clinithink

CLiX ENRICH

White Paper

The Application of CNLP
(Clinical Natural Language Processing)
for Improved Analytics

Clinithink Limited
Version: 1.0 US
Published: November 2014

Contents

Document purpose and intended audience

This document provides an explanation of Clinithink's CLiX ENRICH Data Abstraction Platform. Its intended audiences are potential partners and customers of Clinithink looking to evaluate the technology and understand how it supports a variety of use cases and business models. Some basic knowledge of health informatics is assumed.

The problem

Change before you have to

Globally, healthcare is undergoing profound changes. These changes are driven primarily by the increasingly challenging economics of conventional healthcare delivery. In order to not only survive, but prosper in this new "ecosystem," you have to have access to data. With the adoption of electronic heath records (EHRs) and other digital systems, there has been an explosion of structured data. However, 80% of crucial clinical data is locked up in unstructured documents and reports. Clinical Natural Language Processing (CNLP) solutions such as CLiX ENRICH are a fast, efficient and effective way of unlocking this extremely valuable data.

According to the Economist Intelligence Unit, healthcare costs will increase by 5.3% globally between 2014 and 2017.[1] And as of 2013, these costs consume on average 10-11% of GDP.[2] In the United States, healthcare costs consume roughly 17-18% of GDP and are predicted to grow to almost 20% by 2023.[3]

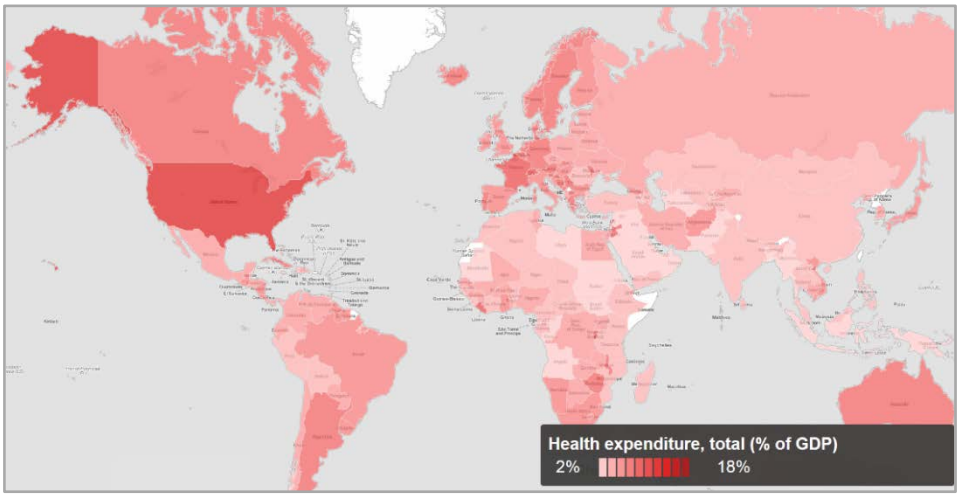

Figure 1

Added to this, aging populations, increases in chronic diseases and increasing life expectancies are increasing the burden on the healthcare systems across the world.

[1] World Healthcare Outlook, Economist Intelligence Unit, August 14, 2013; http://www.eiu.com/industry/Healthcare
[2] World Bank, http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS?display=graph
[3] Center for Medicare Services (CMS); http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/Proj2013.pdf

This rapid growth in cost is simply unsustainable and has to be dealt with in a comprehensive, systematic way. Current practices, processes and technology have to be altered to ensure they optimize quality care delivery, while reducing cost.

Consequently, because of this looming economic crisis, there have been significant changes to the healthcare landscape.

Governments are pushing regulations that increase overall access to healthcare, thus increasing the number of patients moving into the healthcare system. However, physician growth will not be able to keep up with the demand for services. This is forcing caregivers to adopt new processes and technology that improve efficiency.

Healthcare providers are increasing their adoption of information technology used to document patient care and assist with clinical guidance. The adoption of these solutions has led to an explosion of data that can be cultivated to solve an array of problems and answer questions that before were unanswerable. Sharing of data, or "interoperability" among providers is essential to ensuring the best level of care, at the lowest cost.

Reimbursement models are changing due to economic drivers and governmental regulation. The focus is shifting from reimbursement based on the number of times a patient is seen to patient and population outcomes. This forces physicians and hospitals to alter the way they have conventionally treated their patients, placing a greater emphasis on improving outcomes.

In order to survive in the face of challenges posed by this evolving healthcare ecosystem, new technologies are emerging to analyze vast amounts of data. These new technologies adapt business intelligence platforms for use with clinical data for clinical analytics. As Austrian management authority Peter Drucker once said, "You can't manage what you can't measure." Clinical analytics and predictive analytic solutions are emerging as the key to survive and thrive.

