



Artificial Intelligence in Clinical Research and Beyond

Wenjun Bao, Ph.D.

Chief Scientist and Director of Advanced Analysis R&D, JMP

Board of Director, CDISC

Wenjun.bao@jmp.com

Meet the Speaker

Wenjun Bao, Ph.D.



Title: Chief Scientist and Director of Advanced Analysis R&D, JMP

Organization: JMP Statistical Discovery

Title: Board of Director

Organization: CDISC

Title: Member

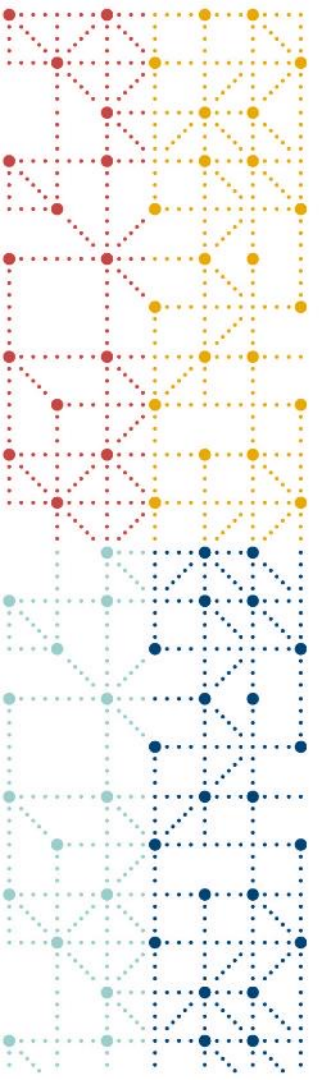
Organization: CDISC Coordinating Committee (C3C)

Dr. Wenjun Bao is a Chief Scientist and Director of advanced analytics for JMP Statistical Discovery, SAS Institute Inc. Before joining JMP, she was an IRTA Fellow at NIH (National Institutes of Health), a professor at Duke University, and a scientist at the US EPA (Environmental Protection Agency). She had rich experiences in clinical, bioinformatics, biochemistry, and molecular biology research. She has expertise in variety data analysis including AI/ML models, clinical trial and genomics data analysis with multiple publications in peer-reviewed journals. Dr. Bao has been a research grant review committee member for NIH since 2005 and a research adviser for scientists at universities and government agencies. Dr. Bao is a Board of Director for CDISC and an adjunct professor at Fudan University.



Agenda

1. FDA Initiatives for Artificial Intelligence (AI)
2. AI Applications in Difference Drug Development Stages
3. CDISC Initiative for AI Ready Dataset: JSON Data



FDA Initiatives for AI

FDA AI Publications List 2019-2023



The screenshot shows the FDA website header with the logo and navigation links. The main heading is 'Artificial Intelligence at FDA: Publications'. Below the heading are social media sharing options for Facebook, Twitter, LinkedIn, Email, and Print. The page content includes a brief introduction to AI's potential in healthcare and a note that the content is current as of 08/15/2024. A sidebar on the left contains a search bar and a link to 'Artificial Intelligence at FDA: Publications'.

<https://www.fda.gov/science-research/artificial-intelligence-and-medical-products/artificial-intelligence-fda-publications>

Date	Title	Document Type
May 2023	Using Artificial Intelligence & Machine Learning in the Development of Drug and Biological Products	Discussion Paper
April 2023	Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions: Draft Guidance for Industry and Food and Drug Administration Staff	Guidance Document
March 2023	FDA Releases FDA Activities for the Safety of Imported Seafood	Constituent Update
February 2023	Artificial Intelligence in Drug Manufacturing	Discussion Paper
October 2022	Distributed Manufacturing and Point of Care Manufacturing of Drugs	Discussion Paper
September 2022	Focus Area: Artificial Intelligence	Report
August 2022	The FDA Moves into Third Phase of Artificial Intelligence Imported Seafood Pilot Program	Constituent Update
August 2022	The FDA Moves into Third Phase of Artificial Intelligence Imported Seafood Pilot Program	Constituent Update
February 2021	FDA Moves into Second Phase of AI Imported Seafood Pilot Program	Constituent Update
January 2021	Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan	Action Plan
November 2019	FDA Extends Comment Period on the New Era of Smarter Food Safety	Constituent Update
September 2019	FDA Announces Public Meeting to Discuss the New Era of Smarter Food Safety	Constituent Update
April 2019	Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback	Discussion Paper

Showing 1 to 13 of 13 entries

Previous **1** Next

FDA AI in Drug Manufacturing: Terminology



<https://www.fda.gov/media/165743/download>

cdisc

Terminology

The definitions below are for the purposes of this discussion paper only.

- **Artificial Intelligence (AI):** A branch of computer science, statistics, and engineering that uses algorithms or models that exhibit behaviors such as learning, making decisions, and making predictions.
- **Machine Learning (ML):** A branch of AI that provides systems with the ability to develop models through analysis of data without being explicitly programmed and to improve based on data or experience.
- **Model:** An abstract description of a physical system in any form (including mathematical, symbolic, graphical, or descriptive) that represents a certain aspect of that physical system.
- **Cloud Computing:** A model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.⁵

