Paper RW05

Using Natural Language Processing to transform real world data

Jane Reed, Linguamatics, Cambridge UK
Thierry Breyette, Novo Nordisk, NYC, USA

ABSTRACT

In pharma and healthcare, understanding the real world impact of therapies on patients is critical. Real world evidence (RWE) can inform all phases of drug development, commercialization, and drug use in healthcare settings. However, many real world data (RWD) sources, like electronic health records, patient forums, social media, news feeds and more, contain unstructured text.

Novo Nordisk provides a success story of using innovative technologies including Natural Language Processing (NLP), Amazon cloud and advanced visualisations to transform these unstructured sources of real world data into actionable structured real world evidence.

INTRODUCTION

In pharma and healthcare, understanding the real world (i.e. outside clinical trials) impact of therapies on patients is critical. Real world evidence (RWE) can inform all phases of drug development and commercialization. RWE enables product development and commercial decision-making, based on a better understanding of disease states and treatment patterns across a broad population. In December 2016, the 21st Century Cures Act was passed directing the FDA to evaluate the applications of RWE in supporting healthcare decision-making. RWE can shed light on real-world clinical effectiveness, on safety profiles of products across a broad patient community; and also be used to assess patient-reported outcomes, to understand product reputation management, for key or digital opinion leader engagement, and more.

However, many real world data (RWD) sources contain unstructured text, which prevents easy analysis. Text analytics can unlock the value from real world sources such as EHRs, claims data, social media, and customer call transcripts. Building a forward-looking analytics framework to tackle these new data challenges requires both extensible and flexible tools, and creative thinking.

Large amounts of RWD on Novo Nordisk product usage are available in multiple formats from medical information requests, field medical affairs and medical scientific liaison reports, customer call center logs, and interactions with healthcare professionals. Mining these disparate structured and unstructured sources using traditional manual scanning and extraction techniques is time consuming and inefficient, and Novo Nordisk wanted to speed up, automate, and scale the process.

In addition, there are other external data streams that can provide valuable insights into impacts of any particular product in the real world. These include news data, information from scientific literature and conferences, and focused social media data (e.g. from patients, carers, HCPs, KOLs, and more). Combining these external data with the internal data sources would provide pharma companies with a comprehensive 360˚ view of what the world saying about their drugs.

In this project, Novo Nordisk use advanced tools and technologies such as natural language processing (NLP) to gain real value from these internal and external RWD sources. The final Medical Patient Dashboard provides interactive displays of trends and key insight categories. These are initially focused on a new oral therapeutic for type 2 diabetes, semaglutide (Rybelsus®), a GLP-1 receptor agonist. Up-to-date information on this new product enables medical affairs, commercial, product, and market access teams to identify macro and micro healthcare market trends in the US and discern patterns in patient sentiment, compliance, routines, behaviors, and overall treatment satisfaction and outcomes.

METHODS and RESULTS

To develop this 360˚ real world data insights system, Novo Nordisk used a combination of three technologies: Linguamatics I2E for natural language processing (NLP)-based text mining, Tableau for dashboard visualisations, and Amazon Web Services (AWS) cloud-based global big data and analytics platform. They have developed a workflow to semi-automatically extract data from real world sources and display these data in dashboards.

The internal (see Figure 1) and external RWD assets were being used across Novo Nordisk but this usage was incomplete, not standardized or systematic, and in places involved considerable vendor spend The various sources are detailed below:

- Medical information requests: Medical information requests provide valuable insights about the information needs of healthcare professionals, particularly for any new drug entering the market. Previously at Novo Nordisk, the Medical Affairs team paid an external vendor to manually review healthcare professional (HCP) interactions with medical science liaison (MSL) teams and create a monthly insights report. These reports provided an analysis of the interactions and macro level healthcare trends. However, application of these static reports to decision-making was not evidenced by clear actions. Additionally, the cost of paying a vendor the equivalent cost of a full time FTE was not considered a good return on investment.
- Field Medical affairs: Medical information team members would manually review their questions from HCPs on a monthly basis and pick out what was judged to be insights and pass these along to medical affairs for their "Monthly Insights" report (curated by an external vendor). This manual curation was not always done by the same people and what constituted an insight was not standardized.
- Patient Centric Customer Care call notes: The Novo Nordisk US Patient Centric Customer Care centers receive over 100,000 calls each year from patients, carers or healthcare professionals. However, Novo Nordisk had no standard way of collecting and communicating insights from the call center interactions to the broader product teams. Additionally, insights that were collected from call center interactions remained mostly confined to call center optimization and not shared more broadly across medical and commercial teams.
- Social Media: Social media is a hugely valuable but noisy source of real world data. Novo Nordisk needed to develop capability to utilize social media to understand the conversation about Rybelsus. What are KOLs saying, where is content being hosted and accessed via social media? What tweets are being shared? What hashtags are being used to discuss Rybelsus? Social media gives us access to non-traditional platforms and generates a "surround-sound" experience of the discussions ongoing about Rybelsus from pre-approval, post approval and market launch.
- News: News feeds provide up-to-date views across the globe. At Novo Nordisk, data from news is being gathered by many different groups but with no comprehensive, centralized way to capture, analyse and share daily news feeds. What news is being published and where? What are the key topics discussed in articles, for example around efficacy, safety, diabetic populations, etc.

