

Effective NLP for unstructured text to power downstream analysis and models in Pharma

*Natural language processing:
Structured data from unstructured real world
data streams*

Jane Z Reed, PhD
Director Life Sciences, Linguamatics, an IQVIA company



Overview

- + Disruptive challenges in the Pharma industry
- + Value of technology for unstructured text
 - Natural Language Processing has come of age
 - NLP for real world data, real world evidence
- + Case studies
 - Use cases across the enterprise, from real world data streams
- + IQVIA NLP Technology
 - NLP ecosystem for broad range of users
 - Value of NLP within data science
- + Summary

Natural Language Processing has come of age, with unstructured text forming a critical component in the pharma data ecosystem

Access to unstructured text unlocks value for business users, data scientists

Why is it so important now?

Highly relevant for today's rich data streams

- More unstructured data is generated than ever, across different touchpoints with Covid accelerating this further
- Emphasis on capturing natural, conversational 'Voice of the Patient' and 'Voice of the Customer'
- Health practitioners cannot capture the complexity and subtlety of patient information within structured forms

A powerful, accepted technology

- Reluctance to capture and use unstructured data due to compliance and other concerns is dissipating or being addressed with the right processes
- Acceptance and even expectation growing to use free text data sources within R&D, real world and commercial

Broader understanding of AI/ML Technology

- The data needed for effective Machine Learning modelling needs to be comprehensive, well labelled and for some critical use cases curated.
- Better understanding of where different types of AI/ML best applied
- Less focus on using innovation budgets for small AI/ML pilots and more towards sustainable value add ML pipelines for the longer terms

Natural Language Processing has come of age, with unstructured text forming a critical component in the pharma data eco-system

NLP normalizes unstructured text for broad range of key applications and use cases

What is NLP?

- Natural Language Processing (NLP) is a family of computational techniques for extracting and normalizing structured data from unstructured sources
- Combination of machine learning and rules-based methodologies
- A critical AI technology for data scientists in 2022 and beyond

Key Applications for Messy Data

- Search, compare, extract
- Topic and emerging topic surfacing
- Data cleaning and normalization for accurate analysis and insights
- Feature engineering for machine learning models
- Identification of relationships between concepts e.g. graph analytics



NLP is a critical component for modern, data-driven Pharma and Healthcare organisations, unlocking the 80% of biomedical information in data ecosystems that is unstructured, for use in analysis, insights generation and data science workflows

Setting the Scene: NLP for real world data sources

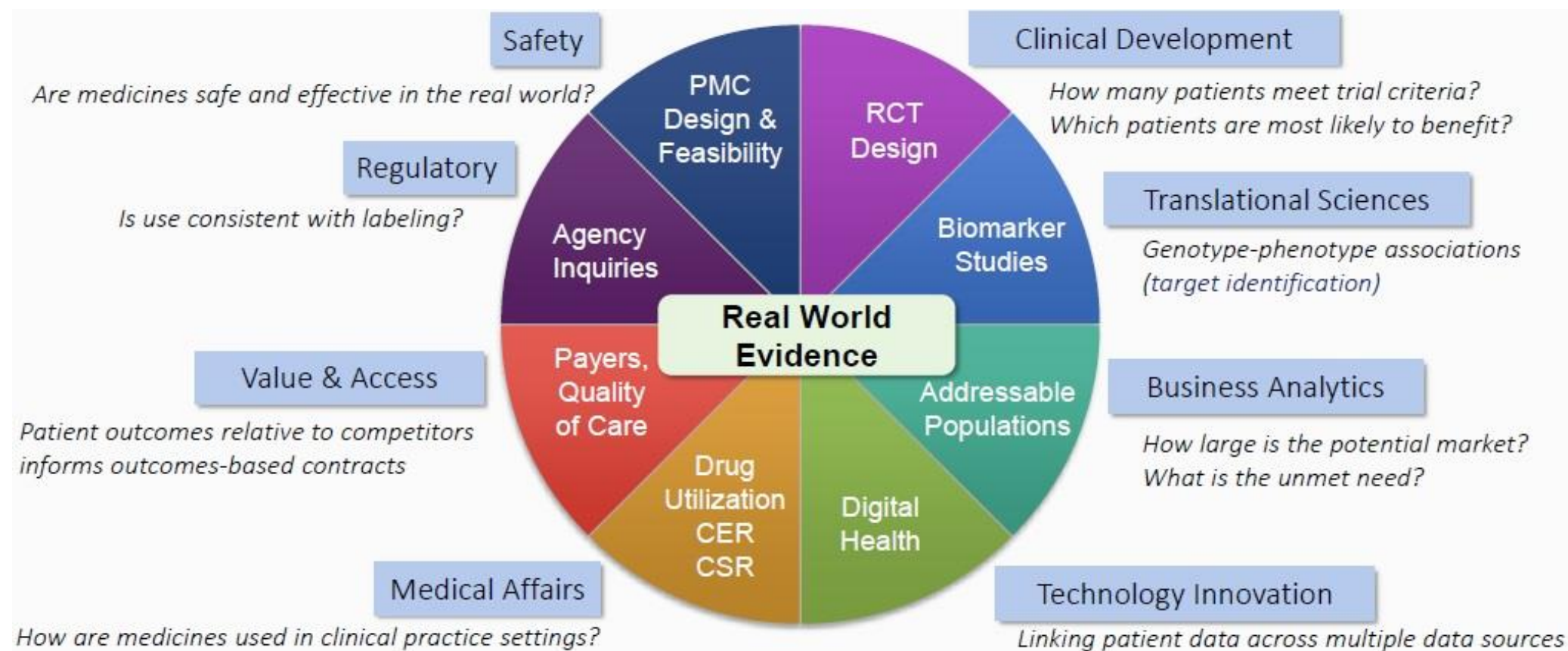
Real World Data, Real World Evidence

- Real world evidence can inform all phases of drug development and commercialization
- Real-world evidence (RWE) is derived from curating and analyzing real-world data (RWD)
- FDA definition: Real-world data are the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources*