FDA AI in Drug Manufacturing: Potential Applications

The potential applications of AI in pharmaceutical manufacturing

- **Process Design and Scale-up:** AI models such as machine learning—generated using process development data—could be leveraged to more quickly identify optimal processing parameters or scale-up processes, reducing development time and waste.
- **Advanced Process Control (APC):** APC allows dynamic control of the manufacturing process to achieve a desired output. AI methods can also be used to develop process controls that can predict the progression of a process by using AI in combination with real-time sensor data. APC approaches that combine an understanding of the underlying chemical, physical, and biological transformations occurring in the manufacturing process with AI techniques are expected to see increasing adoption and have already been reported by several pharmaceutical manufacturers.
- **Process Monitoring and Fault Detection:** AI methods can be used to monitor equipment and detect changes from normal performance that trigger maintenance activities, reducing process downtime. AI methods can also be used to monitor product quality, including quality of packaging (e.g., vision-based quality control that uses images of packaging, labels, or glass vials that are analyzed by AI-based software to detect deviations from the requirements of a product's given quality attribute).
- **Trend Monitoring:** AI can be used to examine consumer complaints and deviation reports containing large volumes of text to identify cluster problem areas and prioritize areas for continual improvement. This offers the advantage of identifying trends in manufacturing-related deviations to support a more comprehensive root cause identification. AI methods integrated with process performance and process capability metrics can be used to proactively monitor manufacturing operations for trends. These methods can also predict thresholds for triggering corrective and preventive action effectiveness evaluations.



<https://www.fda.gov/media/165743/download>

FDA AI & Medical Products: Area of Focus

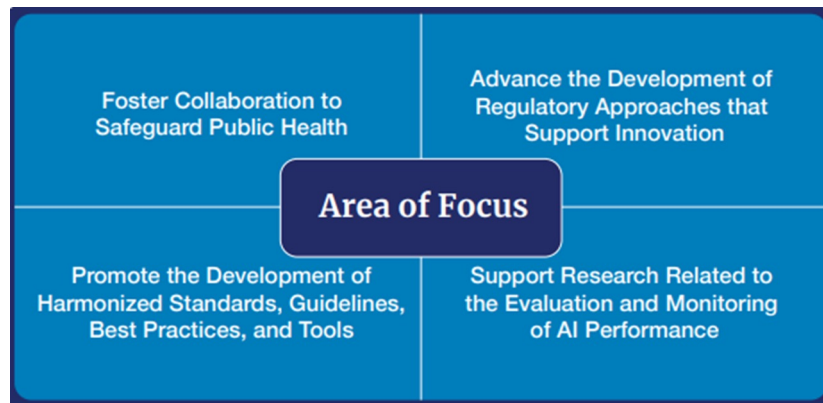


FDA U.S. FOOD & DRUG ADMINISTRATION

March 2024

Artificial Intelligence & Medical Products:
How CBER, CDER, CDRH, and OCP are Working Together

March 2024



<https://www.fda.gov/media/177030/download>

cdisc

FDA precisionFDA: Moving into Cloud

An Introduction to Project PRISM: precisionFDA Regulatory Information Service Module

A Regulatory Cloud Collaborative Initiative

Virginia "Ginny" Hussong
Branch Chief, Data Standards
Center for Biologics Evaluation and Research
US FDA

Oct. 2023

Presentation for RAPS Convergence 2023

Session on "Regulatory Collaboration Through Cloud Platforms"
Montreal, Canada
October 5, 2023

FDA U.S. FOOD & DRUG
ADMINISTRATION

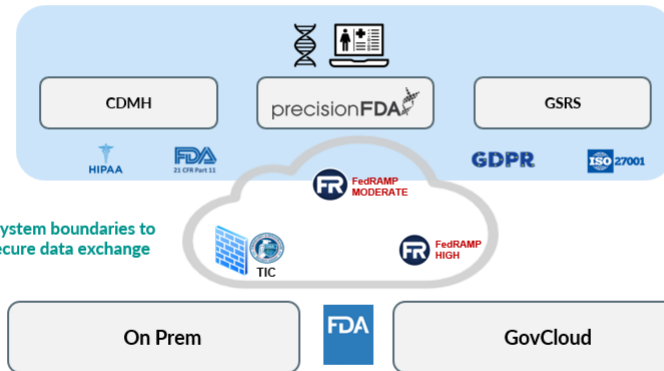


prec

precisionFDA

ATO'd Cloud Platform for Regulatory Science and Review

Regulatory-grade
web service



Bridges system boundaries to
enable secure data exchange

Gold-standard for multi-omics
bioinformatics reference data and
apps

Reference platform for RWD
informatics and RWE in regulatory
submission

Secure platform for interactive
regulatory review

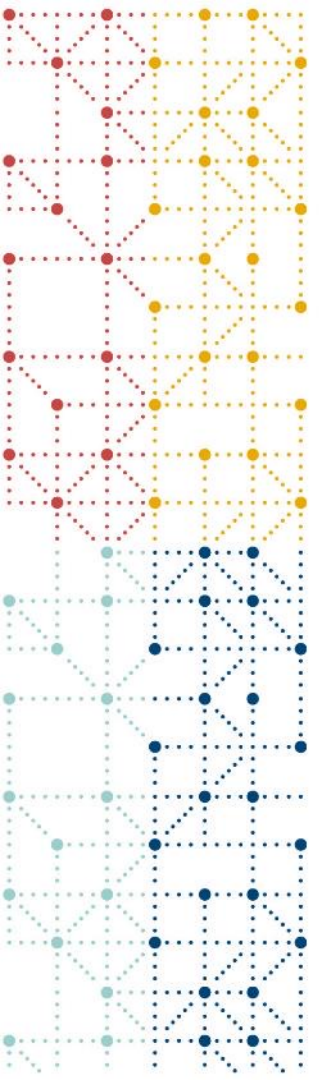
FDA AI & ML in Drug & Biological Products