In addition to the above, two more traditional sources of data are also being included; scientific literature and conference abstracts. Information on health economics and outcomes; diabetes and obesity disease states; new research on targets, pathways, mechanism of action; clinical trial updates and much more is available from scientific research, both from published abstracts and full text papers and also from conference abstracts. An additional benefit of accessing data from conference abstracts is the ability to capture information earlier in time (i.e. before a full scientific publication) and to follow the research trajectory from poster to full paper.

In the Novo Nordisk Medical Patient Dashboard pipeline there are three main components:

1: LINGUAMATICS NLP TO EXTRACT DATA FROM REAL WORLD AND OTHER SOURCES

The RWD sources include medical information requests (20,000 per year), field medical affairs notes (3,300 per year), and customer call center reports (130,000 per year; see Figure 1). Additionally, there are thousands of social media posts, news articles, publications, & conference abstracts. These were linguistically processed with Linguamatics Natural Language Processing (NLP) text analytics solution I2E (https://www.linguamatics.com/products-services/about-i2e) and indexed, using MeSH, NCI, MedDRA, and in-house developed ontologies. I2E queries were developed relating to topics such as safety, efficacy, PK/PD, randomized controlled trials, patient populations, dosing, and devices. The queries were refined in close collaboration with internal subject matter experts (e.g. pharmacists on the team), to tune the I2E queries to the desired level of precision and recall. Once the algorithms and post-processing were optimized, they were embodied in an integrated workflow that could be triggered as needed on a regular basis.

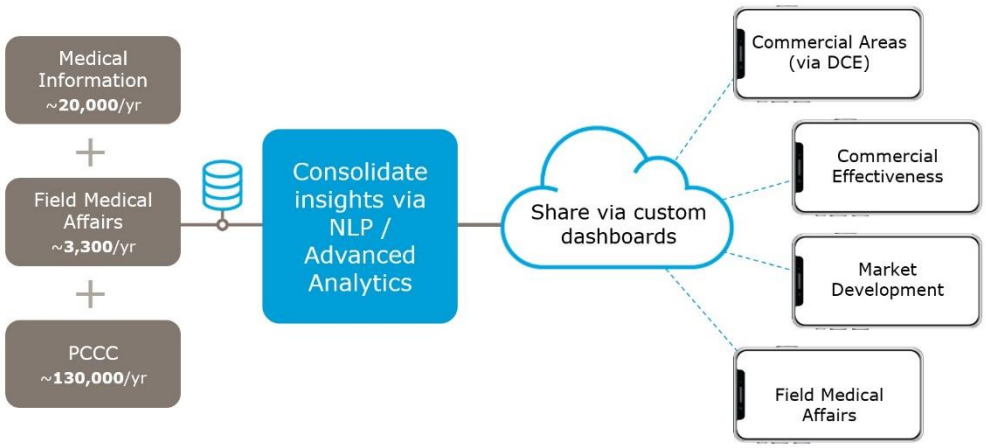


Figure 1: Three real world data sources (medical information requests, notes from field medical affairs, and Patient Centric Customer Care call notes) are combined, insights extracted using Linguamatics's NLP solution, and shared to a variety of teams for Diabetes Care and Education, market development, commercial effectiveness, etc.

2: TABLEAU DASHBOARDS FOR VISUAL ANALYTICS

Tableau (https://www.tableau.com/) is an interactive data visualization tool that can be used to create customized dashboards and generate business insights; providing access to many different and interlinked displays. For the RWD extracted and standardized using I2E (see Figure 2), a range of visual displays were used, including barcharts, geographic maps, and timelines. These allowed product and commercial teams to ask questions such as:

- What questions are healthcare professionals asking about Rybelsus? These might include, for safety issues, does the question pertain to dosage? Renal or hepatic issues? Pancreatitis? Absorption? Or for efficacy issues, impact on weight loss? A1C levels?
- Where types of efficacy questions are being asked regarding Rybelsus? And in what US state or region?
- What questions are being asked about Rybelsus vs. other GLP-1 drugs? And what are the trends over time?

