WHITE PAPER

Data Analytics for Unstructured Clinical Case Notes

Breakthroughs in analysing unstructured clinical data show potential to improve the delivery of integrated, resource-efficient patient care

Julie Price
Acting Director of Performance and Business Intelligence, North East London NHS Foundation Trust

Malcolm Linington
Healthcare Solutions Architect, Intel Corporation

Dr. David Frost
Chief Clinical Information Officer, Santana Big Data Analytics

Executive Summary

Clinical case notes are vital to delivering high-quality, coordinated patient care and optimising resources across the health economy. Yet case notes are unstructured and are often housed in incompatible record-keeping siloes, making the information unavailable for the sophisticated presentation and analysis that structured electronic medical record (EMR) systems can provide.

Advances in big data analytics show the potential to capture clinically and operationally relevant information from diverse sources of unstructured clinical case notes. This analysis, presented in easy-to-grasp, clinical decision-support portals, can support clinical and operational improve ments that may help:

- Provide a more comprehensive longitudinal view of patients' lives and health, enabling clinicians and organisations to deliver more personalised, higher quality care
- Offer system-wide opportunities to enhance operational efficiency, improve resource utilisation, and benefit the nation's health economy
- Increase the satisfaction and safety of patients and clinicians

A collaborative proof-of-concept (POC) amongst the North East London National Health Service (NHS) Foundation Trust (NELFT), Intel Corporation, and Santana Big Data Analytics (Santana BDA) Ltd has demonstrated a practical, affordable approach to extracting relevant information from large volumes of clinical case notes. This paper describes the POC and suggests next steps for healthcare systems that want to use next-generation data analytics to improve healthcare.

Urgent National Issues

In a time of rising demand and falling budgets, the healthcare system must deliver the right care at the right time and the right place in the care continuum. Yet with valuable information housed in clinical case notes and other un-structured data sources, clinicians and administrators must often make planning decisions based on limited data. This can have negative consequences for clinical outcomes, resource planning, and operational efficiencies.

For example, a King's Fund study[1] found that in the NHS system, the elderly ac-count for 68 percent of unplanned acute care admissions annually. This amounts to approximately two million unplanned admissions each year, and the number is growing. In some areas of the country, each adult over 65 spends an average of four days annually as an unplanned admission in a hospital bed, incurring costs to the NHS of GBP 8.3 billion per annum. The Kings Fund study estimates

Table of Contents

that the NHS could save GBP 600 million and eliminate 7,000 acute-care beds across England through more effective elderly care planning.

And, of course, the NHS's costs are dwarfed by the social care costs and the wider personal and economic costs, including the loss of economic productivity due to carer commitments. Unplanned, acute-care admissions also contribute to overcrowded geriatric wards, lack of capacity in social care beds, problems reintroducing patients back into their own home settings, higher than optimal length of stay, and exacerbation of co-morbidities.

The limited visibility into unstructured clinical information means that the design of services such as those for independent living is based on incomplete data. The lack of data visibility can make it difficult to assess the quality and effectiveness of various services and treatments since outcomes may be difficult to define. The result is an economy of new services that are being commissioned, where the return on investment in cost, quality, and clinical safety is smaller than it should be. Case notes can help tell us which services are the most effective, qualitatively initially, and then quantitatively when processed at scale.

Recent innovations in big data analytics provide healthcare leaders with a significant opportunity to reshape this picture by analyzing data from clinical case notes and using it to inform clinical care and institutional planning. NELFT worked with Intel and Santana BDA to explore the technical feasibility of such an approach.

NELFT and the Importance of Clinical Notes

NELFT provides integrated community and mental health services for a diverse population of almost 1.5 million people living in the London Boroughs of Waltham Forest, Redbridge, Barking & Dagenham, and Havering. The Trust also manages community health services in south west Essex and mental health services to young people across the whole of Essex. NELFT has an annual budget of more than GBP 325 million in 2013/2014 and around 5,500 staff. NELFT is a recognised research leader and innovator, partnering with diverse academic and private-sector leaders to explore new approaches to improving the quality of its services.

As a community provider, NELFT provides services that are critical to preventing unplanned admissions and enabling people to live longer, healthier lives by consuming services away from hospitals. These include preventative and rehabilitative services as well as those designed to preserve each individual's quality of life.

High quality, succinct case notes are key to NELFT operations and services. For example, case notes at NELFT are used to:

- **Identify candidates for early screening,** including testing for dementia, before the disease becomes a problem requiring a higher-cost and potentially less-effective treatment
- **Assist in care handover between shifts,** among members of care teams, and across the accountable care continuum, for example encompassing patients on the frail elderly care pathway
- **Optimise referrals through the care system,** to avoid or minimise inappropriate referrals

The Challenge of Clinical Case Notes

Unfortunately, clinical case notes can be challenging to work with. Unstructured in form, case notes can vary widely from one clinician to the next. For patients who may have had complex circumstances or decades of encounters with the healthcare system, the case notes can consist of a stack of file folders several feet high. Relevant patient data may be housed in diverse systems, including GP records, community nursing records, social care support records, A&E records, and other data siloes. Too often, clinicians are overwhelmed with information.

