0092-8615/94
Copyright © 1994 Drug Information Association Inc.

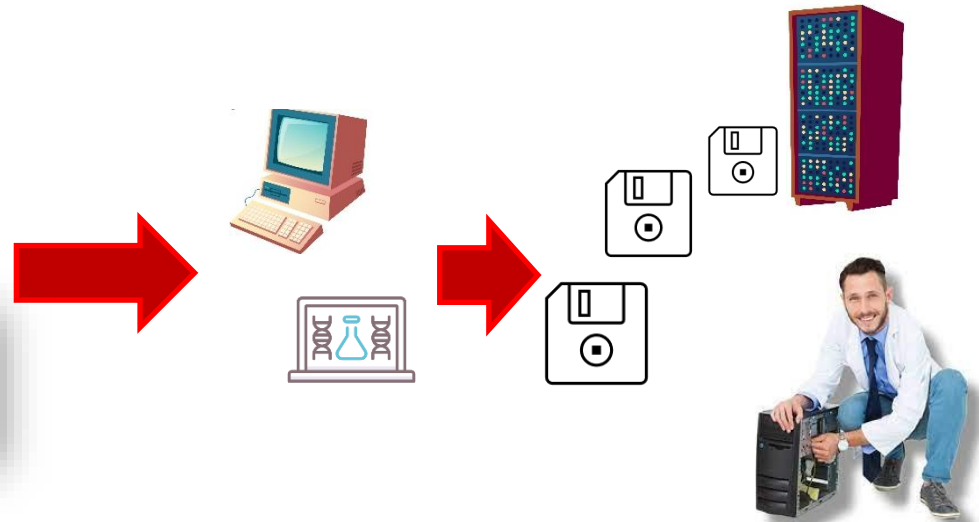
MANAGING CLINICAL LABORATORY DATA FLOW

JUDITH DAVEY, BA, MSC
Praxis PLC., Bath, United Kingdom

A Data Base Approach to Laboratory Computerization

CLEMENT J. McDONALD, M.D., LAWRENCE A. WHEELER, M.D., Ph.D.,
TULL GLAZENER, B.S., AND LONNIE BLEVINS, B.S.

American Journal of Clinical Pathology, 83:6:707–715
<https://doi.org/10.1093/ajcp/83.6.707>



- **Defining the flow of information** (data structure, type of storage, transmission of data from multiple centers)
- **Controlling the flow of information** (workflow management systems, e-mails, imaging systems)

History Timeline - Understanding the flow of "laboratory information"



ANCIENT ERA

Small and basic facilities equipped with relatively simple methods and relied heavily on manual techniques (e.g. microscopes). Manual testing with mainly qualitative assays



ARCAIC ERA

Primarily based on physical examination and qualitative evaluations



MIDDLE ERA

Specialized facilities, subdivided in areas, characterized by automation of main instruments, trained personnel, wide range of tests



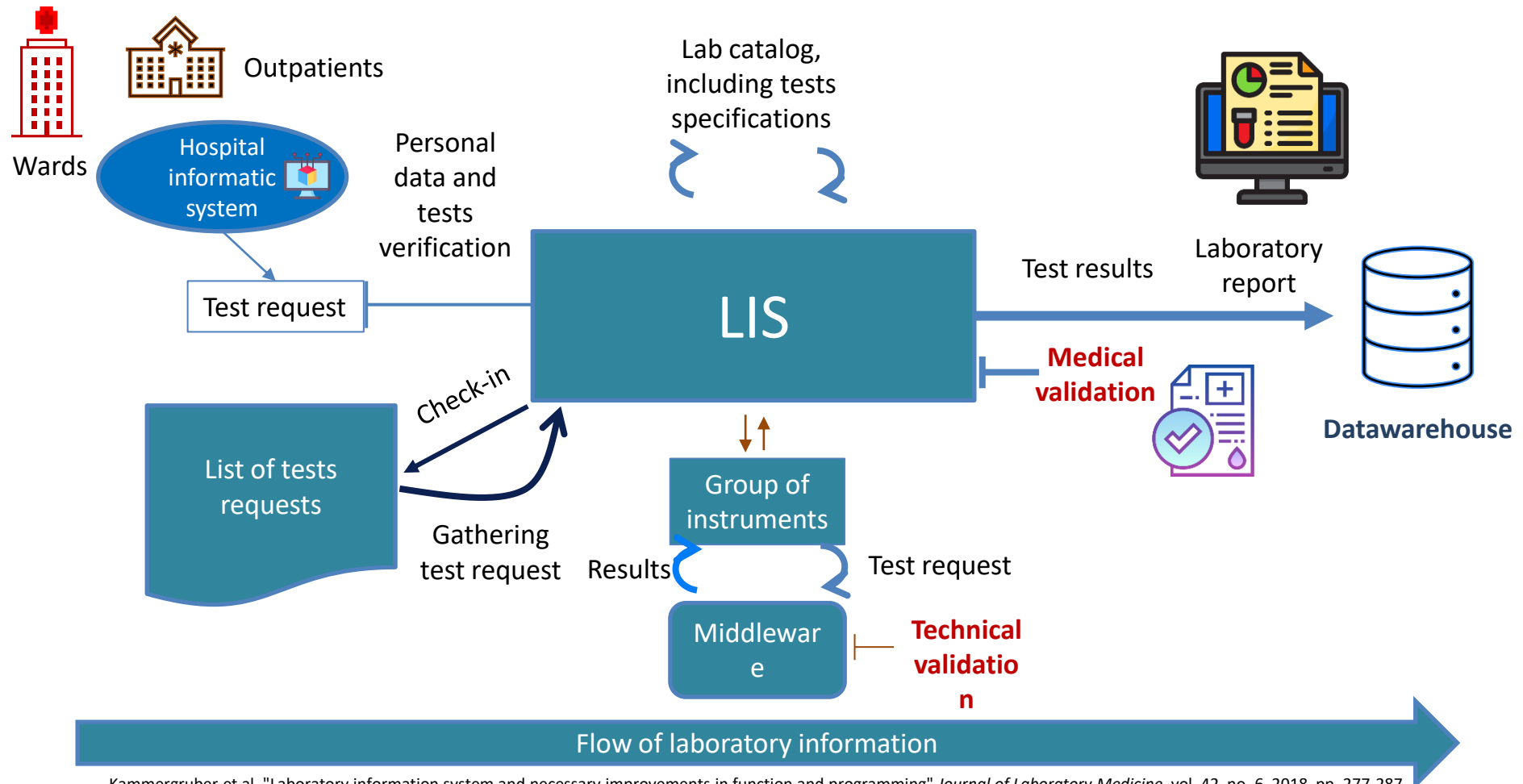
TODAY

Advanced instruments, advanced technology (e.g. omics), informatic support for several processes



Adapted from: Plebani M. Exploring the iceberg of errors in laboratory medicine. Clin Chim Acta 2009;404(1):16–23.

Understanding the flow of "laboratory" information - The modern era



History Timeline - Understanding the flow of "laboratory information"



ANCIENT ERA

Small and basic facilities equipped with relatively simple methods and relied heavily on manual techniques (e.g. microscopes). Manual testing with mainly qualitative assays



FUTURE DIRECTION

Precision medicine, data analytics, miniaturized and portable devices powered by AI



ARCAIC ERA

Primarily based on physical examination and qualitative evaluations



MIDDLE ERA

Specialized facilities, subdivided in areas, characterized by automation of main instruments, trained personnel, wide range of tests