Figure 2: Example dashboards for the NLP-extracted data for Rybelsus. **Figure 2a** shows the medical affairs dashboard, including geographic distribution of field conversations, the key insights HCPs are discussing (e.g. hypoglycaemia, renal issues), trends over time, and HCP type. **Figure 2b** reveals the main trends and topics related to Rybelsus (semaglutide) in key twitter feeds to the end-users. **Figure 2c** is the News dashboard, including a barchart for the publisher source, frequencies of most discussed insight categories, and a treemap of key insights. **Figure 2d** enables users to access conference abstracts, with the ability to select date, conference, insight category or abstract type.

Figure 2a

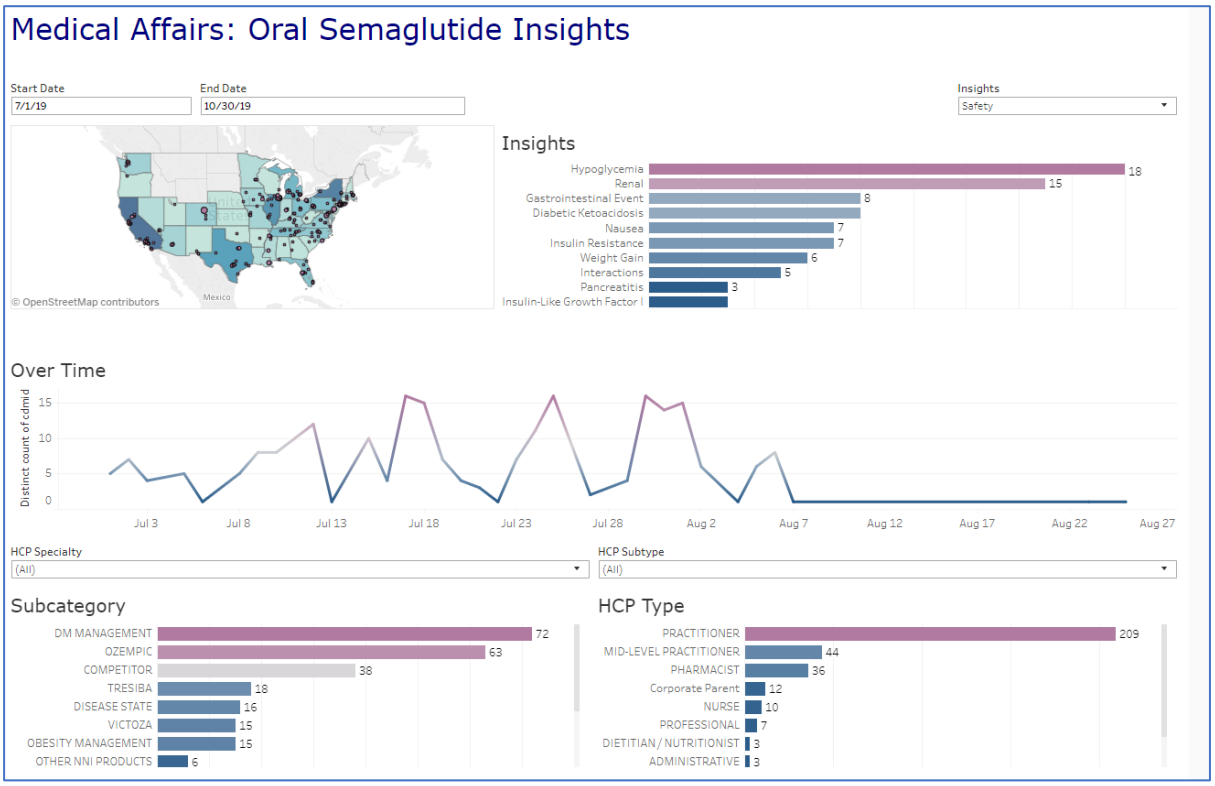

Medical Affairs: Oral Semaglutide Insights
Start Date
7/1/19
End Date
10/30/19
Insights
Safety
© OpenStreetMap contributors
Insights
Hypoglycemia 18
Renal 15
Gastrointestinal Event 8
Diabetic Ketoacidosis
Nausea 7
Insulin Resistance 7
Weight Gain 6
Interactions 5
Pancreatitis 3
Insulin-Like Growth Factor I
Over Time
Distinct count of cdmid
15
10
5
0
Jul 3
Jul 8
Jul 13
Jul 18
Jul 23
Jul 28
Aug 2
Aug 7
Aug 12
Aug 17
Aug 22
Aug 27
HCP Specialty
(All)
HCP Subtype
(All)
Subcategory
DM MANAGEMENT 72
OZEMPIC 63
COMPETITOR 38
TRESIBA 18
DISEASE STATE 16
VICTOZA 15
OBESITY MANAGEMENT 15
OTHER NNI PRODUCTS 6
HCP Type
PRACTITIONER 209
MID-LEVEL PRACTITIONER 44
PHARMACIST 36
Corporate Parent 12
NURSE 10
PROFESSIONAL 7
DIETITIAN / NUTRITIONIST 3
ADMINISTRATIVE 3