The result is that even the most dedicated clinician often faces the patient armed with only the structured data record and a chance to skim the case notes from the patient's most recent visit. As a result, the clinician has a fragmented view of the relevant patient data, and can easily miss incipient problems that become "visible" only when a crisis occurs. For example, the structured EMR will show that a patient has had a fall, broken a bone, and requires A&E services. But what of the underlying contributors to this unplanned admission? Could the event have been prevented if unstructured information in the clinical notes (information about the patient's loneliness, alcohol use, social support network, and other important factors) could be easily and securely shared amongst the extended clinical team?

If we can combine diverse sources of clinical notes, extract relevant information, and present it in clinically relevant ways, we should be able to build a more comprehensive longitudinal record for each patient. This should help care teams better manage the risks that can lead to unplanned admissions. It should also enable service designers to do a better job forecasting, planning, and evaluating the effectiveness of their services. This was the value proposition behind the NELFT POC.

Clinical Requirements for a Big Data Solution

NELFT uses two EMRs, RiO and SystmOne, as well as a robust business intelligence platform that relies on structured and often coded (thus latent) EMR data. Clinical requirements for information are often expressed in terms of need for integrated care records that go beyond the coded EMR and commissioned-care pathway data sets and present data in a more timely way to integrated and yet virtual teams.

Looking at next-generation analytics that could add value to these platforms, NELFT clinicians liked the idea of having a single source of clinical and operational truth in a secure, web-based environment. To optimally impact clinical and operational practice, they wanted such a solution to:

- Provide a complete overview of all the organisation knows about the patient
- Highlight important information about the patient and identify patterns that might not have been visible previously
- Alert them to situations where the patient may be at increasing risk of an adverse event
- Deliver information to clinicians in a consumable format so they can use it to inform care, identify unmet needs, prioritise their caseloads, and optimise their time with patients

The NELFT Performance Team worked with analytics innovator Santana BDA to customise and explore Santana's next-generation clinical analytics solution. Based on natural language programming (NLP), Santana Big Data Analytics combines case notes mining with predictive modelling algorithms. The solution converts text to concise, reliable, and consumable business and clinical intelligence by ingesting case notes, processing them through Santana's advanced analytics engine, and merging the results into an organisation's clinical and operational software, clinical portal, or business intelligence platform for viewing on a desktop, laptop, or mobile device.

The Santana solution thus makes it possible to generate large amounts of coded clinical data without additional data entry requirements. It also provides ways to readily combine information from different electronic records systems, such as the RiO* and SystmOne* platforms in use at NELFT. Santana is headquartered in Lincolnshire and have extensive experience and domain expertise in NLP for healthcare.

POC Description

NELFT want to use case notes mining and predictive modelling algorithms to help predict risk, optimise the design of care protocols based on a deeper understanding of patient needs, and inform the design of integrated care pathways. From the voluminous unstructured case notes, they wanted to surface vital information regarding a patient's social support network, tendency to self-harm, reports of loneliness, coping issues, interactions with primary care colleagues, and other matters. They wanted to gain insights by bringing this information together with the structured data set. To help clinicians focus quickly on what is important, it would be important to present the results in an easily understood fashion in a clinical viewer that could facilitate early clinical and operational interventions.

In discussion with clinicians, several themes emerged and were taken into account when designing the user interface. These included the need to:

- Summarise key words and phrases and cluster them into themes such as suicide risk, to support predictive modelling
- Blend structured and unstructured data into a longitudinal view showing service interventions and fluctuations in the patient's wellness over time
- Assess the influence of near and extended family and friends in wellness
- Apply clinical best practice models to review compliance with good practice

The team decided to analyze Mental Health Service notes to identify significant life events and highlight them for the clinician. Understanding these events could help the clinician provide better care for the patient and better manage risk. The analytics used a variety of clinical scoring algorithms and data classifiers based on recognised, high-quality, academic and clinical research such as the Bournemouth score, Holmes and Rahe stress inventory, the U.S. Veterans suicide risk, and the Health of the Nation Outcome Scales (HoNOS) instruments. Additional analytic descriptors, including terms to help identify indicators of abuse, were drawn from the clinical expertise of NELFT and Santana team members.

The Santana NLP engine can also support machine learning. For example, some frailty markers that were not part of the Bournemouth scoring were seen to appear regularly in patients whose records indicated a high Bournemouth score. Those markers, if clinicians feel they are clinically valid and relevant, can be used to augment the analysis and expand the analytic dictionary.

Putting the Solution to Work to Enhance Clinical Decision Support

Data Analysis

The trust liaises with a number of social care services, GPs, hospitals, and other systems supporting the activities of the trust, and NELFT wanted to better understand the touchpoints patients had with each service. For the POC, the team processed the patient records of 6,335 individuals from three Adult Community Services organisations in Barking & Dagenham, Havering, Redbridge, and Waltham Forest. From the unstructured clinical notes, the Santana NLP engine tagged 1,571 life events, 1,554 indicators of abuse, and 1,377 indicators of frailty. The team used the patient identity numbers as part of the Master Patient Index, and established governance protocols allowing patient-level data to flow at scale from the EMRs to the big data appliance, so they could be processed and the results potentially fed into NELFT's BI portal.