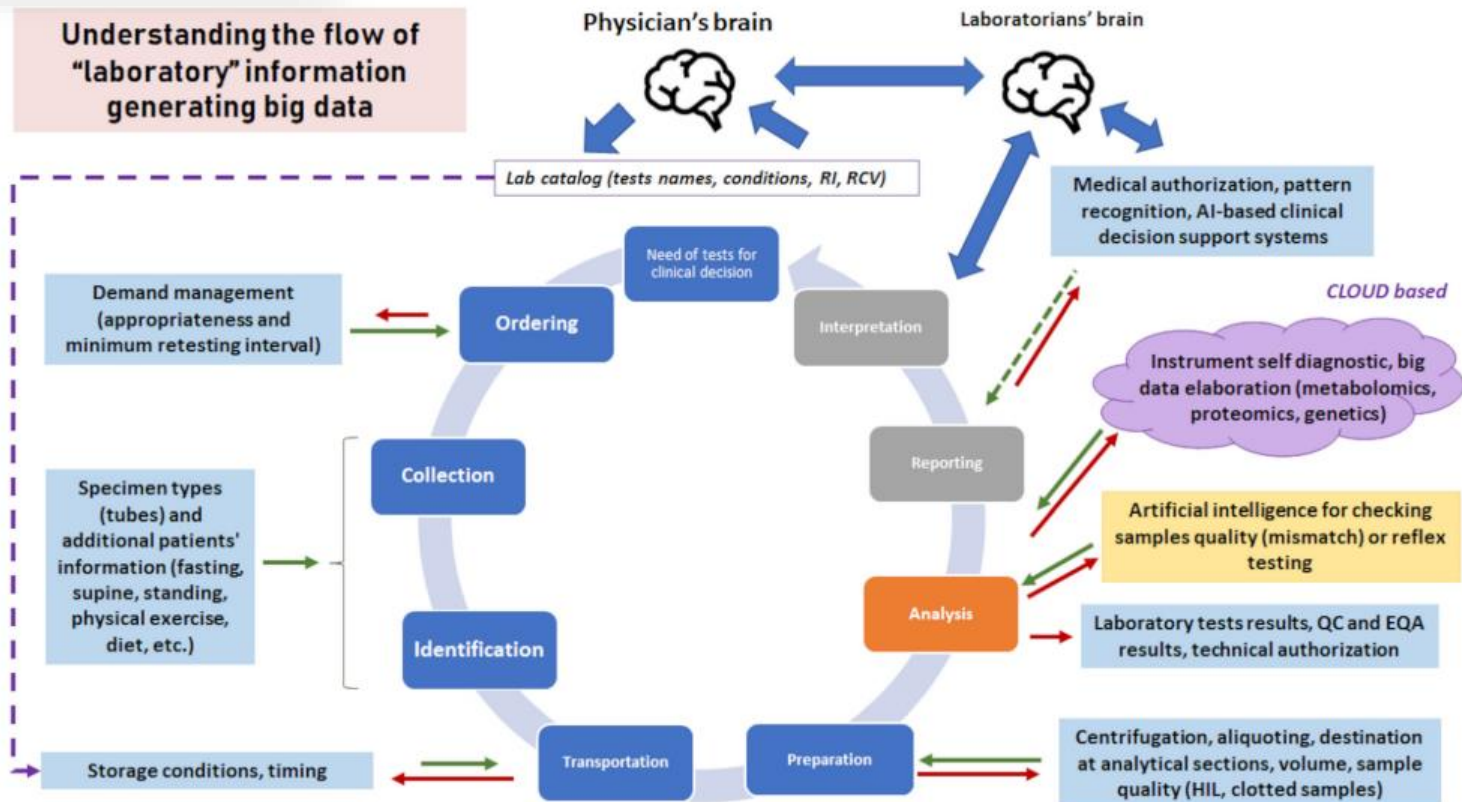
2022

TODAY

Advanced instruments, advanced technology (e.g. omics), informatic support for several processes

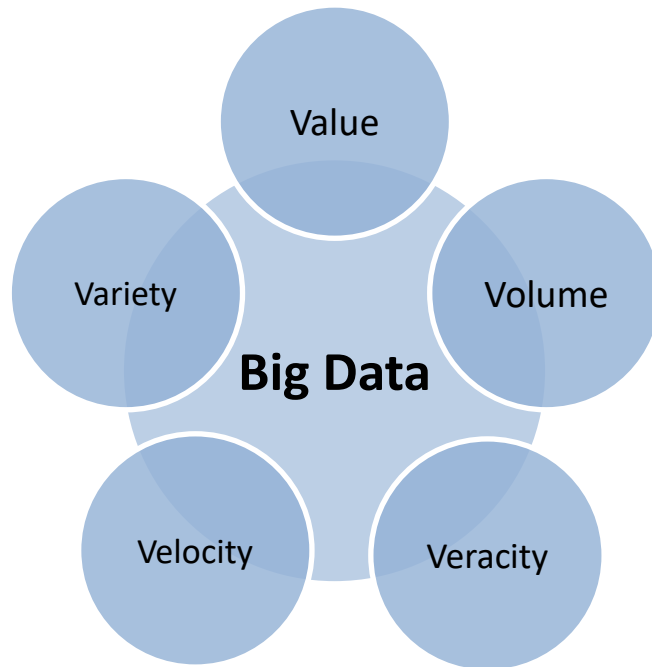


Adapted from: Plebani M. Exploring the iceberg of errors in laboratory medicine. Clin Chim Acta 2009;404(1):16–23.



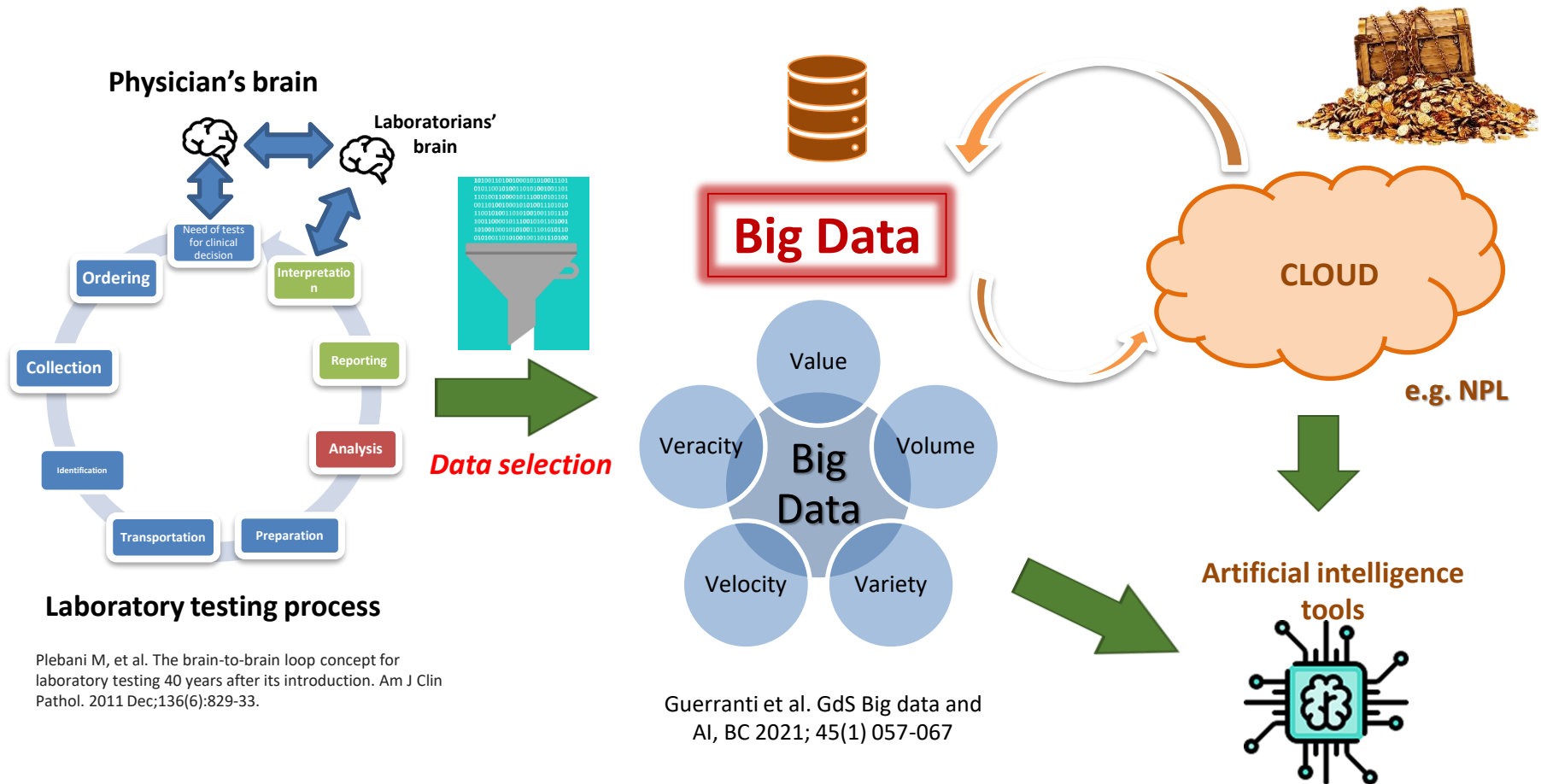


Big data in clinical laboratories



- **Value:** good data have a significant value
- **Volume:** the size is enormous
- **Veracity:** since the data is collected from multiple sources, we need to check the data for accuracy before using it for business insights
- **Velocity:** refers to the high speed of accumulation of data, which could be transient
- **Variety:** structured, semi-structured and unstructured data

