Commentary

Automation With Intelligence in Drug Research

Hanming Tu, MS, MCRP[1]; Zhongping Lin, PhD[1]; and Kevin Lee, MS[2]

[1]*Frontage Laboratories, Inc., Exton, PA, USA; and* [2]*Clindata, Moraga, CA, USA*

ABSTRACT

The industry has adopted Clinical Data Interchange Standards Consortium standards for clinical trial data and the Food and Drug Administration electronic common technical document standard for documents for many years but still faces many challenges. The solutions based on these standards enable integration among solo systems, but the integration needs to be based on business requirements and provides the end-to-end intelligence for the business. The more standards are adopted, the more meaningful and timely metadata are needed to manage the change of the standards and need to be applied in the process. Automation that uses artificial intelligence and machine learning will be the next game changer in the industry to provide data with higher quality and more efficiency. This article discusses the challenges in managing standards adoption, potential approaches for automation through using robotic processes, artificial intelligence, and the maturity model for metadata-driven automation in clinical research. (*Clin Ther*. 2019;41:2436–2444)

Keywords: maturity model of intelligent automation, standard-based integration, metadata-driven automation, automation efficiency matrix, artificial intelligence smart bioanalytics.

INTRODUCTION

The adoption of standards in the life sciences industry is accelerating, and with it pharmaceutical and service companies are demanding greater efficiency and higher quality from standard-based solutions. Automation that is based on artificial intelligence (AI) and machine learning will be the next game changer in the industry to provide higher quality data and more efficient process.

The life sciences industry has adopted Clinical Data Interchange Standards Consortium standards for clinical trial data and Food and Drug Administration (FDA) electronic common technical document standards for document preparation for many years, but it still faces substantial challenges. FDA binding guidance went into effect December 17, 2016, and requested sponsors whose studies start after December 17, 2016, must submit data in FDA-supported formats listed in the FDA Data Standards Catalog.[1,2] The solutions based on these standards enable integration among siloed systems, but integration solutions also need to be based on business requirements and provide end-to-end intelligence for the business. Because some of the standards are still in the process of being developed or being matured, the changes about the standards — the metadata of the standards — need to be collected and managed properly and timely. So the more that standards are adopted, the more that meaningful and timely metadata are needed to manage the process of upgrading those standards and the application of the new standards to the existing processes and documentation.

What are challenges in managing standards adoption (SA)? How to identify potential approaches for automation through using robotic process and AI? What kind of model can we use to assess maturity of adopting standards, improving efficiency in process, and building intelligent automation in drug research? These are the questions that we will explore and try to answer in this article.

STANDARD-BASED INTEGRATION

Delivering quality data is the core goal of clinical trial and data management in drug research. The key for data quality is to adopt a common standard, to

Accepted for publication September 5, 2019
https://doi.org/10.1016/j.clinthera.2019.09.002
0149-2918/$ - see front matter

adhere to the standard practices and processes, and to use industry-strength technology to reduce human errors. Adoption of standards not only increases interoperability and efficiency but also provides a foundation for data integration and increases the degree of code reusability (CR).

Data Integration

Most of the data that are generated in clinical trials are stored and processed in a heterogeneous array of independent, non-interoperable systems. System integration is an engineering concept, the goal of which is to bring together the component subsystems into one system to deliver overarching functionality and to ensure that the subsystems function together as a single system.[3] In the real world, system integration involves integrating existing, often disparate, systems to increase value. Because value equals quality/cost, value can be increased either by enhancing product quality and thereby performance or by reducing the cost to the customer.

As the number of independent systems increases, the number of integrations increases exponentially, whereas with a standard in place, the number of integrations increases linearly. This makes a compelling case for the use of standards. Data integration is a key element of conducting scientific investigations with modern platform technologies, and it is a starting point for the management of increasing complexity of drug discovery and clinical research and for the eventual, full realization of economies of scale in larger enterprises that relies on reusable code that comes from data integration.

