Journal of Information Systems and Telecommunication
Special Issue, Winter,2026, PP 9-24
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Transforming Public Healthcare Supply Chains: A Framework to Measure Efficiency of Heterogeneous Public Healthcare Supply Chains across Nation for Improving Drug Availability.

Abhishek Verma[1*], Dr. Rekha Agarwal[1] , Jitendra Singh[2]

[1].AIIT, Amity University, Noida, Uttar Pradesh
[2].Centre For development of Advanced Computing, Delhi

Received: 25 Jul 2025/ Revised: 13 Oct 2025/ Accepted: 02 Nov 2025

Abstract

Health in Indian scenario is a state subject which means that states usually use their independent respective IT systems based on their specific needs and requirements. This brings a big challenge in terms of diversified nomenclatures used in across Indian states and Union Territories (UTs). Centralized data analysis is required by central agencies like Ministry of Health and Family Welfare (MoHFW) and the National Health Systems Resource Centre (NHSRC) for performance monitoring and policy making. This diversified nomenclature poses a significant hindrance in the process. The aggregation, standardization, deduplication, and visualization of data from these heterogeneous sources is both complex and resource intensive.This paper presents solution for the above challenges and propose a comprehensive framework for analyzing data from heterogeneous sources related to supply chain of drugs and vaccines. The framework incorporates fuzzy logic-based algorithms for deduplication of drug and vaccine nomenclature and supports real-time data analysis through the use of Key Performance Indicators (KPIs). It further enables centralized monitoring and decision-making via a built-in visualization layer, accessible to stakeholders at multiple administrative levels.While the framework has been tailored to the Indian public health context, its modular design makes it broadly applicable to other domains requiring integration of diverse data sources for strategic planning and policy implementation.The use of open-source technologies for development of various configurable and integrated layers like ETL, Deduplication, Standardization and Visualization layers encompassing into a single framework ensuring cost effectiveness in delivering the end-to-end solution makes it a novel and impactful for adoption specially in resource constrained environments.

Keywords: Public Health Informatics; Indian Health Framework; Deduplication; Digital Health; Drugs and Vaccines Availability; Essential Drugs and Vaccines; Extraction Transformation and loading (ETL); Public Health; Supply Chain; Warehouse.

1- Introduction

Health is a state subject in India, with 36 States and Union Territories (UTs) each using independent systems with unique nomenclatures. [1]described India as a "Nation within Nations" due to its vast population. Population of each state is comparable to that of many countries, which makes the implementation of health initiatives really challenging. Effective health improvement requires studying prevalent health systems, disease burden, and risk mitigation.

The Global Burden of Disease (GBD) study highlighted the need for consistent health systems across states [1]. [2] emphasized the importance of state-level health investments and the federal government's role in encouraging such initiatives and increasing public

✉ **Corresponding Author**
abhishekverma@cdac.in

healthcare investment. For effective informed decision-making, standardized data is essential for central agencies like National Health Systems Resource Centre (NHSRC) and the Ministry of Health and Family Welfare (MoHFW). They also require effective visualization on standardized data for analysis and further use in the centralized health system.

India's health system is divided into several major programs, with each state having its own IT-enabled systems to address health issues [3]. Reports indicate a high dependence on central agencies for smooth implementation of health services. (Atreya, 2020) noted that deficiencies in technical strength available with states were highlighted during pandemic, emphasising the role of central government in providing expertise, funds, and policy guidance.

This paper proposes a comprehensive framework for central agencies for Extraction, Transformation, and Loading (ETL) of diverse health data from various sources with a aim to make data usable and ready for analysis at central level along with providing visualization for analysis to them for effective decision making.

The objective is to design and develop a framework for handling large health data for drug and vaccine availability in Indian context. The authors propose utilizing open-source technologies to ensure cost effective and scalable solution for ETL, deduplication, standardization, and visualization, facilitating data analysis for central agencies through Key Performance Indicators (KPIs).

This novel framework involves six components for an end-to-end solution, from remote data extraction to visualization, assisting users in making informed decisions. The implementation uses open-source technologies like J2EE, jQuery, and JavaScript to reduce costs. Each component is detailed in the following sections of this paper

2- Literature Review

[4] describe ETL systems as "resource-intensive, costly, and highly complex," accounting for approximately 70% of the workload in data warehouse development. They highlight the lack of systematic methodologies and supporting tools to meet ETL quality requirements, noting that most implementations rely heavily on developer experience.

[5] and [6]describe ETL (Extract, Transform, Load) as a fundamental process in data warehousing. Extraction involves gathering data from diverse sources; Transformation includes cleaning and restructuring the data for analysis; and Loading refers to inserting the processed data into the target system such as a database, data mart, data lake, or warehouse.

In the ELT approach, data is loaded into the target system before transformation occurs, allowing transformations to leverage the processing capabilities of modern data warehouses. Sivabalan et al. (2021) describe ELT as a process of collecting data from APIs or SQL/NoSQL sources, followed by validation, transformation, and storage—all scheduled systematically [7].

[8] states that ECL-TL method reduces coupling between these stages and enhances adaptability for complex transformation scenarios.

[9] cites various challenges of the ETL process like high velocity data, strive for low latency, security etc and discusses various strategies for addressing these challenges [10] , [11] ETL remains critical for integrating and processing data before analysis. Variants such as ELT (Extract, Load, Transform) and ECL-TL (Extract, Clean, Load – Transform, Load) have emerged to address challenges in traditional ETL pipelines –.

In the healthcare context, Dixit et al. (2019) review the supply chain of drugs and vaccines, emphasizing the importance of IT-based systems to improve efficiency and reliability [12]. Atinga et al. (2019) explore supply chain gaps in Ghana, recommending improvements in human resources, technical infrastructure, and IT-based systems to enhance performance [13].

[14] addressed the scenario of the IFA supplementation program under the Anemia Mukt Bharat (AMB) program, wherein they refer to the problems related to the supply chain of drugs and vaccines and calls for the requirement of supply chain review framework for performance monitoring through measurement of defined Key Performance Indicators

A survey of software used across Indian states and Union Territories revealed varying practices for procurement and distribution of pharmaceuticals [15]. The study aimed to identify a solution that meets regional requirements using available software resources.

Popular ETL tools were reviewed for their applicability in healthcare data integration:

SAS Enterprise Guide (EG) 7.1 is a Windows-based application providing GUI-driven SAS functionality, metadata storage, and automation features. However, it has limitations in consuming web services (SOAP/REST), deduplication, and advanced visualization [16] , [17] , [18].

Microsoft SSIS supports integration from XML, flat files, and RDBMS sources. It offers built-in transformations and a graphical design interface but lacks robust dashboard visualization capabilities [19].

Informatica PowerCenter is noted for its robust ETL features, automated testing, and support for zero-downtime operations, though it comes with premium licensing and optional add-ons [20].

Pentaho Data Integration (PDI), formerly Kettle, is Java-based and supports reporting and OLAP through additional subprojects. Enterprise editions are available [21] , (https://en.wikipedia.org/, 2019).

IBM DataStage is recognized as a leading data integration tool enabling the development and execution of data movement and transformation jobs [23].
Adeptia focuses on reducing integration complexity, offering real-time data access and user-defined transformations for business productivity (Adipati, 2022a, 2022b).
Talend provides ETL solutions for integration, quality control, and big data, with products like Talend Open Studio offering an open-source environment for pipeline management [26] [27].
CloverETL supports automation, scheduling, and integration through a J2EE-compatible platform with SOAP/HTTP APIs for control [28].
Hevo Data emphasizes ease of use, automation, and pricing based on usage. It facilitates integration from various sources into warehouses for analytics readiness (Hevo-Features, 2021).
Oracle Data Integrator is positioned as a comprehensive integration solution that integrates seamlessly with Oracle's enterprise suite [30].
(Khan & Hoque, 2015;2017), [32] and [33] underscore the relevance of ETL tools in healthcare systems –, [10], [34], [35], [36] shares insight about the large resources that are consumed during the ETL processes. Comparative analysis of seven tools across 33 different criteria are provided by (Biplob et al., 2018).

Summary and Research Gap
The literature reveals a consensus on the complexity and centrality of ETL processes in data warehousing, particularly in healthcare applications. Existing tools vary in functionality, integration capabilities, cost, and user experience. Despite the availability of numerous ETL solutions, there is a lack of customizable, open-source platforms tailored for Indian healthcare supply chain management needs.

Proposed Framework
Based on the literature review, this study proposes an open-source, configurable framework designed for aggregating data from all Indian states and Union Territories. The framework supports data cleaning, standardization, centralized storage, and a visualization layer for decision-makers. It aims to bridge the gap between available ETL solutions and the specific needs of regional health departments.

3- Proposed Solution and Methodology

The following components are envisaged by the proposed framework:

Component 1 represents the remote repositories of data where IT-based systems are in place, these can be understood as the respective Indian states and union territories that have their IT-based systems for the supply chain of drugs and vaccines. These states and UTs have their respective rulesets and nomenclature as well that are being used in their respective systems. The respective systems work well to satisfy the needs of respective states / UTs.

Component 2 represents the ETL Layer. It is proposed to be configurable as ETL or ELT or ETLT as per the requirements of the KPIs under consideration.

Component 3 represents the Standardization Layer. This needs to be adaptive to handle the needs of all Indian states and UTs. It can be understood as the mapping process to ensure common nomenclature for drugs/vaccines/health facilities/suppliers etc.