Table of Contents

I. Background and Scope	1
II. Current and Potential Uses of AI/ML in Drug Development	3
A. <u>Drug Discovery</u>	3
1. Drug Target Identification, Selection, and Prioritization	3
2. Compound Screening and Design	4
B. <u>Nonclinical Research</u>	5
C. <u>Clinical Research</u>	5
1. Recruitment	6
2. Selection of Trial Participants	6
3. Dose/Dosing Regimen Optimization	6
4. Adherence	7
5. Retention	7
6. Site Selection	7
7. Clinical Trial Data Collection, Management, and Analysis	7
8. Clinical Endpoint Assessment	8
D. <u>Postmarket Safety Surveillance</u>	9
1. Case Processing	9
2. Case Evaluation	10
3. Case Submission	10
E. Advanced Pharmaceutical Manufacturing	10
1. Optimization of Process Design	11
2. Advanced Process Control	11
3. Smart Monitoring and Maintenance	12
4. Trend Monitoring	12
F. FDA Experience with AI/ML for Drug Development	12
III. Considerations for the Use of AI/ML in Drug Development	15
A. Overarching Standards and Practices for the Use of AI/ML	15
B. Discussion of Considerations and Practices for AI/ML in Drug Development ..	17
IV. Next Steps: Engagement and Collaboration	23
Glossary	24
References	27

<https://www.fda.gov/media/167973/download?attachment>



AI Applications in Different Drug Discovery & Development Stages

AI Applications in Drug Discovery & Development

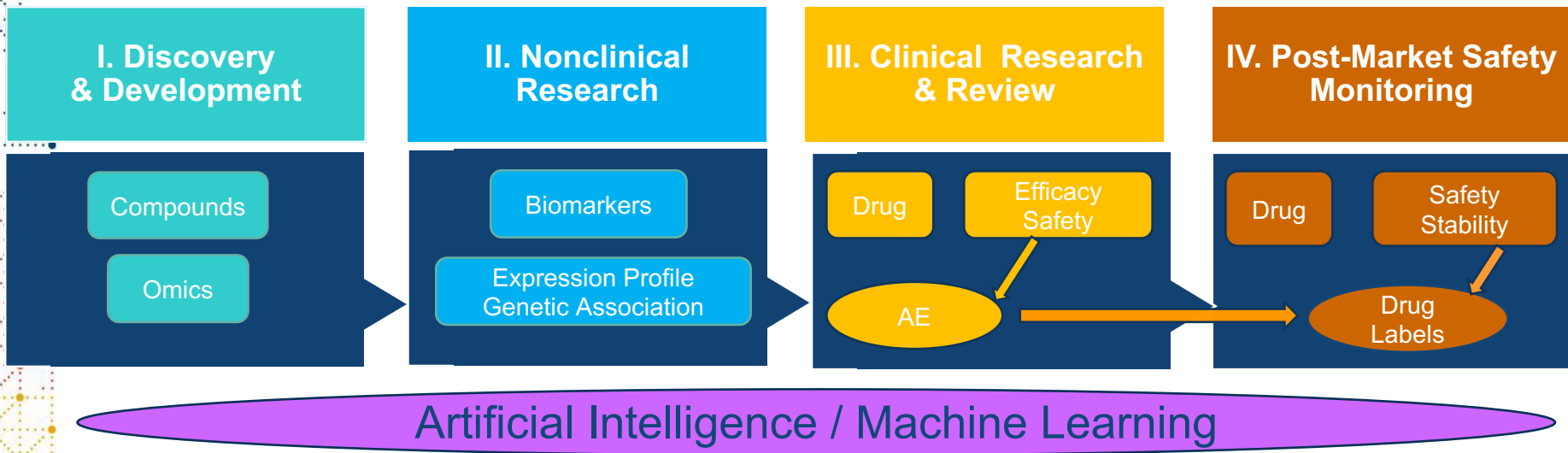


Figure 1 Overview the advantages of AI/ML can take for the data generated from the different stages of drug discovery and development. The omics data in the stage I and II can help screening compounds and biomarkers. Clinical trial data in stage III not only can assist to predict clinical outcomes but also support to generate drug labels. Drug labels in the stage IV provide the important drug information for the public health.

AI Empowers the Assessment of Hepatotoxicity Potential and Clinical Endpoints in Different Stages of Drug Discovery and Development

Minjun Chen, Zhichao Liu, Joshua Xu, Thomas J Pedersen, Byron Wingerd, Meichen Dong, Nicholas Mann, Weida Tong, Russell D Wolfinger, and Wenjun Bao*

Accepted in the book: Advance of Artificial intelligence (AI) in Biomedical Science 2024.