Figure 2b

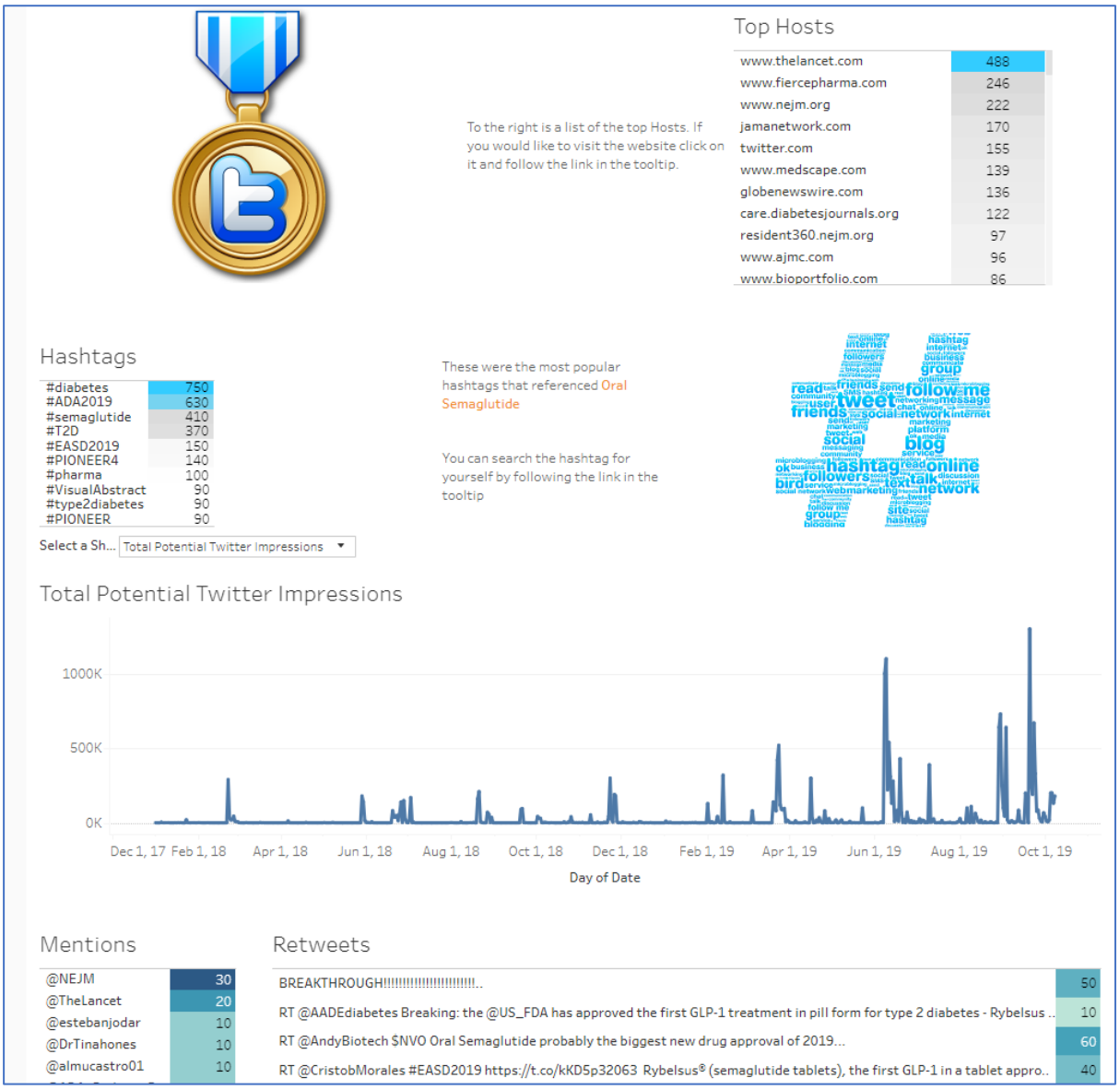

Top Hosts
www.thelancet.com 488
www.fiercepharma.com 246
www.nejm.org 222
jamanetwork.com 170
twitter.com 155
www.medscape.com 139
globenewswire.com 136
care.diabetesjournals.org 122
resident360.nejm.org 97
www.ajmc.com 96
www.bioportfolio.com 86
To the right is a list of the top Hosts. If you would like to visit the website click on it and follow the link in the tooltip.
Hashtags
#diabetes 750
#ADA2019 630
#semaglutide 410
#T2D 370
#EASD2019 150
#PIONEER4 140
#pharma 100
#VisualAbstract 90
#type2diabetes 90
#PIONEER 90
These were the most popular hashtags that referenced Oral Semaglutide
You can search the hashtag for yourself by following the link in the tooltip
Select a Sh... Total Potential Twitter Impressions
Total Potential Twitter Impressions
1000K
500K
0K
Dec 1, 17 Feb 1, 18 Apr 1, 18 Jun 1, 18 Aug 1, 18 Oct 1, 18 Dec 1, 18 Feb 1, 19 Apr 1, 19 Jun 1, 19 Aug 1, 19 Oct 1, 19
Day of Date
Mentions
@NEJM 30
@TheLancet 20
@estebanjodar 10
@DrTinahones 10
@almucastro01 10
Retweets
BREAKTHROUGH!!!!!!!!!!!!!!!!!!!!!!!!.. 50
RT @AADEdiabetes Breaking: the @US_FDA has approved the first GLP-1 treatment in pill form for type 2 diabetes - Rybelsus .. 10
RT @AndyBiotech $NVO Oral Semaglutide probably the biggest new drug approval of 2019... 60
RT @CristobMorales #EASD2019 https://t.co/kKD5p32063 Rybelsus® (semaglutide tablets), the first GLP-1 in a tablet appro.. 40