Component 4 represents the Deduplication Layer. The authors propose the ruleset specific to the drugs and vaccine's nomenclature and corresponding weights. Multiple levels of fuzzy logic functions are included in it for predicting the duplicity chances and then presenting the probability of 2 entities being duplicates or not. The final Decision for the removal of such entities is left to the central users, as of now. However automatic removal can also be configured in this framework.

Component 5 represents the Visualization Layer. Support for multiple and configurable options for visualization using libraries from Google Charts, Bootstrap, Hi-charts, etc. is proposed for this component.

Component 6 represents the final users for whom the system will generate the visualizations based on the required Key Performance Indicators (KPIs). This can be understood as the States / Union Territories / Central Agencies like NHSRC, CHI, MoHFW, etc.

The following figure presents the components of the proposed framework.

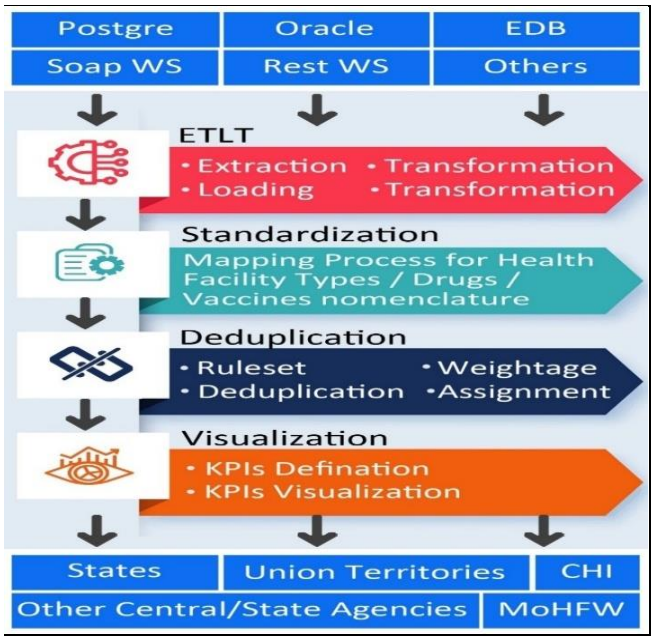


Figure 1: Components of the Proposed Framework.

[37] in his work mentions that data errors increase once the integration of data across the system is sought due to heterogeneity in the data, He outlines the value of clean and quality data and stresses the role of technology in providing quality data. In the health systems, data formats, and nomenclatures vary significantly between states and union territories. The states / UTs have their IT-based solutions for Health Systems for the distribution of drugs and vaccines. A detailed study of such systems used in India was carried out by the authors. The systems have their own databases which store the data of specific states / UTs [15]. Although the systems were similar the data across the systems is heterogeneous both in terms of structure and nomenclature. The problem of heterogeneity of the data is resolved by the pre and post-processing components and the standardization component of the proposed framework. The various components of the proposed framework are described in the following sections.

3.1 ETLT Component:

The authors adopted a configurable approach wherein the ETL can be configured as ETL ELT or ETLT depending on a case-to-case basis. The proposed 5 Layered Approach as follows:

Layer 1: Remote Layer – represents remote data repositories like Postgres, EDB, Oracle, SQL Server, MySQL, etc.

Layer 2: Extraction Layer – represents the extraction part. The extraction can be through REST-based Web Service APIs, SOAP-based Web Service APIs, direct DB connection, Stub-based DB connection, etc.

Layer 3: Transformation Layer –This layer represents pre-processing requirements. This can be an optional layer that is used for pre-processing for cleaning of main repository/table, versioning, DSS creation, or for implementation of misc. logic which may be required to be performed before loading.

Layer 4: Loading Layer – represents loading activity. The loading can be simple loading in which data is simply pushed into the repository as it is fetched.

Layer 5: Transformation Layer – This layer represents post-processing requirements that may include data standardizations, partitioned tables, multi-table inserts, and implementation of misc. logic which may be required to be performed after loading.

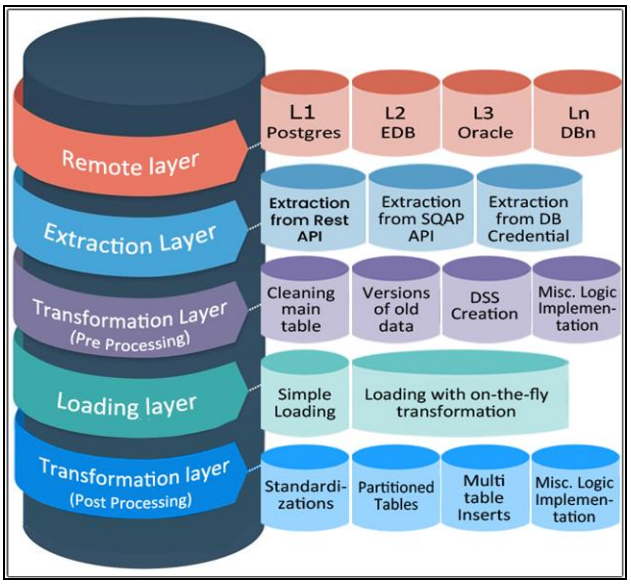


Figure 2: Proposed Architecture for ETL Component.

The Authors designed a utility system based on the designed framework which consists of screens for configuration of remote locations, configuration of jobs on those locations, and provision of one-time, manual, or scheduled execution of created jobs. The Proposed ETL Solution consists of the following steps.

Step 1: Configuration for connection and extraction (State / UT Configuration)

i. When the Destination DB details can be provided to the utility
ii. When the Destination DB details cannot be provided to the utility, Through Web Services (APIs)

The normal flow is presented in the following figure:

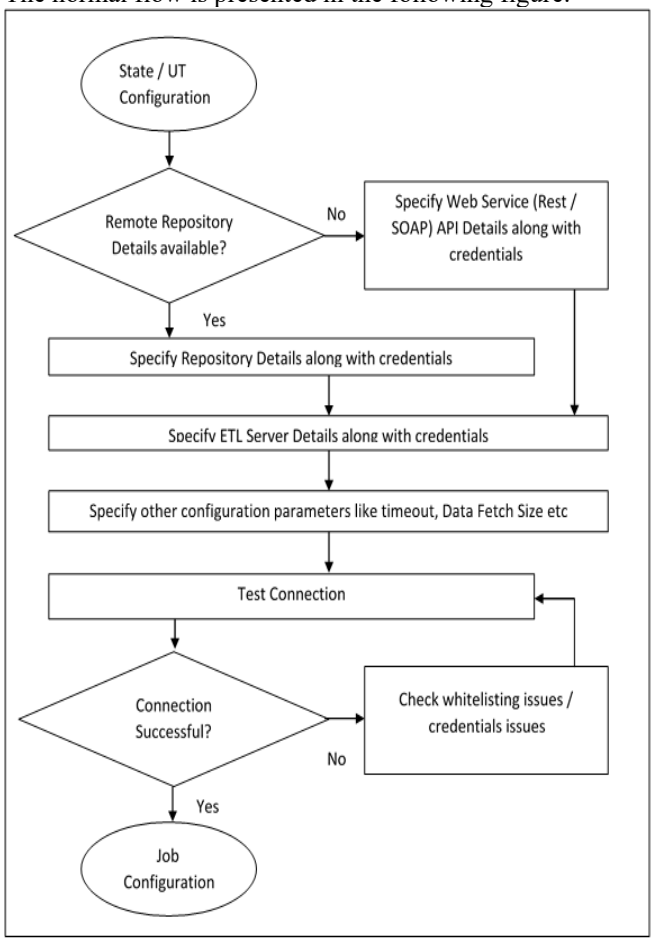


Figure 3: Configuration for connection and extraction (State / UT Configuration)

Step 2: Creation of Jobs for states for the ETL Process which includes.

i. Select Fetch Query [Extraction]
ii. Preprocessing Activity [Transformation] [Optional Transformation BEFORE Loading. It can be used for the creation of history tables / DSS tables or the creation of versions of past data and cleaning of the main table.]
iii. Insert into the Central Warehouse [Loading]
iv. Post Processing Activity [Transformation] [Optional Transformation AFTER Loading. It can be used for standardization via updating of codes from the mapping tables of the centralized Repository. It can also be used for creating partitioned tables or different tables based on some Key parameters like location etc.]

The normal flow is presented through the following figure:

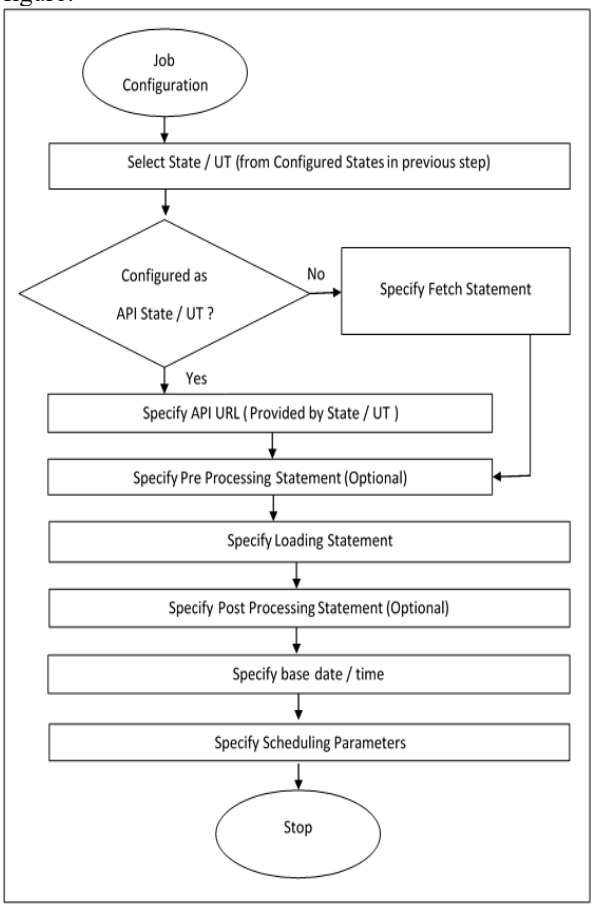


Figure 4: Configuration for ETL Jobs.