AI Applications in Drug Discovery & Development

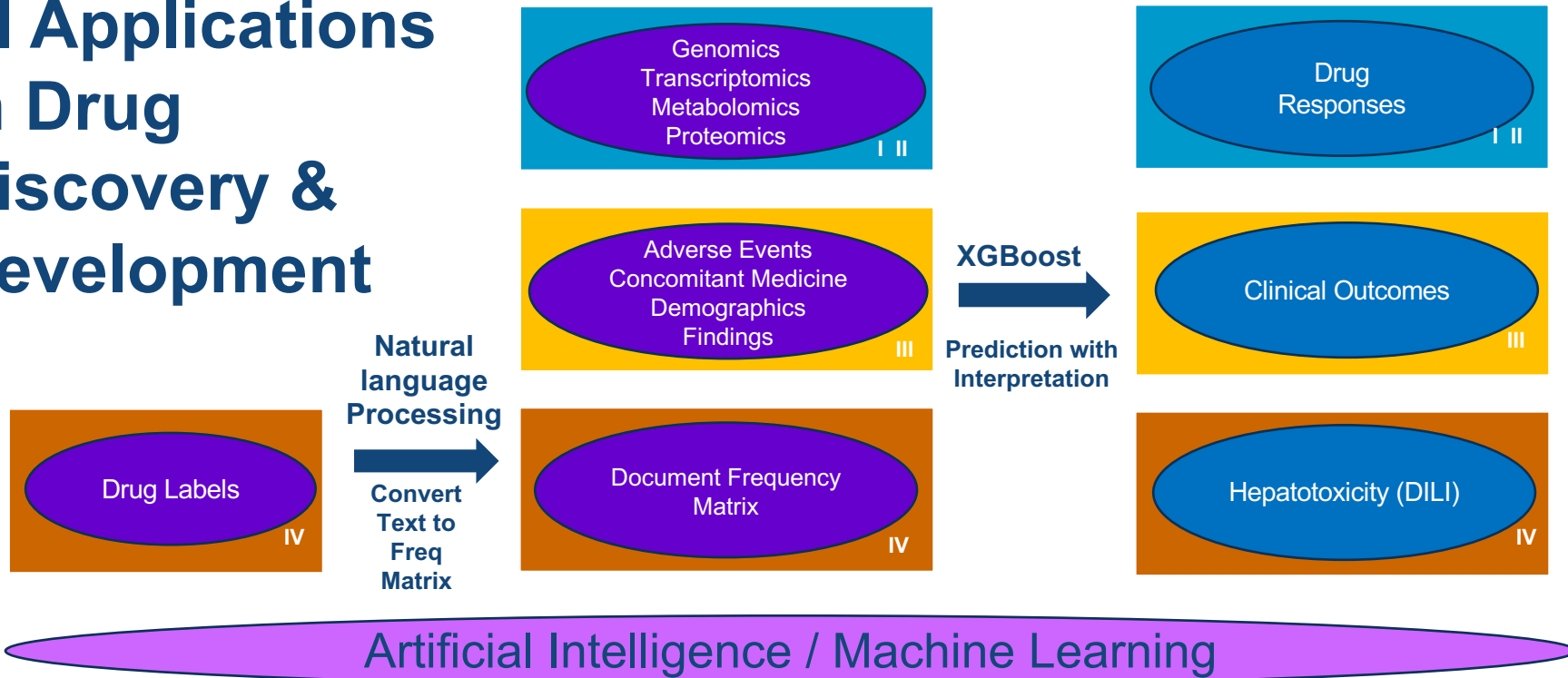


Figure 2. Overview AI/ML tools in different drug discovery and development. Natural language processing to convert drug labels in stage IV to document frequency matrix. XGBoost tools employ omics data in stage I and II, information in stage III, and document frequency matrix in stage IV to predict drug responses, clinical outcomes and hepatotoxicity (DILI) respectively.

AI Empowers the Assessment of Hepatotoxicity Potential and Clinical Endpoints in Different Stages of Drug Discovery and Development

Minjun Chen, Zhichao Liu, Joshua Xu, Thomas J Pedersen, Byron Wingerd, Meichen Dong, Nicholas Mann, Weida Tong, Russell D Wolfinger, and Wenjun Bao*

Accepted in the book: Advance of Artificial intelligence (AI) in Biomedical Science 2024.

Post-Market Safety with Drug Labels for Classifying DILI

Text Explorer + XGBoost in JMP Pro



Products Solutions Industries For You

Shop Drivers Support



What is XGBoost?

<https://www.nvidia.com/en-us/glossary/xgboost/>

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

It's vital to an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon: supervised machine learning, decision trees, ensemble learning, and gradient boosting.



Products Solutions Industries For You

Shop Drivers Support



Why XGBoost?

XGBoost gained significant favor in the last few years as a result of helping individuals and teams win virtually every Kaggle structured data competition. In these competitions, companies and researchers post data after which statisticians and data miners compete to produce the best models for predicting and describing the data.

Initially both Python and R implementations of XGBoost were built. Owing to its popularity, today XGBoost has package implementations for Java, Scala, Julia, Perl, and other languages. These implementations have opened the XGBoost library to even more developers and improved its appeal throughout the Kaggle community.

XGBoost has been integrated with a wide variety of other tools and packages such as scikit-learn for Python enthusiasts and caret for R users. In addition, XGBoost is integrated with distributed processing frameworks like Apache Spark and Dask.

In 2019 XGBoost was named among InfoWorld's coveted Technology of the Year award winners.