Data is the fuel that runs this new healthcare engine. And clinical and predictive analytics solutions need a lot of fuel to run smoothly. But it's not just about large volumes of data. It's about getting quality data and the right data to see the accurate picture.

Clinical information systems typically capture data in structured formats, derived from template-driven user input. They also contain unstructured narrative within progress reports, discharge summaries and operative reports, which typically goes unused. This unstructured content is a rich vein of essential data critical to realizing the benefits of "new healthcare." Most systems lack the ability to analyze this unstructured data which comprises approximately 80% of the meaningful data contained within the patient record.

In order to maximize the value of healthcare changes, organizations need a solution that can create meaning from the unstructured data and is simple to integrate to produce the data in a usable format.

The solution: CLiX ENRICH

Accessing and querying unstructured clinical data

Dealing with the new complexities of healthcare requires access to data, which typically resides in both structured and unstructured forms. Accessing both is essential. Most computer systems can effectively handle structured data, but struggle with access to unstructured data.

Natural Language Processing (NLP) takes this unstructured content and translates it into structured data. NLP solutions have been around for over two decades and have been used to solve very difficult problems in a variety of business areas. With the recent explosion of healthcare data over the last five years, NLP has now established a role within healthcare.

However, healthcare has a vocabulary that is unique and complex. It's not enough just to parse unstructured data using NLP. To extract true insights from the healthcare data, an NLP solution has to associate this newly structured data to a standardized clinical classification and coding system, such as SNOMED CT. This gives way to true Clinical Natural Language Processing solutions, or CNLP.

Lastly, providing access to the data is essential and has paved the way for technology solutions that can handle massive volumes of data. Storing and persisting this data for retrospective analysis, predictive modelling and the like is a 'must have' for any solution.

To address some of these tough business and clinical problems in healthcare today, Clinithink has developed the CLiX ENRICH solution enabling users to import unstructured clinical narrative data, process it using CLiX CNLP (Clinical Natural Language Processing) platform and then interrogate the processed output using Clinithink's powerful abstraction and query platform, CLiX Query. The resulting data is then stored for use by the consuming system or can be integrated into a business intelligence platform or enterprise data warehouse.

In addition, because CLiX CNLP provides default output in SNOMED, a standard clinical terminology, the input narrative can be normalized as a basis for comparison across patients and populations. Users can develop and run their own queries against the data to meet individual business needs.

Understanding that the healthcare landscape can change rapidly, Clinithink has taken a modular approach to CLiX ENRICH. This provides Clinithink and its partners with the ability to customize a solution based on their needs. Customization is at the heart of CLiX ENRICH comprised of six modules which can be configured to respond to specific customer use cases.

A diagrammatic representation of CLiX ENRICH is provided in Figure 2.

