Paper PP25

UNLOCKING THE POTENTIAL OF UNSTRUCTURED DATA IN REAL-WORLD EVIDENCE

Puri Maria, **Pfizer Inc.,** Athens, Greece
Kountouris Nikolaos, **Pfizer Inc.,** Athens, Greece
Panagiotakis Georgios, **Pfizer Inc.,** Athens, Greece

ABSTRACT

The utilization of unstructured data in Real-World Evidence (RWE) has emerged as a promising avenue for extracting deeper insights into patient health, diagnosis, and treatment outcomes. [1]. While structured data from Electronic Medical Records (EMR) and claims provide a foundation for analysis, the richness of unstructured, free text data remains largely untapped. By developing in house capabilities such as Natural Language Processing (NLP) techniques, researchers can extract meaningful information from diverse textual sources such as clinical notes, free-text reports or physician narratives, and patient-reported outcomes. [2] [3] In this poster, we aim to delve into the potential of unstructured data and its pivotal role in augmenting patient classification for regulatory submissions through validation studies as well as the current methodological trends in utilizing unstructured text data collected in real world and its future usage in driving evidence-based decision-making and advancing patient care.

1. INTRODUCTION

There has been a relatively long historical precedent for using Real World Data (RWD) for safety purposes, while the Food and Drug Administration (FDA) has more recently established a framework for the use of RWD for generating evidence of effectiveness. The FDA defines Real-World Evidence (RWE) as "clinical evidence concerning the use and potential benefits or risks of a medical product, obtained from the analysis of Real-World Data." [4] The objectives of the new opportunity are to improve the quality and/or use of RWD, promote better understanding of RWE study designs and develop specific tools to evaluate aspects of RWE generation. [5] Sources of RWD include electronic health records (EHRs), claims activities, product or/and disease registries and patient-reported outcomes. Utilizing RWD, experts and healthcare providers can acquire insights into treatment trends and outcomes, as well as safety profiles which can aid in making well informed regulatory judgments and implementing personalized medical strategies. Additionally, healthcare policies are leveraging advanced methodologies like Natural Language Processing (NLP) and machine learning techniques to analyze unstructured data for extracting critical information. For instance, NLP can help to identify specific conditions, symptoms, and treatment responses from clinical notes, which can then be used to refine patient stratification and treatment plans. [6]

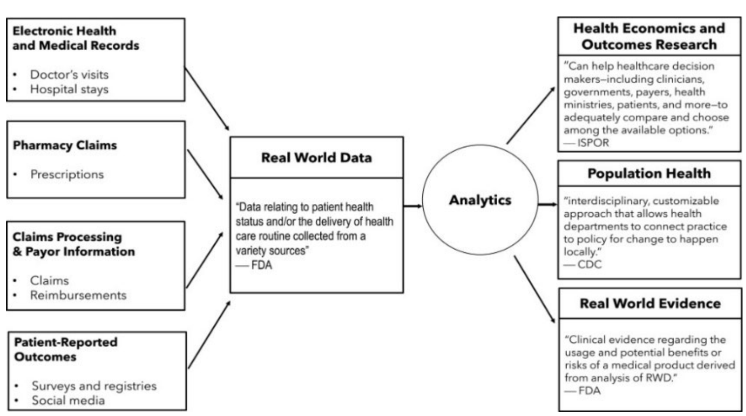

Figure 1 : Real World Data to Real World Evidence [2]

2. DATA SOURCES

RWE research relies on sources such as, health records (EHRs) insurance claims and billing records lab results, hospital information, registries of products and diseases, patient generated information. Analysis of RWE plays a role in evaluating the effectiveness, efficiency, and safety of products or electronic devices.[7]

Below is a brief description of each data source, categorized into structured and unstructured formats:

Structured Data Sources

1. Electronic Health Records (EHRs): These records include clinical notes, medications, diagnoses, and treatment plans, which are part of patient encounters, physician observations or progress reports offering a longitudinal view of patient care across different settings.
2. Insurance Claims and Billing Records: These records capture detailed information on healthcare services rendered, costs, and payer information, allowing researchers to assess healthcare utilization patterns, costs, and treatment pathways.
3. Laboratory Results: Lab data includes critical biomarkers and diagnostic test outcomes, helping to assess disease progression, treatment response, and adherence to clinical guidelines.
4. Hospital Information Systems: These systems capture inpatient and outpatient encounter data, providing insights into procedures, hospitalizations, and care delivery within acute and chronic care settings.
5. Product and Disease Registries: Registries collect standardized data on specific diseases or treatments, enabling large-scale tracking of outcomes, disease prevalence, and the real-world performance of medical products.

Unstructured Data Sources

1. Patient Generated Data: Data from wearable devices, mobile apps, and home-use settings provides real-time insights into a patient's daily health status, behaviors, and adherence to treatments outside of clinical environments. Examples include monitoring heart rate, blood pressure, body temperature, and habits such as smoking or drinking.

2. Unstructured clinical data: Free-text clinical notes, radiology and pathology reports, and patient- generated text, which capture detailed but unformatted information about patient conditions, treatments, and observations. This data is often inconsistent and lacks standardization, making it challenging to analyze without advanced techniques.

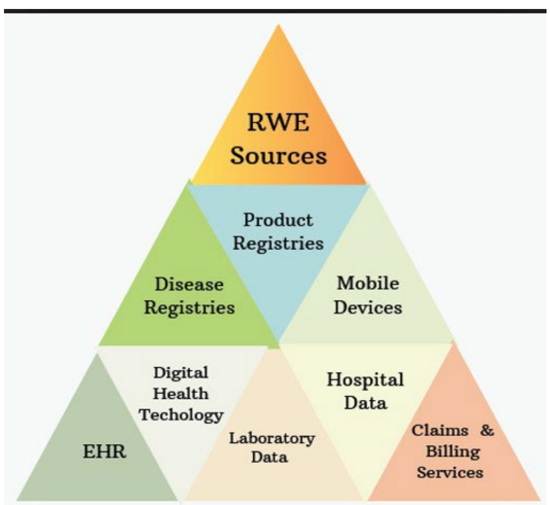

Figure 2 : Real World Evidence Sources

3. MACHINE LEARNING TECHNIQUES FOR UNSTRUCTURED HEALTHCARE DATA

Machine Learning Algorithms: Machine learning in healthcare is broadly categorized into super- vised and unsupervised learning, each serving different purposes based on the type of data and the task at hand:

a. **Supervised Learning:** This approach is used when labeled datasets are available, meaning that the model is trained on data where the outcomes are known. It is commonly applied in healthcare for tasks such as disease diagnosis, predicting treatment outcomes, and patient risk stratification. For instance, supervised learning can be used to classify whether a patient is likely to respond to a certain therapy based on past data. [8]

b. **Unsupervised Learning:** This is applied when working with unlabeled data, helping to discover hidden patterns within large datasets. It is especially valuable in healthcare for exploratory tasks such as clustering patients into subgroups based on similar characteristics or detecting anomalies in patient populations. Unsupervised learning is often employed when dealing with unstructured data, such as clinical notes, where predefined labels may not exist. [8]

The key difference between these two types lies in the availability of labeled data: supervised learning requires labeled data for training, while unsupervised learning does not. Supervised learning excels in prediction and classification tasks where the relationships are known, while unsupervised learning is more suited to exploratory analysis, uncovering patterns in unstructured data without predefined outcomes. [8] [9]

In healthcare, integrating both approaches with Natural Language Processing (NLP) allows for robust analysis of structured and unstructured data, enabling deeper insights into patient care and outcomes. However, the richest insights and highest data volume are often found in unstructured data.

The following tables outline the main machine learning algorithms commonly used in healthcare data analysis, including both supervised and unsupervised learning algorithms.