Data Presentation

Analytic results were presented in a clinical decision-support viewer using a variety of formats. For example, a tag cloud summary (Figure 1) provides a weighted view of key words and phrases that appeared in the patient's clinical notes. Clicking on an element in the tag cloud shows the clinician the note or notes that generated the tag. Figure 1 also shows a timeline that provides a quick mapping of a patient's significant life events against important structured metrics, as well as client contacts in the RiO record.

Clinicians can set filters and bump items up or down in importance depending on whether they judge them to be valid and significant for this patient. This flexibility can aid in avoiding false positives and false negatives. The viewer can be customised and can become a potential component of an enterprise clinical portal.

Summary tables allow easy viewing of aggregate data. Tabular presentations for the POC offer summaries of data by patients, services, clinicians, classifiers, and other categories. Figure 2 shows a patient summary produced by the Santana NLP solution. The bars identify areas of significant risk, such as a serious risk of falling, abuse, suicide, or mental health incident, calling them to the clinician's attention. Other views allow for drilling down into the indicators that led to the risk score.

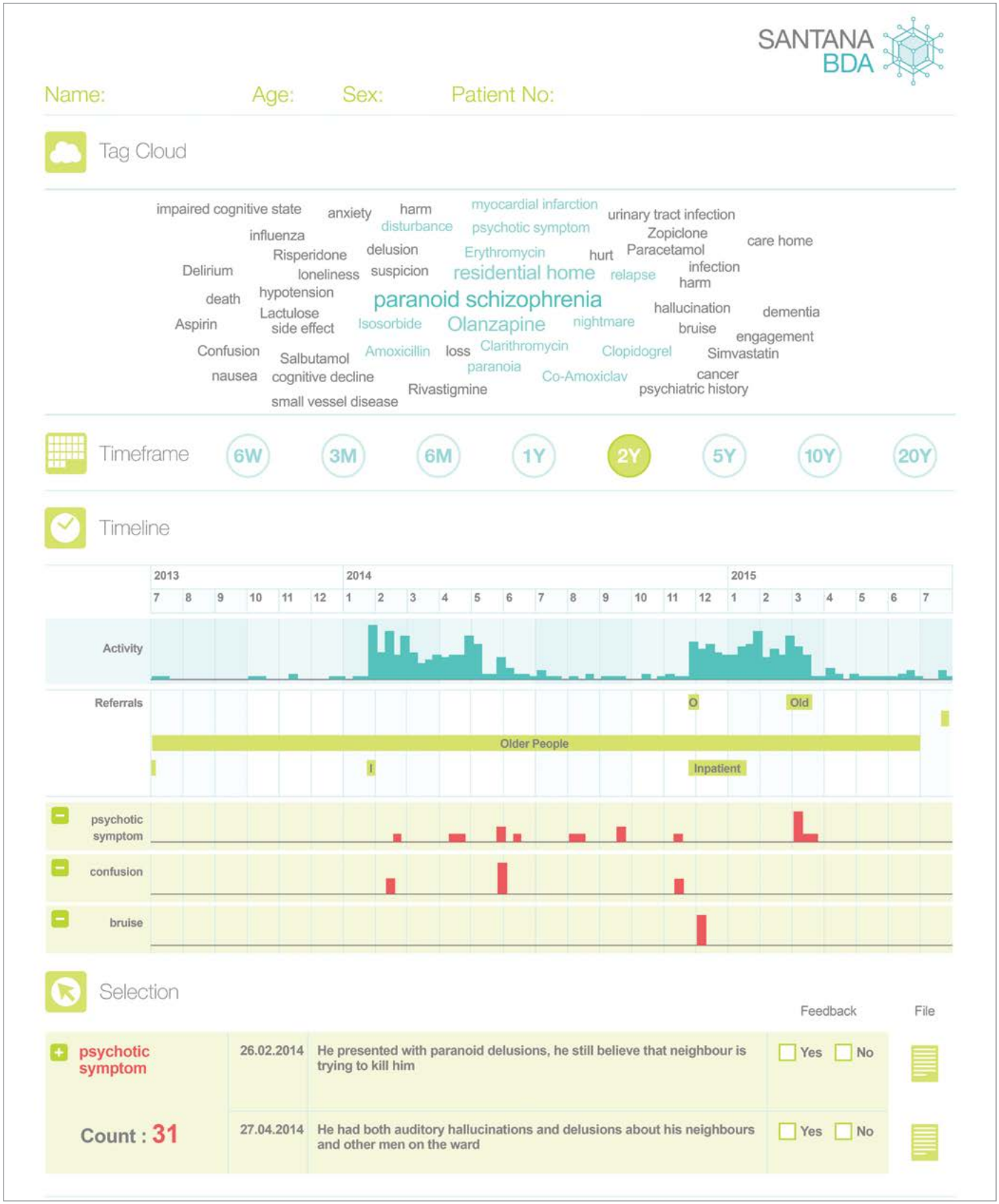
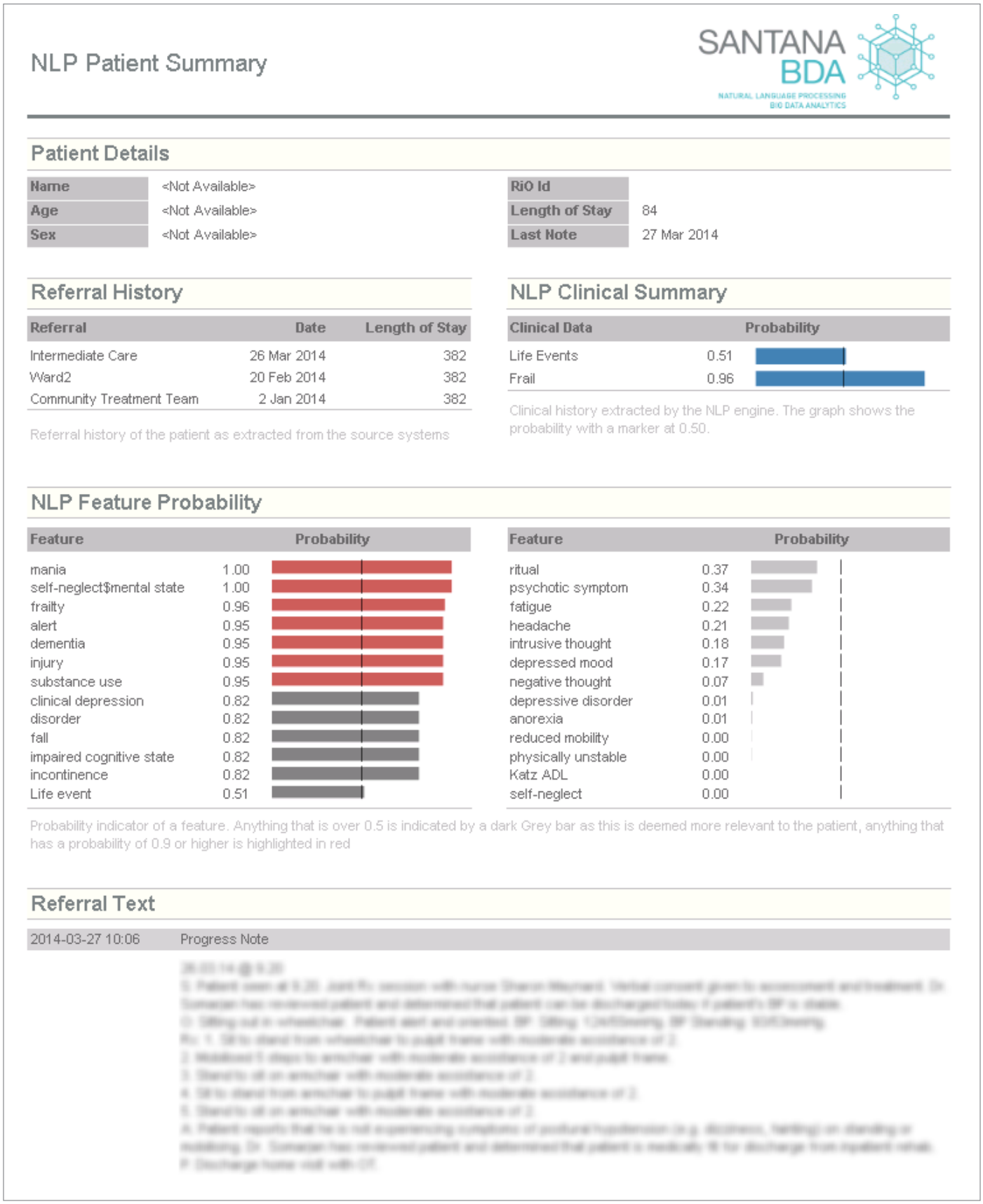