The authors found the Extraction Transformation Load Transformation (ETLT) approach most useful.

3.2 Standardization Component:

Different states / UTs have implemented different systems as per their respective needs. There are different masters available for the same object. For instance, the nomenclature for Drugs is different for every state / UT e.g. Paracetamol 500 mg tablet is present with different codes/IDs and different names at respective states. So, the need was to standardize such objects. For standardization, the Authors decided to create a master set at a central location with unique instances of objects from every state. Secondly, a mapping process was envisaged to enable states / UTs to map their drug with the drug created in the central master. The following figure illustrates this concept in greater detail:

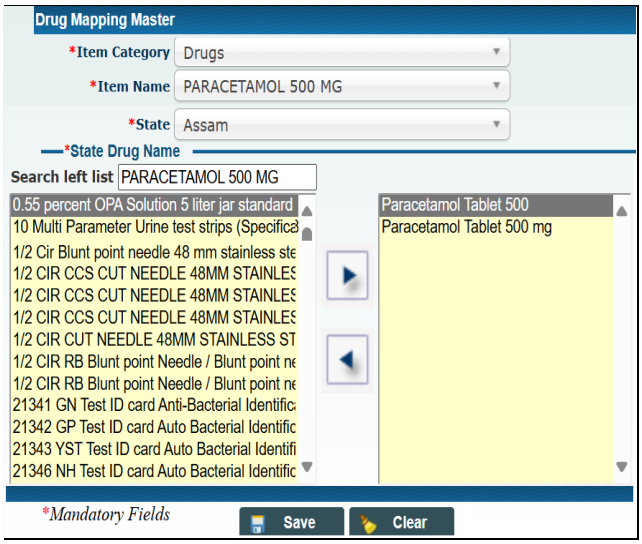


Figure 5: Standardization component.

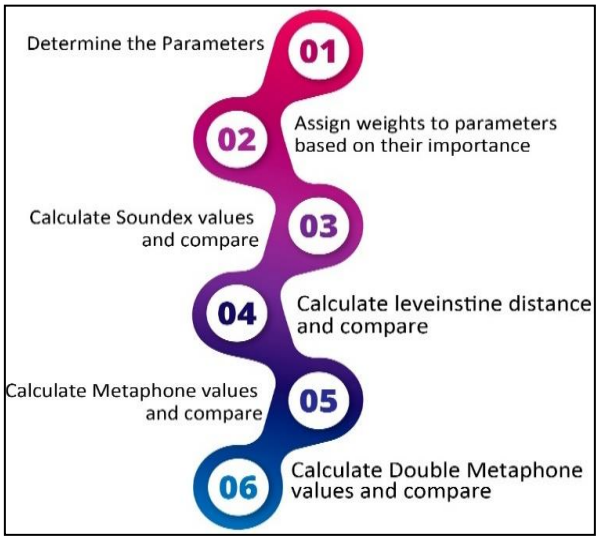


Figure 6: Steps followed for deduplication of drugs.

So here a screen is divided into three parts (Top Pane, Left Pane, and Right Pane). The top Pane represents the drugs from Central Master, Left pane shows the unmapped drugs of the selected state. Now when any state wants to map its drug with a central drug, it selects the drug in the Left pane and moves it to the Right Pane. This way drug nomenclature is standardized for all states / UTs. Similarly, provision for standardization of Health Facilities and Suppliers, etc. is conceptualized in our framework.

3.3 Deduplication Component:

Authors faced the challenge of the absence of exact salt information in the data repositories of state / UTs, authors devised an algorithm for finding duplicate records via probabilities. The following steps are proposed:

3.3.1 Decide upon the parameters for determining the probability of duplicity.

3.3.2 Decide upon the initial weights for parameters.

3.3.3 Storage of two compared data.

3.3.4 Execution of the algorithm to calculate the probable percentage that any 2 drugs are probably duplicated by 'x' percent. It includes implementations of Fuzzy Logic for string comparison.

This calls for gathering the drug data and assigning weights to related parameters which will eventually help in deduplication.

For finding duplicate drugs, weights need to be given to the various parameters based on our experience (initially) like:

a. Parameter - Drug Name – Initial Weight - .75
Drug nomenclature is the systematic naming of drugs, mainly pharmaceutical drugs. The name given by the producing company or based on the active ingredient a drug produces of the brand name drug [38], [39], [40], [41], [42].

b. Parameter - Strength – Initial Weight - .05
Strength is the amount of drug in a given dosage form, for example, 500 mg/tablet [12].

c. Parameter - Group – Initial Weight – .02
Each drug is classified by its presence of active substances which are divided into different like Anaesthetic agents, Adrenocortical steroids, Analgesics Antipyretics Nonsteroidal Anti-inflammatory Medicines & Anti-Rheumatic Drugs, Anti-Depressants, Immunological, Psychotropic Drugs, etc.

d. Parameter - Subgroup – Initial Weight - .01
Each drug is classified by its presence of active substances and divided into groups based on certain pharmacological properties [13].

e. Parameter - Drug_Type – Initial Weight - .02
Based on different types of packaging and the way it can be consumed. For example, injections, inhalators, oral drugs, vials, etc.

f. Parameter - Drug_VED– Initial Weight - 0
Classification of any drug belonging to the Vital, Essential, or Desirable category. Vital drugs are those which are expected to be present in all types of health facilities, they can be termed as “Life-Saving Drugs”. Essential are those that are relatively less important as

Lifesaving but are required to be present for treatment of common illnesses, The Desirable category is for those drugs that are not required to always be in stock, they can be understood as "Good to Have" drugs.

g. Parameter - Drug_CAT – Initial Weight – 0
The Public Health System in India is categorized into three levels.
Primary level (sub-centers and Primary Health Centres (PHCs)),
Secondary level (Community Health Centres (CHCs) and smaller Sub-District hospitals) and
Tertiary level (Medical Colleges and District/General Hospitals).
Classification of any drug belonging to Primary, Secondary, or Tertiary category. Accordingly.

h. Parameter - Drug_Soundex – Initial Weight – .08
Soundex is a phonetic algorithm that helps in indexing names by sound the way they are pronounced in English. This algorithm aims to match homophones despite minor differences in spelling. Also, it encodes consonants and vowels will not be encoded till it is the first letter [43]. The Soundex function is used in this framework to find the matches between two drug names and to address the issue of spelling mistakes. Limitations of Soundex include sensitivity to spelling variations, dependence on the initial letter, noise intolerance (mistyping, extra consonants, swapped consonants), and differing transcription systems. Other potential errors can be the use of initials, particles in names, perceptual differences, and silent consonants.

i. Parameter - Drug_Double_Metaphone – Initial Weight – .02
Metaphone is an improved version of the Soundex algorithm as it uses information about variations and inconsistencies in pronunciation of English spellings to produce a more accurate encoding [44]. The Double Metaphone provides an improved algorithm over the original Metaphone algorithm. [45]. Rudrappa, Agarkhed, and Vaidya (2019) while describing the implementation of Double Metaphone mentions it as a second-generation algorithm, which contains fundamental design improvements [42]. To handle ambiguous cases as well as multiple variants of surnames with common ancestry, it returns two codes for a string i.e. a primary and a secondary code and hence the name double is attached to the algorithm. [46]