Artificial intelligence meets big data in laboratory medicine



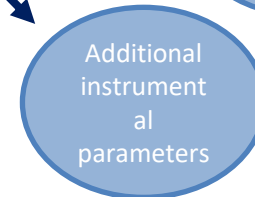
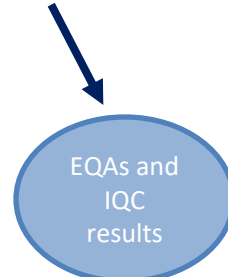
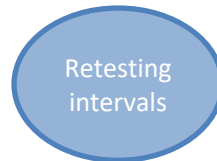
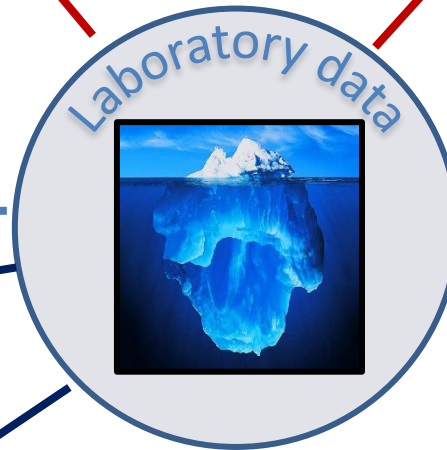
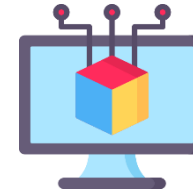
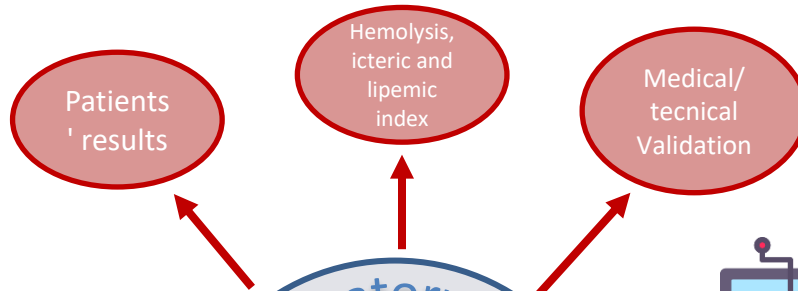
Plebani M, et al. The brain-to-brain loop concept for laboratory testing 40 years after its introduction. Am J Clin Pathol. 2011 Dec;136(6):829-33.

Guerranti et al. GdS Big data and AI, BC 2021; 45(1) 057-067



What types of LIS data are readily available?

Readily available LIS data



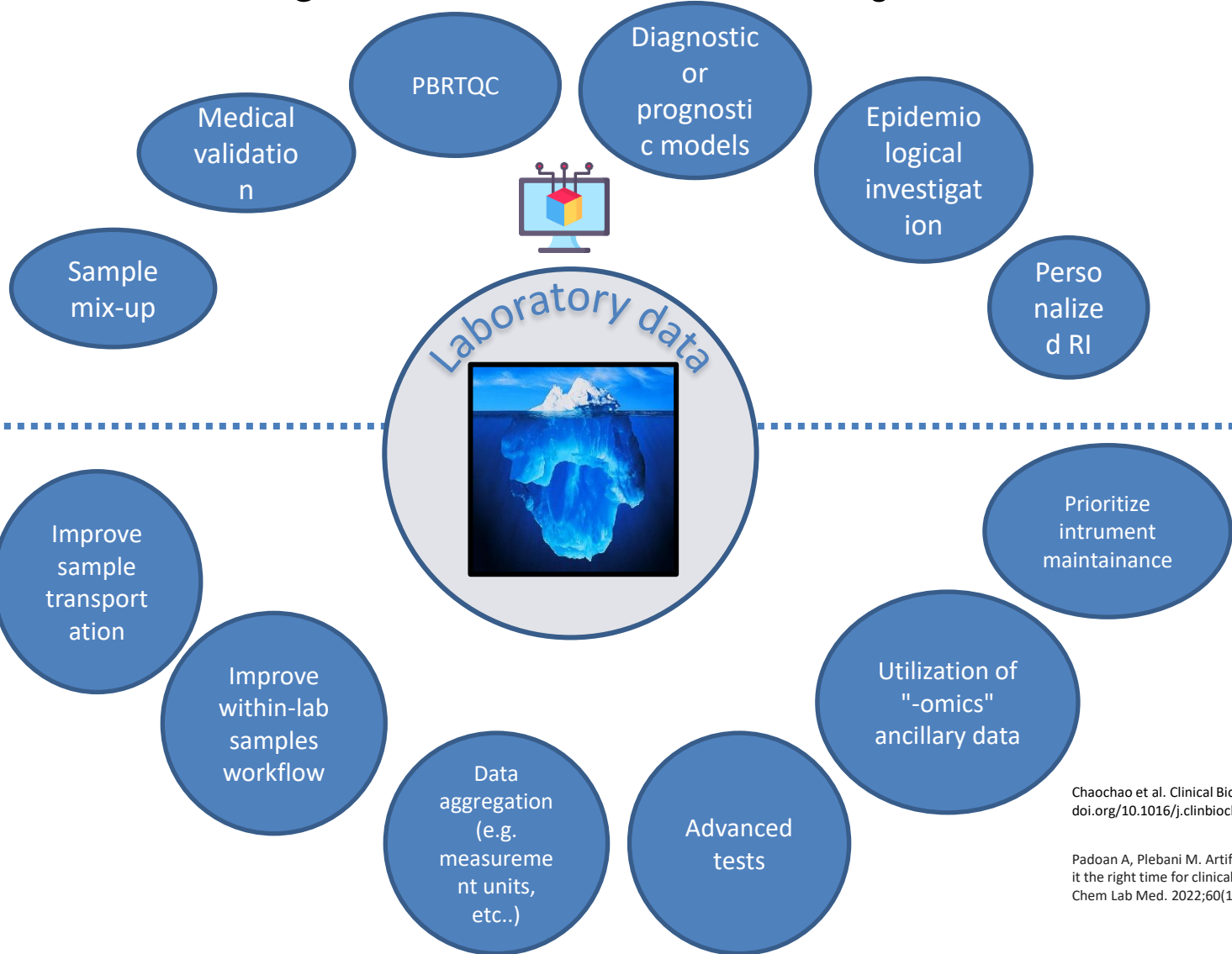
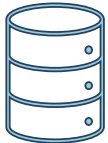
Not-readily (or not recorded) LIS data





Real-world big-data studies in laboratory medicine

Readily available LIS data



Chaochao et al. Clinical Biochemistry: 2020
doi.org/10.1016/j.clinbiochem.2020.06.014.

Padoan A, Plebani M. Artificial intelligence: is it the right time for clinical laboratories? Clin Chem Lab Med. 2022;60(12):1859-1861.

Metadata

- **Test Metadata:** information about the specific tests conducted, such as the test name, code (LOINC), analytical system, and reference ranges. It provides details about the purpose of the test, the analytes measured, and the units of measurement.
- **Sample Metadata:** It includes details such as the sample type (e.g., blood, urine, tissue), unique identifiers, and any pre-analytical treatments or processing steps performed on the sample.
- **Laboratory Metadata:** specific to the laboratory itself, including the laboratory name, location, accreditation or certification details, other equipment details, calibration information, reagents details, and other operational parameters.
- **Data Provenance and Audit Trail Metadata:** These metadata types capture information about the origin, history, and changes made to the data (e.g. whether a result is changed). They include timestamps, data entry or modification details, and any data transformations or conversions performed.



Available guidelines for metadata collection ISO 11179, ISO 15926, and ISO 19763

Metadata

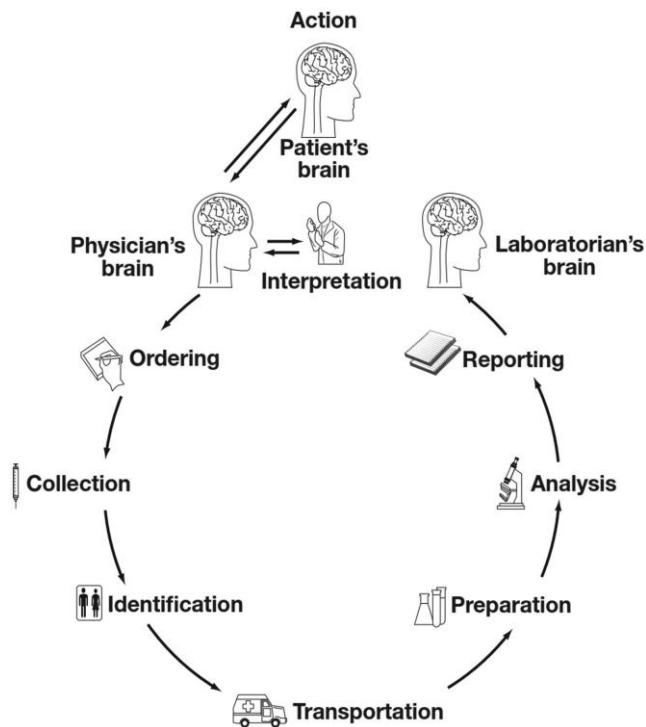
- **Test Metadata:** e.g. information about the specific tests conducted, such as the test name, code (LOINC), etc.
- **Sample Metadata:** It includes details such as the sample type (e.g., blood, urine, tissue), unique identifiers, and any pre-analytical treatments or processing steps performed on the sample.
- **Laboratory Metadata:** specific to the laboratory itself, including the laboratory name, location, accreditation or certification details, etc...
- **Data Provenance and Audit Trail Metadata:** These metadata types capture information about the origin, history, and changes made to the data (e.g. whether a result is changed). They include timestamps, data entry or modification details, and any data transformations or conversions performed.