Figure 1. Tag cloud and timeline example.



NLP Patient Summary

SANTANA BDA
NATURAL LANGUAGE PROCESSING BIO DATA ANALYTICS

Patient Details

Name	<Not Available>	**RiO Id**	
Age	<Not Available>	**Length of Stay**	84
Sex	<Not Available>	**Last Note**	27 Mar 2014

Referral History

Referral	Date	Length of Stay
Intermediate Care	26 Mar 2014	382
Ward2	20 Feb 2014	382
Community Treatment Team	2 Jan 2014	382

Referral history of the patient as extracted from the source systems

NLP Clinical Summary

Clinical Data	Probability
Life Events	0.51
Frail	0.96

Clinical history extracted by the NLP engine. The graph shows the probability with a marker at 0.50.

NLP Feature Probability

Feature	Probability
mania	1.00
self-neglect$mental state	1.00
frailty	0.96
alert	0.95
dementia	0.95
injury	0.95
substance use	0.95
clinical depression	0.82
disorder	0.82
fall	0.82
impaired cognitive state	0.82
incontinence	0.82
Life event	0.51

Feature	Probability
ritual	0.37
psychotic symptom	0.34
fatigue	0.22
headache	0.21
intrusive thought	0.18
depressed mood	0.17
negative thought	0.07
depressive disorder	0.01
anorexia	0.01
reduced mobility	0.00
physically unstable	0.00
Katz ADL	0.00
self-neglect	0.00

Probability indicator of a feature. Anything that is over 0.5 is indicated by a dark Grey bar as this is deemed more relevant to the patient, anything that has a probability of 0.9 or higher is highlighted in red

Referral Text

2014-03-27 10:06	Progress Note

Figure 2. Patient summary report.