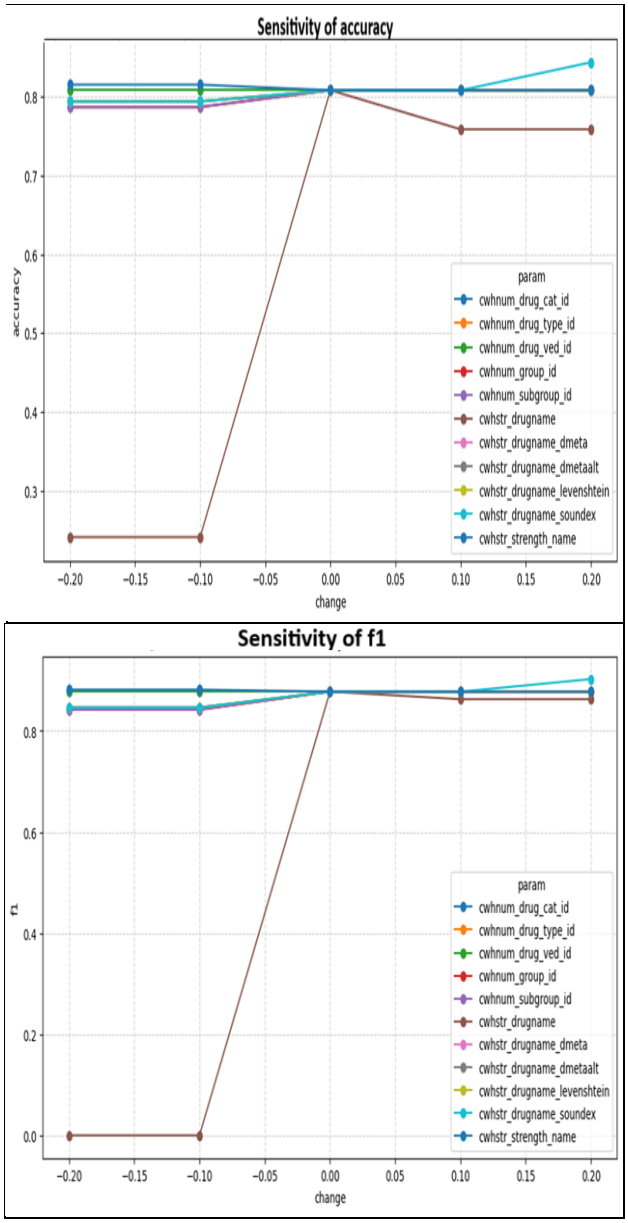
j. Parameter - Drug_Double_Metaphone_Alt – Initial Weight – .02
This is the secondary code for a string in double metaphone.

k. Parameter - Drug_Levenshtein – Initial Weight - .03
The Levenshtein distance is a string metric that measures the difference between two sequences to convert one into the other by the minimum number of single-character edits (insertions, deletions, or substitutions) [47].

The total weights need to be 1 for ease of comparison of results.

Sensitivity Analysis

Sensitivity Analysis of the Initial Weights is as follows:



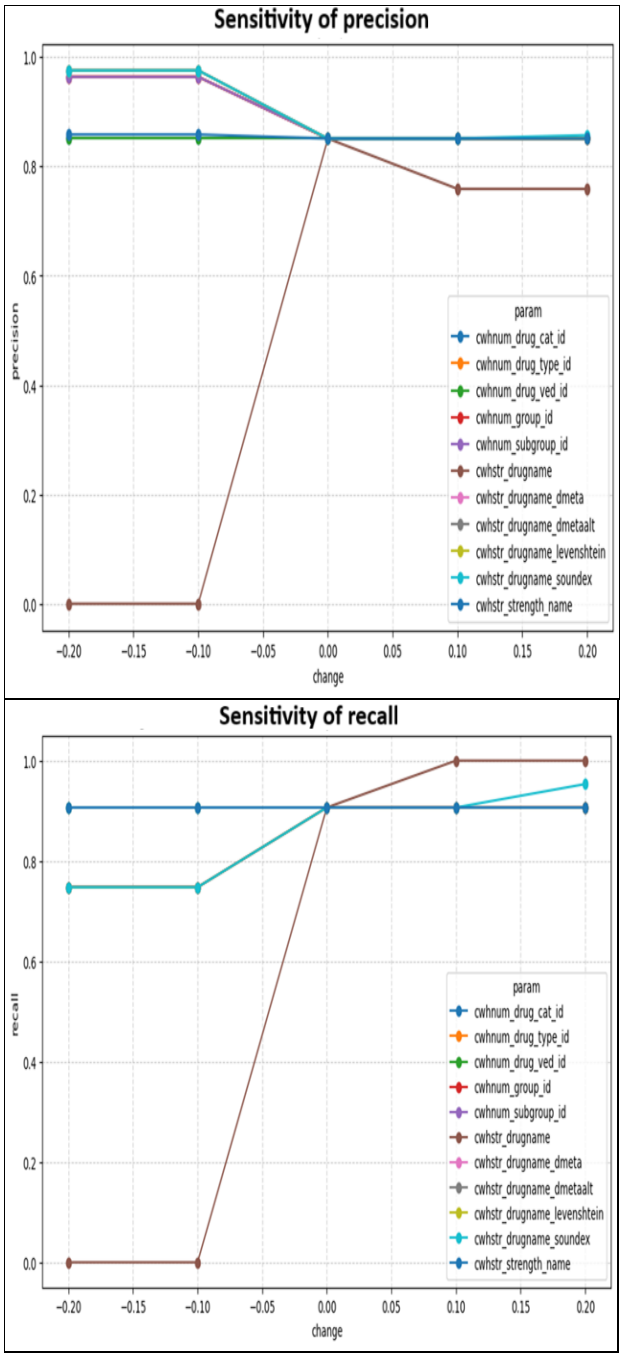


Figure 7: Sensitivity Analysis with original weights.

Summary:

1. Drug name exact match (cwhstr_drugname) dominates.
   Too low = model breaks.
   Too high = recall increases, precision decreases.
   Current baseline looks balanced.
2. Phonetic + Levenshtein features are valuable.
   Increasing their weight slightly improves recall and F1.
   Recommend boosting these to catch spelling variations / typos.
3. group-level features (group_id, subgroup_id, drug_type_id, etc.) are less sensitive.
   They stabilize the model but don't strongly shift performance.
   Safe to keep their weights moderate.

Weights Optimization:

Although the weights provided accurate results, models were tested with multiple values of weights and minor optimizations were done in the weights
After various rounds, optimized weights are as follows:

param_weights_optimized =

```
{
   "cwhnum_group_id": .02,
   "cwhnum_subgroup_id": .01,
   "cwhnum_drug_type_id": .02,
   "cwhnum_drug_cat_id": 0,
   "cwhnum_drug_ved_id": 0,
   "cwhstr_strength_name": .04,
   "cwhstr_drugname": .72, # keep dominant but slightly reduced
   "cwhstr_drugname_soundex": .10,  # boosted
   "cwhstr_drugname_dmeta": .03,    # boosted
   "cwhstr_drugname_dmetaalt": .03, # boosted
   "cwhstr_drugname_levenshtein": .05  # boosted
}
```

Sensitivity analysis shows that although the weights which were chosen initially based on purely experience basis were sufficient, still they have minor scope of improvements wherein minor boosting is recommended in parameters related to fuzzy matching of the stings like Soundex, Metaphone and Levenshtein while the weights for the drug name can be slightly reduced to have better performance of the model.

Sensitivity Analysis shows following improvements:

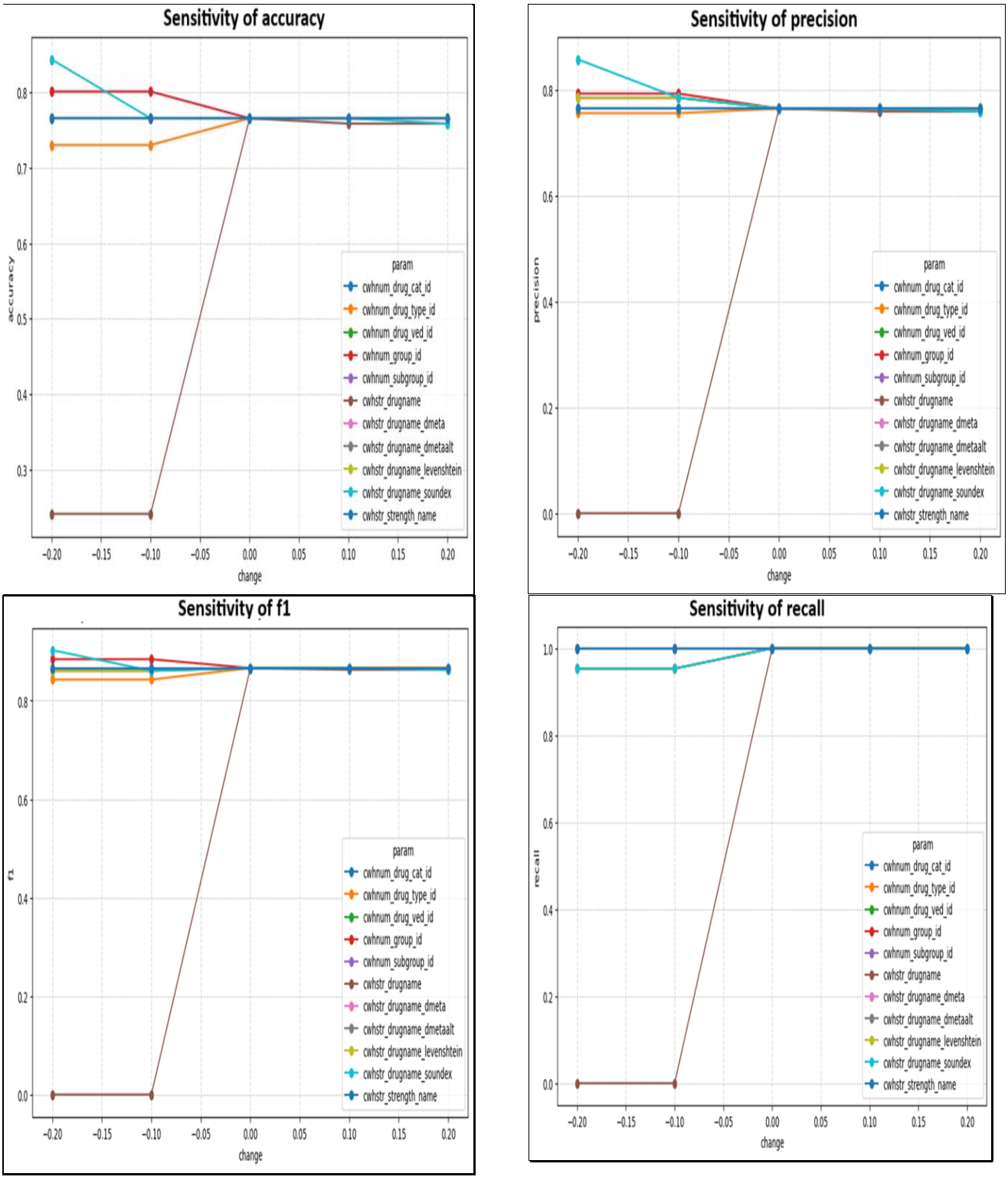


Figure 8: Sensitivity Analysis with optimized weights.