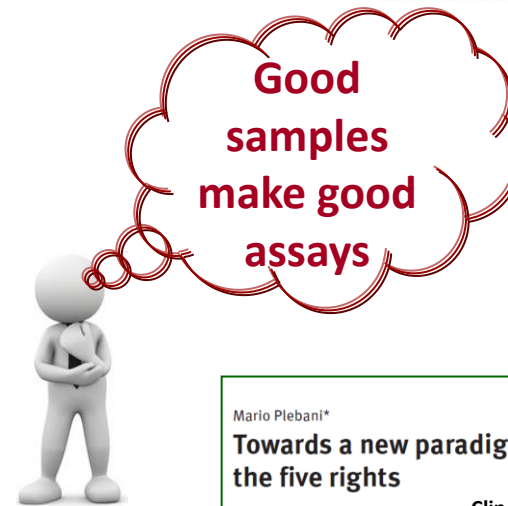
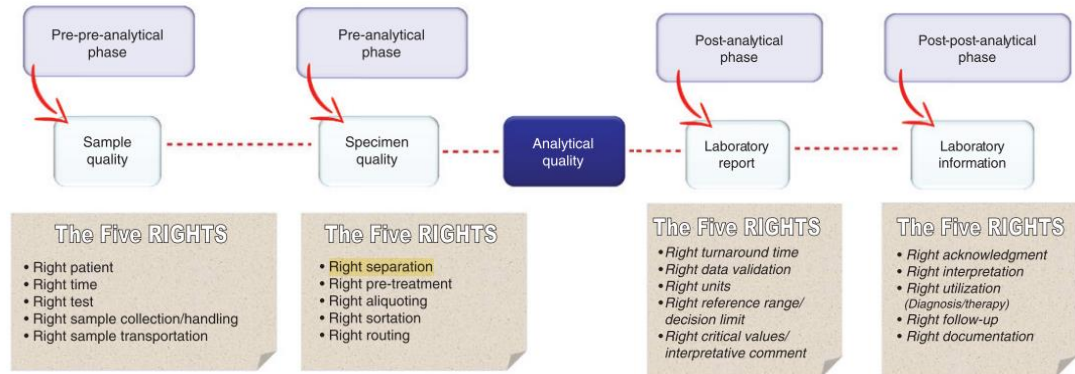
Why is so important to collect Laboratory Metadata?

- **Data Integration:** Big data often originates from diverse sources and formats. Metadata helps in integrating and combining data from multiple sources (e.g. multiple labs) by providing details about the data's origin, format, structure, and relationships. It facilitates the process of mapping and aligning data from various sources, enabling data integration and aggregation.
- **Data Quality and Reliability:** Metadata plays a crucial role in assessing the quality and reliability of big data. It includes information about data provenance, collection methods, data transformations, and data lineage. By understanding these aspects through metadata, users can assess the trustworthiness and accuracy of the data and make informed decisions about its usability.
- **Data Reusability:** Metadata enhances the reusability of big data by capturing information about the data's structure, format, and meaning. It allows users to understand the data without having to delve into its underlying intricacies. With proper metadata, data assets become more discoverable, understandable, and accessible, facilitating their reuse across different projects, departments, or organizations.

As we've already well understood...



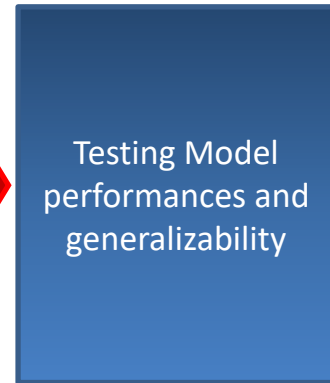
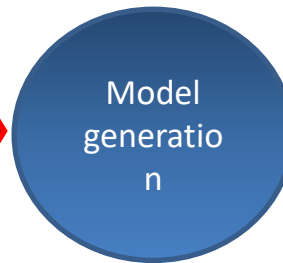
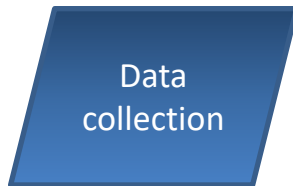
Plebani M, et al. The brain-to-brain loop concept for laboratory testing 40 years after its introduction. Am J Clin Pathol. 2011 Dec;136(6):829-33.



Mario Plebani*
Towards a new paradigm in laboratory medicine: the five rights
 Clin Chem Lab Med 2016;54:1881-91

Pre-ML steps on the AI algorithms are important for quality of results

Real-world laboratory results



Artificial intelligence

Garbage in, garbage out !



Bad performances in real data

data reliability is a critical factor !!!










Quality of Laboratory data

Letter to the Editor
Clinical Chemistry

ANNALS OF
LABORATORY
MEDICINE

 Ann Lab Med 2023;43:104-107
<https://doi.org/10.3343/alm.2023.43.1.104>
ISSN 2234-3806 eISSN 2234-3814


Proposed Model for Evaluating Real-world Laboratory Results for Big Data Research

Sollip Kim , M.D., Ph.D.¹, Eun-Jung Cho , M.D., Ph.D.², Tae-Dong Jeong , M.D., Ph.D.³, Hyung-Doo Park , M.D., Ph.D.⁴, Yeo-Min Yun , M.D., Ph.D.⁵, Kyunghoon Lee , M.D., Ph.D.⁶, Yong-Wha Lee , M.D., Ph.D.⁷, Sail Chun , M.D., Ph.D.^{1*}, and Won-Ki Min , M.D., Ph.D.^{1*}

Clinical Chemistry 00:0
1–9 (2023)

Special Report

Machine Learning in Laboratory Medicine: Recommendations of the IFCC Working Group

Stephen R. Master ^{a,b,*} Tony C. Badrick,^c Andreas Bietenbeck ^d and Shannon Haymond^{e,f,*}

IEEE SA
STANDARDS
ASSOCIATION

IEEE Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence

Clinical Chemistry 68:3
392–395 (2022)

Opinion

How Can We Ensure Reproducibility and Clinical Translation of Machine Learning Applications in Laboratory Medicine?




Shannon Haymond^{a,b,*} and Stephen R. Master^{c,d}

 diagnostics



Review

Big Data in Laboratory Medicine—FAIR Quality for AI?

Tobias Ueli Blatter ^{1,*}, Harald Witte ¹, Christos Theodoros Nakas ^{1,2} and Alexander Benedikt Leichtle ^{1,3}

Blatter TU, Witte H, Nakas CT, Leichtle AB. Big Data in Laboratory Medicine-FAIR Quality for AI? Diagnostics 2022;12(8):1923.

Some of the major issues arising with laboratory-based datasets for machine learning (AI) use

1. **Insufficient data** (the number of observation is too low)
2. **Analytical bias** (loss of calibration or calibrator change by manufacturers)
3. **Missing data**, especially when missingness rate is associated to subjects' class (e.g. missing values are mostly in controls)
4. **Imbalanced classes** (e.g. controls subjects are much more than individuals with diseases)
5. **Precision significance digit** (e.g. these two values, 0.1 and 0.2 mmol/L, presented a CV of 47%).
6. **Analytical values recorded as "below LOD" alphanumeric characters** (e.g. values < 2 ng/L).
7. **Duplicate case entries** (e.g. multiple measurements of the same patients should be included?)
8. **Domain generalization** (data differ from underlying patterns, distributions, and relationships with the outcome) (e.g. ML models generated using data from emergency departments, but applied thereafter in general medicine wards)