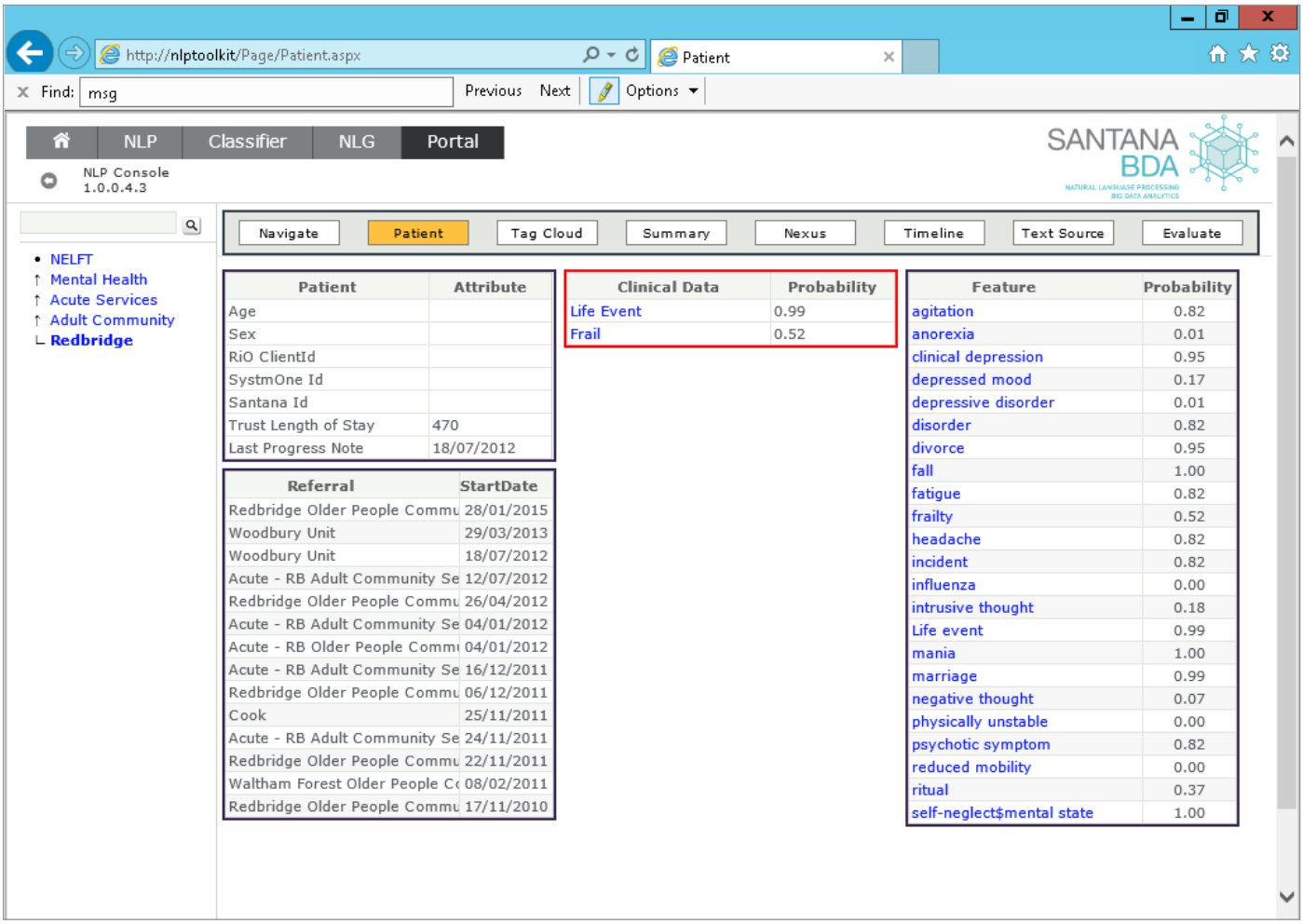


Figure 3. Clinical classifications.

Figure 3 depicts one patient's clinical classifications with the Bournemouth attributes identified by the NLP engine on the right side and data from the structured records on the left.

Santana used Microsoft SQL Server Reporting Services (SSRS) to generate a report and dashboard that can be included with NELFT's standard reporting tools and used for planning purposes (Figure 4).

Technology Behind the Analysis

The Santana Big Data Analytics engine (Figure 5) is architected to run in a secure cloud or server cluster running on premises or externally. The initial implementation of Santana NLP engine used SQL Server* technology to process the data. This worked well at NELFT for processing batches of 100,000 patient records. To create a solution that can process larger volumes of historical data, the Santana team are working with Cloudera to utilise the power of Apache* Hadoop.* They have implemented the NLP engine as a scalable appliance running on Cloudera Distribution for Apache Hadoop (CDH) Enterprise. Both implementations run on scalable infrastructure powered by Intel® Xeon® processors.

