DATA GLOSSARY AS AN ASSET TO OPTIMIZE DATAOPS IN PHARMACEUTICAL INDUSTRY

Prepared for the partial fulfilment for the award of the degree of MBA in Hospital and Health Management

By

Dr Pravie

Under the Supervision and Guidance of

Dr Dhirendra Kumar

(Professor, IIHMR University, Jaipur)

Overview

- Background: The amount of clinical data is compounding at an astonishing rate in Pharmaceutical Industry. Largely, this is due to increased ability to generate and collect vast amounts of data with the help of latest technology. Pharmaceutical companies rely heavily on huge amounts of data generated during R&D processes across many functions and are looking for ways to derive fact-based and informed decisions based on that data. There is a need for organizations to optimize their data operations to enhance their business outcomes while maintaining regulatory compliance.

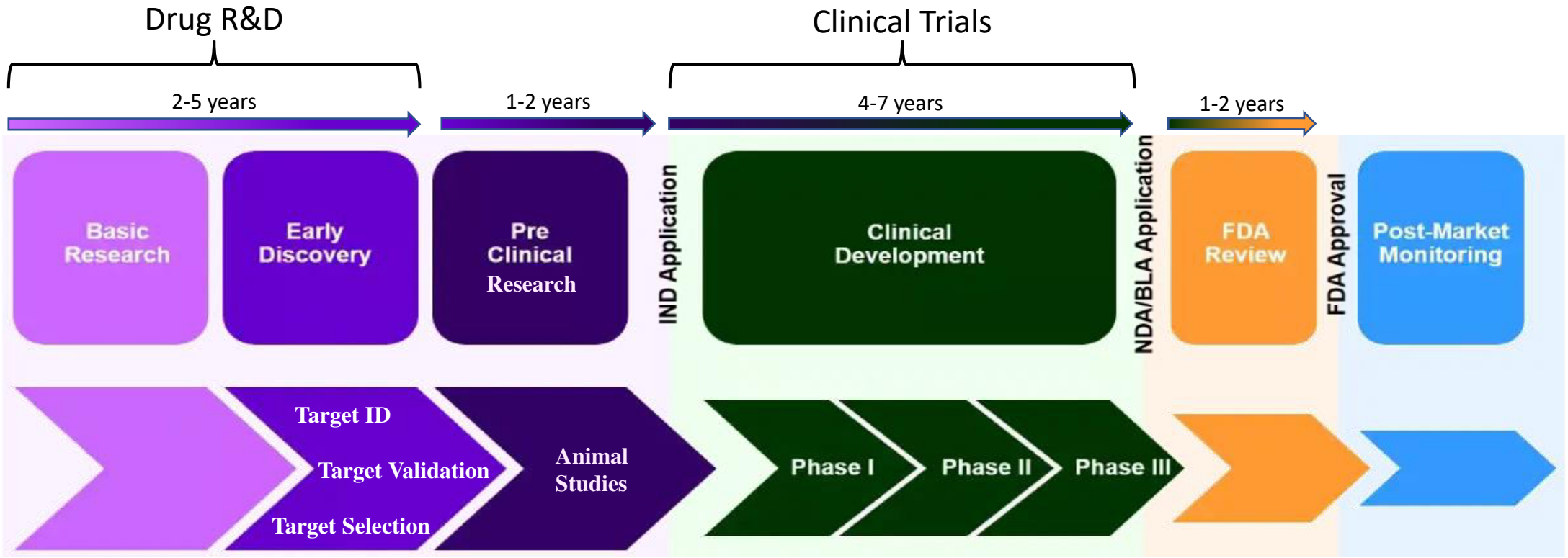
- Aim: To determine how data glossaries can be built as an asset that can help the pharmaceutical companies to optimize their data operation.

- Methodology: Conducted exhaustive secondary research on 43 organizations from pharmaceutical industry to understand their usage of data standards for clinical research so as to get an idea about what standards are being majorly adopted by pharmaceutical industry and build data glossary on the basis of that.

- Results:
  - Out of 43 organizations, 40 are adopting CDISC standards while 3 are using standards other than CDISC, such as SNOMED CT, LOINC, Hl-7, etc.
  - CDISC standards terminology is being adopted as the global language for data sharing in the pharmaceutical industry.