*<https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>

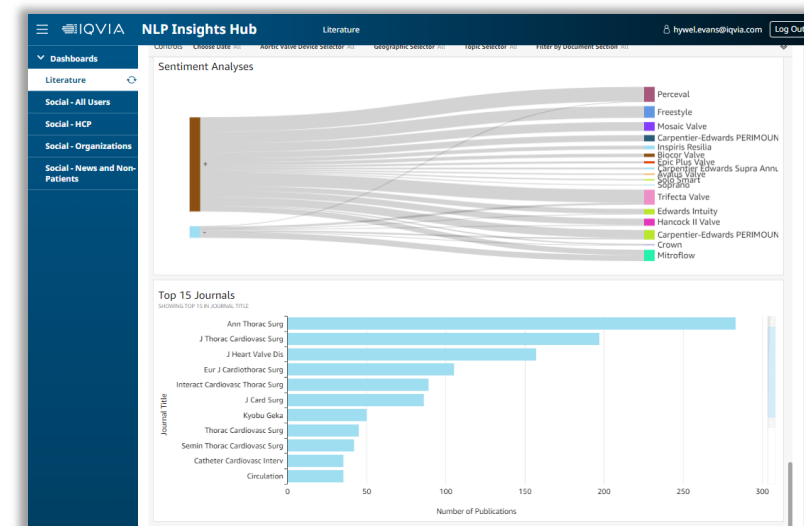
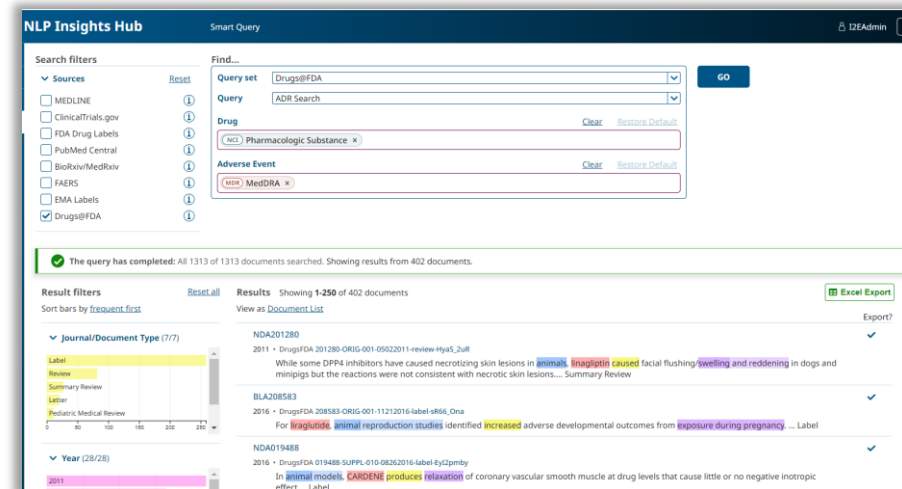
Real World Evidence brings broad business value



NLP for real world data sources

NLP enables more effective access to Real World Data sources

- There are **a multitude of sources** available externally that can be used to answer specific questions, track trends or performance, scan the market and power research
- Sources that can provide real world data, like MEDLINE, social media, medical records, may be familiar but **effectively mining them is a challenge** and Natural Language Processing can help with
- IQVIA tackle the problem of **unlocking the value in large scale, unstructured or text based sources** by building robust extraction and indexing processes and applying NLP to them
- Experts such as **Data Scientists** can be provided with access to data via our NLP platform allowing them to develop Python and other pipelines taking advantage of the speed and power of our NLP
- **Business and scientist users** can access these sources using our interactive search and dashboarding tools to more effectively find and extract what they need



NLP case studies in real world

Value of IQVIA NLP: Coding of topics, concepts, metrics and questions areas

Example concepts identified and categorised with IQVIA NLP

Burden of Disease
▪ Apply disease definitions ▪ Incidence ▪ Prevalence ▪ Mortality ▪ Hospitalisation

Patient Populations
▪ Epidemiology of target sub-groups ▪ Patient population profiling including factors such as age, socio-demographics

Patient Pathways
▪ Treatment pathways ▪ Disease pathways & progression

Standard of Care
▪ Standard of care for relevant subgroups ▪ Variation in care provision

Unmet Needs
▪ Underserved sub-populations

Treatment Costs
▪ Costs and cost variation ▪ Value for money

Economics
▪ Resource use and cost ▪ Economic impact ▪ Cost of standard of care

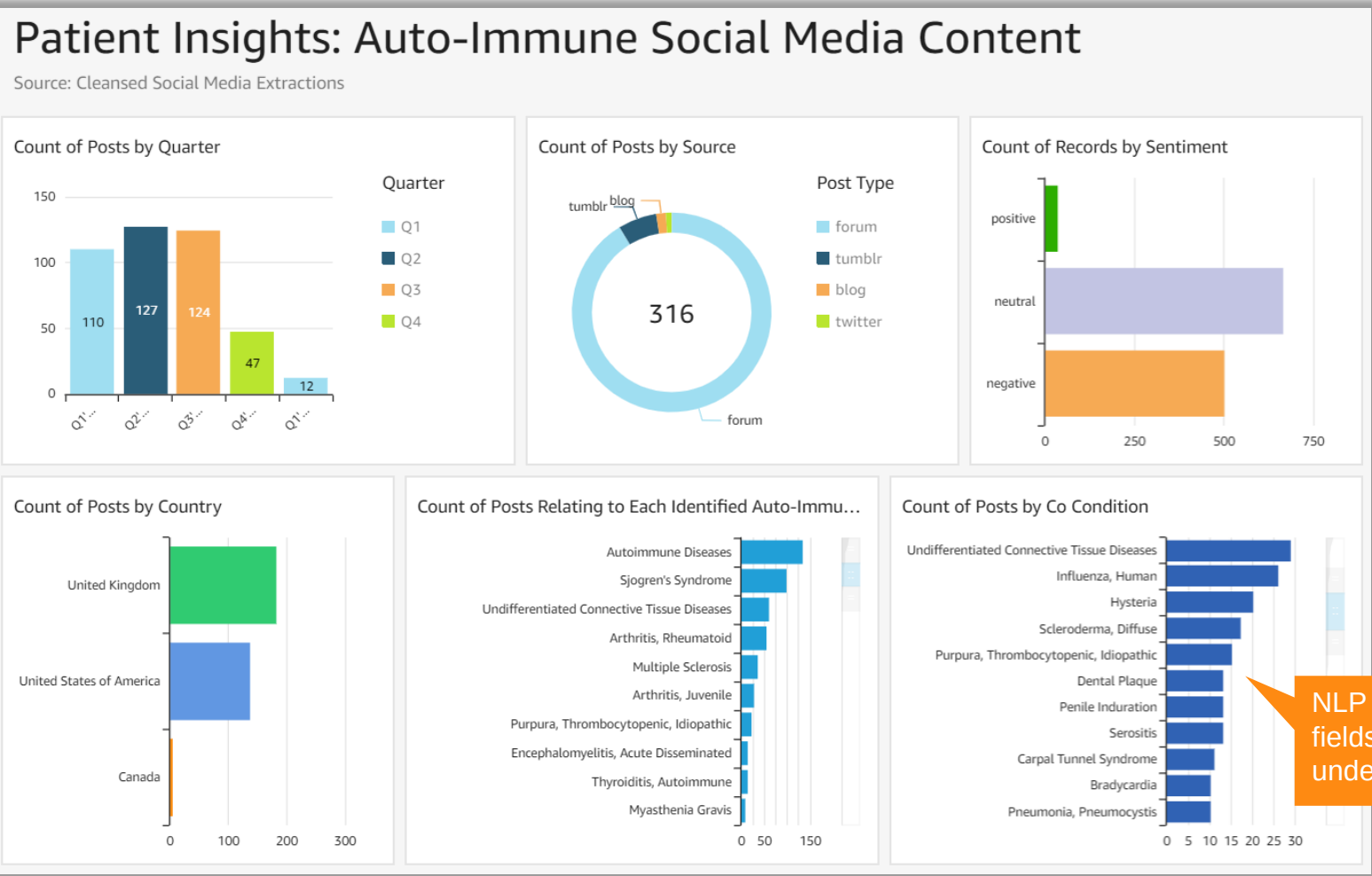
Clinical data
▪ Efficacy, safety, QoL ▪ Target therapy and comparators ▪ Comparator identification