Apache Hadoop is an open-source software framework that allows massive data sets to be distributed and processed across clusters of computers. Hadoop offers a flexible, affordable, and scalable platform for analyzing unstructured data, as well as for other

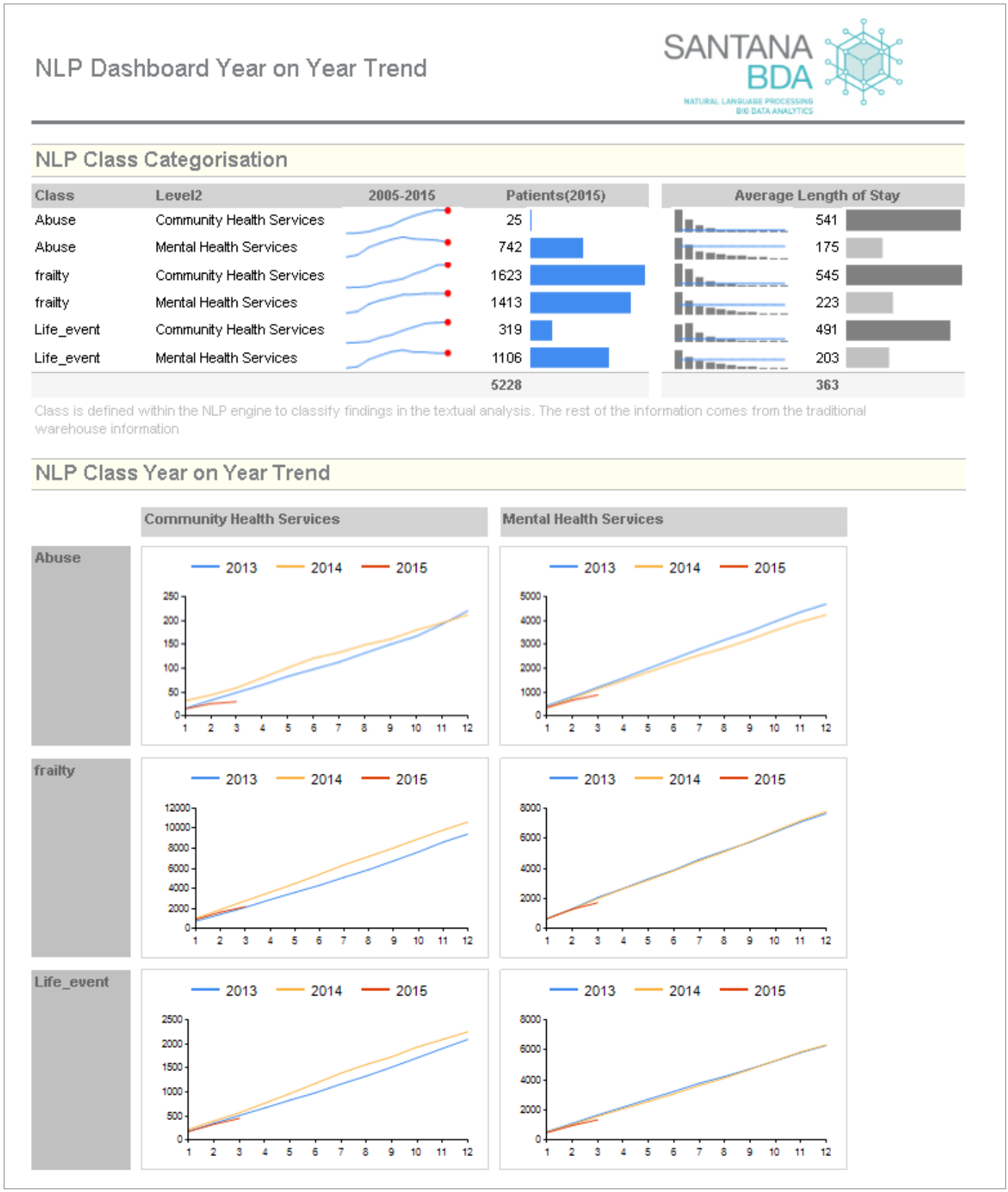
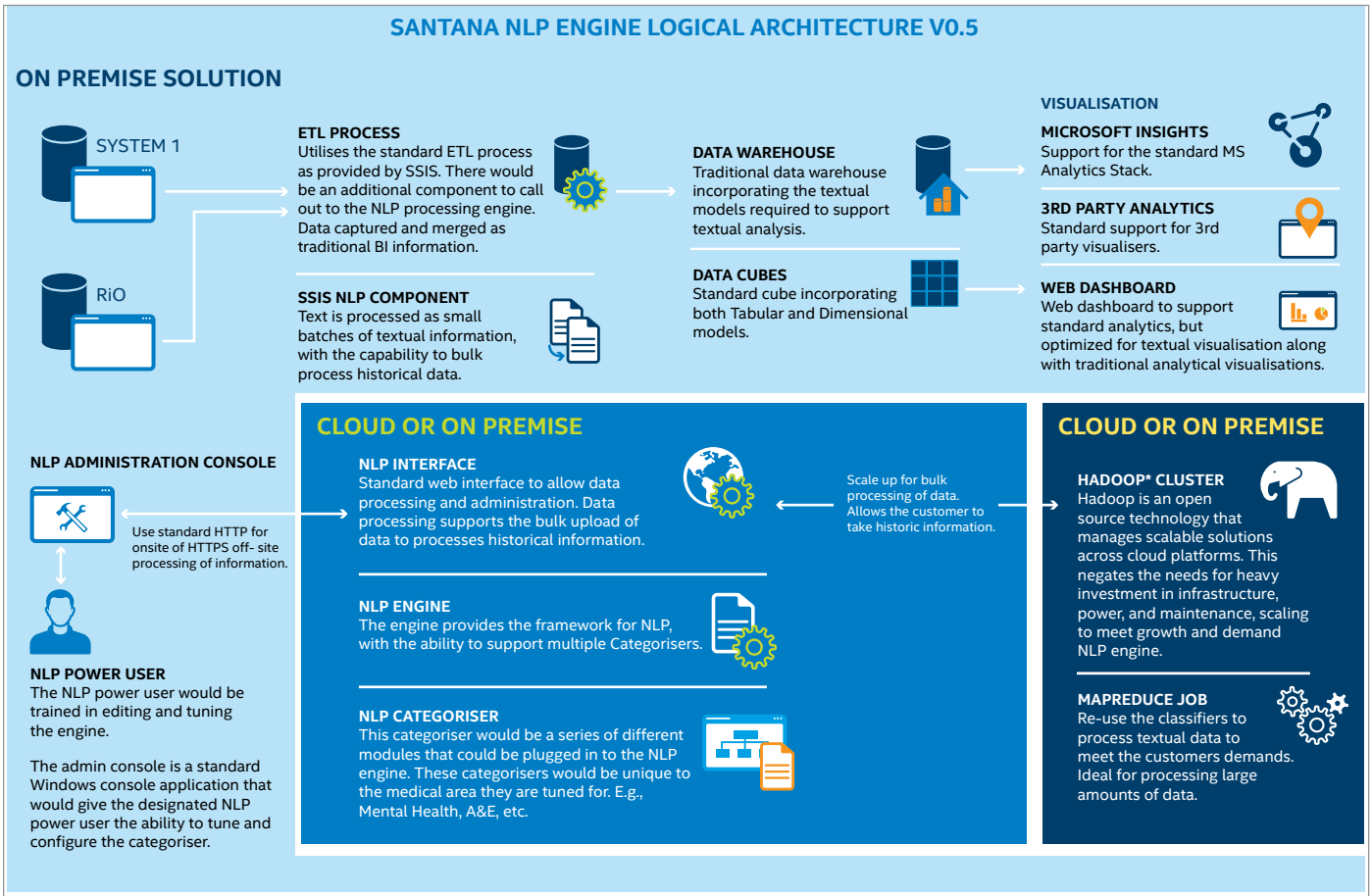