Introduction

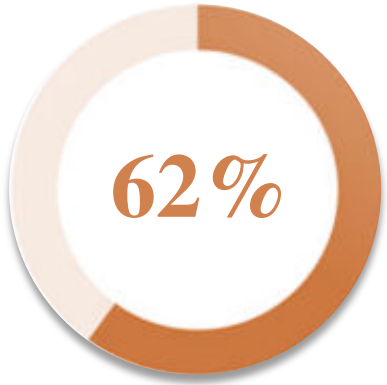
- The pharma R&D process requires years of research and huge data gathering.
- It takes an average of **12-15 years** from basic research to FDA approvals to bring a new drug to market.
- Through the whole process from discovery and development to clinical trials, **a massive amount of data** is generated.
- A colossal sum of **USD 1 billion** is required for drug discovery and development.



ORGANIZATIONS DON'T MEASURE THE ANNUAL FINANCIAL COST OF POOR-QUALITY DATA

Source: Gartner's Data Quality Market Survey, January 2018

Impact of Bad Data Quality



of 65 Life Sciences companies surveyed reported having insufficient data governance processes and data standards for regulatory compliance

Increased Risk

- Non-compliance risks
- Information integration risk
- Reputation risk

Increased Costs

- Frequent data remediation
- Increased resource cost
- Health authority penalties
- Potential rework

Decreased Revenues

- Lost revenues
- Increased cost/volume
- Lost opportunities

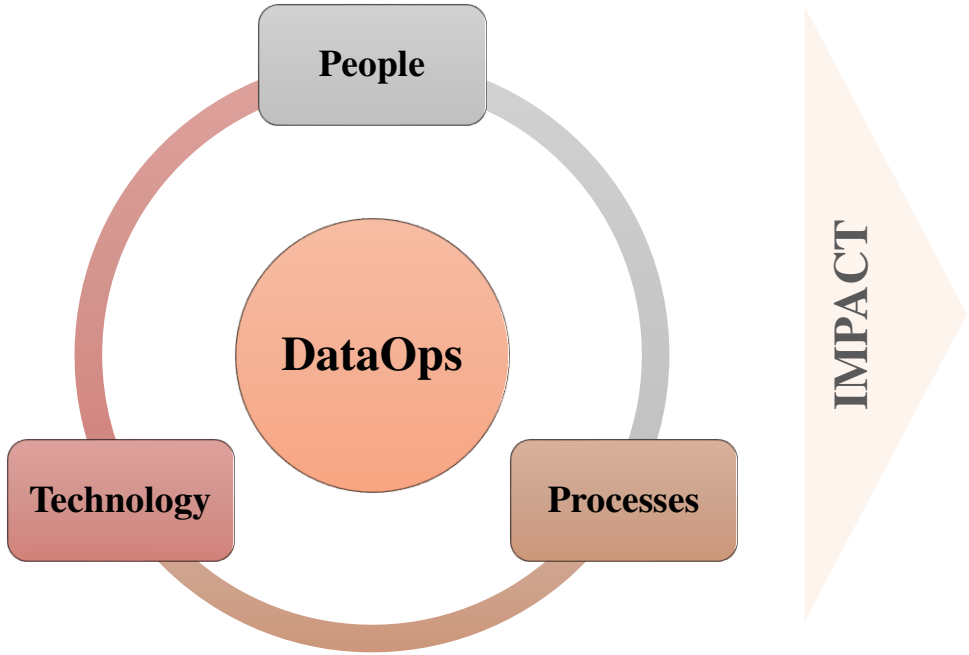
Less Confidence

- Impaired decision making
- Lowered forecasting
- Organizational trust issues
- Inconsistent management reporting

Literature Review

According to Gartner

> *DataOps is a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization. The goal of DataOps is to deliver value faster by creating predictable delivery and change management of data, data models and related artifacts. DataOps uses technology to automate the design, deployment and management of data delivery with appropriate levels of governance, and it uses metadata to improve the usability and value of data in a dynamic environment.*



Implementation of DataOps across entire organization results in:

- **Improved quality** of data that is ready to be consumed by the end users
- **Rapid and secure** ingestion/migration of data
- **Reliable** flow of data throughout data lifecycle
- Supports integration of data science and machine learning for improved and fact-based business decision making
- Establishment of **standardized data** that enables compliance and standardized data security policies.