Adverse Events
▪ Concomitant diseases & treatments ▪ Side effects and adverse events

- Rapid identification of relevant documents and sources
- Surfacing potentially useful data points
- Landscaping and alerts

Forums can be a rich source of Voice-of-the-Patient insights

Patients provide a range of facts, opinions and reactions through forums



Combining Social Media structured variables such as date and time with machine learning derived variables, such as sentiment, and NLP derived variables such as Condition, Co-Condition, Topic and more, increases the richness of the analysis

Social Media for Patient Voice in Clinical Trials

Situation

- Top 10 pharma wanted to use NLP to make sense of social media for patient-centered drug development (FDA initiative).
- Needed solution to develop clinical endpoints relevant to patients in early Parkinson's disease

Solution

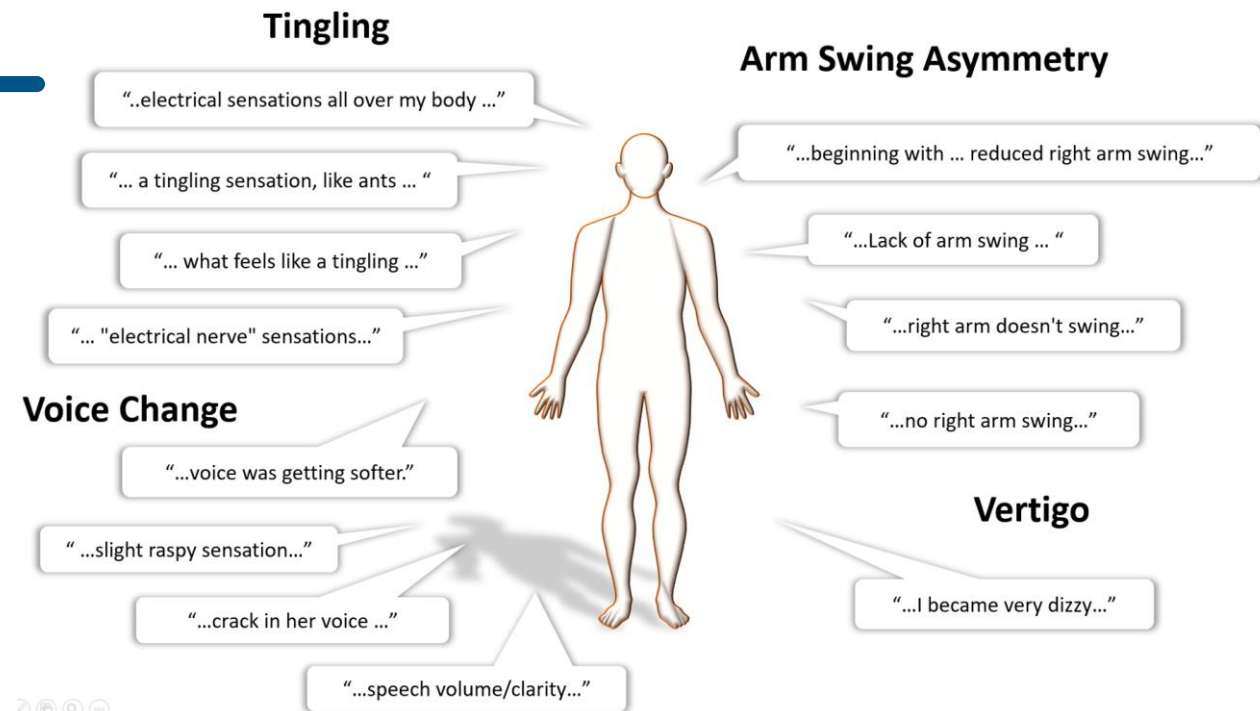
- Selected “highly trusted” social media sources (i.e. patient-focused communities rather than more diverse Twitter or Facebook posts)
- Used IQVIA NLP to discover and categorise comments into symptom or impact categories.