Post-Market Safety with Drug Labels for Classifying DILI

Text Explorer + XGBoost in JMP Pro

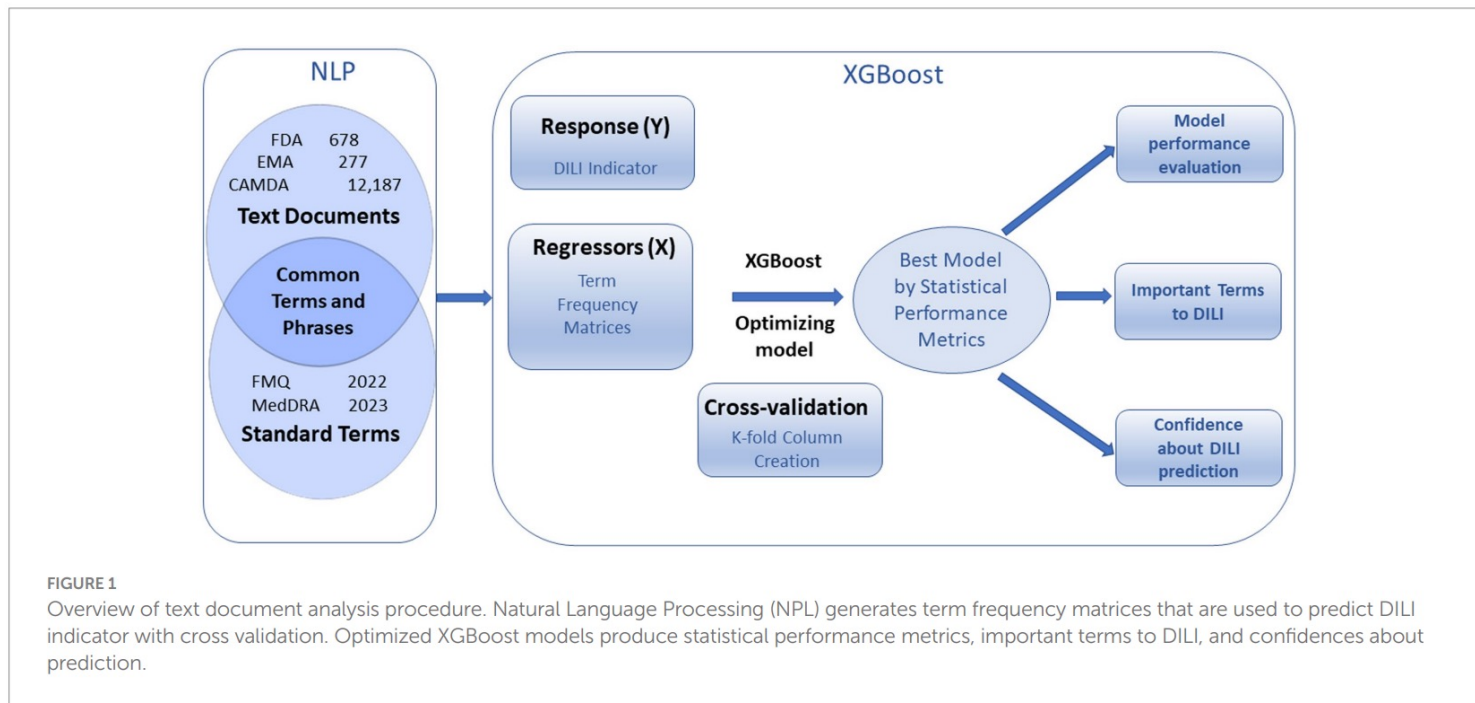
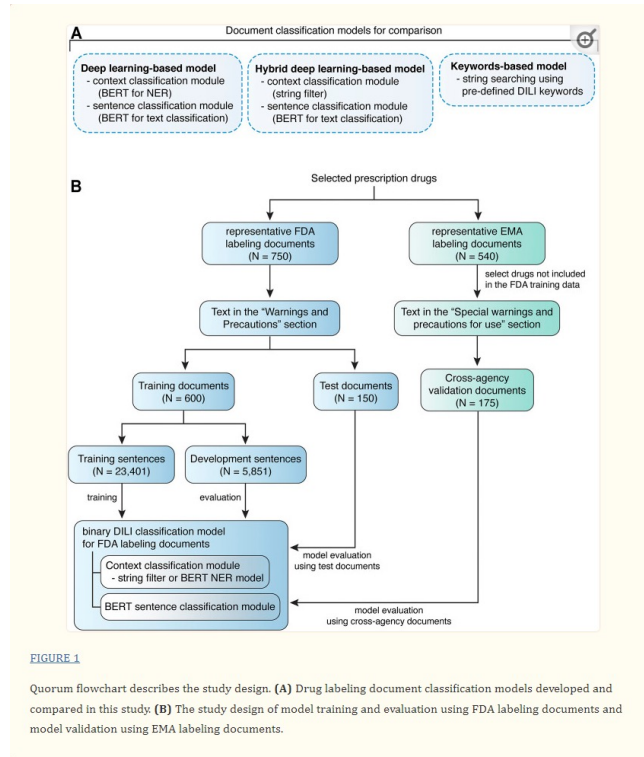


FIGURE 1
Overview of text document analysis procedure. Natural Language Processing (NLP) generates term frequency matrices that are used to predict DILI indicator with cross validation. Optimized XGBoost models produce statistical performance metrics, important terms to DILI, and confidences about prediction.



Post-Market Safety with Drug Labels for Classifying DILI

BERT (Bidirectional Encoder Representations from Transformers) by Google



Val ACC 0.81

FIGURE 1

Quorum flowchart describes the study design. (A) Drug labeling document classification models developed and compared in this study. (B) The study design of model training and evaluation using FDA labeling documents and model validation using EMA labeling documents.