Results of Weights Optimization:

Overall F1 range: ~0.84 → 0.90 (pretty stable, good performance).

Accuracy: ~0.73 → 0.84.

Recall: Mostly 1.0 (catching all true matches), except some cases where it drops to 0.95.

Precision: Varies 0.75 - 0.97 depending on the parameter - main driver of changes in accuracy/F1.

So the model is recall-heavy (it rarely misses matches).

Algorithm: The designed algorithm below briefs out the fundamental steps to be followed which pull out the duplicate records. The first procedure is to be followed recursively till all the drugs are processed and above 90% probability is achieved.

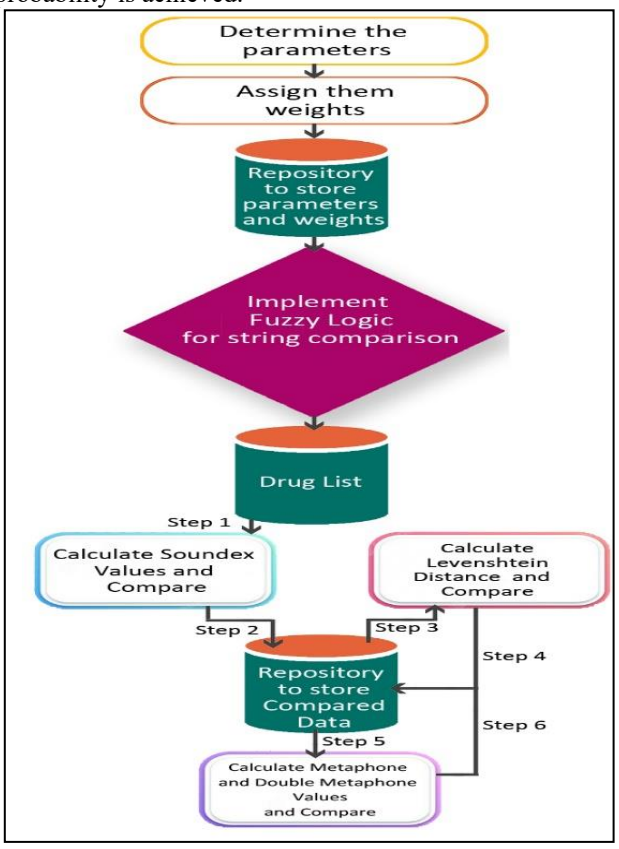


Figure 9: Algorithm for Data Deduplication.

Results: After execution of the above algorithm across every row, the results came into the picture with the following implementation:

```
CREATE OR REPLACE FUNCTION cwht_drug_match(p_drug_id1 numeric, p_drug_id2 numeric)  RETURNS numeric AS $BODY$
declare
v_group_id1 numeric(4,0);v_subgroup_id1 numeric(4,0);v_drug_type_id1 numeric(4,0);v_drugname1 character varying(500);
v_drug_cat_code1 character varying(20);
v_drug_ved_code1 numeric(4,0);v_strength_name1 character varying(200);v_group_id2 numeric(4,0);v_subgroup_id2 numeric(4,0);
v_drug_type_id2 numeric(4,0);
v_drugname2 character varying(500);v_drug_cat_code2 character varying(20);v_drug_ved_code2 numeric(4,0);v_strength_name2
character varying(200);
v_weight NUMERIC(10,2);v_param_name character varying (150);v_param_weight numeric(10,2);

begin
 v_weight :=0;
 select cwhnum_group_id,cwhnum_subgroup_id,cwhnum_drug_type_id, upper(cwhstr_drugname), cwhstr_drug_cat_code,
cwhnum_drug_ved_code, upper(cwhstr_strength_name)
 into v_group_id1,v_subgroup_id1,v_drug_type_id1,v_drugname1,v_drug_cat_code1,v_drug_ved_code1,v_strength_name1
 FROM cwht_drug_mst  where cwhnum_drugid=p_drug_id1;

 select cwhnum_group_id,cwhnum_subgroup_id,cwhnum_drug_type_id, upper(cwhstr_drugname), cwhstr_drug_cat_code,
cwhnum_drug_ved_code, upper(cwhstr_strength_name)
 into v_group_id2,v_subgroup_id2,v_drug_type_id2,v_drugname2,v_drug_cat_code2,v_drug_ved_code2,v_strength_name2
 FROM cwht_drug_mst  where cwhnum_drugid=p_drug_id2;

 if v_group_id1   = v_group_id2      then v_weight := v_weight +
 (select cwhnum_param_weight from dwh.cwht_drug_match_weight_mst where upper(cwhstr_param_name)='CWHNUM_GROUP_ID'); end if;
 if v_subgroup_id1  = v_subgroup_id2   then v_weight := v_weight +
 (select cwhnum_param_weight from dwh.cwht_drug_match_weight_mst where upper(cwhstr_param_name)='CWHNUM_SUBGROUP_ID'); end if;
 if v_drug_type_id1 = v_drug_type_id2  then v_weight := v_weight +
 (select cwhnum_param_weight from dwh.cwht_drug_match_weight_mst where upper(cwhstr_param_name)='CWHNUM_DRUG_TYPE_ID'); end if;
 if v_drug_cat_code1= v_drug_cat_code2 then v_weight := v_weight +
 (select cwhnum_param_weight from dwh.cwht_drug_match_weight_mst where upper(cwhstr_param_name)='CWHNUM_DRUG_CAT_ID'); end if;
 if v_drug_ved_code1= v_drug_ved_code2 then v_weight := v_weight +
 (select cwhnum_param_weight from dwh.cwht_drug_match_weight_mst where upper(cwhstr_param_name)='CWHNUM_DRUG_VED_ID'); end if;
 if v_strength_name1= v_strength_name2 then v_weight := v_weight +
 (select cwhnum_param_weight from dwh.cwht_drug_match_weight_mst where upper(cwhstr_param_name)='CWHSTR_STRENGTH_NAME'); end if;
 if v_drugname1   = v_drugname2    then v_weight := v_weight +
 (select cwhnum_param_weight from dwh.cwht_drug_match_weight_mst where upper(cwhstr_param_name)='CWHSTR_DRUGNAME'); end if;

 return v_weight;
end $BODY$  LANGUAGE edbsql VOLATILE  COST 100;
```

Figure 10: Implementation of Algorithm for Data Deduplication.

Output received from the deduplication process where the drug ID at Left-Hand-Side (LHS) is compared with the drug ID at Right-Hand-Side (RHS). The value is the probability that the percentage of the drug at LHS is a duplicate of the drug at RHS.

Table 1: Result of Algorithm for Data Deduplication.

Count i	Count j	DrugId LHS	DrugId RHS	Value
1	1	21191000	21191001	0.1
1	2	21191000	21191002	0.05
1	3	21191000	21191003	0.03
1	4	21191000	21191004	0.03
…	…	…	…	…

Processing these raw results, the following illustrations came into consideration that around 1.5% of duplicate records are found with a 90% possibility of duplicity.

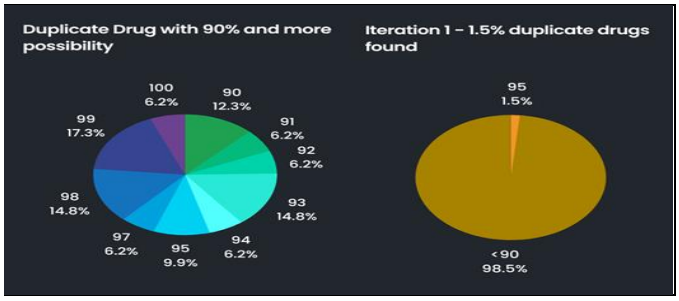


Figure 11: Distribution of duplicate records in Iteration-1

After removing these records, above-mentioned procedures are executed on the remaining rows which lead to new values. Iteration continued till the achievement of 90% accuracy was discontinued resulting in the following graph:

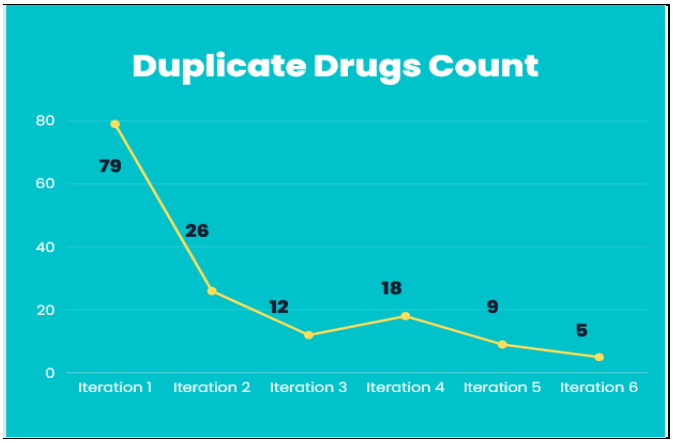


Figure 12: Count of duplicate records across total records.