Staunton *et al*; Journal of Parkinson's Disease, 2022.
<https://pubmed.ncbi.nlm.nih.gov/34657850/>

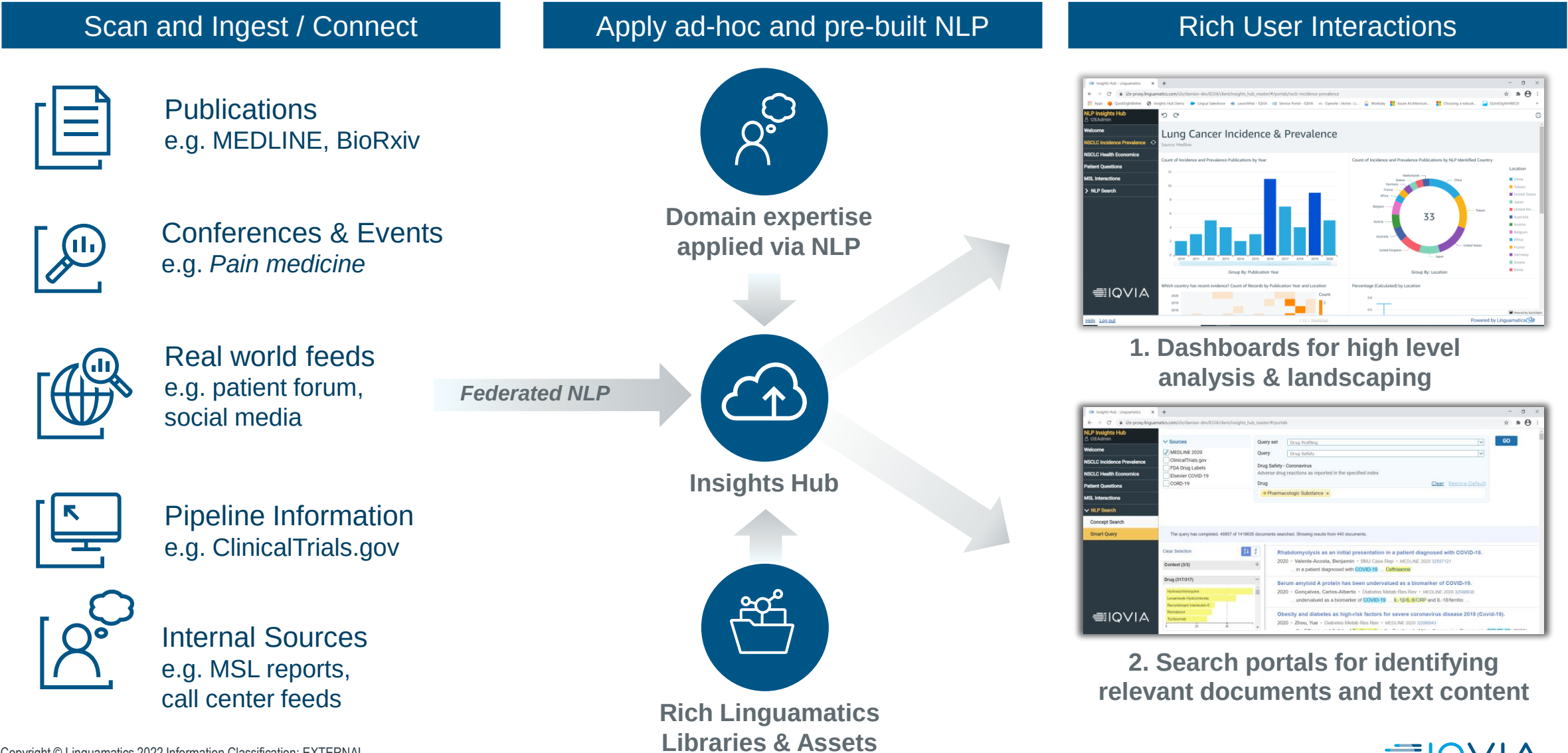


Results

- NLP extracted symptoms and impacts confirmatory of the clinical trial endpoints
- In addition, discovered new signals (e.g. tingling, voice change, vertigo)
- These specific recommendations have been taken forwards with clinical teams for additional patient-centric clinical endpoints



NLP Insights Hub allows users to interact intuitively with unstructured data from across different sources



Actionable insights from diverse real world data streams

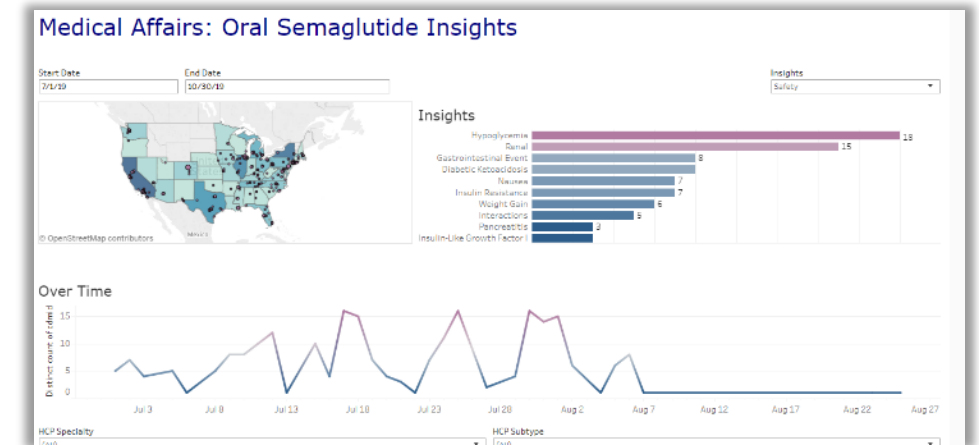
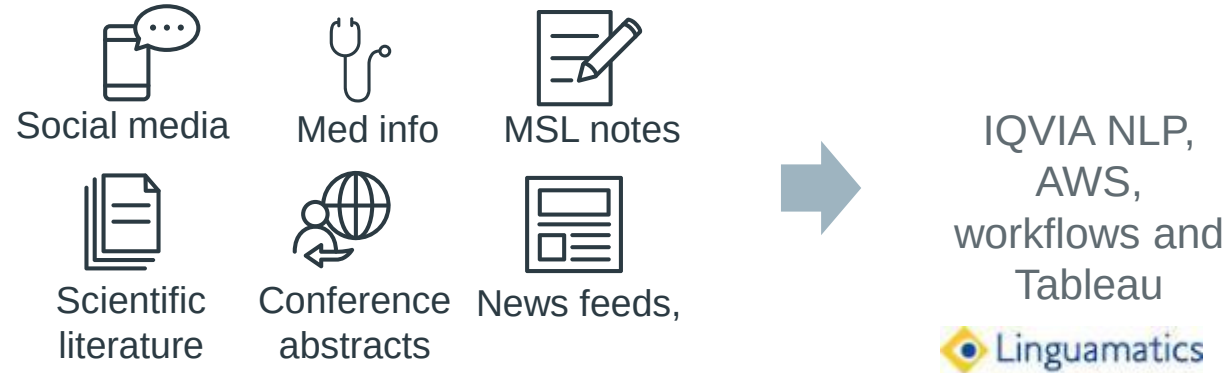


Top 10 Pharma

Situation

- Top 10 Pharma needed to efficiently scan and extract from disparate RWD sources to identify market trends and patterns
- Many sources contained valuable information. Manual scanning and extraction processes are labour intensive and inefficient

Solution



Results

- Rapid repeatable up-to-date insights; Reduced vendor spend; Broader access to insights across the enterprise
- Customer quote: "This gives us a comprehensive 360° view of what the world is saying"

Why Use Voice of the Customer Feeds?