Post-Market Safety with Drug Labels for Classifying DILI

Text Explorer + XGBoost in JMP Pro

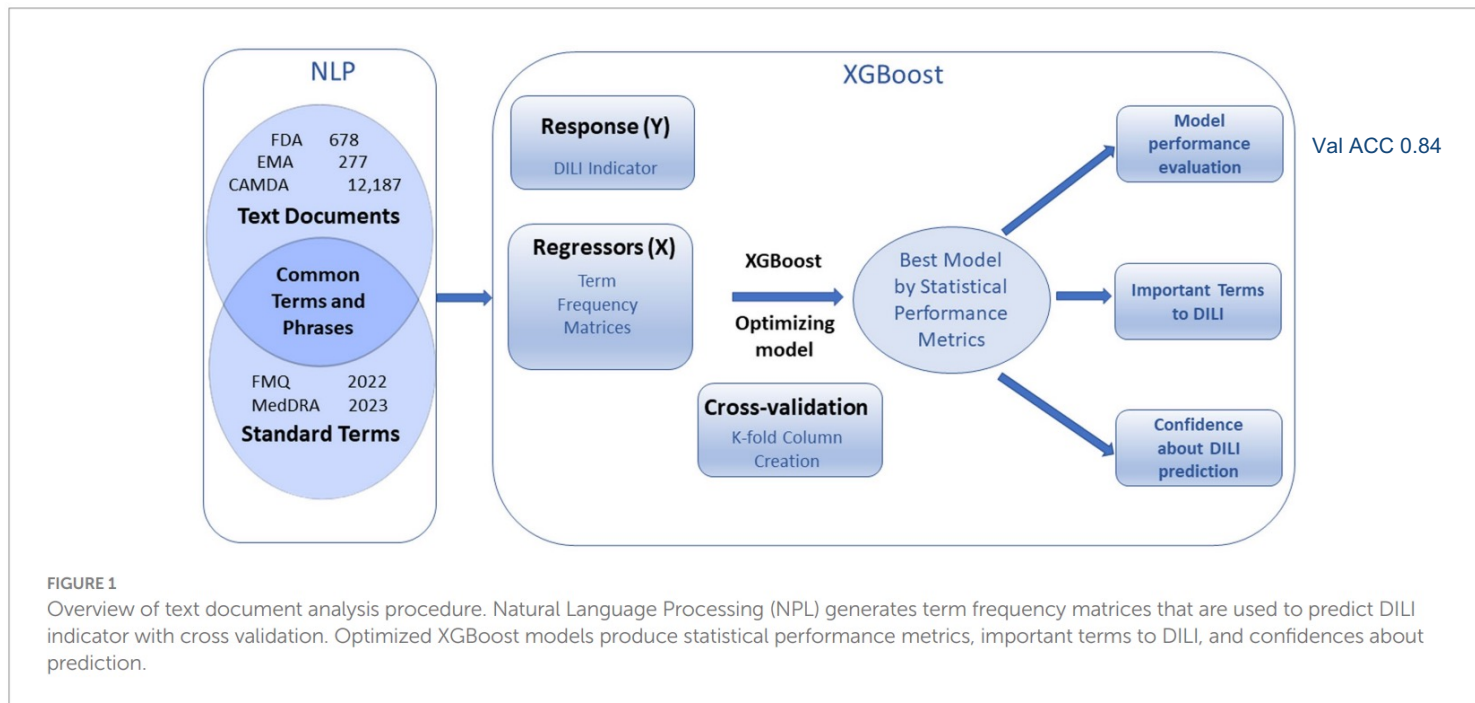
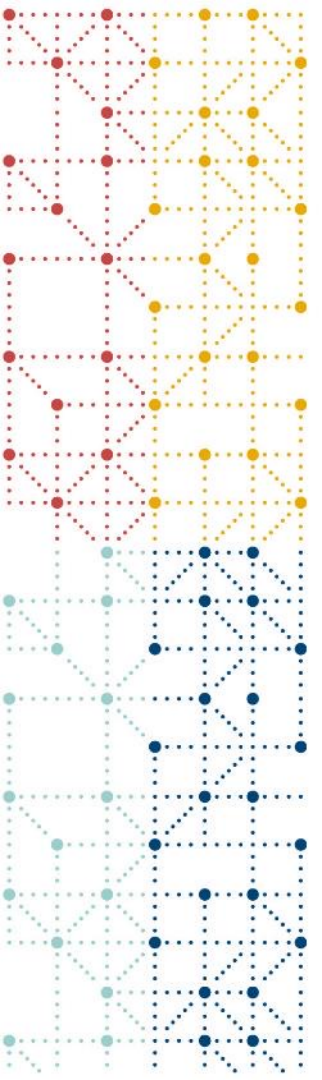


FIGURE 1
Overview of text document analysis procedure. Natural Language Processing (NLP) generates term frequency matrices that are used to predict DILI indicator with cross validation. Optimized XGBoost models produce statistical performance metrics, important terms to DILI, and confidences about prediction.





CDISC Initiatives for JSON



Dataset-JSON

[Pilot](#) [Specification](#)

CDISC and PHUSE are delighted to announce a new pilot project aimed at accelerating the adoption of Dataset-JSON submissions. This pilot builds upon the considerable amount of work done by CDISC and PHUSE with regulatory authorities.

The pilot will be split into short-term goals of the acceptance of Dataset-JSON and the future strategy relating to the adoption of advanced Dataset-JSON.

Milestone 1: Short-Term

- Pilot submissions using the JSON format with existing XPT imports
- Same content, different suitcase, no disruption to business processes
- In parallel, evaluate with the FDA how their toolset can support Dataset-JSON

Milestone 2: Development of Future Strategy

- Evaluate how current and future industry standards can benefit from Dataset-JSON
- Evaluate combining metadata with data (e.g., Define-XML / Define-JSON)
- Enhanced conformance rules
- Collaborate with the FDA to develop plans to retool their environment

<https://www.cdisc.org/dataset-json>

Dataset-JSON

[Pilot](#) [Specification](#)

Dataset-JSON was released as part of ODM v2.0 in 2023. Dataset-JSON version 1.1 is currently under development and will be published as an independent standard.

Dataset-JSON was adapted from the Dataset-XML Version 1.0 specification but uses JSON format. Like Dataset-XML, each Dataset-JSON file is connected to a Define-XML file that contains detailed information about the metadata. One aim of Dataset-JSON is to address as many of the relevant requirements from the PHUSE 2017 [Transport for the Next Generation](#) paper as possible, including the efficient use of storage space.