Algorithm	Description	Formula	Variable Explanations
Linear Regression	Predicts the relationship between a dependent variable and one or more independent variables. It assumes a straight-line relationship and is used for continuous outcomes.	$y = \beta_0 + \beta_1 x + \epsilon$	y: Dependent variable, x: Independent variable, β_0: Y-intercept, β_1: Slope, ϵ: Error term
Logistic Regression	Predicts the probability of a binary outcome using the logistic function, modeling the relationship between a dependent binary variable and independent variables.	$P(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$	$P(y=1)$: Probability of y being 1, x: Independent variable, β_0, β_1: Coefficients
Classification Trees	Predicts categorical variables using decision trees, which start with a root node, pass through internal nodes, and end with a leaf node.	Gini Impurity: $G = 1 - \sum_{i=1}^{n} p_i^2$	G: Gini Impurity, p_i: Probability of class i, n: Number of classes
Naïve Bayes	Uses Bayes' theorem to predict the likelihood of an event, assuming variables are independent. Often applied for text classification in healthcare.	$P(A\|B) = \frac{P(B\|A)P(A)}{P(B)}$	$P(A\|B)$: Probability of A given B, $P(B\|A)$: Probability of B given A, $P(A), P(B)$: Probabilities of A and B
K-Nearest Neighbors (K-NN)	Predicts outcomes based on the closest data points in the dataset (neighbors) using distance metrics.	Euclidean distance: $d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$	d: Distance between two points, x_i, y_i: Coordinates of points, n: Number of dimensions

Figure 3: Supervised Learning Algorithms [8]

Algorithm	Description	Formula	Variable Explanations
Apriori Algorithm	Identifies association rules in datasets to find frequently co-occurring items (e.g., diagnoses or treatments).	Support: $supp(X) = \frac{\lvert t \in T; X \subseteq t \rvert}{\lvert T \rvert}$	$supp(X)$: Support of itemset X, T: Number of transactions, t: Transaction containing X
K-means Clustering	Clusters data points into K groups based on similarity, calculating centroids for each cluster.	$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \lvert\lvert x_i^{(j)} - c_j \rvert\rvert^2$	J: Objective function, k: Number of clusters, n: Number of data points, $x_i^{(j)}$: i-th point in j-th cluster, c_j: Centroid
Principal Component Analysis (PCA)	Reduces the dimensionality of data by identifying the principal components that capture the most variance.	Covariance: $Cov(X) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$	$Cov(X)$: Covariance matrix, n: Number of observations, x_i: i-th observation, $\bar{x}$: Mean of observations

Figure 4: Unsupervised Learning Algorithms [8]

NLP techniques help convert unstructured text into structured data that can be analyzed using machine learning algorithms. By processing and interpreting human language, NLP enables the extraction of valuable insights from vast amounts of textual data, which would otherwise be challenging to analyze using traditional methods. [10]
Among the various artificial intelligence capabilities, the main relevant and highly applicable to the generation of RWE are shown below:

Natural Language Processing (NLP):

a. Text Extraction / Classification:
   i. Automatically extracting relevant information (e.g. symptoms) from unstructured text sources, such as patient records and clinical notes to identify common patterns.

b. Named Entity Recognition (NER):
   i. Identifies and classifies entities within text, such as diseases, medications, and patient attributes to track their prevalence.
   ii. Supports proactive healthcare by forecasting disease progression, treatment responses, and potential complications.