Code Reusability

There are two basic approaches to building reusable code: either opportunistic or planned. An opportunistic approach is an ad hoc way of finding or creating reusable code. A planned approach offers the opportunity to create reusable code in a systematic way that is the most efficient way possible. The code developed for datasets in one phase of a clinical study could be used for all phases or from one study that could be used for all studies in the same therapeutic area. It is a strategy for increasing productivity and improving quality in data transformation and standardization. Although simple in concept, successful code reuse implementation is difficult in practice.

Code reuse traditionally is achieved through leveraging code libraries or by adding custom code into the libraries. Oracle Warehouse Builder (OWB) is a single, comprehensive tool for data integration that maintains data quality while providing data auditing. OWB also provides fully integrated relational and dimensional modeling and full lifecycle

Table I. Comparison among selected ETL tools.

Traditional Approach	OWB Approach	AutoDCD
ETL using custom programming such as SAS, PL/SQL, JAVA, Perl, etc.	ETL with user interface	Web-based user interface
High-paid programmers	Users do not need to know the programming language PL/SQL	No PL/SQL programming is needed
No audit trail	In an audited environment	In an audited environment and AutoDCD-validated product
No security	Built-in security: database and OWB security	Authenticated and authorized users only with audit trails
No consistency among coding	Consistency with all the users	Automatic and consistent coding
Difficult to manage and support	Easy to manage and support	Easy to manage and support
Scalability: silo and not scale; through adding more manpower	Scalability through hardware and software	Very scalable

DCD = data conversion development; AutoDCD = automated DCD; ETL = extract, transform, and load; JAVA = general-purpose computer programming language; OWB = Oracle Warehouse Builder; PL/SQL = procedural language extension of Structured Query Language; SAS = computer programming language.

management of data and metadata. OWB includes a simple graphic user interface to allow users to drag and drop lines to build up the source-to-target mapping. Although it does not require much programming skill for users, it does need a bigger screen and take a lot of time to draw those lines. After using OWB for 3 years, we built a much faster web-based system to automate the data conversion development (DCD). Table I provides a comparison among the traditional approach, a commercial product (OWB), and our company's custom, web-based approach, DCD.[4]

Important considerations for increasing CR include the following. (1) Adoption of standards is the key for CR, which involves the following aspects: train people to understand the standards, define standard templates, build public libraries for code snippets and public transformation, group code snippets and functional transformation into modular mapping and transformation, and define workflow to govern the process. (2) Metadata-driven processes are the key to automation because metadata is machine readable and makes the individual datasets meaningful by connecting them to provide new insights. (3) Automation helps to facilitate creation of utilities that replicate processes for project set up, mapping specification, and mapping creation.

Once reusable components have been identified, we can follow the following steps to build reusable modules or even a base project that can be used to start a new project. Extract common components by (1) building a public code library that includes transformation and utilities (functions, procedures, packages, pluggable maps, and workflows); (2) building a metadata repository that includes study data tabulation model data modeling, controlled terminologies, and specification look-up tables for mapping intelligence; and (3) creating a base project to build common modules and public locations (database links). Build follow-on projects by creating location links to metadata repositories, importing public utilities through transformation and use of data rules and subject matter experts, and copying the base project and associated modules.

There are many ways that we can increase the CR of an extract, transform, and load (ETL) process in the data integration. However, viewing data integration simply as a data issue or a technical issue underestimates the unique, scientific, management challenges that it embodies: challenges that could require significant process reengineering, methodological, and even cultural changes in our approach to data, particularly metadata.