Insufficient data: the example of indirect RI estimation



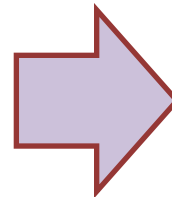
refineR (R tool for indirect RI estimation)

method: refineR (v1.6.0)

model: BoxCox

N data: 508

N bootstrap: 200



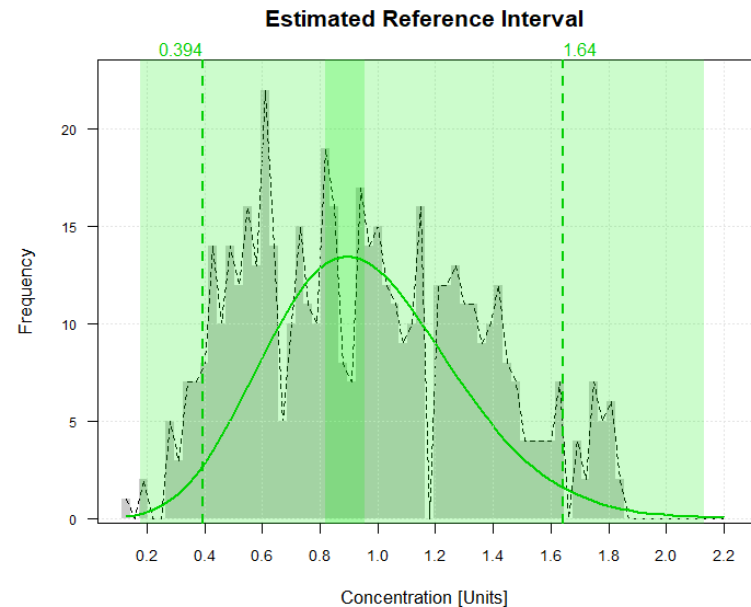
Reference Intervals

lower limit [2.5% perc]: **0.39** (0.18; 0.95)

upper limit [97.5% perc]: **1.64** (0.82; 2.13)

Estimated RI

Estimated confidence intervals of RI



Real world data, collected from 10 years of Triglycerides Lab results in children with age < 1 (duplicated removed)

Analytical bias: the effect on ML algorithms, a dummy example

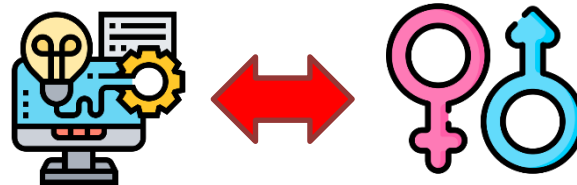


5579 patients' data on **gender, age, RBC, WBC, total cholesterol, Creatinine, AST, ALT and GLU**

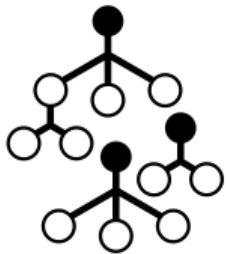
Padova
LIS, April
2023

Will be ML able to predict gender using the following lab parameters:

1) age, 2) RBC, 3) WBC, 4) total cholesterol, 5) Creatinine, 6) AST, 7) ALT and 8) GLU ?



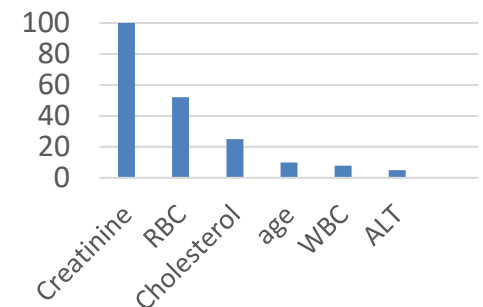
```
rfmodel <- train( sex ~., tuneLength =
1, data = training, method = "rf",
trControl = trainControl( method =
"cv", number = 5, repeats = 10,
verboseIter = TRUE ))
```



1. Split database in training (75% of observation) and testing (25% of observation) sets
2. Auto tune Random forest algorithm hyperparameters by using CV for tuning using the training set
3. Define the "variable importance"
4. Obtain the performances using the testing set



Variables Importance (%)



R and R studio, Caret Package

ML algorithm's performance could be greatly impacted by bias

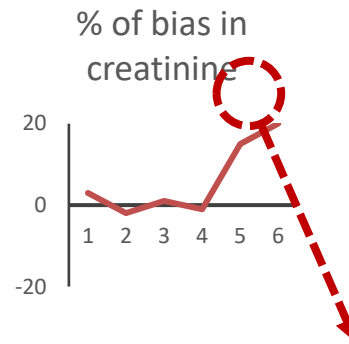
**Model
Performances**



Parameters	Testing set without bias
Accuracy	77.4%
Sensitivity	75.9%
Specificity	78.8%

ML algorithm's performance could be greatly impacted by bias

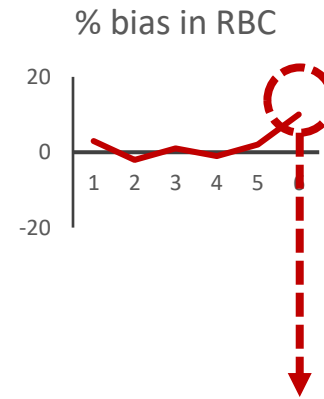
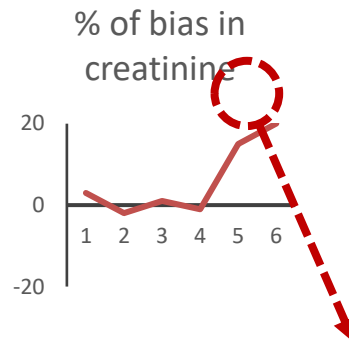
Model Performances



Parameters	Testing set without bias	Testing set with 15% bias for creatinine
Accuracy	77.4%	73.3%
Sensitivity	75.9%	84.5%
Specificity	78.8%	62.5%

ML algorithm's performance could be greatly impacted by bias

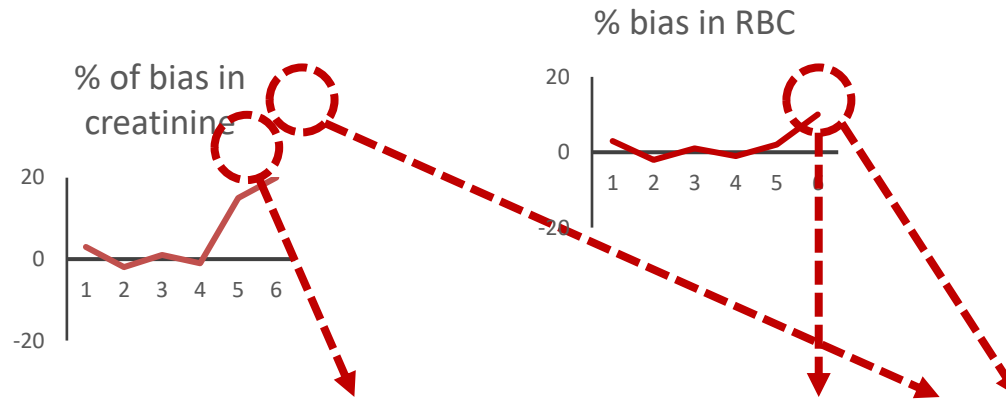
Model Performances



Parameters	Testing set without bias	Testing set with 15% bias for creatinine	Testing set with 15% bias for creatinine and 10% bias for RBC
Accuracy	77.4%	73.3%	70.0%
Sensitivity	75.9%	84.5%	87.8%
Specificity	78.8%	62.5%	52.7%

ML algorithm's performance could be greatly impacted by bias

Model Performances



Parameters	Testing set without bias	Testing set with 15% bias for creatinine	Testing set with 15% bias for creatinine and 10% bias for RBC	Testing set with 20% bias for creatinine, 10% bias for RBC
Accuracy	77.4%	73.3%	70.0%	67.0%
Sensitivity	75.9%	84.5%	87.8%	88.4%
Specificity	78.8%	62.5%	52.7%	46.4%












Overall performances decreased, especially specificity

Letter to the Editor
Clinical Chemistry

ANNALS OF
LABORATORY
MEDICINE

Ann Lab Med 2023;43:104-107
<https://doi.org/10.3343/alm.2023.43.1.104>
ISSN 2234-3806 eISSN 2234-3814

Proposed Model for Evaluating Real-world Laboratory Results for Big Data Research

Sollip Kim , M.D., Ph.D.¹, Eun-Jung Cho , M.D., Ph.D.², Tae-Dong Jeong , M.D., Ph.D.³, Hyung-Doo Park , M.D., Ph.D.⁴, Yeo-Min Yun , M.D., Ph.D.⁵, Kyunghoon Lee , M.D., Ph.D.⁶, Yong-Wha Lee , M.D., Ph.D.⁷, Sail Chun , M.D., Ph.D.^{1*}, and Won-Ki Min , M.D., Ph.D.^{1*}