Bio BERT for Enhanced NLP:

Bio BERT (Bidirectional Encoder Representations from Transformers) for Biomedical Text Mining, is highly specialized for biomedical data. Trained on large datasets like PubMed and PMC, Bio BERT enhances performance in tasks such as Named Entity Recognition and relation extraction. It offers more accurate identification of medical terms, symptoms, diseases, and drug-related entities, making it especially useful for extracting clinical insights from unstructured data. [11]

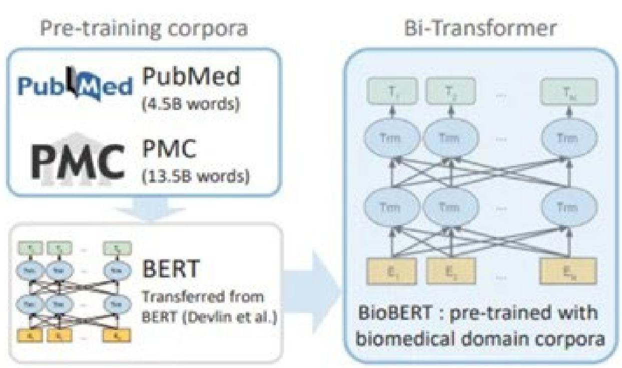

Figure 5: Bio-Bert Architecture [12]

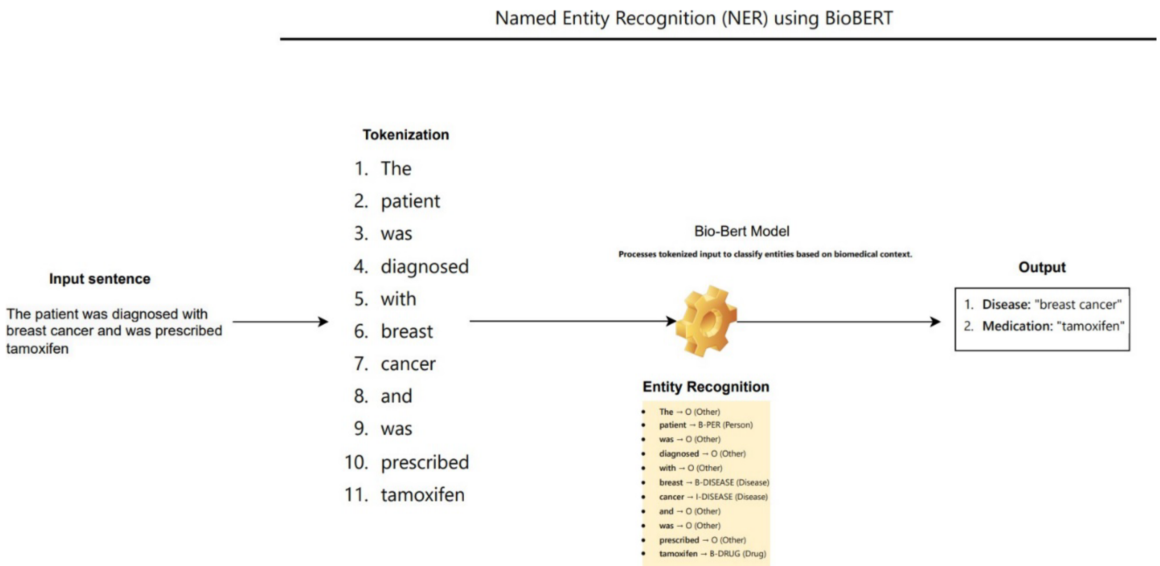

Figure 6: Named Entity Recognition (NER) using Bio BERT

4. BENEFITS WITH REAL-WORLD EVIDENCE AND AI FOR REGULATORY DECISION

Artificial Intelligence (AI) and Real-World Evidence (RWE) play a transformative role in enhancing regulatory decision-making by streamlining the analysis of vast and complex datasets. These tech- nologies enable regulatory agencies to use large amounts of data—from electronic health records (EHRs), insurance claims, and patient registries to data generated by wearable devices and mobile applications—to provide a comprehensive picture of treatment outcomes in real-world settings.

AI may detect complex patterns and relationships that traditional analytical tools may miss. This skill enables regulators to more accurately assess the safety and efficacy of medicines, simplifying the discovery of possible hazards and advantages associated with novel therapies. Furthermore, the integration of RWE enables ongoing monitoring of treatment performance after approval, ensuring that safety signals are noticed quickly, and regulatory measures are performed as needed.