Figure 2c

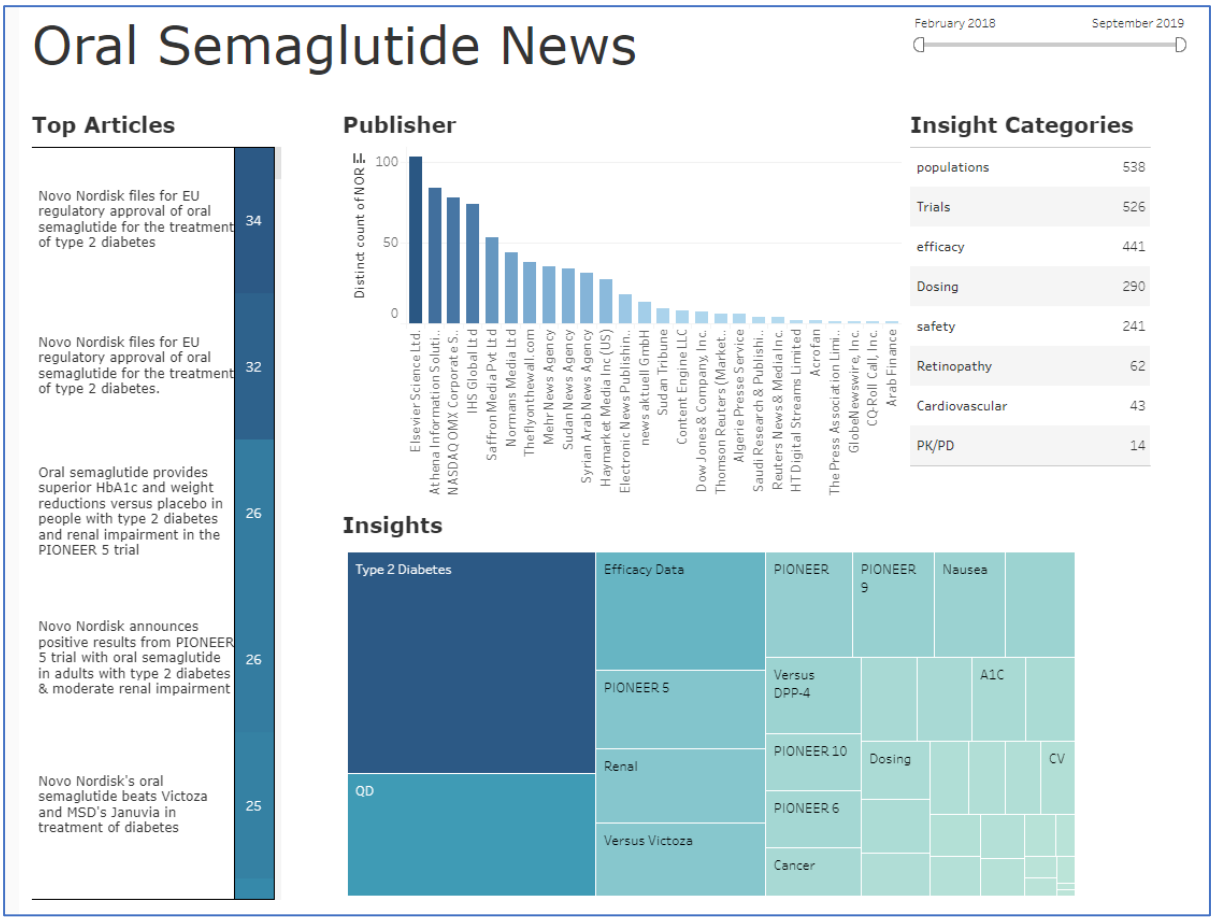