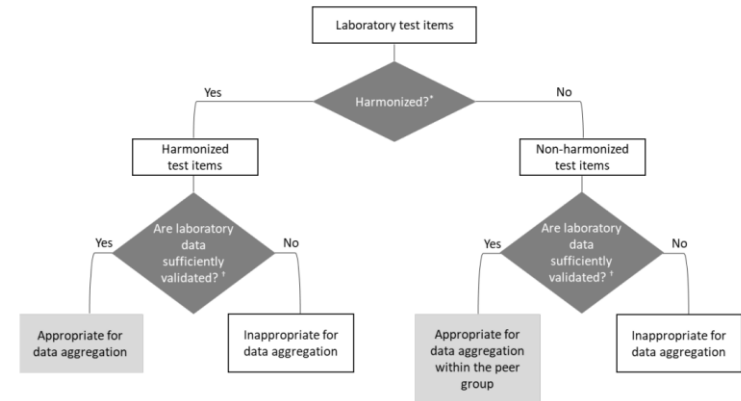
Principle 1: Cumulative EQA data (e.g., data collected over several years) should be used for evaluation to reflect the laboratory's reliability over time, because EQA reflects only the performance of a laboratory at a certain time point, whereas big data analysis is based on longitudinally collected data

Principle 2: Set the acceptance criteria as the total error. The total error is derived from bias and imprecision

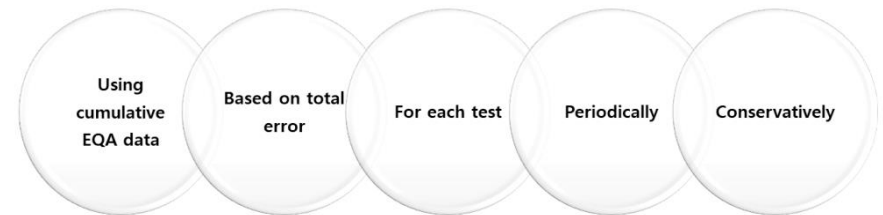
Principle 3: Evaluate each test item. The performance of each test differs for each test item even in the same laboratory.

Principle 4: Evaluate periodically (e.g., annually); even in the same laboratory, instruments/reagents can change over time, or the level of quality control may vary

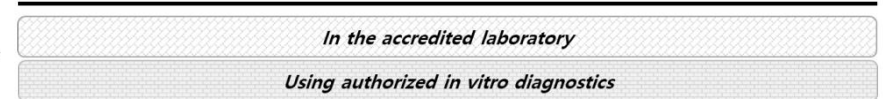
Principle 5: Take a conservative approach for evaluation (i.e., evaluate strictly). If big data research is biased by including unreliable results, significant side effects can occur in the long run.



Evaluation Principles



Prerequisites



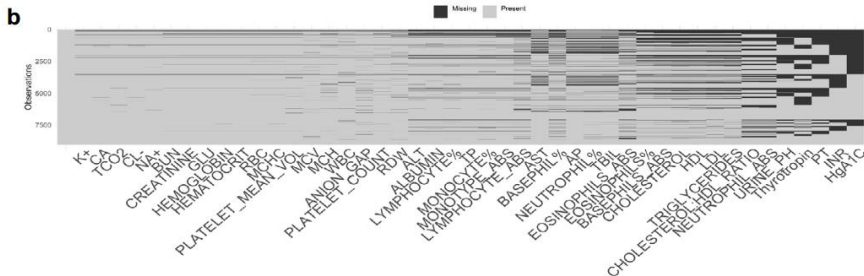
Missing data issue in clinical ML studies

npj | Digital Medicine www.nature.com/npjdigitalmed

ARTICLE OPEN Check for updates

Imputation of missing values for electronic health record laboratory data

Jiang Li¹, Xiaowei S. Yan², Durgesh Chaudhary¹, Venkatesh Avula¹, Satish Mudiganti², Hannah Husby², Shima Shahjouei¹, Ardavan Afshar^{3,7}, Walter F. Stewart⁴, Mohammed Yeasin⁵, Ramin Zand⁶ and Vida Abedi^{1,6,8}



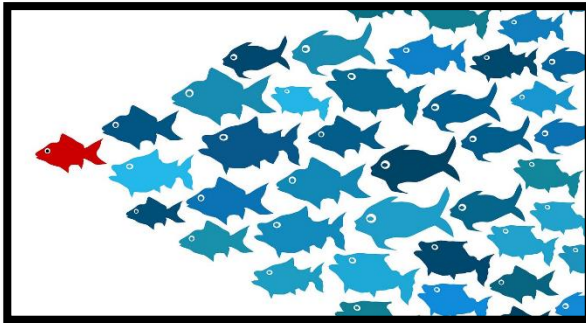
Sex	WBC	HB	PLT	TRIG	COL	HDL	LDL
Maschio	6.64	102	329	1.59	5.54		
Maschio	6.51	101	308				
Maschio	7.52	103	260				
Maschio	6.81	99	291				
Maschio	4.14	94	206				
Maschio	3.03	92	226				
Maschio	15.17	96	316				
Maschio	24.34	119	300				
Maschio	15.07	120	313				
Maschio	10.1	118	454				
Maschio							
Maschio	4.76	112	341				
Maschio				0.97	1.99		
Femmina	6.17	128	252	0.64	4.5	1.65	2.65
Maschio	10.49	140	308	0.67	3.94		
Maschio	6.6	139	264				
Maschio				0.8	4.32		
Maschio	6.39	102	267				
Maschio	6.29	90	217				
Maschio	5.21	81	175				
Maschio	8.7	84	220	2.29	4.51		
Maschio	8.22	88	261	1.59	4.33		
Maschio	10.37	83	339				
Maschio	12.01	87	390	2.63	5.65		

The pattern of missingness in EHR laboratory variables was not random and was highly associated with patients' comorbidity data.



The missing pattern and mechanism for a given dataset should first be recognized. Whether the competition is favoring a certain method or procedure has to be determined in the “real-world” data with “real-world” missingness by considering recognized and unrecognized missing pattern/mechanism, as well as the plausible distribution of missing data.

Imbalanced classes




- Relevant for ML studies with rare diseases
- Relevant for pediatric studies
- Relevant for ML studies using laboratory tests, which are usually not requested in all individuals



Deal with: study design, statistical methods, etc...

Precision significance digits

Insufficient digits			A correct # of digits		
rep 1	rep 2	CV %	rep 1 fd	rep 2 fd	CV % fd
0.1	0.2	47.1	0.19	0.21	7.1
0.2	0.2	0.0	0.22	0.25	9.0
0.3	0.1	70.7	0.3	0.19	31.7
0.2	0.1	47.1	0.22	0.18	14.1
0.1	0.2	47.1	0.18	0.23	17.2
0.2	0.1	47.1	0.21	0.18	10.9
0.2	0.1	47.1	0.24	0.17	24.1

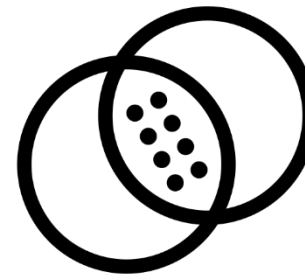
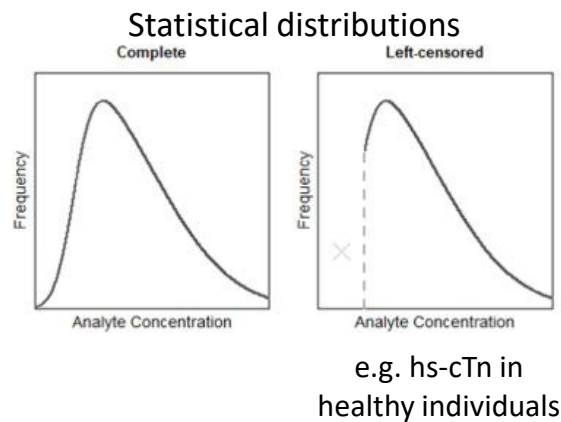
Mean CV = **44%**  Mean CV = **16%**



Deal with: Collect middleware or instrumental data

Analytical values recorded as "< LOD"