Figure 4. NELFT planning dashboard from the Santana analysis.



ETL PROCESS
ETL utilises standard out of the box SSIS functionality, with the ability to hook in to the NLP engine and process textual information along with traditional BI enhancing the standard ETL toolkit.

NLP INTERFACE
The NLP interface is designed to support on site and offsite configuration. The interface can be used for both processing text and NLP administration.

NLP CATEGORISER
The categorisers have been defined to fit in to functional areas of health, with the functionality to tune and optimise for an individual customers needs.

WEB DASHBOARD
The web based reporting solution is tuned for reporting against Word analytics. With the ability to be integrated in to 3rd party application enhancing the existing analytics capabilities.

SSIS NLP COMPONENT
Bespoke NLP component used to transform textual information through the NLP engine endpoint providing the access to the various text categorisers.

NLP ENGINE
The NLP engine is a host for the NLP interface and Categorisers, with the capability to switch in different categories and different engines to manage speed and efficiency of the Engine.

NLP ADMINISTRATION
This would be a dot net application used to configure the NLP engine and add Categorisers. This allows Power Users to be trained in tuning and optimising the solution.

HADOOP* CLUSTER
The Apache Hadoop software library is essentially a framework that allows for the distributed processing of large datasets across clusters of computers using a simple programming model. Hadoop can scale up from single servers to thousands of machines, each offering local computation and storage.

Figure 5. Santana Big Data Analytics information architecture.

analytics scenarios where the velocity, volume, and variety of data make them impractical for traditional databases. Cloudera CDH provides enterprise capabilities for Apache Hadoop processing along with system management capabilities that make it well-suited to deployment in healthcare and other enterprise environments.

The close collaboration among NELFT, Santana, and Intel, coupled with Santana methods and tools, meant that the appliance could be installed quickly in the NELFT infrastructure. Within a few weeks, the team had gathered the data, and the Santana engine on the Cloudera and Intel appliance churned through records in seconds that would have taken months of labor to read and analyse manually. The appliance was placed within a controlled sub-domain, with NELFT controlling the domain boundaries. Intel also supported the POC by providing hardware, working with Santana to optimise data analytics throughput on the Cloudera solution stack, and advising on matters such as user interface design and clinical vocabulary.

POC Results

The proof of concept validated the ability to derive meaningful decision-support data from clinical case notes and combine it with information from the EMR. When the POC team demonstrated the results to clinicians and other NELFT staff, the responses were positive and enthusiastic regarding both the analytic results and the ways of presenting the results to the clinician or planner. The timeline was viewed as a quick and compelling way of tracing patterns, for example, seeing the point at which a patient went into crisis and identifying previously unidentified factors that might have influenced the crisis and its outcome. Bringing together significant social and medical events in a patient's life provided the potential for greater insight and, in turn, for care that would be more personalised, comprehensive, and coordinated across services.