NLP finds insights within call center verbatims, social media, medical insights

- Text analytics enables:
 - Identification of relevant metadata within text
 - Allocation of verbatims to appropriate topics
 - Deep analysis of drug-disease associations by understanding context
- Topic and trend analysis can highlight issues:
 - Product concerns/issues in clinical practice
 - Scientific unmet needs/wants
 - Potential areas of development
 - Niche areas for the product

Simple

1. Unexpected adverse events
2. Users splitting tablets and opening capsules
3. Counterfeit/change in tablet/change in packaging
4. Demographic information on callers

Mid-complex

1. Contraindicated medications among concomitant drugs
2. Switching
3. Off label use
4. Lack of efficacy questions (at what dose?)

Complex

1. Dose related questions
2. Distinguish side effect from pre-existing condition
3. Misspelled words and other human errors

NLP and Machine Learning for Medical Information Requests

Top 10 Pharma

Situation

- Top 10 Pharma needed to understand topics and trends from Medical Information Requests and Customer Requests coming from call centers
- NLP already in use to map verbatims to known categories
- Key need: How to identify a Medical Information Request (MIR) that is being asked for the very first time?

Solution

- Workflow was developed that uses data features extracted and cleaned in NLP process to feed Machine learning classifier
- Features such as drug, disease, dosage, clinician type, time period
- Lemmatization, stopword removal, vectorization
- Unsupervised ML algorithm clusters verbatims based on similarity
- Output is visualised to see new topics

Results

System enables users to ask:

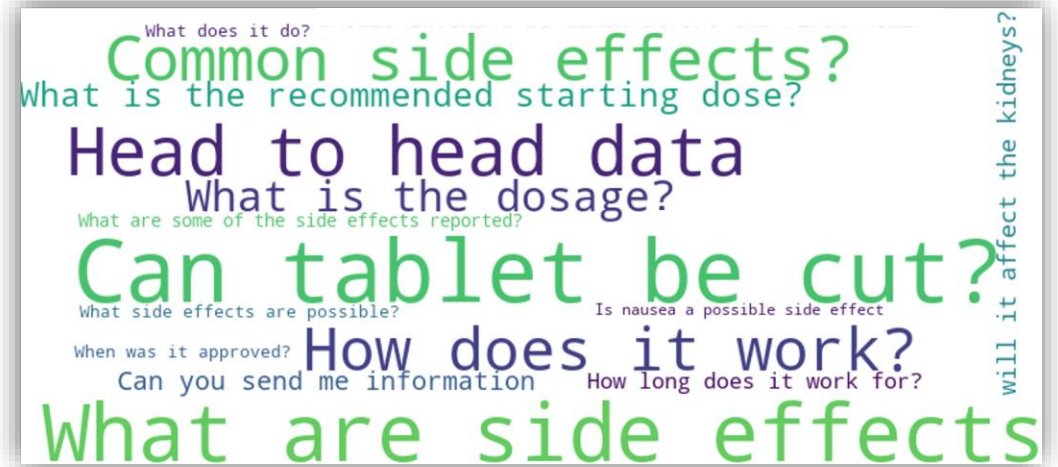
What MIRs are new? Completely unique or asked x times before?

Time periods for new topics?

Spikes in MIRs?

Allows effective identification of new and emerging signals across suite of brands in cardiovascular and respiratory franchises

Enables users to find novel information, insights



AI/ML for drug safety signal predictions

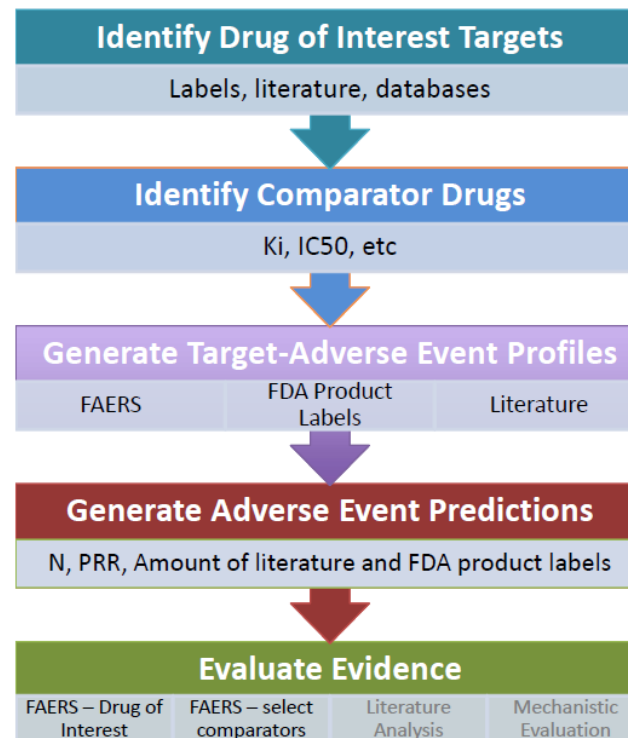
Situation

- FDA use a variety of AI/ML tools to develop models that can **predict adverse events** (AEs) using post-market safety data, for new drugs coming on the market.
- AEs for particular drugs are linked by inference to the drug target, which then enables a **level of risk prediction for new drugs** targeting the same protein

<https://pubmed.ncbi.nlm.nih.gov/32349656/>
<https://pubmed.ncbi.nlm.nih.gov/33090463/>

Solution

- FDA used Linguamatics NLP to **generate MedDRA mappings** from FDA drug labels, to compare with AEs from other sources (FAERS, literature)