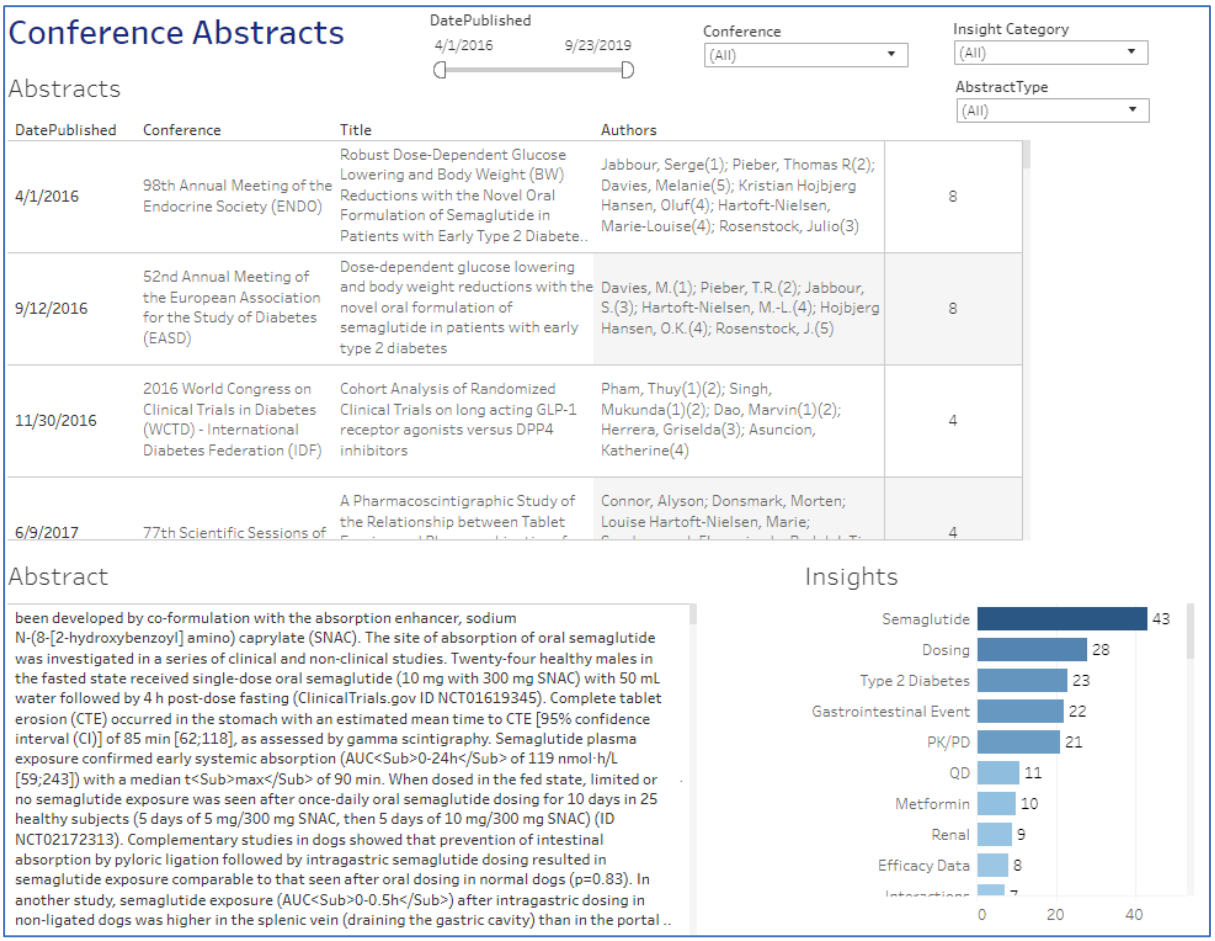