Domain generalization



ML within a domain, generalized to another domain

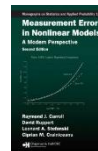


Deal with: substitute < LOD with $LOD/\sqrt{2}$ or by $E(x|X)$

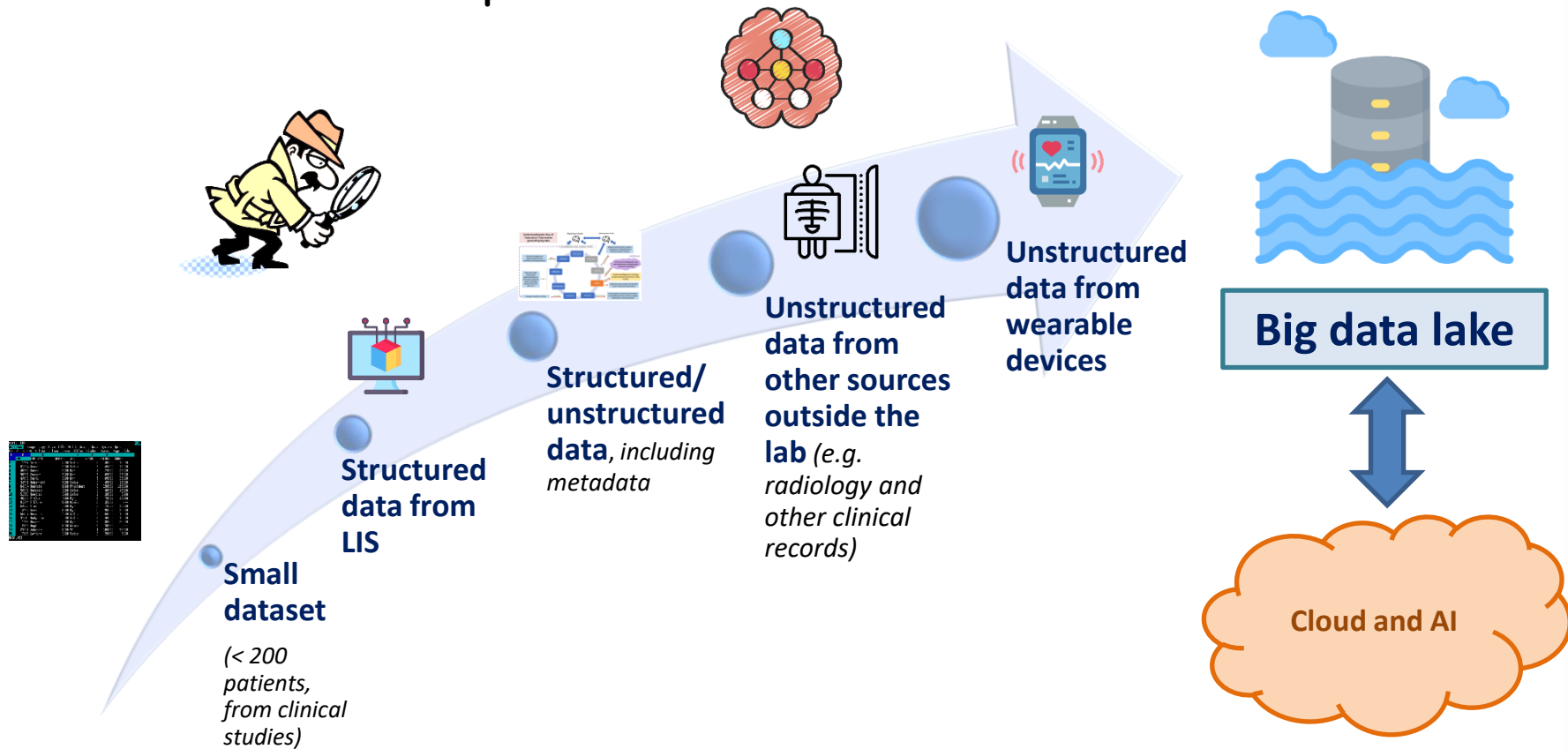


Use the correct domain with ML

Measurement Error in Nonlinear Models,
Raymond J. Carroll et al. 2006, Chapman and Hall

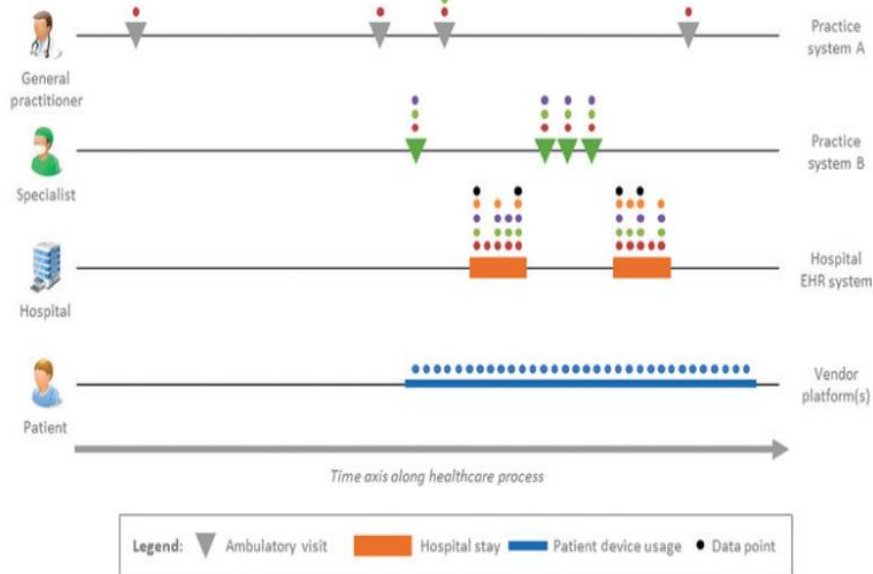


The future scenario: the role of Medical labs for the explosion of healthcare data



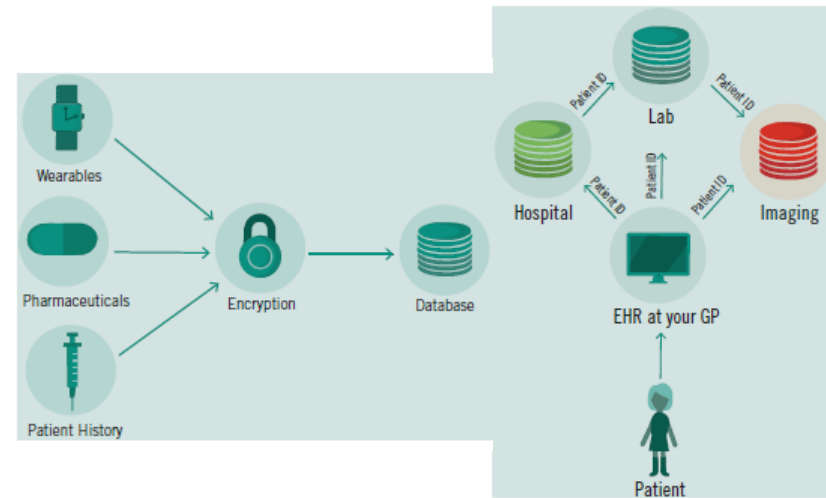
Infrastructures and data lake: the issue of data integration

Current scenario



Ganslandt T & Neumaier M. *Clin Chem Lab Med.* 2019;57(3):336–42.

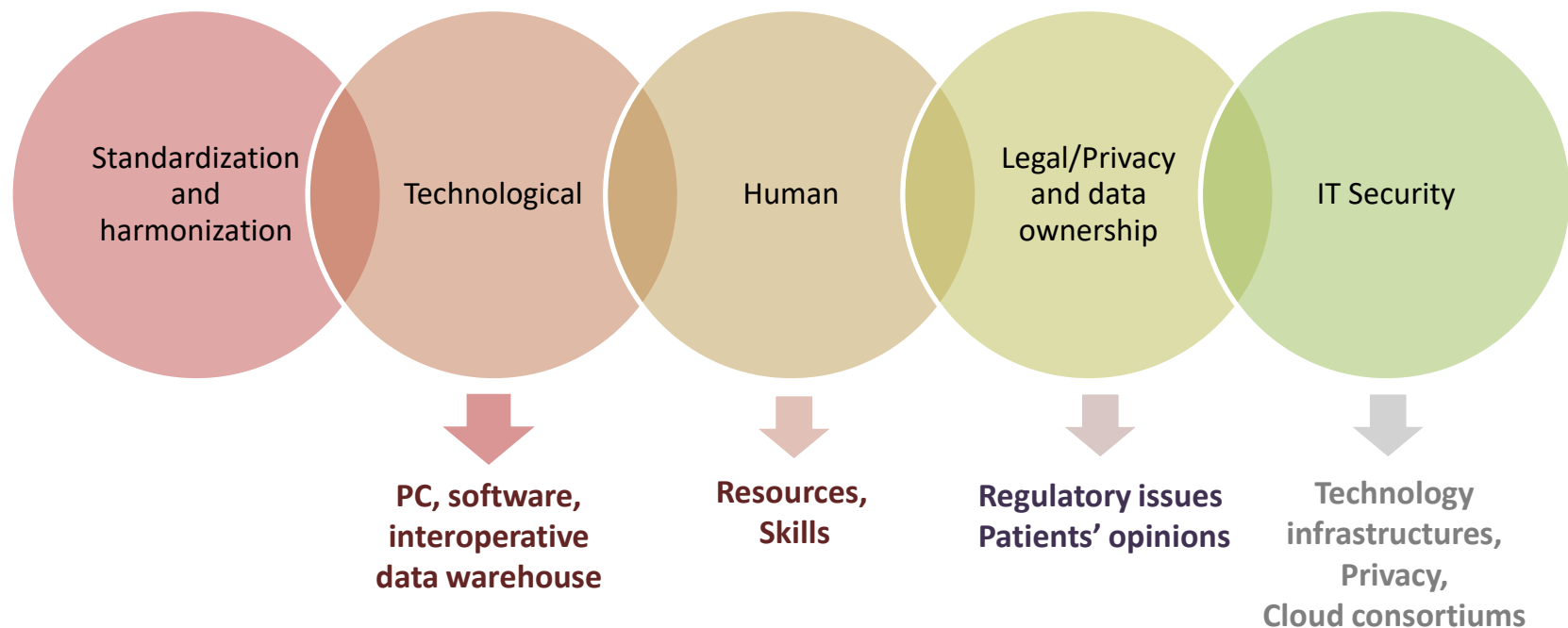
Future scenario



Agrawal, et al. *Heredity* **124**, 525–534 (2020). <https://doi.org/10.1038/s41437-020-0303-2>