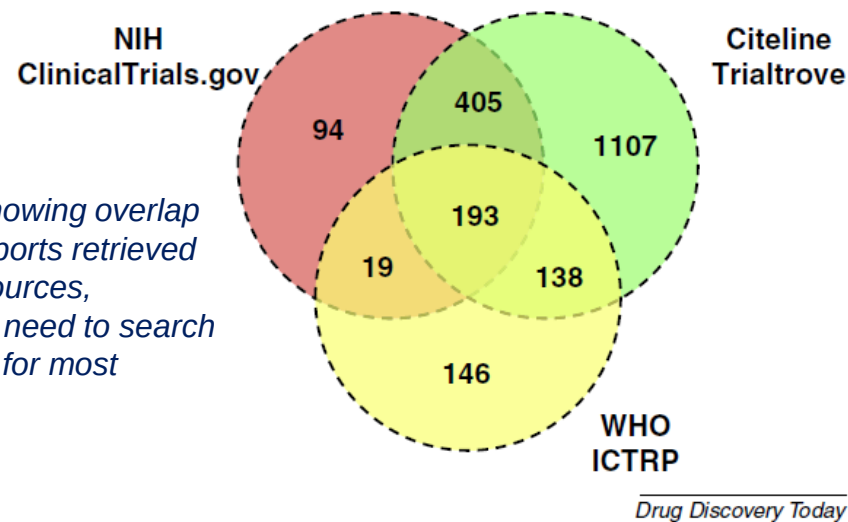
Results

- Features were fed into an **ensemble machine learning model**, which predicted post-market AEs (F-score 0.71)
- FDA demonstrated that **pharmacological target AE profiles, based on marketed drugs, can be used to predict signal of potential AEs for a new drug at the time of approval**
- Model has the potential to **speed safety assessments** by FDA evaluators

Comparative Effectiveness Research

Situation

- Comparative Effectiveness Research (CER) identifies what treatments work best for improving patient health
- Top 10 pharma wanted to identify all clinical trials that included specific Merck drugs within a comparator framework to assess evidence for these drugs
- Goals:
 - > To proactively identify opportunities and risks
 - > To make faster informed decisions on inline and development products



Solution

- NLP workflow developed to query ClinicalTrials.gov, WHO International Clinical Trials Registry Platform (ICTRP), and Citeline Trialtrove
- Broad range of information extracted (e.g. sponsor, drug and comparator, sample size, study duration, indications, patient information, utilisation/adherence)

Results

- Resulting data fed dashboards and alerts for Product teams; with considerable speed-up compared to manual process
- **“The timely information alerts generated by this system provide the earliest and most complete picture of emerging clinical research”**

<https://pubmed.ncbi.nlm.nih.gov/26854423/>

Text mining EMRs for Patient stratification of Heart Failure Risk

Top 10 Pharma



Challenge

- Heart failure (HF) clinical heterogeneity is problematic for risk stratification and treatment
- Improved phenotyping might help characterize HF subpopulations and improving risk stratification
- Top 10 pharma wanted to extract suitable parameters from EMRs to cluster patients for diagnostic metrics and disease outcomes



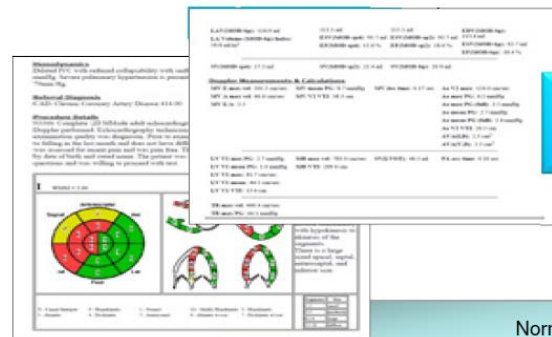
Solution

- Acquired EMR and imaging data from Geisinger for 907 patients
- Used IQVIA NLP to extract and normalise 39 demographic, clinical and echocardiogram variables
 - E.g. ejection fraction, grade of diastolic function, LV mass, relative wall thickness, etc.



Result

- Statistical clustering to identify subgroups associated with different outcomes
 - > 4 distinct groups found (for clinical, ECG and mortality)
- Approach enables HF risk stratification with potential implications for clinical trial design and identification of unmet needs



Physician summary
from diagnostic images
is extracted as text

abnormal
normal
abnormal Stage 1
mildly abnormal (grade I)
not assessed
Stage I
Moderately abnormal (stage II)
Grade II
moderately abnormal (grade II).
The left ventricular diastolic function is mildly abnormal (grade I).
There is no significant aortic regurgitation

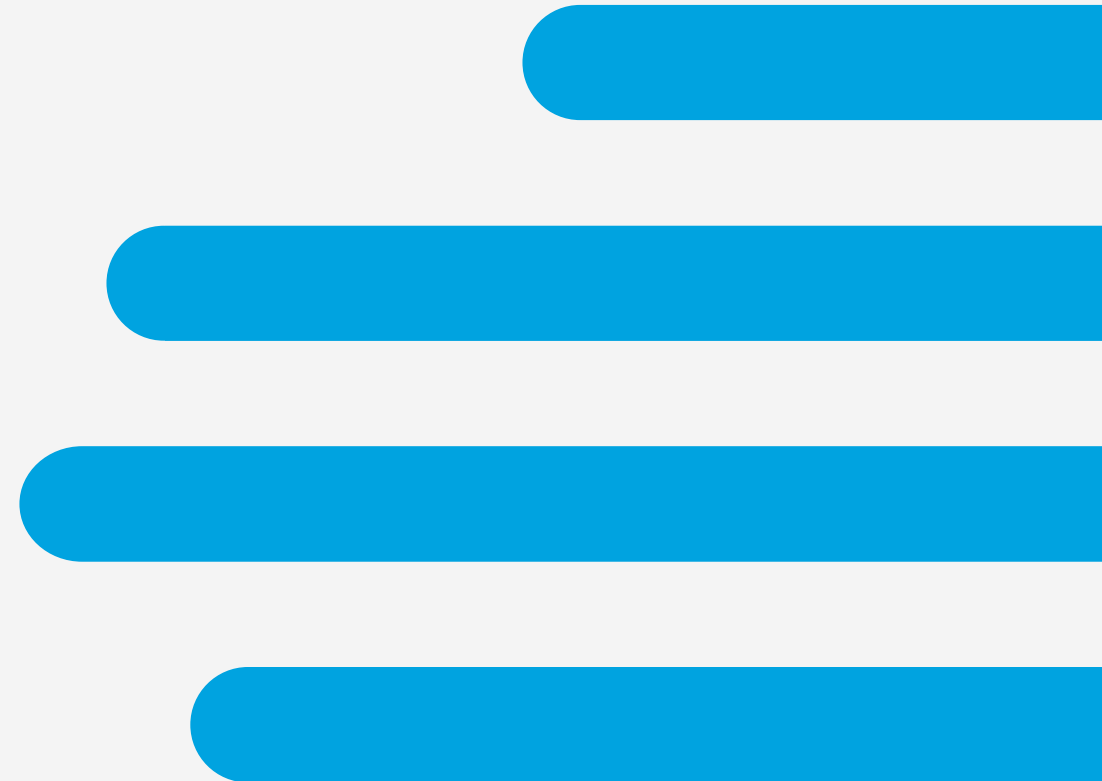
Unstructured text is converted to
standardized values for analysis