Hence in total around 149 duplicate drug entries were detected in 7 iterations which says around 4% of total drugs were duplicates. Stating that under one system, entries of 4% duplication are found then this percentage will be much higher in the case of big data and their analysis leading to much higher business issues.

3.4 Visualization Component:

The visualization component takes the input of the data that is stored by the ETL Utility, standardized and de-duplicated. The visualization component creates dashboards with charts/graphs / tabular data etc. through Key Performance Indicators (KPIs). This framework proposes different layers of Visualization Components i.e.: Widget Layer, Tab Layer, Dashboard Layer. These three layers can be implemented through three different processes i.e. Widget Master, Tab Master, and Dashboard Master.

Widget Master is the basic unit that contains the API or the Query or the Procedure call to the repository to get data for visualization. Customizations like Font size, color, background size, color, etc. are also handled at this level. The widget can be of type Tabular, Graph, KPI, Map, News Ticker, Another link, or Iframe. Option for Caching the data is also required at this level along with Refresh time. Various visualization libraries like Google Charts, HI Charts, etc. must be available here for selection. Tab Master is the collection of widgets that are required to be displayed on the page. It must provide the location on which a particular widget is to be displayed along with the order in which it must be displayed to the user. Dashboard Master is the collection of tabs that the dashboard must contain. It presents the tabs as pages along with their menus. Al Zoubi, S., Gharaibeh, L., Jaber, H. M, et al. (2021) pointed out the panic behavior at the time of COVID-19 which resulted in shortages of sanitizers and other related medicines. This framework provides a way to early detect such sudden increases in stockouts through visualizations so that decision-makers can take a call to control the situation [48].

Present Deployment of the integrated framework used open-source Java, React and Python with the following technologies:

a) GIT for code synchronization and version control
b) Jenkins for continuous integration and deployment
c) Nginx server for internal routings
d) WildFly Application Server and
e) Postgre Database Server

4- Result:

The proposed data integration and analytics framework have been successfully deployed at a centralized data center, effectively serving as a backbone for healthcare supply chain management operations. The deployment integrates a total of thirty-one geographically distributed remote data sources, each of which transmits data continuously into the centralized system.

During the system evaluation phase, the framework demonstrated substantial scalability and performance. It consistently managed over 5 GB of data ingestion per remote source per day, culminating in a cumulative daily data volume exceeding 155 GB. This high throughput underscores the system's ability to manage large-scale, real-world datasets with minimal latency and data loss.

Data Ingestion Strategy

Data ingestion operations are tailored according to predefined business rules and domain-specific Key Performance Indicators (KPIs). Based on the analytical requirements:

Incremental loading is used for time-sensitive data marts where only recent changes are captured, ensuring efficiency.

Full data refreshes are scheduled for scenarios requiring complete synchronization, particularly in rapidly evolving datasets such as vaccine inventories and drug distributions.

This dual-mode ingestion strategy balances performance with completeness, adapting dynamically to evolving data needs.

Key Advantages of the Proposed Framework

The framework introduces several innovations that collectively address the limitations of traditional ETL tools and commercial platforms, particularly in low-resource or high-scale public health environments.

Cost-Effectiveness

The entire solution stack is based on open-source

technologies, eliminating licensing fees and reducing total cost of ownership. This makes it accessible and scalable.

High-Customizability

The proposed solution can be configured in multiple ways which can coexists with each other e.g. for one state ETL can be used and for other ETLT can be used, it makes solution very convenient to use as per specific data load of a particular state.

Flexible Deployment Configuration

The system supports configurable deployments for transformations e.g. transformations can be performed during load time itself or can be deferred to after completion of load activity.

Advanced API Integration Capabilities with Deduplication

To support both legacy and present-day systems, solution provides data intake through variety of methods like through REST API, SOAP API or if network and credentials are available then through direct DB connections along with deduplication algorithm unlike commercial platforms such as SAS Enterprise Guide,

In-Memory Data Transformation and Standardization

To address ETL latency, solution facilitate on-the-fly data cleaning, standardization, and structuring.

Integrated Deduplication Mechanism

The proposed solution uses user-configurable parameters, weights, rules and fuzzy matching algorithms to address data redundancy. Solution provides probable duplicate data sets as per the defined weight on parameters and related rules.

Low-Code/No-Code Visualization Layer

The solution consists of an integrated visualization module for various herarchy of users to build and customize interactive dashboards using predefined Key Performance Indicators (KPIs). Software coding or development knowledge is not required to create the KPIs in the dashboards.

Rapid Adaptability and Iterative Enhancement

The modular design and scriptable components allow for quick customization, versioning, and deployment of updates. This shortens turnaround times for new feature implementation compared to traditional monolithic ETL systems.

Role-Based Access Control (RBAC)

Controlled access to system functionalities and data visualizations ensures data security and compliance.

Comprehensive Data Security

Security of any system is an important aspect to consider. The framework adopts layered encryption combining RSA (asymmetric) and AES (symmetric) encryption algorithms. Sensitive health data is encrypted during storage and transmission. OWASP guidelines have been adhered to for ensuring that top ten vulnerabilities are properly addressed. Only users with valid roles and credentials can access the information.

The following figure shows the decrease in stockout percentages in case of particular state stores. It can be seen that the average stockout percentages for essential drugs defined by states have been reduced by more than 10%

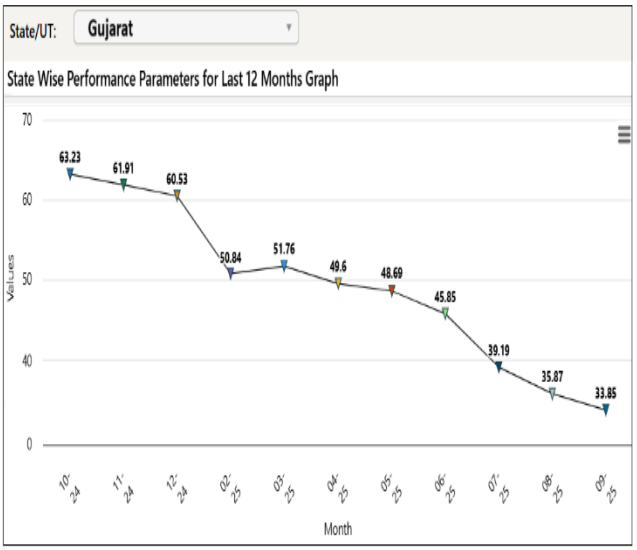


Figure 13: Reduction in stockout percentages in District Hospitals of an Indian State.

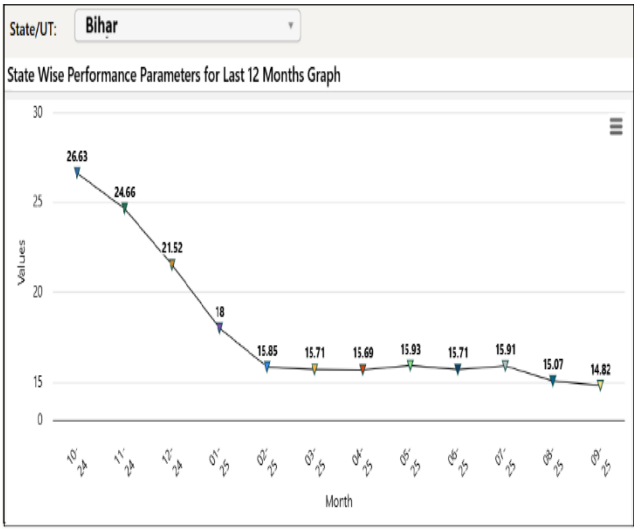


Figure 14: Reduction in stockout percentages in District Hospitals of another Indian State

Below figure presents the average reduction of stockout percent across particular stores of Indian states and UTs. It is evident that the stockout % is reduced from approx. 67.16 % to approx. 60.69%.

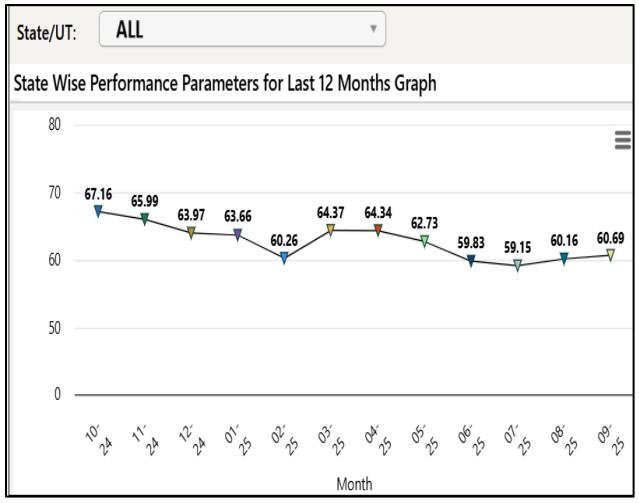


Figure 15: Reduction in stockout percentages in District Hospitals across integrated Indian State and UTs

Overall, the results confirm that the proposed framework is not only technically viable but also scalable, secure, and aligned with the operational demands. Use of opensource technologies, combined with advanced integration and transformation capabilities, makes it a compelling alternative to commercial ETL platforms, especially in scenarios where availability of resources is critical.