METADATA-DRIVEN AUTOMATION

Data About Data

The key to reusability and automation is metadata. Metadata is data that describes other data[5] and that provides information about other data.[6] In clinical research, the focus is to collect subject and drug data and in preclinical studies, the focus is to analyze sample data; however, the metadata about clinical research is critical as well. Strategic, business, and technical considerations are necessary to the implementation of a metadata repository.[7] For instance, for a clinical study, it is necessary to collect information about the person and methods used to develop the data, as well as the timing, system, integrity, security, and quality of the data. According to the National Information Standard Organizations,[8] metadata can be categorized into three main types: descriptive metadata, structural metadata, and administrative metadata. The role of metadata for Life Sciences & Healthcare has changed[9]: It is not only increasing the desire for integration across patient-centric business processes but also driving automation and collaboration within drug research among sectors in the life science industry.

Process Repeatability

Another important aspect of metadata-driven automation is process repeatability (PR). Efficiency can be gained through building repeatable workflows. A defined and repeatable workflow also ensures work quality. However, in reality, there are many isolated systems and single purposed codes in the real world. How could reusable, silo codes be linked together into a repeatable process?

In the AutoDCD project, our company built an automatic data conversion system that links data integration with standardization. AutoDCD provides the workflow to link maps to form controlled data flows, from vertical CR to horizontal PR. The AutoDCD portal provides a web-based software tool for the Data Integration and Standardization Department to use. The portal improves and accelerates the DCD process.

AutoDCD is a three-tier web-based application: Oracle database, Oracle application server (Apache web server), and a browser. In AutoDCD design, a SAS (computer programming language) service is used to integrate with SAS scripts, including upload (import) and download (export) SAS macros. The AutoDCD project is an example that uses source and target metadata to build reusable codes and to drive repeatable process automation.[10] This type of automation can be seen in the various phases of drug research as depicted in Figure 1.

INTELLIGENT AUTOMATION MODELS

Adoption of standards and metadata collection about the clinical study and the process of conducting the study enable automation; metadata-driven automation not only improves data quality but also increases efficiency. Collecting and classifying study-related metadata are the first steps to building AI about clinical study automation.

Conceptual Model of Intelligent Automation

Intelligent automation is a more advanced form of what is commonly known as robotic process automation with contextual metadata. The robotic process automation is driven by predefined contextual metadata such as how to log into various systems, when to conduct pivot transformation, how to merge data from different domains, etc. This type of operation may be overwhelming to end users, but machines have different strengths and capabilities that complement their human supervisors. Together, they are changing what is possible.

Intelligent automation brings fundamental changes to how drug research is conducted, how data is explored, and how decisions are made by individuals