Normal
Mildly Abnormal (Grade I)
Moderately Abnormal (Grade II)
Severely Abnormal (Grade III)
Abnormal (other)

http://www.onlinejacc.org/content/69/11_Supplement/778



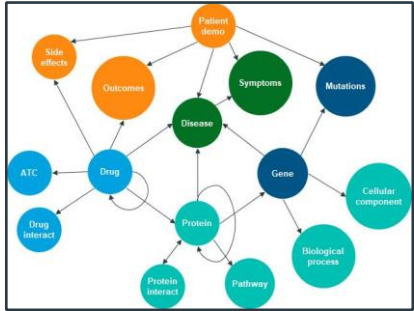
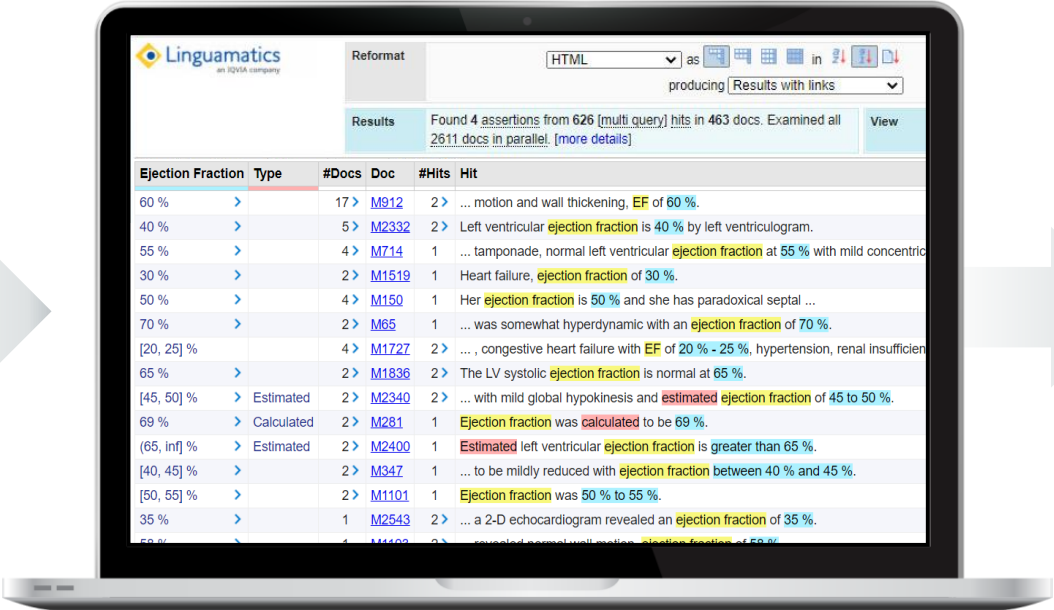
IQVIA NLP technology to transform RWD



IQVIA NLP transforms text into structured information and metadata for a wide range of uses

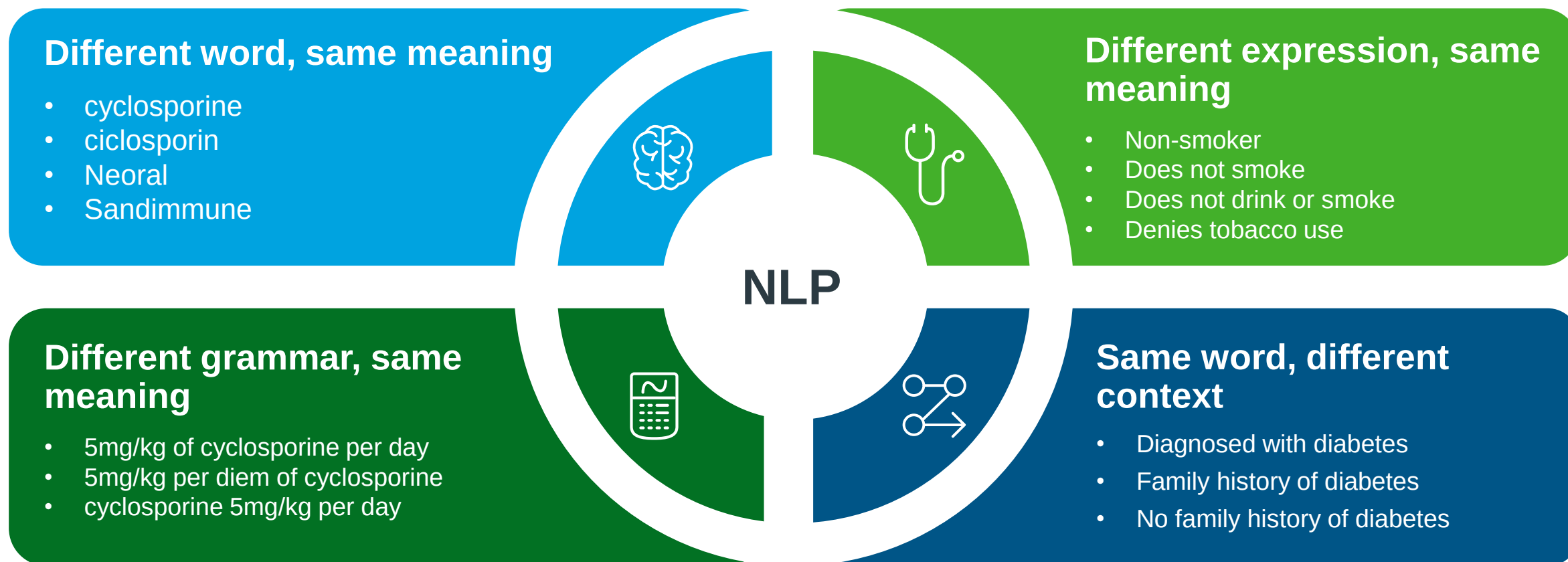
Free insights trapped in scientific and clinical text for decision support, bench to bedside

Search and transform text into... ... structured data using powerful queries to drive analytics and outcomes



Natural Language Processing – Ontologies – Statistical Methods – Machine Learning – Chemistry – Regular Expressions – etc.