Oral Semaglutide News
February 2018
September 2019
Top Articles
Novo Nordisk files for EU regulatory approval of oral semaglutide for the treatment of type 2 diabetes
34
Novo Nordisk files for EU regulatory approval of oral semaglutide for the treatment of type 2 diabetes.
32
Oral semaglutide provides superior HbA1c and weight reductions versus placebo in people with type 2 diabetes and renal impairment in the PIONEER 5 trial
26
Novo Nordisk announces positive results from PIONEER 5 trial with oral semaglutide in adults with type 2 diabetes & moderate renal impairment
26
Novo Nordisk's oral semaglutide beats Victoza and MSD's Januvia in treatment of diabetes
25
Publisher
Distinct count of NOR...
100
50
0
Elsevier Science Ltd.
Athena Information Soluti..
NASDAQ OMX Corporate S..
IHS Global Ltd
Saffron Media Pvt Ltd
Normans Media Ltd
Theflyonthewall.com
Mehr News Agency
Sudan News Agency
Syrian Arab News Agency
Haymarket Media Inc (US)
Electronic News Publishin..
news aktuell GmbH
Sudan Tribune
Content Engine LLC
Dow Jones & Company, Inc.
Thomson Reuters (Market..
Algerie Presse Service
Saudi Research & Publishi..
Reuters News & Media Inc.
HT Digital Streams Limited
Acrofan
The Press Association Limi..
GlobeNewswire, Inc.
CQ-Roll Call, Inc.
Arab Finance
Insight Categories
populations 538
Trials 526
efficacy 441
Dosing 290
safety 241
Retinopathy 62
Cardiovascular 43
PK/PD 14
Insights
Type 2 Diabetes
QD
Efficacy Data
PIONEER 5
Renal
Versus Victoza
PIONEER
PIONEER 9
Nausea
Versus DPP-4
A1C
PIONEER 10
Dosing
CV
PIONEER 6
Cancer

Figure 2d



3: INFRASTRUCTURE BASED ON OASIS, NOVO NORDISK'S CLOUD BASED GLOBAL DATA & ANALYTICS PLATFORM

OASIS is built on Amazon Web Services and managed in collaboration with services from Accenture. An "Open AnalyticS & InsightS" platform, OASIS is focused on making data readily available from essential repositories so that Novo Nordisk staff can derive insights from their ever-growing data in the most efficient way possible. OASIS leverages Amazon simple storage solution S3 as a backbone of the data lake to store a wide variety of data, both unstructured and structured. In addition to data storage, the OASIS platform also provides data engineers, data scientists and analysts the tools to work with their data (see Figure 3).

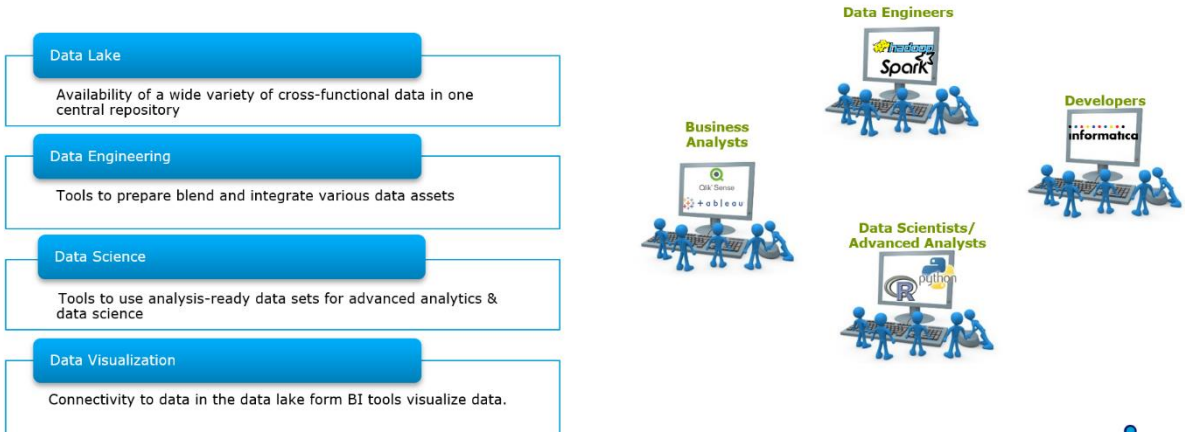


Figure 3: Schematic of OASIS, Novo Nordisk's cloud based global Data & Analytics platform. Hosted on AWS, the central data lake hosts a wide variety of cross-functional data in one central repository. These data are available to a variety of groups, for data engineering, advanced analytics, business intelligence tools, and more.

An I2E server was migrated to AWS, and a data pipeline build using Knime. This controls the data flowing in and out of I2E, with a specific end-user interface for an analyst to specify names of indexes and queries to run the workflow (see Figure 4).