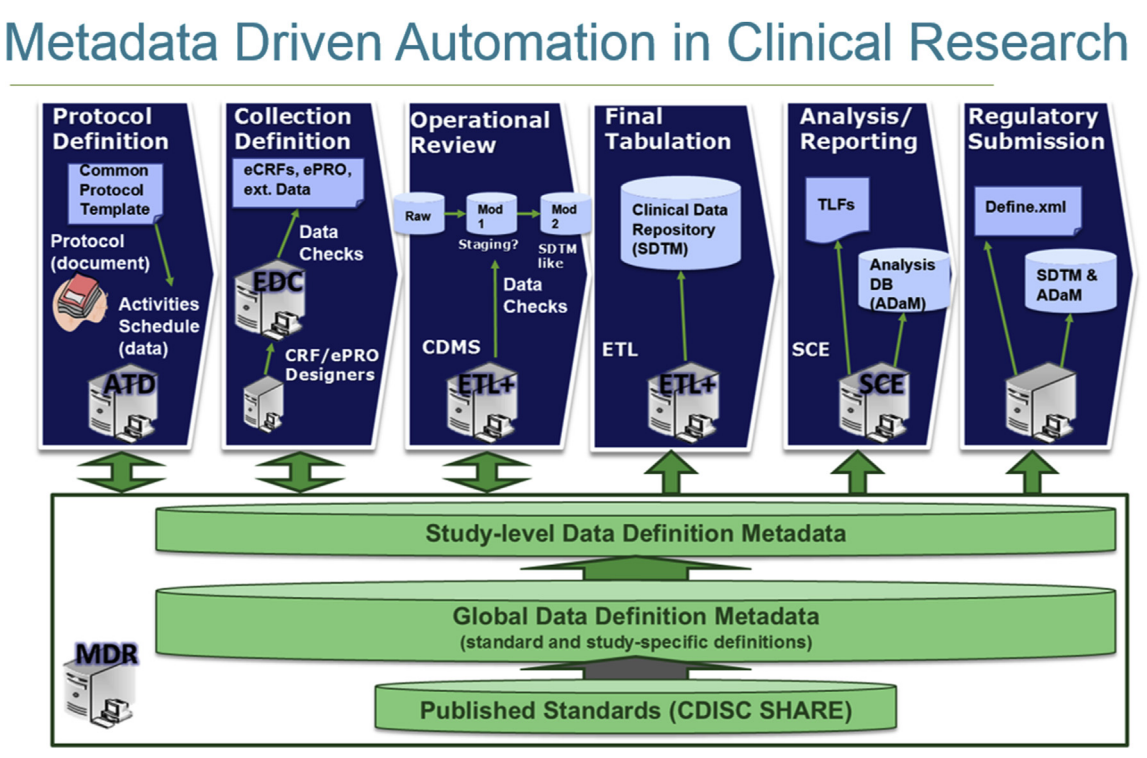

Figure 1. Metadata-driven automation in clinical research. ATD = adaptive trial design; CDISC = Clinical Data Interchange Standards Consortium; CDMS = clinical document management system; CRF = Case Report Form; eCFR = electronic Case Report Form; DB = database; ETL = extract, transform, and load; MDR = metadata registration system; SDTM = study data tabulation model; xml = Extensible Markup Language; ePRO = electronic Patient Reported Outcome; SCE = Statistical Computing Environment; ADaM = Analysis Data Model; TLFs = Table, Listing, and Figures; EDC = electronic data capture; SHARE = Shared Health and Clinical Research Electronic Library.

working with data. Figure 2 shows the model for automation with intelligence.

The conceptual model depicts how data are processed, classified, analyzed, modeled, visualized, and used in gaining new insightful knowledge and making intelligent decisions. It has four components The first component is data processing to make sure the data are clean, classified, and ready to be used. Second is machine learning to build meaningful algorithms or models such as decision tree, deep neural network, convolutional neural network, recurrent neural network, etc.[11] Third is data presenting to present data in different ways and different dimensions to allow the user to understand the meanings and insights of the data and to help make decisions. Kelly Lautt at Microsoft in 2007 created the term *data presentation architecture* and recognized that it requires a much broader skill set than data visualization. Data visualization is just one element of data presentation architecture.[12] The fourth component is intelligence building to feed more diverse data, collect more metadata, gain more insights, and make better decisions.

Maturity Model of Intelligent Automation

The conceptual model is used in different industries but was just explored in the life science industry recently. What are key factors that need to be focused and how to assess maturity of intelligent automation in a company? A maturity model of intelligent automation is proposed to evaluate a company's maturity level (ML) with the following three suggested key indicators: (1) SA could be the potential *integration indicator*, (2) CR makes clinical research more efficient and can be used as an *efficiency indicator*, and (3) PR is to make conducting clinical research more intelligent and can be used as an *intelligence indicator*. Table II shows the definition of indicator levels.

From the answers to the questions proposed, each key indicator can be divided into three levels. These three indicators can be used as X (PR), Y (CR), and Z (SA) axes to form a cube to create a matrix for efficiency, based on metadata-driven automation — level of PR (X), standard-based integration — level of CR (Y), and adoption of standards — level of intelligent automation (Z). It indicates that standard-based systems allow for higher level of integration measured by CR, whereas metadata-driven systems enable higher level of automation measured by PR. The more that standards are adopted, the more that meaningful metadata could be applied in the process and the more efficient the process could become.