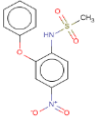
NLP finds information however it is expressed...



... and normalizes to the format required

The key to generating normalized structured results is to always treats the same concept in the same way

Increasingly differentiated capability

Concept	Text	Normalized Value
Diseases	breast cancer	Breast Neoplasm
	carcinoma of the breast	
Genes	Raf-1	RAF1
	Raf I	
Dates	27th Feb 2014	20140227
	2014/02/27	
Measurements	0.2g	200 mg
	Two hundred milligrams	
Mutations	Val 158 Met	V158M
	Val by Met at codon 158	
Behaviours	denies alcohol and tobacco use	Non-smoker
	is not a cigarette smoker	
Relationships	...nimesulide, a selective COX2 inhibitor, ...	 inhibits 

Data normalization

Overview

- Converts text into a standard format
- Is a fundamental component in transforming text into structured data and driving actionable insights
- IQVIA NLP provides broad, differentiated capability

Key benefits

- Find concepts however they are expressed
- Cluster or facet results by concept or quantity
- Compare measurements with different units e.g. kg vs. lbs
- Join results to discover new indirect relationships
- Integrate results into target database

IQVIA offers a comprehensive powerful NLP content and technology ecosystem to suit all users

Proprietary Content
Clinical Research Reports
Regulatory Documents

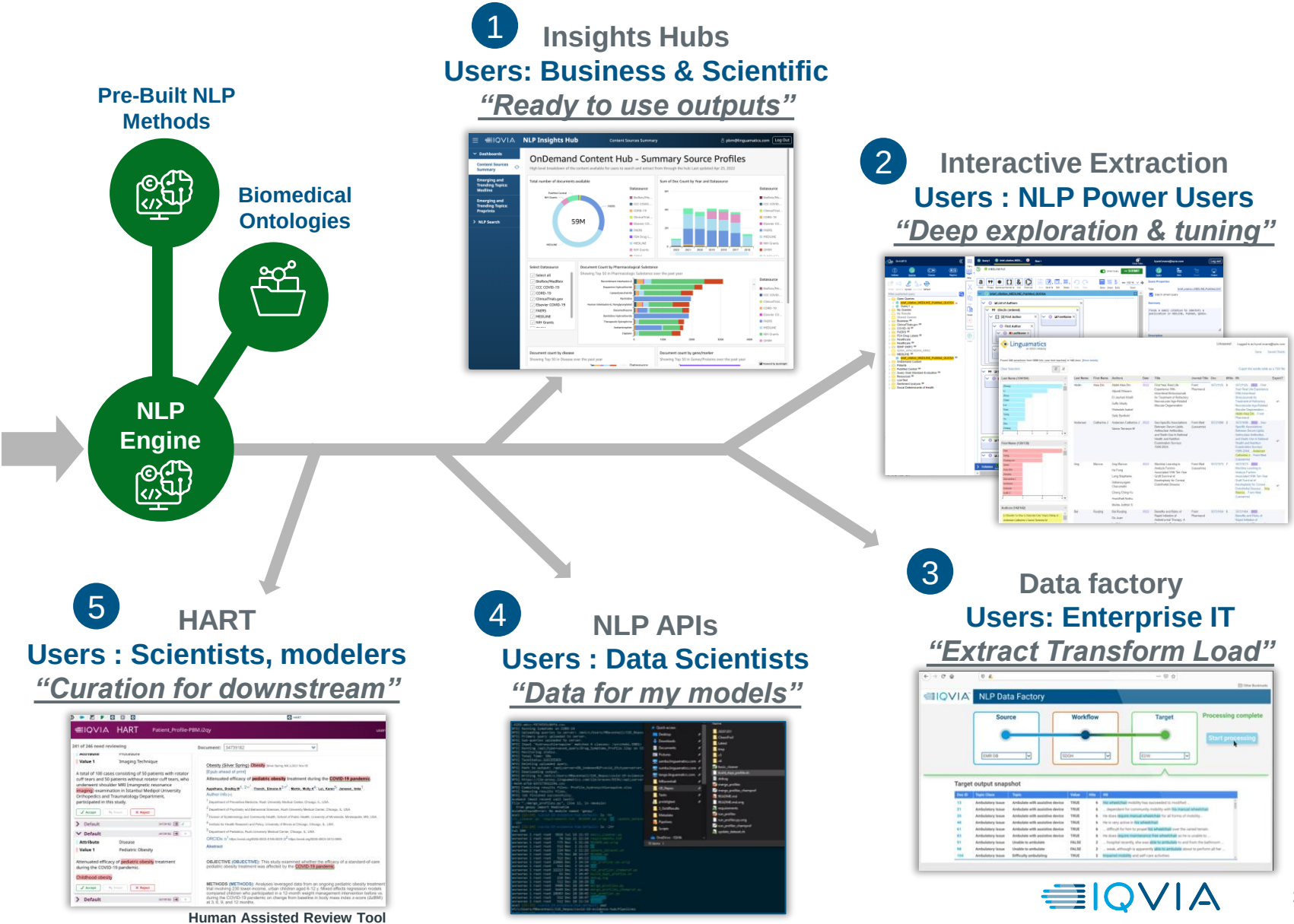
Interaction Data
e.g. CRM, MSL Notes
Knowledge Repositories
e.g. SharePoints
Primary Market Research

External Content
Partnerships & 3rd Party Providers

Social Media & News
Web, Congress, Grey Literature

IQVIA Content Store
MEDLINE
ClinicalTrials.gov
PubMed Central
FDA Drug Labels
OMIM

medRxiv & bioRxiv
FAERS
CCC COVID-19
CORD19 preprints (arXiv)
Elsevier Coronavirus Dataset
NIH Grants



In summary



Where NLP brings most value for real world data sources:

- Where you need information from rich but unstructured text from proprietary and external data sources
- Where your teams spend too much time sifting through data and documents
- Where you need up to date information from disparate sources

Broad ecosystem of NLP solutions, for range of users

NLP for rapid, systematic, comprehensive access to critical insights from unstructured text

Thank you for listening!

Questions?

For more information:

- Visit iqvia.com/nlp
- Contact me (jane.reed@iqvia.com)