Fragmentation of data points along the healthcare process acquired by different providers in shared care as well as the patients themselves, and captured in different practice or hospital

Challenges and pitfalls for healthcare data integration



Claudia Bellini*, Andrea Padoan, Anna Carobene and Roberto Guerranti, on behalf of the Italian Society of Clinical Biochemistry and Clinical Molecular Biology Big Data and Artificial Intelligence Working Group

A survey on Artificial Intelligence and Big Data utilisation in Italian clinical laboratories

It was designed by the members of the WG and the SurveyMonkey platform (SurveyMonkey Inc.) was used to administer it. 1,351 SIBioC participants were invited to take part in the survey by email through the distribution of special newsletters (between April and July 2021)

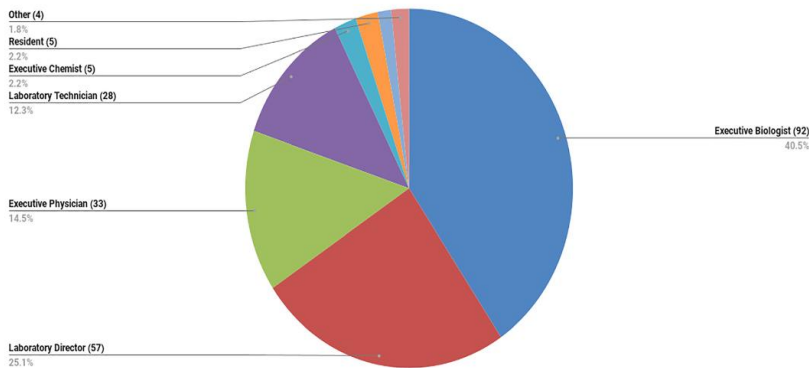


Figure 1: Professional profiles of respondents.

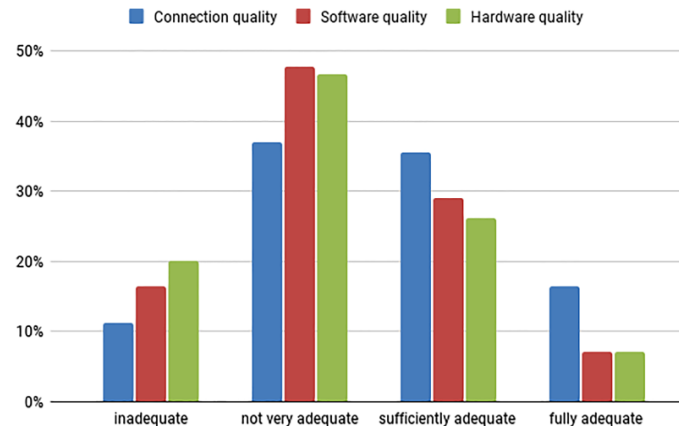


Figure 3: Judgement of the quality of connections and adequacy of software and hardware.

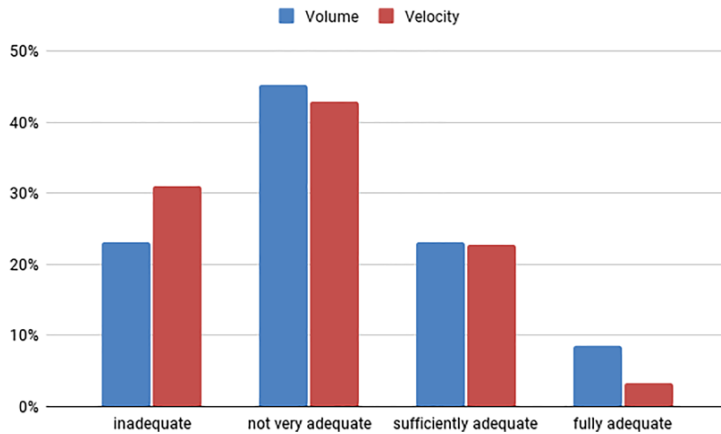


Figure 5: Judgements on speed and volume of data extraction from the Laboratory Information System.

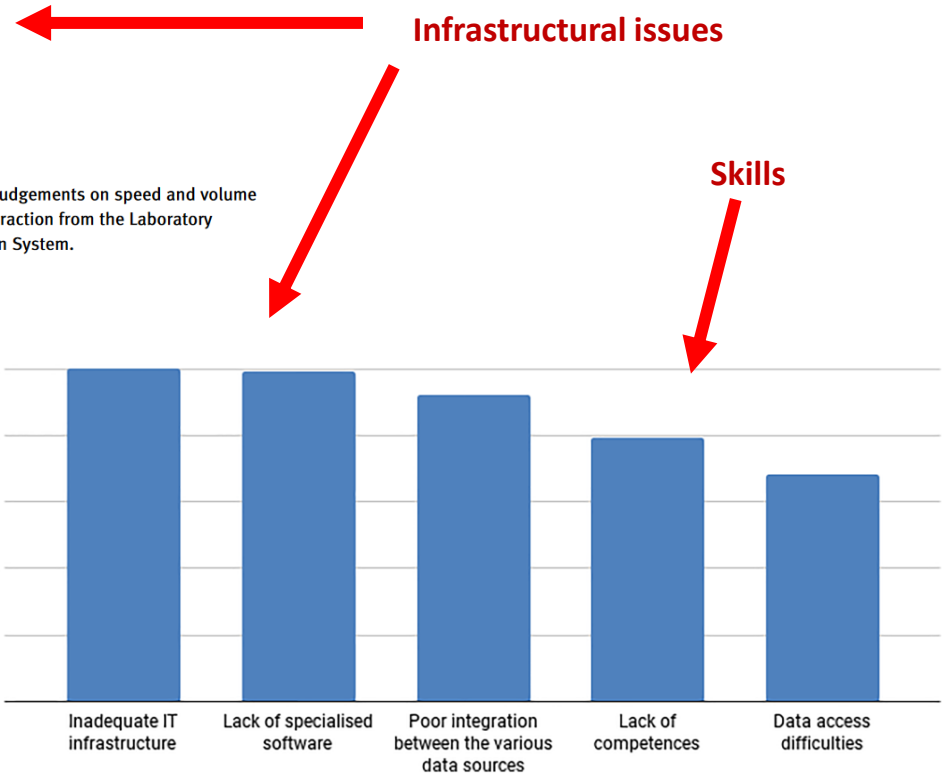


Figure 6: Opinion on the major barriers to the implementation of Big Data and Artificial Intelligence.

In conclusion, the opinions gathered show that none of the obstacles to the development of BAI in LM stand out more than the others, emphasizing the need to improve many aspects that prevent the use of these new methodologies: from the adaptation of IT infrastructures (data warehouses that combine the various data sources, acquisition of specialised software for BAI analysis, the resolution of the limitations on accessibility and use of data in respect of privacy), to the management of training and the acquisition of new skills.

Take home messages

- Artificial intelligence and Big data can help precision medicine applications, using laboratory medicine data
- Within laboratories, multiple sources of information exist, including data readily available on LIS and metadata
- By leveraging metadata, laboratories can ensure that their data is reliable, reusable and can easily be integrated with other sources of information.
- Quality of data is essential and several issues should be carefully considered when preparing the study and when using ML application
- Future applications will use structured and unstructured data from several sources, whilst data integration is of the utmost importance for obtaining big data sources.
- Different limitation to data integration are present; within the laboratories, the most common challenges were inadequate infrastructure and a shortage of skilled personnel to handle and interpret data.

Thanks for your attention

**Big data for artificial intelligence
applications in laboratory
medicine: challenges and
opportunities**

Andrea Padoan

Department of Medicine (DIMED), University of Padova, Italy