Incorporating AI and RWE into regulatory frameworks not only expedites the review and approval process but also fosters a more adaptive regulatory environment. This adaptability is crucial in responding to emerging health challenges and ensuring that patients have timely access to innovative therapies that meet their needs. Moreover, the insights derived from RWE can inform clinical guidelines and policy decisions, ensuring that healthcare practices are aligned with the latest evidence on treatment effectiveness.

Ultimately, the synergy between AI and RWE enhances transparency and trust in regulatory processes, as decisions are increasingly based on robust, real-world data rather than solely on controlled clinical trials. This shift toward data-driven regulatory practices not only improves patient outcomes by facilitating faster access to effective treatments but also contributes to a more efficient healthcare system that prioritizes patient safety and therapeutic efficacy. [12]

Benefit	Description
Improved Patient Outcomes	To promote customized treatment regimens, regulators can better understand how medications function in a variety of patient populations outside of controlled clinical trials when RWE and AI are coupled.
Faster Approvals	AI helps rapidly analyze large amounts of real-world data, identifying patterns and potential risks much faster than traditional methods. This can lead to quicker decision-making, allowing beneficial treatments to reach patients sooner while still ensuring they are safe and effective.
Enhanced Decision Making	The integration of RWE provides regulators with a more comprehensive understanding of the efficacy and safety of a treatment. It incorporates data from various real-world sources such as insurance claims and electronic health records. AI assists by quickly processing this massive volume of data, spotting patterns and insights that might not be apparent otherwise, enabling data-driven decisions.
Broader Population Coverage	RWE utilizes data from real-world sources such as electronic health records (EHRs), insurance claims, and patient registries, covering a broader range of people with varied demographics and health conditions. This helps researchers and regulators understand how treatments affect different populations, leading to more inclusive assessments of a drug's effectiveness and safety.
Transparency	"Transparency" in AI and RWE refers to making regulatory choices more understandable by ensuring they are supported by clear, data-driven evidence. Regulators employing RWE and AI consider real-world data, such as patient outcomes, making the regulatory process more open and based on observable patient experiences, thus increasing public trust.
Post-Market Surveillance	With AI, post-market surveillance becomes more efficient, as it can quickly analyze large amounts of real-world data to detect emerging safety issues, such as unexpected side effects. This continuous monitoring helps ensure that risks or problems are identified and addressed timely, maintaining patient safety.

Figure 7: Benefits of AI and RWE in Regulatory Decision-Making

5. CASE STUDY

I. Patient Stratification Using Unstructured Clinical Text Data and LLMs

Unstructured clinical text data, such as physician notes, laboratory reports, and narratives by patients, are rich in granular information that often hides in structured data. This will shed light on subtle differences in the presentation of diseases, variations in the responses to treatment, and side effects. Using NLP and Deep Learning, more specifically Large Language Models, this stratification process would be further enunciated to execute the precise and personalized care of patients.

A new research examines how large language models (LLMs) might extract significant information from unstructured electronic health records (EHRs) or clinical text-free data for patient stratification. There has been usage of models which are using over 90 billion words, achieves cutting-edge performance in clinical NLP tasks like as idea and connection extraction. This development helps healthcare practitioners to better grasp patient narratives and personalize treatment strategies, hence improving patient outcomes. [13]

II. Case study's core theoretical framework

The core theoretical framework guiding this case study is based on integrating patient stratification models with unstructured clinical text analysis. Leveraging Large Language Models (LLMs) and advanced Natural Language Processing (NLP) techniques, this framework focuses on extracting valuable insights from clinical narratives. By categorizing patients into subgroups based on clinical factors such as commodities, symptoms, and treatment responses, it aims to enhance personalized care.

The following outlines the key steps that illustrate the process flow and interactions within the framework presented above.