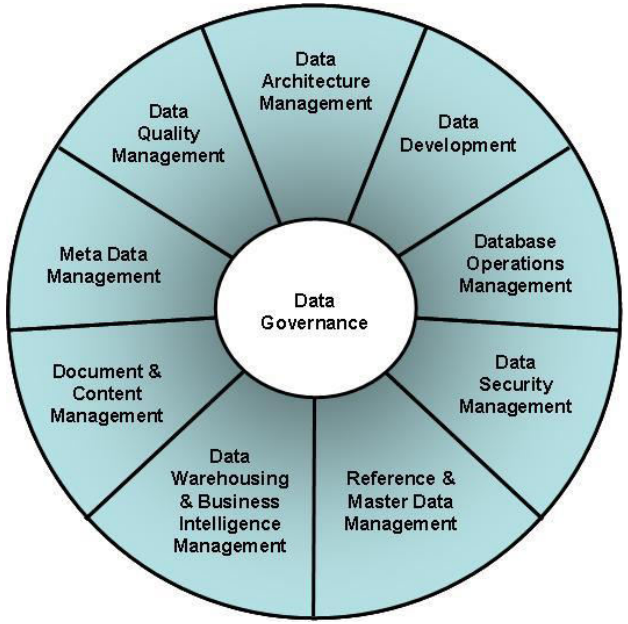
According to Indian Economic Survey 2021

- By 2023, the global pharmaceutical market is expected to reach USD 1.5 trillion
- Indian pharmaceutical market is currently estimated at USD 41 billion
- Projected to reach USD 65 billion by 2014 with a further rise of USD 120-130 billion by 2030
- India's expenditure on medicine to rise by 9-12% in next 5 years.

Challenges faced by Pharmaceutical Industry

Pressure to lower costs to bring new drug to market	Speed up the R&D processes to quickly bring new drugs	Stricter regulatory and compliance requirements	Demand for improved and more effective data management and governance	Reputational losses due to counterfeit drugs

DataOps Catering to these Challenges



DAMA-DMBOK Functional Framework

Need to build and implement assets and accelerators to optimize these data management functional areas.

What is a Data Glossary

It is an easy-to-search compilation of all the terms that defines multiple data types and their key characteristics used within the organization as well as the industry so that everyone may refer and utilize the standard definitions and impart harmony around the system's guiding concepts.

Need of Data Glossaries in Pharmaceutical Industry

Poor Quality of Data

The data to be ingested is unorganised, fragmented, unstructured, missing values, inconsistent with potential biases. Organizations are compelled to spend their valuable time onboarding, cleaning and arranging the data.

Poor Performance

Need to process vast quantities of data easily and rapidly to obtain timely and relevant insights. Lack of efficient consolidation, authentication and mining of clinical research data.

Lack of Collaboration

Standardized vocabulary of clinical data terms enables collaboration and ease of data exchange to enhancing decision-making.

Poor Compliance

Adherence to ever evolving regulations and data quality guidelines for the appropriate usage, storage and disposal of sensitive data.

Data Glossary as a Solution

Standardized Data Onboarding & Data Transfers

Standardized and structured data is easy to process and flow

Safeguard Data

Data glossaries can assist organizations in handling their data more efficiently and avoid legal and regulatory violations.

Cost Reduction

By eliminating data silos, companies can better understand their data ecosystem, uncover redundancies, streamline processes resulting in decreased costs.

Risk Mitigation

Data glossaries can help in identification of poor quality of data if the data is not at par with the defined standards hence highlighting the potential issues and preventing crisis.

Regulate Change

By ensuring operational consistency, minimizing risks and reducing costs, comprehensive data glossaries enable enterprises in their attempts to ease the pressure of regulations.

Methodology

The main objective of this research is to determine how can we build a standardized data glossary that is specific to pharmaceutical industry to optimize the data operations. It is better to follow the existing industrial standards rather than building the data glossary from scratch or creating new ones. This will make the data glossaries standardized throughout the organization as well as industry no matter the data type or a new use case. We conducted exhaustive secondary research on a total of 43 companies that in one way or other constitutes pharmaceutical industry such as, top 15 big pharmaceutical companies (based on their R&D spending, 2021), top 10 CROs, 5 big consulting companies, 5 big Biotech companies, and few government organizations to identify the answers to following research questions:

i. Are there any type of standards that companies are either using or offering services that utilize these standards to standardize their data operations?
ii. If they are using standards, then what are those standards.