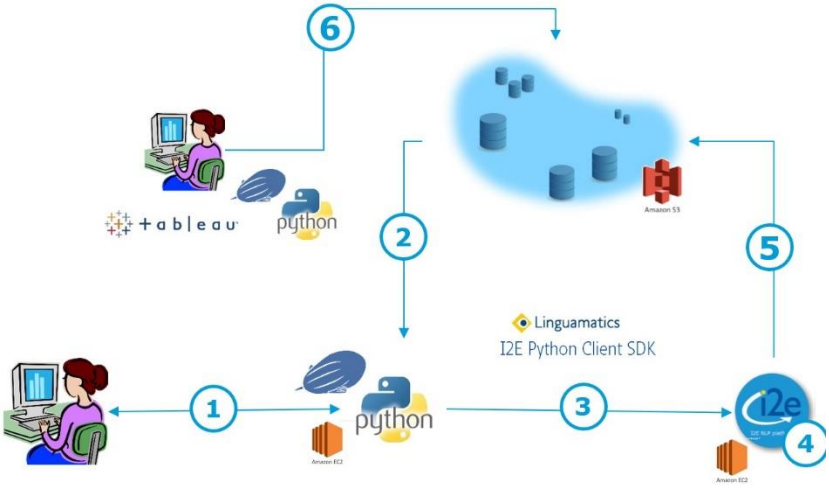


Figure 4: The data pipeline developed to enable I2E to access the real world data sources, extract key information, and provide this for visualization in Tableau. The steps in the pipeline are (1) connect via web browser; (2) source files sourced from data lake; (3) Python program to run workflow in I2E using KNIME nodes; (4) results saved back to the data lake; and (5) business end-users can visualize and analyze the data via Tableau dashboards.

The resulting system is now implemented for Novo Nordisk US, and is being rolled out globally. Critically, the pipeline developed has been able to overcome a number of challenges, such as reducing the time needed for the overall workflow; enabling much more frequent updates; automating the integration of data from multiple sources; enabling access to legacy data, and accommodating demand from a broader set of Novo Nordisk affiliates.

CONCLUSION

To summarize, Novo Nordisk's application of Linguamatics text mining software, combined with Tableau visualizations and AWS cloud-based pipeline, enables effective and comprehensive analysis across six valuable sources of information combining real world data and scientific research. I2E enables high levels of recall and increased efficiency of topic and trend identification across the different data feeds (medical information requests, field medical affairs notes, customer call center reports, news feeds, scientific literature and conference abstracts). This data extraction process is otherwise highly time-consuming via standard manual methods.

It is difficult to quantify exactly the significant value of a robust process that provides a comprehensive view of real world patient and HCP responses, news updates and scientific research, to the commercial teams at Novo Nordisk. Pharmaceutical companies are driven to better utilize RWD by pressure to demonstrate value for market access and drug reimbursement, the need to better understand the patient journey, and greater acceptance of RWD among regulators. It is estimated that a big pharmaceutical company spends nearly USD 20 million annually for generating RWE-based insights (ReportLinker, 2018).

By bringing together six data sources in one unified and dynamic dashboard Novo Nordisk are able to cut out vendor spend, automate and standardize the process of generating and communicating insights, reduce manual work by FTEs (allowing them to focus on actual analysis and communication of trends), and ensure that insights can be acted upon in context. The system has broadened how and where insights are communicated and ties together medical and patient insights collected through multiple and previously siloed channels.

Being able to respond rapidly to trends seen in clinical practice, concerns from patients or HCPs, or indications of competitor product advantage, enables Novo Nordisk to stay at the forefront of diabetes care.

With the migration to the AWS platform, the data pipeline saves significant time, and provides stakeholders with on-demand insights to enable evidence-based decision making. In this respect, Novo Nordisk is exemplifying the increased use of cloud infrastructure for big data performance, connectivity, flexible provisioning, and powerful compute and data analytics tools; a trend we are seeing across the pharma industry.

REFERENCES

ReportLinker (October 2018); "Pharmaceutical and Life Sciences Real World Evidence: Market Landscape and Competitive Insights, 2018-2030". https://www.reportlinker.com/p05723260/Pharmaceutical-and-Life-Sciences-Real-World-Evidence-Market-Landscape-and-Competitive-Insights.html

ACKNOWLEDGMENTS

We would like to acknowledge the efforts and expertise of the Novo Nordisk team involved in developing these workflows and dashboards. In particular:

- Tom Horan, Senior Analytics Specialist
- Lauren D'Amato, Systems Analyst III

CONTACT INFORMATION

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

- Jane Z Reed, Director Life Science; Linguamatics, an IQVIA company. jane.reed@linguamatics.com; www.linguamatics.com.
- Thierry Breyette, Director, Scientific Analytics; Clinical Development, Medical & Regulatory Affairs; Novo Nordisk Inc. syby@novonordisk.com; www.novonordisk.com.