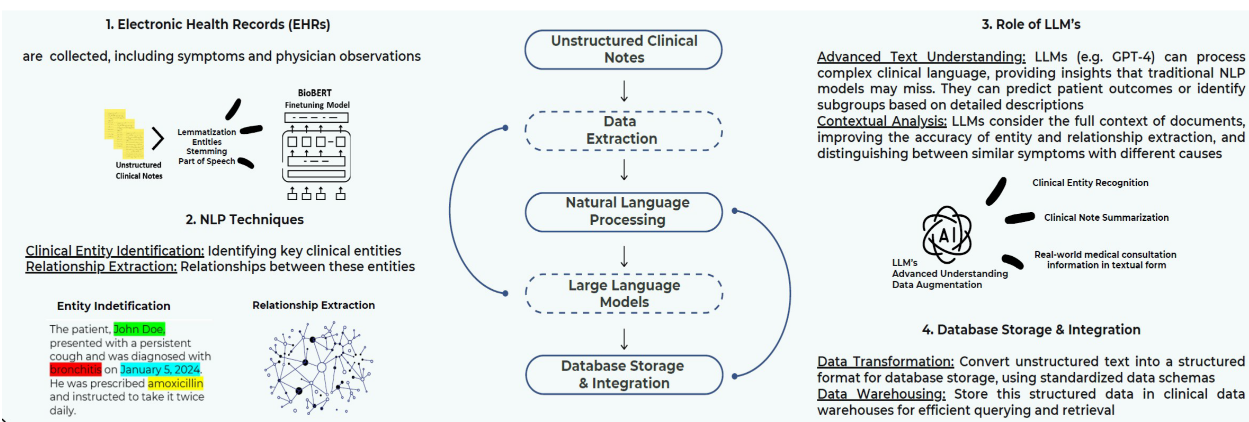

Figure 8: Sequential steps involved in leveraging unstructured data for patient stratification

1. Data Extraction Sources Unstructured data is sourced from Electronic Health Records (EHRs) and patient self-reports. This data encompasses a variety of information, including symptoms, patient history, physician observations, and treatment regimens.

2. NLP Techniques
    a. Entity Recognition: This involves identifying and classifying key clinical entities such as diseases, symptoms, and medications. For instance, recognizing terms like “diabetes” and “insulin resistance” within clinical notes.
    b. Relationship Extraction: This technique seeks to understand the relationships between iden- tified entities, such as noting that “the patient exhibits symptoms of diabetes due to insulin resistance.”

3. Role of LLMs
    a. Advanced Text Understanding: Large Language Models (LLMs) like GPT-4 excel in process- ing complex clinical language, providing insights that traditional NLP models might overlook. They can hypothesize about patient subgroups or predict outcomes based on detailed clinical descriptions.
    b. Contextual Analysis: LLMs can evaluate context throughout entire documents, which enhances the precision of entity and relationship extraction. They are adept at distinguishing between similar symptoms that arise from different underlying causes based on contextual clues.
4. Database Storage and Integration
    a. Structured Data Transformation: Unstructured insights are converted into a structured for- mat suitable for database storage. This transformation involves developing standardized data schemas that include essential fields such as patient ID, phenotype, genotype, and treatment response.
    b. Data Warehousing: The structured data is then stored in clinical data warehouses, facilitating efficient querying and retrieval. This structured information can be integrated with other clinical data for comprehensive analysis.

6. CONCLUSION AND FUTURE WORK

Integrating unstructured data into Real-World Evidence (RWE) efforts is a crucial advancement in utilizing healthcare data for clinical decision-making. While this integration offers substantial benefits, it also presents complexities due to the various forms of unstructured data, including clinical notes and patient feedback. To maximize its potential, we need robust methodologies, such as Large Language Models (LLMs), to effectively extract insights. As future work, collaborative efforts among data scientists, healthcare providers, and regulatory experts are essential for developing tools that are both effective and ethically sound. Additionally, protecting patient privacy and ensuring data security is critical, as this data often contains sensitive information. Overall, successfully integrating unstructured data into RWE can enhance healthcare quality and outcomes, making it imperative to prioritize collaboration and ethical considerations in this effort.