Sr. No.	Name of the Company	Company Type	Any Standards for Data Operations	Type of Standards
1	AbbVie	Pharmaceutical	Yes	CDISC Standards
2	Accenture	Consulting	Yes	CDISC Standards
3	Amgen	Pharmaceutical	Yes	CDISC Standards
4	AstraZeneca	Pharmaceutical	Yes	CDISC Standards
5	Bayer Healthcare	Pharmaceutical	Yes	CDISC Standards
6	Biogen	BioTech	Yes	CDISC Standards
7	Boehringer Ingelheim Pharmaceutical	Pharmaceutical	Yes	CDISC Standards
8	Bristol-Myers Squibb	Pharmaceutical	Yes	CDISC Standards
9	Cognizant	Consulting	Yes	SNOMED CT, LOINC, RxNorm, ICD-10
10	Deloitte	Consulting	Yes	CDISC Standards
11	Eli Lily	Pharmaceutical	Yes	CDISC Standards
12	EY	Consulting	Yes	HL7 FHIR, SNOMED CT, LOINC
13	F. Hoffmann-La Roche	Pharmaceutical	Yes	CDISC Standards
14	Food & Drug Administration	Government	Yes	CDISC Standards
15	Genmab	BioTech	Yes	CDISC Standards
16	Gilead Sciences	Pharmaceutical	Yes	CDISC Standards
17	GlaxoSmithKline	Pharmaceutical	Yes	CDISC Standards
18	IBM (IBM Watson)	Consulting	Yes	CDISC Standards
19	ICON Clinical Research	CRO	Yes	CDISC Standards
20	IQVIA	CRO	Yes	CDISC Standards
21	Janssen (J&J)	Pharmaceutical	Yes	CDISC Standards
22	KCR	CRO	Yes	CDISC Standards
23	KPMG	Consulting	Yes	SNOMED CT
24	Labcorp (Covance)	CRO	Yes	CDISC Standards
25	Medpace Inc	CRO	Yes	CDISC Standards
26	Merck & Company	Pharmaceutical	Yes	CDISC Standards
27	National Cancer Institute	Government	Yes	CDISC Standards
28	National Institute of Allergy & Infectious Diseases	Government	Yes	CDISC Standards
29	National Library of Medicine	Government	Yes	CDISC Standards
30	Novartis Pharmaceutical Corporation	Pharmaceutical	Yes	CDISC Standards
31	Novo Nordisk	BioTech	Yes	CDISC Standards
32	Parexel	CRO	Yes	CDISC Standards
33	Pfizer	Pharmaceutical	Yes	CDISC Standards
34	PPD	CRO	Yes	CDISC Standards
35	Pharmaceuticals and Medical Devices Agency	Government	Yes	CDISC Standards
36	PSI CRO AG	CRO	Yes	CDISC Standards
37	Sanofi	Pharmaceutical	Yes	CDISC Standards
38	Syneos Health Inc	CRO	Yes	CDISC Standards
39	Takeda	Pharmaceutical	Yes	CDISC Standards
40	TCS Life Sciences ADD	Consulting	Yes	CDISC Standards
41	US Centres for Disease Control (CDC)	Government	Yes	CDISC Standards
42	Vertex Pharmaceuticals	BioTech	Yes	CDISC Standards
43	WuXi Clinical Development, Inc.	BioTech	Yes	CDISC Standards

Result and Discussion

Out of 43 organizations, 40 of them are using CDISC standards while 3 are using other standards such as SNOMED CT, LOINC, etc. From the above outcomes it can be concluded that CDISC Standards are becoming the universal language for information exchange in the pharmaceutical industry R&D processes. Majority of the organizations are now adopting CDISC Standards to homogenize their clinical research data. Building data glossaries based on these standards will impart a significant impact on streamlining and optimizing the data operations.

CDISC Standards

The Clinical Data Interchange Standards Consortium (CDISC) is an open, non-profit organization that develops and supports global data standards to improve the quality and interoperability of medical research and healthcare. CDISC standards are widely used for study planning and data collection, tabulation, analysis, and submissions to the U.S. Food and Drug Administration (FDA), Japanese Pharmaceuticals and Medical Devices Agency (PMDA), and other regulatory agencies internationally.

To create and foster controlled terminology for all CDISC core standards, CDISC collaborates with NCI Enterprise Vocabulary Services (EVS).