5- Conclusion:

The proposed framework marks a significant step in creating a centralized, scalable, and budget-friendly solution for integrating and analyzing data in India's public healthcare supply chain. The system offers comprehensive functionality from setting up remote data sources and scheduling ETL processes to real-time data transformation, smart deduplication, and dashboard-based analytics to improve drug availability across nation.

The architecture uses open-source technologies like Java 2 Platform, Enterprise Edition (J2EE) for the frontend and PostgreSQL for managing data on the backend. A key strength of this framework is its ability to standardize and clean data in real time, even when dealing with a wide variety of data sources. This ensures consistency and reliability, even at scale. The dynamic ETL pipelines are particularly flexible, allowing for quick adjustments to transformation rules as data environments or policy needs shift. This includes automatic deduplication when needed. Meanwhile, built-in analytics and visualization tools offer user-specific dashboards that help stakeholders monitor performance and gain operational insights.

Currently, the system has been deployed across 31 Indian states and union territories, processing more than 5 GB of data per state each day on average. This level of adoption highlights how scalable and reliable the framework is. Health authorities at both central and state levels use it to view KPIs, identify delays or issues, and measure the efficiency of supply chains across different states and UTs. These insights have led to faster response times, better transparency, and more informed decision-making in public health logistics and improved drug availability.

Beyond performance, the framework also scores high on adaptability and security. Its modular setup means new states / UTs or data sources can be added with minimal hassle, making it a viable long-term solution.

Ultimately, the results show that a smart, unified platform like this can support data-driven governance at both national and state levels. This framework offers a model that could be replicated in other sectors aiming for digital transformation.

Sample implementation of ETL component based on the proposed framework:

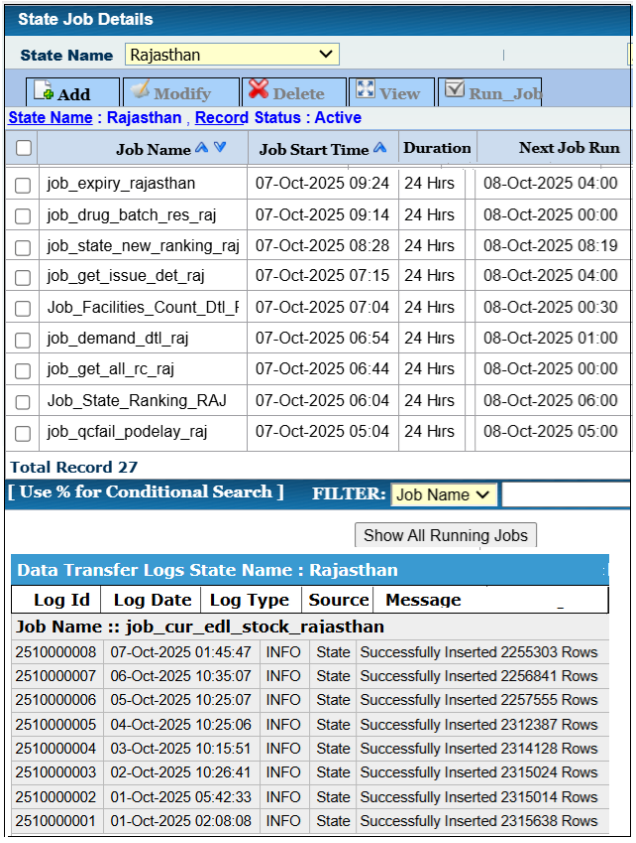


Figure 16: Sample implementation of ETL Component.

Sample implementation of the Visualization Component based on the Proposed Framework:

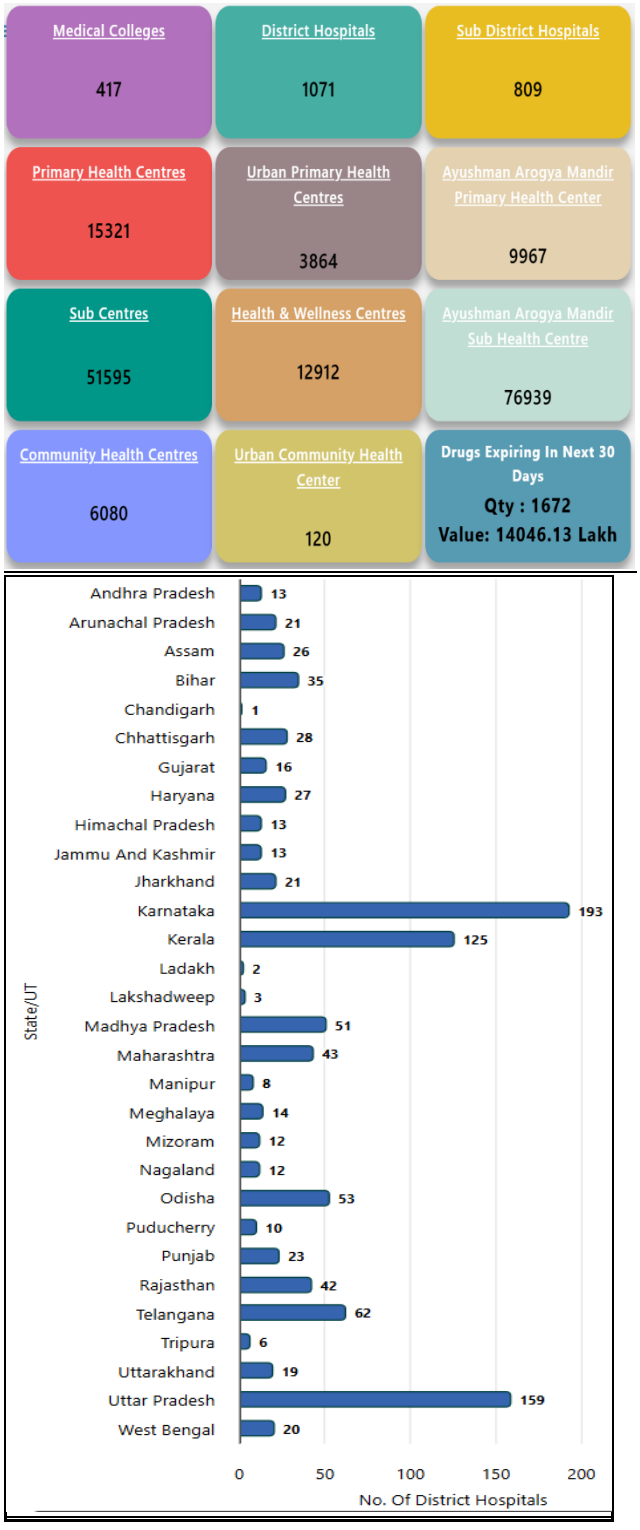


Figure 17: Sample implementation of Visualization Component.

Limitation: The proposed framework covers all aspects from ETL to visualization. Authors think that as the data size increases, horizontal/vertical scaling may be required according to the size of the data. Further, in the Visualization component, various types of charts/graphs can be added to provide more visualization options.

Future Work: The authors are planning to include more features in the visualization component like automatic outliers' detection, based on the selection of mathematical functions like Average and standard Deviation, Automatic visualization of data into the quartiles, automatic identification of candidates for mapping, and deduplication through machine learning techniques. Further enhancement of the visualization component by inclusion of predictive analysis is also planned for future work.

Conflict of interest:

The authors have no conflicts of interest regarding this investigation**.**

Acknowledgments:

The authors would like to thank Ministry of Health and Family Welfare and NHSRC for their kind support during formulation and implementation of the proposed framework.

References

[1] L. Dandona *et al.*, "Nations within a nation: variations in epidemiological transition across the states of India, 1990–2016 in the Global Burden of Disease Study," *The Lancet*, vol. 390, no. 10111, pp. 2437–2460, Dec. 2017, doi: 10.1016/S0140-6736(17)32804-0.

[2] The Lancet, "India—a tale of one country, but stories of many states," The Lancet, vol. 390, no. 10111, p. 2413, Dec. 2017, doi: 10.1016/S0140-6736(17)32867-2.

[3] MoHFW, "MINISTRY OF HEALTH AND FAMILY WELFARE Major Schemes and Programmes Government of India New Delhi," 2000. [Online]. Available: http://mohfw.nic.in

[4] S. Saebao, S. Matayong, and N. trakulmaykee, "QoX based ETL Design for Business Intelligence System of Lecturers' Qualifications Analysis," in 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), IEEE, Jun. 2020, pp. 539–542. doi: 10.1109/ECTI-CON49241.2020.9158113.

[5] S. Zhao, "What is ETL? (Extract, Transform, Load)," www.edq.com

[6] T. Pott and lain Thomson, "Extract, transform, load? More like extremely tough to load, amirite?," www.theregister.com.