Maturity Levels

It may not be possible to use quantitative measurements to clearly define each ML, but with the three levels in each key indicator, automation

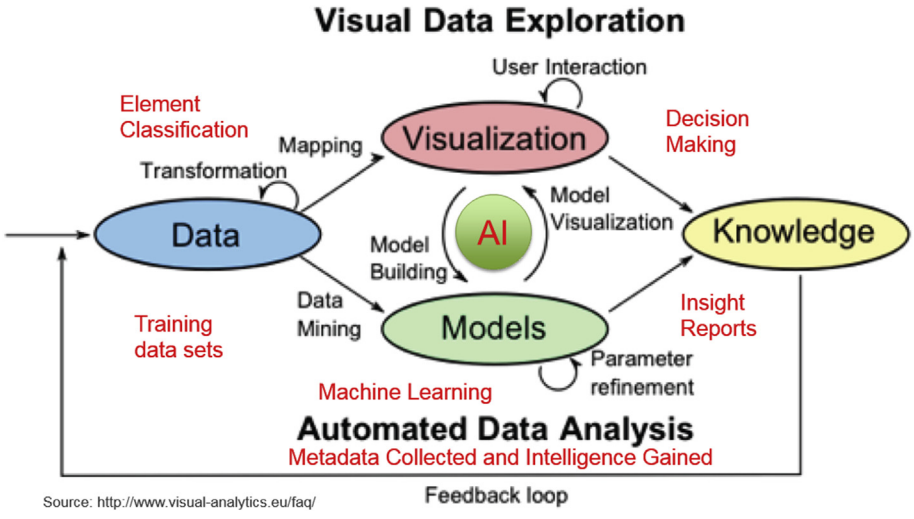

Figure 2. Conceptual model of automation with intelligence. AI = artificial intelligence.

Table II. Definition of indicator levels in the maturity model.

Variable	Questions	Indicator Level 1	Indicator Level 2	Indicator Level 3
Standards adoption (SA)	Is there at least one dedicated resource for SA in the company? Is there a periodical training to people involved in data processing and analyses? Is there a permanent governance body for SA and implementation? Is there metadata registration system implemented to manage SA? Is there any robotic process automation (RPA) and artificial intelligence (AI) technology being used to gain insights and help with decisions?	2 Yeses	3—4 Yeses	5 Yeses
Code reusability (CR)	Is there a common function library? Could the codes be modularized? Are code metadata being collected?	1 Yes with 30% CR	2 Yeses with 30%—60% CR	3 Yeses with >60% CR
Process repeatability (PR)	Is there a defined workflow? Could workflow be grouped and classified? Are the process metadata being collected?	1 Yes with 30% PR	2 Yeses with 30%—60% PR	3 Yeses with >60% PR

efficiency can be divided into 9 MLs based on the combination of SA, CR, and PR.

ML1

No common standard is adopted. No codes are reused, and double programming is used for every study. It is difficult to reproduce the process in another study.

ML2

Departmental standards exist in the company. Small set of functions is being developed and codes are reused at the functional level. Processes used in different studies are evaluated but PR is not addressed.

ML3

Coding standard and styles exist, but no common data and process standard is being adopted. Common functions and macros are developed, code reuse occurs at the module level, and companywide code standard and reusable code library exist. PR is fragmented and very manual.

ML4

Begin standards education and adopt standards in part of process in siloed systems. Still conduct double programming for every study with small percentage of codes being reused. Data flow in some phases of the study have been defined and about one-half of process can be repeated cross studies in the same therapeutic areas.

ML5

A data standard coordinator is pointed, and data standard is adopted in some parts of the whole study

life cycle. Some reusable functions and modules are built and put into the code library and are used across studies. Some common workflows are defined and used across studies.

ML6

Dedicated standard personnel has been hired to manage SA. Most functions that conduct data transformation, standardization, and classification are put into a shared common library and can be used for most of the studies. Some common workflows are defined and used across studies in multiple therapeutic areas.

ML7

Standard governance body is formed and standard department is established to coordinate all the standard implementation in the company. Double programming for every study is conducted with some CR; some algorithmic metadata have been collected but no metadata registration system is built. Collected metadata help to automate the most of the workflows being defined and used across studies.