CDISC has developed controlled terminology for its following foundational standards:

- **Study Data Tabulation Model** – a world-wide standard for clinical research data approved by the FDA as a standard electronic submission format. SDTM terminology consists of vocabulary for CDISC Questionnaires, Ratings and Scales.
- **Clinical Data Acquisition Standards Harmonization** – The CDASH Terminology is a subset of SDTM Terminology. In order to enable unambiguous traceability of submission data into the SDTM and increase visibility for regulators and those who execute data review, CDASH establishes a standard approach to gather data consistently across studies and sponsors.
- **Analysis Data Model** – ADaM controlled terminology defines standards supporting efficient generation, replication, review, and submission of analysis results from clinical trial data.
- **Standard for the Exchange of Nonclinical Data** – Controlled terminology for SEND defines the standards for collection and presentation of nonclinical data in a consistent format. SEND is an implementation of the SDTM standard for nonclinical studies.
- **Protocol Representation Model (PRM)** – Protocol terminology elaborates the standards for planning and designing a research protocol with focus on study characteristics such as study design, eligibility criteria, and requirements from the government bodies such as ClinicalTrials.gov, WHO registries, and EudraCT registries.
- **CDISC Glossary** – This initiative of CDISC tends to harmonize definitions as well as acronyms, abbreviations, and initials used in CDISC standards for clinical research.

According to CDISC 2014 Business Case Highlights

> the use of CDISC standards at project initiation can save 70-90% of time and resources spent prior to first patient enrolled and approximately 75% of the non-patient participation time during the study conduct and analysis stages.

Implementing CDISC standards in pharmaceutical R&D settings has a number of important advantages such as:

- Ready to plug and play – CDISC standards due to their maturity can be applicable to any type of clinical research studies. Standards are also offered for a wide range of therapeutic areas.
- Promoting Interoperability – CDISC standards can aid interoperability between clinical systems and EHRs, lowering the amount of data re-entry and improving data quality.
- Supporting Regulatory Submission – Data standardization facilitates effective and high-quality data review. In today's environment regulatory bodies have made it compulsory for the organizations to standardize their clinical trials data to facilitate timely reviews.
- Enabling Collaborations – Data sharing is greatly aided by CDISC standards, which improve R&D efficiencies both internally as well as with external partners.

- To build a holistic and comprehensive data glossary terminologies associated with other clinical data standards can also be considered such as SNOMMED CT, LOINC, HL7-FHIR (Fast Healthcare Interoperability Resources), etc.
- No duplication of terms and their definitions. The idea is to make a data glossary that has terminology from all the possible sources but industry specific and each term in the glossary is unique.
- To speed up the transfer of data from Electronic Health Records to datasets that are ready for CDISC submission, the HL7 FHIR standard for electronically transferring healthcare information partners with the CDISC Standards CDASH, SDTM, and Lab Data Exchange.

Why Not Just Simply Use SNOMED Clinical Terms

- CDISC is an open platform. Anybody can easily access CDISC terminologies. Whereas on other hand there is license requirement for use of SNOMED. Though governments of multiple countries have arranged licences for users but there are still some parts of the world where use of SNOMED is not free.
- SNOMED terms lack definitions. Though the terms are very-well documented, they are not defined.

Conclusion

In conclusion, though the idea of data glossary is very simple, but it plays a very important role in optimizing data operations and ultimately imparts a huge positive impact on business. There is no doubt that there are endless advantages of establishing a comprehensive and well-maintained data glossary such as – quicker time-to-market of drugs and therapeutics; streamlined processes, decreasing the error rate and time-saving; reduced costs and resources; increase in operational productivity; improved data and processes transparency and visibility across all the verticals of the organization; improved data access provisioning and monitoring, improved ease of data exchange and onboarding; enhanced data quality facilitating data governance, and many more. CDISC terminologies certainly aid in building standardized data glossaries. CDISC has proved to be a success in standardizing clinical research data for regulations worldwide. It aids in streamlining regulatory submission, review, and approval along with enabling data sharing and metadata analysis. For the purpose of creating data glossaries, it is strongly advised to take into account CDISC standards as they will offer advantages in terms of cost, time, and high data accuracy. Additionally, these guidelines serve as a catalyst to speed up the review process and assist businesses in releasing new medications onto the market. But to build a holistic and comprehensive Data Glossary all the terminology from the existing standards should be reviewed and considered from all the possible sources while making sure that each data term is unique.

THANK YOU