[7] S. Sivabalan and R. I. Minu, "Heterogeneous Data Integration with ELT and Analytical MPP Database for Data Analysis Application," in 2021 Innovations in Power and

Advanced Computing Technologies (i-PACT), IEEE, Nov. 2021, pp. 1–5. doi: 10.1109/i-PACT52855.2021.9696841.
[8] B. Pan, G. Zhang, and X. Qin, "Design and realization of an ETL method in business intelligence project," in 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), IEEE, Apr. 2018, pp. 275–279. doi: 10.1109/ICCCBDA.2018.8386526
[9] Shiva Kumar Vuppala and Manohar Reddy Sokkula, "Challenges in Streaming ETL Pipelines for High-Frequency Data Ingestion and Real-Time Processing," International Journal For Multidisciplinary Research, vol. 6, no. 6, Dec. 2024, doi: 10.36948/ijfmr.2024.v06i06.33506
[10] A. Kabiri, F. Wadjinny, and D. Chiadmi, "Towards a Framework for Conceptual Modeling of ETL Processes," 2011, pp. 146–160. doi: 10.1007/978-3-642-27337-7_14.
[11] T. Jörg and S. Dessloch, "Near Real-Time Data Warehousing Using State-of-the-Art ETL Tools," 2010, pp. 100–117. doi: 10.1007/978-3-642-14559-9_7.
[12] A. Dixit, S. Routroy, and S. K. Dubey, "A systematic literature review of healthcare supply chain and implications of future research," Int. J. Pharm. Healthc. Mark., vol. 13, no. 4, pp. 405–435, Nov. 2019, doi: 10.1108/IJPHM-05-2018-0028.
[13] R. A. Atinga, S. Dery, S. P. Katongole, and M. Aikins, "Capacity for optimal performance of healthcare supply chain functions: competency, structural and resource gaps in the Northern Region of Ghana," *J. Health Organ. Manag.*, vol. 34, no. 8, pp. 899–914, Oct. 2020, doi: 10.1108/JHOM-09-2019-0283.
[14] K. Ahmad *et al.*, "Public health supply chain for iron and folic acid supplementation in India: Status, bottlenecks and an agenda for corrective action under Anemia Mukt Bharat strategy," *PLoS One*, vol. 18, no. 2, p. e0279827, Feb. 2023, doi: 10.1371/journal.pone.0279827.
[15] A. Verma, A. Rana, H. Monga, A. Chaudhary, and J. Singh, "Distribution Management of Drugs/medicines and vaccines vis-a-vis Free Drugs Service Initiative (FDSI) of Ministry of Health and Family Welfare (MoHFW), Government of India in the Indian States," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, IEEE, Sep. 2021, pp. 1–5. doi: 10.1109/ICRITO51393.2021.9596365.
[16] SAS, "SAS Enterprise Guide," https://support.sas.com/documentation/onlinedoc/guide/tut8/en/m0_2.htm. Accessed: Jun. 18, 2025. [Online]. Available: https://support.sas.com/documentation/onlinedoc/guide/tut8/en/m0_2.htm
[17] E. Kodhai, K. Divakar, A. Andrews, Y. Pachipala, and J. Alzubi, "MANAGING THE CLOUD STORAGE USING DE-DUPLICATION AND SECURED FUZZY KEYWORD SEARCH FOR MULTIPLE DATA OWNERS," *International Journal of Pure and Applied Mathematics*, 2018.
[18] R. Kumar, J. Lachure, and R. Doriya, "Use of Hybrid ECC to enhance Security and Privacy with Data Deduplication," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, Aug. 2021, pp. 934–941. doi: 10.1109/ICESC51422.2021.9532948
[19] SQL Server, "SQL Server Integration Services," https://learn.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver15. Accessed: Jun. 18, 2025. [Online]. Available: https://learn.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver15
[20] Informatica, "Informatica.com," https://www.informatica.com/de/products/data-integration/powercenter.html. Accessed: Jun. 18, 2025. [Online]. Available: https://www.informatica.com/de/products/data-integration/powercenter.html
[21] Pedersen T and Mohania M, *Data Warehousing and Knowledge Discovery*, vol. 5691. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. doi: 10.1007/978-3-642-03730-6.
[22] https://en.wikipedia.org/, "Pentaho," https://en.wikipedia.org/wiki/Pentaho. Accessed: Jun. 18, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Pentaho
[23] Ibm, "Datastage," https://www.ibm.com/products/datastage. Accessed: Jun. 18, 2025. [Online]. Available: https://www.ibm.com/products/datastage
[24] www.adeptia.com, "www.adeptia.com," https://www.adeptia.com/products/connect. Accessed: Jun. 18, 2025. [Online]. Available: https://www.adeptia.com/products/connect
[25] www.adeptia.com, "www.adeptia.com," https://www.adeptia.com/solutions. Accessed: Jun. 18, 2025. [Online]. Available: https://www.adeptia.com/solutions
[26] tutorialspoint.com, "Talend," https://www.tutorialspoint.com/talend/index.htm. Accessed: Jun. 18, 2025. [Online]. Available: https://www.tutorialspoint.com/talend/index.htm
[27] talend.com, "TalendSolutions," https://www.talend.com/products/talend-open-studio/. Accessed: Jun. 18, 2025. [Online]. Available: https://www.talend.com/products/talend-open-studio/
[28] Clover ETL Reference Manual," https://docs.huihoo.com/cloveretl/server/CloverETLServer-ReferenceManual-4_7_0_016M1.pdf. Accessed: Jun. 18, 2025. [Online]. Available: https://docs.huihoo.com/cloveretl/server/CloverETLServer-ReferenceManual-4_7_0_016M1.pdf
[29] Hevodata, "Hevo-features," https://docs.hevodata.com/introduction/hevo-features/. Accessed: Jun. 18, 2025. [Online]. Available: https://docs.hevodata.com/introduction/hevo-features/
[30] "Data-integrator," https://www.oracle.com/in/middleware/technologies/data-integrator.html. Accessed: Jun. 18, 2025. [Online]. Available: https://www.oracle.com/in/middleware/technologies/data-integrator.html
[31] S. I. Khan and A. S. Md. L. Hoque, "Towards development of health Data Warehouse: Bangladesh perspective," in *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, IEEE, May 2015, pp. 1–6. doi: 10.1109/ICEEICT.2015.7307514.
[32] Md. B. Biplob, G. A. Sheraji, and Khan Shahidul Islam, *2018 International Conference on Innovations in Science, Engineering and Technology : ICISET 2018 : International*

Islamic University Chittagong, Chittagong, Bangladesh : 27-28 October 2018. IEEE, 2018. doi: 10.1109/iciset.2018.8745574.

[33] A. R. Chaturvedi, A. K. Choubey, and Jinsheng Roan, "Scheduling the allocation of data fragments in a distributed database environment: a machine learning approach," *IEEE Trans. Eng. Manag.*, vol. 41, no. 2, pp. 194–207, May 1994, doi: 10.1109/17.293386

[34] A. Simitisis, P. Vassiliadis, S. Skiadopoulos, and T. Sellis, Data Warehouses and OLAP. IGI Global, 2007. doi: 10.4018/978-1-59904-364-7.

[35] P. Vassiliadis and A. Simitsis, "EXTRACTION, TRANSFORMATION, AND LOADING," 2009.

[36] C. Joe and K. Ralph, The data warehouse etl toolkit : practical techniques for extracting, cleaning, conforming, and delivering data. Wiley, Hoboken, N.J., 2013, 2011.

[37] W. W. Eckerson and R. Sponsors, "Achieving Business Success through a Commitment to High Quality Data TDWI REPORT SERIES DATA QUALITY AND THE BOTTOM LINE." [Online]. Available: www.dw-institute.com

[38] Drug_nomenclature," https://en.wikipedia.org/wiki/Drug_nomenclature. Accessed: Jun. 18, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Drug_nomenclature

[39] First Nations Health Authority, "Generic vs Brand-name prescription drugs faq," https://www.fnha.ca/about/news-and-events/news/generic-vs-brand-name-prescription-drugs-faq. Accessed: Jun. 18, 2025. [Online]. Available: https://www.fnha.ca/about/news-and-events/news/generic-vs-brand-name-prescription-drugs-faq

[40] "Understanding expressions of drug amounts," http://courses.washington.edu/pharm309/calculations/Lesson2.pdf. Accessed: Jun. 18, 2025. [Online]. Available: Understanding expressions of drug amounts

[41] www.who.int, "ATC Classification," https://www.who.int/tools/atc-ddd-toolkit/atc-classification. Accessed: Jun. 18, 2025. [Online]. Available: https://www.who.int/tools/atc-ddd-toolkit/atc-classification

[42] S. Rudrappa, D. V. Agarkhed, and S. S. Vaidya, "Healthcare Systems: India," in *Quality Spine Care*, Cham: Springer International Publishing, 2019, pp. 211–224. doi: 10.1007/978-3-319-97990-8_13.

[43] Soundex," https://en.wikipedia.org/wiki/Soundex. Accessed: Jun. 18, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Soundex

[44] Metaphone," https://en.wikipedia.org/wiki/Metaphone. Accessed: Jun. 18, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Metaphone

[45] "Double_Metaphone," https://en.wikipedia.org/wiki/Metaphone#Double_Metaphone. Accessed: Jun. 18, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Metaphone#Double_Metaphone

[46] Lawrence Philips, "The double metaphone search algorithm," *C/C++ Users Journal*, vol. 18, no. 6, pp. 38–43, 2000.

[47] Levenshtein_distance," https://en.wikipedia.org/wiki/Levenshtein_distance. Accessed: Jun. 18, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Levenshtein_distance

[48] S. Al Zoubi, L. Gharaibeh, H. M. Jaber, and Z. Al-Zoubi, "Household Drug Stockpiling and Panic Buying of Drugs During the COVID-19 Pandemic: A Study From Jordan," *Front. Pharmacol.*, vol. 12, Dec. 2021, doi: 10.3389/fphar.2021.813405.