Dataset-JSON uses lowerCamelCase notation for attribute names, compared to Dataset-XML PascalCase (e.g., `clinicalData` vs `ClinicalData`).

JSON format does not allow to specify or control order of attributes. Despite that, as most JSON engines allow to control the order of attributes, it is strongly recommended to follow the attribute order specified in detail. Due to a possible large size of Dataset-JSON files, following the specified order will enable a software using streaming approaches to read the file to work in an efficient and fast way.

Dataset-JSON must contain only one dataset per file.

Top Level Attributes

At the top level of Dataset-JSON object, there are technical attributes and two main optional attributes: `clinicalData` and `referenceData`, corresponding to Dataset-XML elements. At least 1 of the main attributes must be provided. Subject data is stored in `clinicalData` and non-subject data is stored in `referenceData`.



ChatGPT: What is json?

- JSON (JavaScript Object Notation) is a lightweight data interchange format that's easy for humans to read and write and easy for machines to parse and generate.
- JSON is commonly used to transmit data between a server and a web application, serving as an alternative to XML.

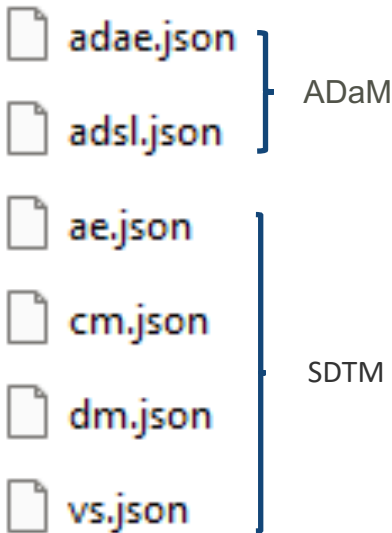
ChatGPT: What are advantages of json?

- **Human-Readable:** JSON is easy for humans to read and write due to its simple syntax, which is based on key-value pairs and arrays. This makes it easier to debug and understand compared to other formats like XML.
- **Lightweight:** JSON is more concise and less verbose than other data formats like XML, making it faster to transmit over networks. Its lightweight nature is particularly beneficial for web applications where performance and speed are critical.
- **Language Independence:** JSON is text-based and can be easily parsed and generated by most programming languages, including Python, Java, JavaScript, C#, and many others. This makes it highly interoperable across different systems and platforms.
- **Easy Parsing:** Most programming languages have built-in support for JSON, allowing developers to easily parse and convert JSON data into native data structures (e.g., objects, arrays). This reduces the complexity of working with data.
- **Structured Data:** JSON can represent complex data structures, including nested objects and arrays, making it versatile for a wide range of applications, from simple configurations to complex datasets.
- **Widely Used in Web Development:** JSON is the standard format for data exchange in web APIs (Application Programming Interfaces). It is commonly used for sending and receiving data between a client (like a web browser) and a server.
- **No Need for Schema:** Unlike XML, JSON does not require a predefined schema. This makes it more flexible, allowing data structures to evolve without breaking compatibility.
- **Supports Unicode:** JSON natively supports Unicode, making it ideal for international applications that need to handle text in multiple languages.

Summary: These advantages contribute to JSON's popularity in modern software development, particularly in web technologies and APIs.

CDISC Initiative for Adoption of Dataset-JSON JSOM Dataset

CDISC Sample Data



```
adsljson
File Edit View

{"clinicalData":{"studyOID":"CDISCPIL0T01","metaDataVersionOID":"CDISC.ADaM.2.1","itemGroupData":{"ADSL":{"records":254,"name":"ADSL","label":"Subject-Level Analysis","items":[{"OID":"ITEMGROUPDATASEQ","name":"ITEMGROUPDATASEQ","label":"Record Identifier","type":"integer"},{"OID":"ADSL.STUDYID","name":"STUDYID","label":"Study Identifier","type":"string","length":12},{"OID":"ADSL.USUBJID","name":"USUBJID","label":"Unique Subject Identifier","type":"string","length":11},{"OID":"ADSL.SUBJID","name":"SUBJID","label":"Subject Identifier for the Study","type":"string","length":4},{"OID":"ADSL.SITEID","name":"SITEID","label":"Study Site Identifier","type":"string","length":3},{"OID":"ADSL.SITEGR1","name":"SITEGR1","label":"Pooled Site Group 1","type":"string","length":3},{"OID":"ADSL.ARM","name":"ARM","label":"Description of Planned Arm","type":"string","length":20},{"OID":"ADSL.TRT01P","name":"TRT01P","label":"Planned Treatment for Period 01","type":"string","length":20},{"OID":"ADSL.TRT01PN","name":"TRT01PN","label":"Planned Treatment for Period 01 (N)","type":"integer","length":8},{"OID":"ADSL.TRT01A","name":"TRT01A","label":"Actual Treatment for Period 01","type":"string","length":20},{"OID":"ADSL.TRT01AN","name":"TRT01AN","label":"Actual Treatment for Period 01 (N)","type":"integer","length":8},{"OID":"ADSL.TRTSDT","name":"TRTSDT","label":"Date of First Exposure to Treatment","type":"integer","length":8},{"OID":"ADSL.TRTEDT","name":"TRTEDT","label":"Date of Last Exposure to Treatment","type":"integer","length":8},{"OID":"ADSL.TRTDUR","name":"TRTDUR","label":"Duration of Treatment (days)","type":"integer","length":8},{"OID":"ADSL.AVGDD","name":"AVGDD","label":"Avg Daily Dose (as planned)","type":"float","length":8},{"OID":"ADSL.CUMDOSE","name":"CUMDOSE","label":"Cumulative Dose (as planned)","type":"float","length":8},{"OID":"ADSL.AGE","name":"AGE","label":"Age","type":"integer","length":8},{"OID":"ADSL.AGEGR1","name":"AGEGR1","label":"Pooled Age Group 1","type":"string","length":5},{"OID":"ADSL.AGEGR1N","name":"AGEGR1N","label":"Pooled Age Group 1 (N)","type":"integer","length":8},{"OID":"ADSL.AGEU","name":"AGEU","label":"Age Units","type":"string","length":5},{"OID":"ADSL.RACE","name":"RACE","label":"Race","type":"string","length":32},{"OID":"ADSL.RACEN","name":"RACEN","label":"Race (N)","type":"integer","length":8},{"OID":"ADSL.SEX","name":"SEX","label":"Sex","type":"string","length":1},{"OID":"ADSL.ETHNIC","name":"ETHNIC","label":"Ethnicity","type":"string","length":22},{"OID":"ADSL.SAFFL","name":"SAFFL","label":"Safety Population Flag","type":"string","length":1},{"OID":"ADSL.ITTFL","name":"ITTFL","label":"Intent-To-Treat Population Flag","type":"string","length":1},{"OID":"ADSL.EFFFL","name":"EFFFL","label":"Efficacy Population Flag","type":"string","length":1},{"OID":"ADSL.COMP8FL","name":"COMP8FL","label":"Completers of Week 8 Population Flag","type":"string","length":1},{"OID":"ADSL.COMP16FL","name":"COMP16FL","label":"Completers of Week 16 Population Flag","type":"string","length":1},{"OID":"ADSL.COMP24FL","name":"COMP24FL","label":"Completers of Week 24 Population Flag","type":"string","length":1},{"OID":"ADSL.DISCONFL","name":"DISCONFL","label":"Did the Subject Discontinue the Study?","type":"string","length":1},{"OID":"ADSL.DSRAEFL","name":"DSRAEFL","label":"Discontinued due to
```