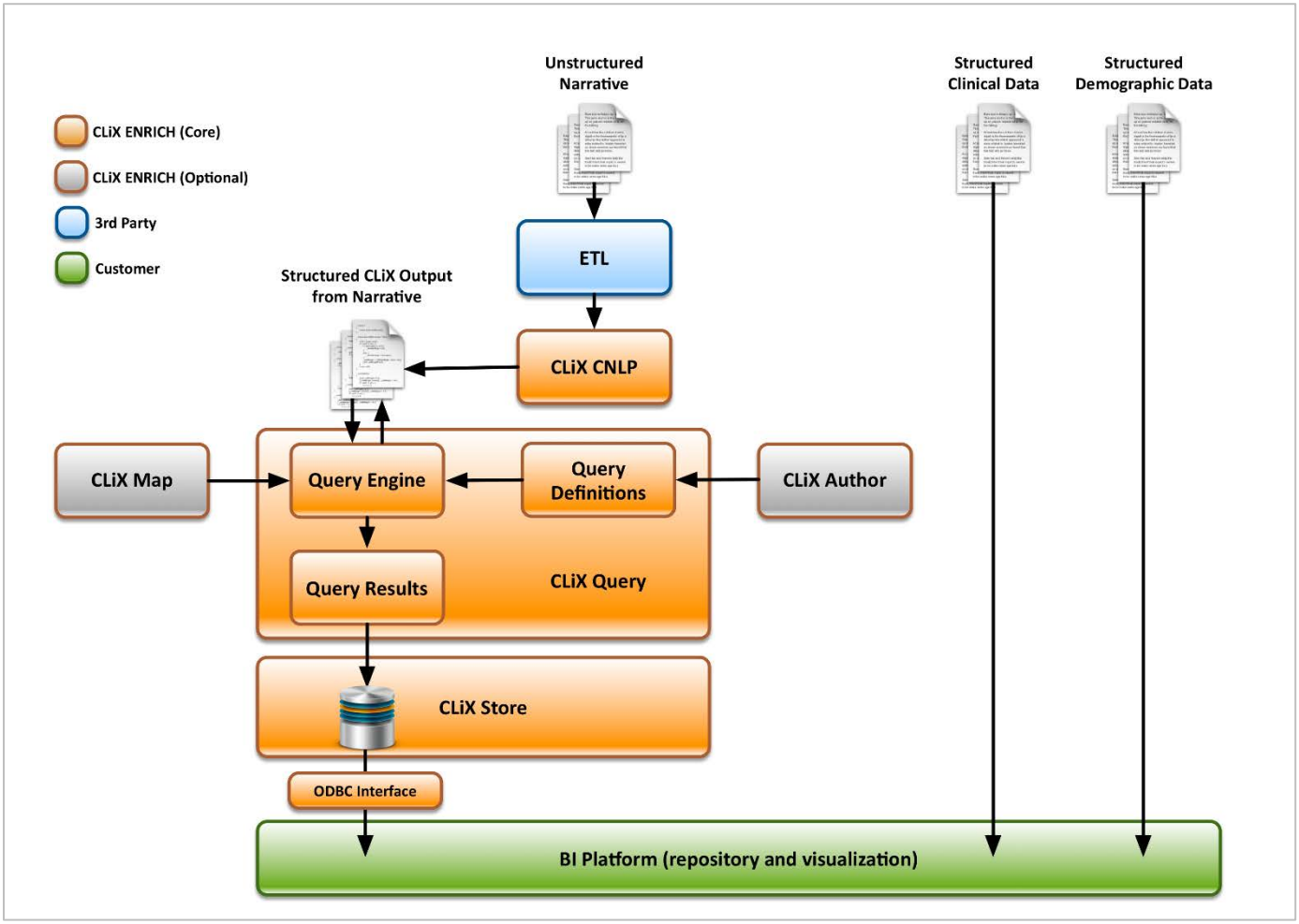

Figure 2

Modules

CLiX CNLP

At the heart of CLiX ENRICH is Clinithink's market-leading CLiX CNLP platform that processes the input narrative and outputs rich, granular SNOMED coding corresponding to the clinical concepts described in the narrative. CLiX CNLP utilizes the SNOMED standard which models 1.8 million granular clinical concepts, combined with Clinithink's extensible library of over 100,000 acronyms, synonyms and abbreviations to enhance concept recognition. Optimization tools enable users to tune the solution to improve baseline recognition by mapping acronyms and colloquialisms unique to the originating source of the data.

For more information on CLiX CNLP technology, refer to Clinithink's White Paper, "Interpreting and Analyzing Clinical Language."

CLiX Query

CLiX Query is a powerful module that allows the user to filter vast amounts of data in order to answer specific questions. CLiX Query does this via the execution of pre-defined queries against the structured output from the CLiX CNLP module. The resulting data is presented to the user or consuming system based on the criteria of those queries. However, in many instances the clinical concepts abstracted from the unstructured narrative contains valuable data points that are essential to rich clinical analysis. CLiX Query abstracts all clinical concepts and values associated with those concepts such as blood pressure, A1C values, etc., from the unstructured narrative.

CLiX Store

As the name implies, CLiX Store provides the document storage capability for data at various stages of processing. CLiX Store saves the original source data, pre-processed text formats, data encoded by CLiX CNLP and the analyzed data from the CLiX Query module.

The data is then published from CLiX Store to the customer's database, which is based on the requisite data schema needed by CLiX ENRICH and provided prior to CLiX ENRICH implementation. There are multiple tables in the destination schema and the diagram below shows the most important data sets. These data sets are extensible by creating new queries designed around specific customer requirement.

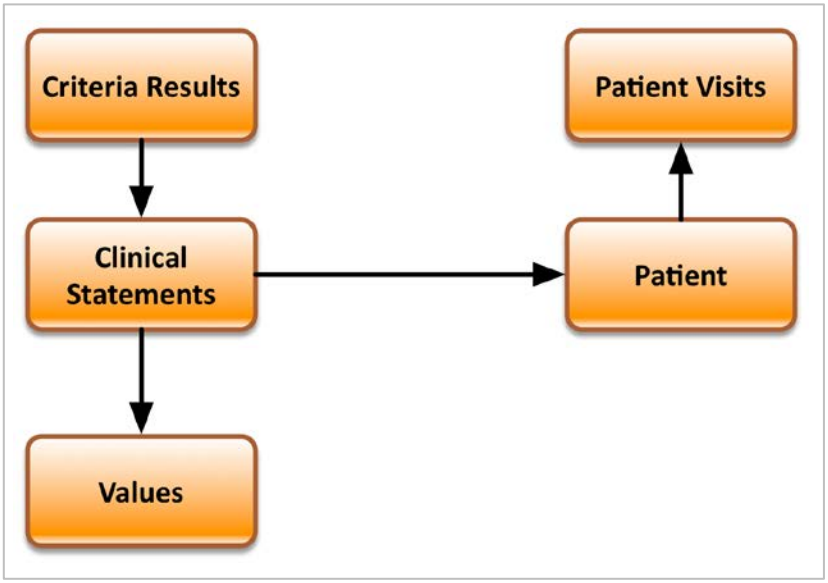

Figure 3

CLiX Store also includes the ability to publish the query results via ODBC (Open Database Connectivity) to relational database management systems (RDBMS) platforms. This RDBMS can be integrated with other data sources in a business intelligence stack.