ML8

Standard governance body is formed and standard department is established to coordinate all the standard implementation in the company; a metadata registration system is implemented. Some reusable functions and modules are built and put into the code library and are used across studies; some visualization tools are used to explore the data. Collected metadata help to automate most of the workflows being defined and used across studies.

ML9

SA is championed by the senior executives in the company; the governance body is embedded in all parts of the process. Training on SA is conducted periodically; a metadata repository system is implemented to manage all versions of the standards, code libraries, workflow definitions, and their usage in various studies. One hundred percent CR is achieved through building modules, models, and services; 100% PR is achieved through collecting metadata and using metadata to drive the whole process.

Further automation can be achieved through intelligent data flow. CR and PR make data conversion very fast; the compliance check ensures the quality is good; thus, intelligent data flow makes the whole clinical data lifecycle smart! So it makes the overall project relatively cheap.

Automation Tests

A key element of application-centric AI is context. As Zeichick[13] has pointed out, smart classification, smart recognition, and smart predictions are three big buckets that encompass many cutting-edge AI and machine learning capabilities.

Many contract research organizations have turned to technologies to build a smart laboratory with various systems designated to acquire, process, and analyze laboratory data. Features of smart laboratories are shown in Figure 3. Commonly deployed systems include (1) Laboratory Information System for processing and reporting data related to individual patients in a clinical setting; (2) Laboratory Information Management System for processing and reporting data related to batches of samples from biology laboratories, water treatment facilities, drug trials, and other entities that handle complex batches of data; (3) Electronic Laboratory Notebook for documenting laboratory research work with electronic notes that are searchable, shareable, and have safeguards of security and backup; (4) Lab Execution System for quality control, quality assurance, and compliance in laboratories; and (5) Scientific Data Management System for acting as a document management system, capturing, cataloging, and archiving data generated by laboratory instruments (HPLC, mass spectrometry), and applications (Laboratory Information Management System, Electronic Laboratory Notebook). It is a gatekeeper, serving platform-independent data to informatics applications and/or other consumers.[14]

These systems are designed to gather, store, and analyze large volumes of data, and some drug discovery units in particular have taken the lead in automating and robotizing their research laboratories. It may take a long time for a fully automated laboratory to be developed, but it is possible to have micro robots in a chip to do complicated experiments and analysis with the use of microelectromechanical system techniques. There are in place smart kiosks or box configurations, such as bioanalytical wet chemistry kiosk, bioanalytical wet chemistry in a box, and the bioanalytical laboratory in a box.[15]