Readable ✓

Understandable ?

CDISC Initiative for Adoption of Dataset-JSON JSON Imports into JMP Clinical

Add Study - JMP Clinical

*Study Name

SDTM Folder **.JMP**
 .JSON

Selected SDTM Domains **.SAS7dbat**
 .XPT

ADaM Folder

Selected ADaM Domains

Advanced

- Make this the current study
- Close this window after adding study

Warning: To prevent name collisions, close all JMP data tables that have same names as the files you are trying to add.

ae.json cm.json dm.json vs.json

```
File Edit View

{"clinicalData":{"studyOID":"cdisc.com/CDISCPIL0T01","metaDataVersionOID":"MDV.MS
pData":{"IG.VS":{"records":1414,"name":"VS","label":"Vital
Signs","items":[{"OID":"ITEMGROUPDATASEQ","name":"ITEMGROUPDATASEQ","label":"Reco
Identifier","type":"integer"}, {"OID":"IT.VS.STUDYID","name":"STUDYID","label":"St
Identifier","type":"string","length":12}, {"OID":"IT.VS.DOMAIN","name":"DOMAIN","l
Abbreviation","type":"string","length":2}, {"OID":"IT.VS.USUBJID","name":"USUBJID"
```

Parsing js1 Public Available

ae - JMP Clinical

ae	STUDYID	DOMAIN	USUBJID	AESQ	AETERM	AEDECOD	AEHLT	AEHLGT
----	---------	--------	---------	------	--------	---------	-------	--------

cm - JMP Clinical

cm	STUDYID	DOMAIN	USUBJID	CMSEQ	CMTRT	CMDECOD	CMCAT
----	---------	--------	---------	-------	-------	---------	-------

dm - JMP Clinical

dm	STUDYID	DOMAIN	USUBJID	SUBJID	RFSTDTCT	RFENDTCT	SITEID
----	---------	--------	---------	--------	----------	----------	--------

vs - JMP Clinical

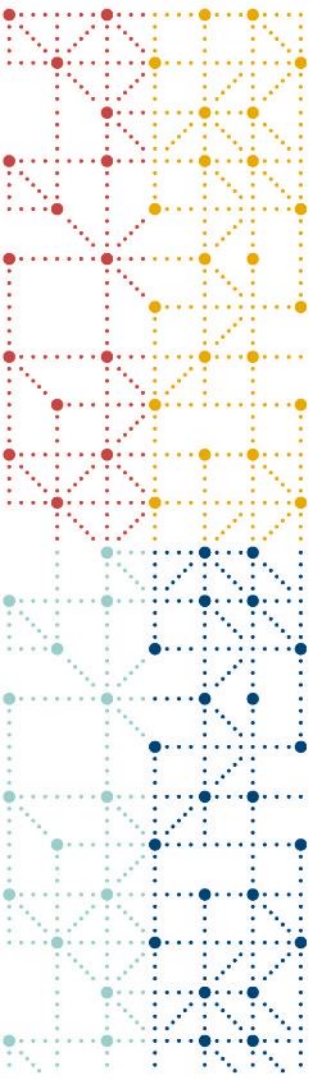
vs	STUDYID	DOMAIN	USUBJID	VSSEQ	VSTESTCD	VSTEST
----	---------	--------	---------	-------	----------	--------

Easy Parsing ✓



Conclusions

1. FDA values AI for drug discovery and development (DDD) with official documents and many research papers.
2. AI can apply to difference DDD stages: Text Explorer + XGBosst in JMP Pro can generate a similar or slight better ACC than BERT from Google without coding.
3. CDISC promotes JSON data format for its flexibility, readability, and broad support across different programming environments.



Thanks!
Wenjun.bao@jmp.com

cdisc