7. REFERENCES

[1] Zou, K.H. et al. (2020) 'Harnessing Real-World Data for Regulatory Use and Applying Innovative Applications', Journal of Multidisciplinary Healthcare, Volume 13, pp. 671–679A' ' [Online] (Accessed: 15 June 2024). Available: https://www. tandfonline.com/doi/full/10.2147/JMDH.S262776#abstract/

[2] Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning,' '[online] (Accessed: 29 June 2024) Available: https://pubmed.ncbi.nlm.nih.gov/22276193/

[3] FDA's Historical Use of "Real World Evidence" - Food and Drug Law Institute (FDLI),' ' [Online] (Accessed: 13 August 2024).Available: https://www.fdli.org/2018/08/update-fdas-historical-use-of-real-world-evidence/

[4] FDA to Support Potential Use of Real-World Data to Generate Real-World Evidence for Regulatory Decision-Making (biopharminternational.com),'' [Online] (Accessed: 21 August 2024). Available: https://www.biopharminternational.com/view/ fda-to-support-potential-use-of-real-world-data-to-generate-real-world-evidence-for-regulatory-decision-making

[5] Garvin, E. (2019) The state of EHR data in the midst of the real-world evidence evolution, Healthcare IT News,'' (Accessed: 21 August 2024) [Online]. Available: https://hitconsultant.net/2019/02/27/state-of-ehr-data-real-world-evidence-rwe/

[6] Schad, F. and Thronicke, A. (2022) Real-world evidence-current developments and perspectives, International journal of environmental research and public health.,'' [Online] (Accessed: 21 August 2024). Available: https://www.ncbi.nlm.nih.gov/ pmc/articles/PMC9408280/

[7] How Healthcare NLP Taps Unstructured Data's Potential (2019) Health Catalyst." [Online] (Accessed: 21 August 2024). Available At: https://www.healthcatalyst.com/insights/how-healthcare-nlp-taps-unstructured-datas-potential.

[8] Mustafa, A. and Rahimi Azghadi, M. (2021) 'Automated Machine Learning for Healthcare and Clinical Notes Analysis', Computers, 10(2), p. 24.'' [Online] (Accessed: 15 October 2024): Available at: https://doi.org/10.3390/computers10020024.

[9] Bansal, S. (2019) Supervised and Unsupervised learning, GeeksforGeeks.'' [Online] (Accessed: 18 October 2024) Available at: https://www.geeksforgeeks.org/supervised-unsupervised-learning/.

[10] Zou, K.H. et al. (2020) Harnessing real-world data for regulatory use and applying innovative applications, Journal of multidisciplinary healthcare,'' [Online] (Accessed: 26 September 2024). Available: https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC7383026/

[11] Lee, J. et al. (2019) BioBERT: A pre-trained biomedical language representation model for biomedical text mining, OUP Academic.,''[Online] (Accessed: 26 September 2024). Available: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506

[12] May 25 (1970) Ai, Big Data and real world evidence – the challenges and opportunities, pharmaphorum. Available at: https://pharmaphorum.com/views-and-analysis/ai-big-data-and-real-world-evidence-the-challenges-and- opportunities,'' [Online] (Accessed: 26 September 2024). Available: https://pharmaphorum.com/views-and-analysis/ ai-big-data-and-real-world-evidence-the-challenges-and-opportunities

[13] (PDF) Gatortron: A large clinical language model to unlock patient information from Unstructured Electronic Health Records, '[Online] (Accessed: 06 October 2024). Available: https://arxiv.org/abs/2203.03540

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Puri Maria
Email: (Maria.Puri@pfizer.com)

Co-Author Name: Kountouris Nikolaos
Email: (Nikolaos.Kountouris@pfizer.com)

Co-Author Name: Panagiotakis Georgios
Email: (Georgios.Panagiotakis@pfizer.com)