Discussions with clinicians, nurses, senior staff, and other personnel identified a wide variety of potential benefits, both operational and clinical. These include potential improvements to:

- **Quality of care and outcomes.** Building a more complete longitudinal record and alerting the clinician to significant risks puts the clinician in a better position to identify unmet needs, and provide proactive care that optimises both outcomes and the use of system resources. The analytics solution helps bring the patient's story to life, allowing clinicians to better understand the richness of the individual's life, make better use of each patient encounter, and deliver more individualised care.
- **Productivity and peace of mind.** Clinicians and staff have better information for patient care decisions without spending hours laboriously reviewing past case notes. They can quickly see which patients in their caseloads are most at risk, helping them prioritise their caseloads and feel more confident that they are serving those most in need. Giving clinicians a clearer picture of patients most at risk may also increase clinician and patient safety, helping clinicians identify when they may be about to encounter a patient that is in a highly vulnerable state.
- **Patient satisfaction.** Patients appreciate spending less time repeating the details of their history while also receiving more comprehensive care. If this helps patients remain in their homes and avoid A&E visits, there is a potentially large impact on quality of life for patients and their carers.
- **Operational savings.** Increased productivity and better use of resources can generate operational savings. A more complete longitudinal patient record also provides more comprehensive data for research and planning purposes, making it possible to assess the effectiveness of varied treatment options and services.
- **Health economy savings.** Avoiding A&E visits and unplanned hospital admissions has the potential to improve the use of costly or scarce system resources and generate cost savings for the health economy.

Next Steps

With the initial POC complete, NELFT is working with clinical teams to confirm the validity and utility of the NLP engine's outputs and identify additional use cases. Future directions include exploring the ability to incorporate unstructured information from social care, GP, and acute care services. Building a more complete picture of different events in the patients' lives and a more comprehensive picture of patients' encounters with the various services when incorporated into operational workflows, has the potential to improve planning and patient care for each service.

NELFT sees potential uses of unstructured data analytics to expand areas such as case summaries, performance monitoring, researching the effectiveness of various clinical approaches and treatments, risk screening, and others. Diverse fields may benefit, including mental health, substance abuse, social care, and the care of the frail elderly.

NELFT, Santana, and Intel plan further work to load and process larger volumes of data, add more classifiers, fine-tune the NLP engine, and expand the solution's capabilities. There are also rich opportunities to customise and train the search

engine, adapting it to provide new insights into issues of particular importance to NELFT. The team are also interested in exploring a dashboard-like clinician tool that would overlay the analytic results with the consultant's daily diary, highlighting high-risk areas for the patients being seen that day.

NELFT are also eager to work with clinicians and operational teams to incorporate the Santana BDA solution into their daily operations and build the business case for widespread use. It is through changes to clinical and operational practice, built atop technology advances such as the Santana BDA NLP engine, that NELFT will fulfill their continued commitment to provide the best care possible for the populations they serve.

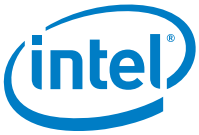
Moving Healthcare Forward

The inability to automate the analysis of clinical case notes and unstructured data has forced clinicians and planners to operate with incomplete information, or to undertake time-consuming and costly manual review of voluminous data. The rapid advances in big data analytics and affordable technology infrastructure provide significant new opportunities to change that situation. Combining those advances with the urgent pressures on the healthcare system, we believe it is time for leading healthcare organisations to explore and deploy next-generation big data analytics solutions to improve clinical care. Based on the NELFT POC, the authors offer the following suggestions to organisations that are joining us on this journey:

- Involve clinicians from outset, and keep them foremost in the planning. Everyone on the team should have the vision that they are engaged in a clinical project, not an IT project. At each decision point, evaluate the impact for clinicians and ultimately for patients. This focus will help ensure the resulting solution meets clinical and organisational objectives.
- Closely related to the preceding bullet, recognise that clinical expertise will be crucial in the utility of a clinical search engine. Choose solutions that have had close clinical involvement in their development, and can be readily customised to match local terminology. Include your own clinicians in expanding the search dictionary and grammar and "training" the engine to pull out terms of interest.
- Plan for the infrastructure impact. The volumes of data to store and process are considerable, and will only grow, so performance, scalability, throughput, and security must be high priorities. A secure private cloud, whether deployed on premises or housed with an external cloud service provider, can provide the necessary scale and performance. Ensure that the cloud uses robust, enterprise-grade security technology and processes to analyse records on-the-fly and safeguard the privacy of confidential data.
- Collaborate. No one should have to reinvent the wheel. Let's all talk with our peers, learn from each other, and move forward together. Our patients will be the biggest beneficiaries.

Learn More

- Find out more about NELFT
- Read about Santana Big Data Analytics
- Read about Intel® technologies for big data in healthcare.
  Join the conversation in the Intel Health and Life Sciences community.



[1] Candace Imison, Emmi Poteliakhoff and James Thompson. Older People and Emergency Bed Use: Exploring Variation. The King's Fund, 2012, http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/older-people-and-emergency-bed-use-aug-2012.pdf.

Printed in USA 1215/DW/HBD/PDF ♻ Please Recycle 333434-001US