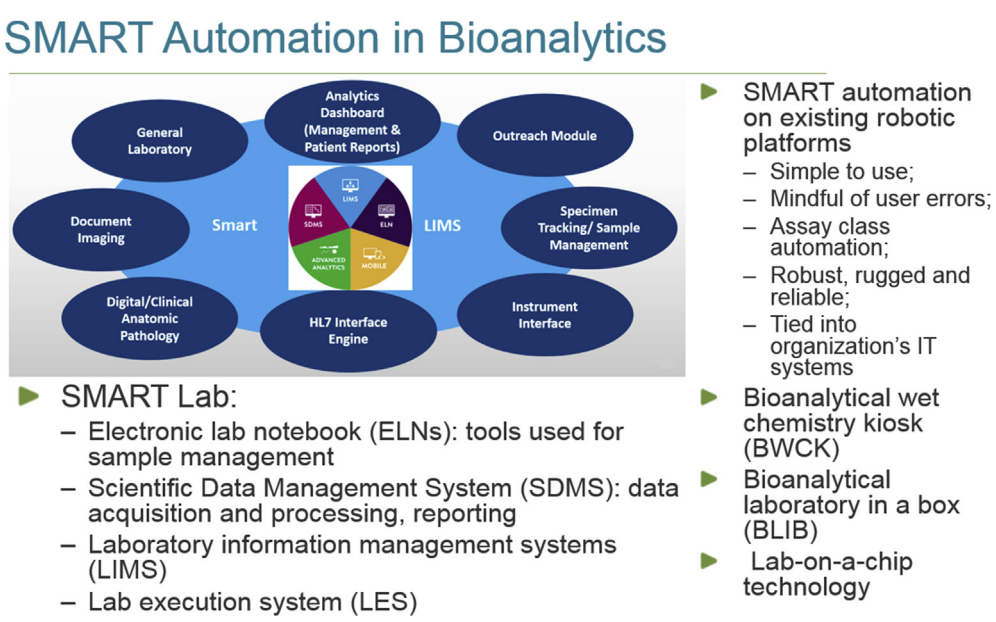

Figure 3. SMART (Simple to use; Mindful of user error; Assay class automation; Robust, rugged, and reliable; Tied into organization's information technology [IT] systems) automation in bioanalytics.

In general, laboratory automation refers to the use of technology to streamline or substitute manual manipulation of equipment and processes. This field of laboratory work comprises different automated laboratory instruments, devices, software algorithms, and methodologies used to enable, expedite, and improve efficiency and to enhance effectiveness of scientific research and test analysis. The bioanalytical wet-chemistry automation of existing robotic platforms can be described as SMART systems: Simple to use; Mindful of user error; Assay class automation; Robust, rugged, and reliable; Tied into organization's IT (information technology) systems.[16,17] These systems run in a validated environment and are in compliant with 21 CFR Part 11 and good laboratory practice.[18] As Bungers[19] has pointed out, the smart laboratory of the future will be made of flexible, digital integration, automation and robotics, integrated functional surfaces, and modular systems.

CONCLUSIONS

Clinical Data Interchange Standards Consortium study data tabulation model is a stable data model recommended for interchange between sponsors and FDA. A standard-based approach enables CR and PR to gain greater efficiency and consistency. An ETL tool such as AutoDCD provides security and scalability in an audited environment. High data quality relies on standard processes, mature technologies, and trained people. The conceptual model of automation with intelligence indicates the key components and their relationships and how data become information, knowledge, and insight through collecting metadata, testing algorithms, and building intelligence. The 9 levels of efficiency metrics provide a framework for measuring the degree of intelligent automation. Metadata-driven automation with AI creates SMART laboratories. FDA Commissioner Scott Gottlieb January 8, 2019, announced at the J.P. Morgan Healthcare Conference in San Francisco that the agency will create a new office tasked with developing more standardized, streamlined, and structured processes for reviewing and evaluating new drugs in the cloud.[20] Pharmaceutical companies may have to build the necessary metadata to feed into the FDA's cloud services sooner than anticipated.

AUTHOR CONTRIBUTIONS

Hanming Tu: Conceptualization; Data curation; Visualization; Methodology; Roles/Writing - original draft. Hanming Tu/Kevin Lee: Formal analysis. John Lin/Kevin Lee: Investigation. John Lin: Validation. Hanming Tu/John Lin/Kevin Lee: Writing - review & editing.

ACKNOWLEDGMENTS

Thanks to my colleagues at Octagon Research and Accenture who worked in the data conversion projects and developed AutoDCD system out of frustration in the old way of conducting ETL (extract, transform, and load) and love of CDISC (Clinical Data Interchange Standards Consortium) data standards! Thanks to Accenture and Frontage Lab allow me to present part of research in Pharmaceutical User Software Exchange (PHUSE) and Drug Information Association (DIA) conferences. Special thanks to Dave Evans, my mentor and supervisor who encouraged me to explore and express.

DISCLOSURES

None declared.

REFERENCES

1. CDISC. *FDA Binding Guidance Goes into Effect December 17th, 2016*. CDISC; 2016. https://www.cdisc.org/news/fda-binding-guidance-goes-effect-december-17th, 2016. Accessed September 24, 2019.
2. FDA. *Study Data Technical Conformance Guide*; 2018. https://www.fda.gov/media/88173/download.
3. Gilkey Herbert T. *New Air Heating Methods", New Methods of Heating Buildings: A Research Correlation Conference Conducted by the Building Research Institute, Division of Engineering and Industrial Research, as One of the Programs of the BRI Fall Conferences, November 1959*. Washington: National Research Council (U.S.). Building Research Institute; 1960:60. OCLC 184031.
4. Tu Hanming. *"CDISC SDTM Data Conversion: Reusability and Repeatability", the 45th DIA Annual Meeting, at the San Diego Convention Center*. 2009. San Diego, California, USA.
5. TechTarget. *Metadata*; 2014. Retrieved from https://whatis.techtarget.com/definition/metadata. This definition is part of our Essential Guide: Guide to managing a data quality assurance program, Last updated in July 2014.
6. Wikipedia. *Metadata*. FL: Wikimedia Foundation, Inc.; 2018. Retrieved October 10, 2018 from https://en.wikipedia.org/wiki/Metadata.
7. Tu Hanming. *Poster - "Considerations for Implementing Cdisc Metadata Repository"*. Baltimore, MD: CDISC Interchange North America; 2011. This poster is one of the two session winners for the poster session.
8. Riley Jenn. *Understanding Metadata: What Is Metadata and what Is it for?*; 2017, 2017 https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf.
9. Healthcare, Life Sciences, September 30, 2015: https://xtalks.com/webinars/metadata-for-life-sciences/.
10. Tu Hanming. *Paper DH08 - "Efficiency Comes from Reusability and Repeatability", PhUSE Annual Conference*. 2016. Barcelona, Spain.
11. Faggella Daniel. *What is machine learning? December 21, 2018*. 2018. published by Daniel Faggella.
12. Wikipedia. *Data Visualization*. FL: Wikimedia Foundation, Inc.; 2018. https://en.wikipedia.org/wiki/Data_visualization. Accessed December 15, 2018.
13. Zeichick Alan. *"Want a Bigger bang from AI? Embed It Into your apps", Forbes, Nov 27, 2018*; 2018. https://www.forbes.com/sites/oracle/2018/11/27/want-a-bigger-bang-from-ai-embed-it-into-your-apps/#66d5384d4e2d.
14. LiMSWiki. *Scientific Data Management System*; 2017. last modified on 29 September 2017, at 17:33 https://www.limswiki.org/index.php/Scientific_data_management_system.
15. Li Ming. Automation in the bioanalytical laboratory: what is the future? *Bioanalysis (2013)*. 2013;5(23):2859–2861.
16. Li M, Chou J, Jing J, et al. MARS (2012): bringing the automation of small-molecule bioanalytical sample preparations to a new frontier. *Bioanalysis*. 2012;4(11):1311–1326.
17. Li M, Chou J, King K, Yang L. ASPECTS: an automation-assisted SPE method development system. *Bioanalysis*. 2013;5:1661–1676, 2013.
18. Lin Z (John), Moyer M. In: Li W, ed. *Handbook of LC-MS Bioanalysis: Best Practices, Experimental Protocols, and Regulations*. Wiley; 2013. ISBN 978-1-118-15924-8; Chapter 10: Current Understanding of Bioanalysis Data Management and Trend of Regulations on Data Management.
19. Bungers Simon. *smartLAB 2017: For Tomorrows Intelligent Laboratories*; 2017. https://www.technologynetworks.com/tn/news/smartlab-2017-for-tomorrows-intelligent-laboratories-209685.
20. DiGrande Samantha, FDA's Scott Gottlieb. *Highlights Biosimilars Initiatives in J.P. Morgan Keynote Address*. 2019.

Address correspondence to: Hanming Tu, MS, MCRP, Frontage Lab, 700 Pennsylvania Drive, Exton, PA 19341, USA. E-mail: htu@frontagelab.com