CLiX Connect

CLiX ENRICH requires that input data be an HL7 message with the subject text embedded within it. Also, CLiX ENRICH can work with a customer's existing ETL capability. There are however instances where a customer might not have the need nor the desire to integrate with their existing ETL platform. In this case Clinithink can provide CLiX Connect to facilitate the extraction of free text from source input systems for use by CLiX ENRICH.

CLiX Map

CLiX ENRICH outputs data using the SNOMED CT clinical classification. CLiX ENRICH can also support representation of data mapped to other terminologies. While SNOMED CT has broad application in representing clinical concepts, there are instances where clinical data needs to be represented in codes used for other purposes, such as ICD-9 and ICD-10.

For example, certain patient safety indicators from the US Agency for Healthcare Research and Quality are based on ICD-9 CM. By using CLiX Map, users are able to extend existing data sets created by CLiX Query to ICD-9 CM data sets. Combining the power of CLiX Author and CLiX Map, users can interrogate the unstructured clinical narrative to calculate patient safety indicators found in discharge summaries, operative reports, progress notes, etc.

CLiX Author

CLiX Author provides the capability to manage query libraries and includes the ability to author new queries. CLiX Author, working in concert with CLiX Query, allows users to create their own custom data sets based on parameters they define.

For example, using the CLiX Author capability in conjunction with CLiX Query, a user can create a query set that defines all diseases associated to the patient. This query set is then used by CLiX Query to create a data set of all diseases, which would be available via CLiX Store and includes a standard ODBC connection to import this new data set into their business intelligence platform. As new data is added to CLiX ENRICH, the abstracted data from CLiX Query would also appear in the existing data set.

If the need arose for a more granular analysis of just heart failure rather than an analysis of all cardiac diseases, using CLiX ENRICH the user would simply create a new query set based on a subset of the parameters of the "all cardiac disease" query set to create a heart failure data set. Again, this new data set would be available via CLiX Store to the customer's business intelligence platform.

Examples

There are any number of use cases that require an approach to unlocking the value of unstructured data. Outlined below are some of these use cases to illustrate how CLiX ENRICH can be used to solve real world problems. This is by no means an exhaustive list of its potential use cases.

Clinical Trials Recruitment

Gathering patients for trials is a costly, time-consuming and predominantly manual process which has an enormous impact on time to market once products have reached Phase 3 clinical trials. The average Phase 3 trial costs $30 mm.[4] On average 13%, or $3.9 mm per study is allocated to trial recruitment.[5] And in some instances, recruitment costs are as high as 50% of the overall budget.

In a 2014 report, Cincinnati Children's Hospital Medical Center conducted an extensive study to evaluate the impact of using a CNLP solution in combination with a data abstraction (query) and machine learning platform to identify candidates for 13 different drug trials.[6] Using structured data for demographic and lab results, and unstructured data from clinical notes, they evaluated the criteria from 13 different drug trials against more than 200,000 patients who visited the emergency department.

When they compared the results of the CNLP solution to a "gold standard," they found that the CNLP solution reduced the workload by 92%, with a 63% accuracy rate. This lead to a 450% increase in trial screening efficiency.

As this shows, leveraging unstructured clinical narrative to identify candidates can dramatically reduce the time needed to gather sufficient patients to make the trial effective thus increasing speed to market for pharmaceutical companies and the contract research organizations running the trials on their behalf.

Population Health Management and Accountable Care

Population Health has been defined as the health outcomes of a group of individuals, including the distribution of outcomes within the group. Improving the health of populations has been the cornerstone of many recent healthcare innovations.

---

[4] "Clinical Development and Trial Operations"; Cutting Edge Information 2013; http://cuttingedgeinfo.com/research/clinical-development/trial-operations/

[5] "Clinical Trial Patient Recruitment"; Cutting Edge Information 2010; http://www.cuttingedgeinfo.com/2010/clinical-trial-budget/

[6] "Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department"; Journal of the American Medical Informatics Association, July 2014; http://www.ncbi.nlm.nih.gov/pubmed/25030032

In addition, as funding and financial models evolve to a value-based rather than a volume-based structure, the importance of identifying and managing high risk patient groups likely to develop complications increases dramatically. By using CNLP to identify high risk groups, based on granular clinical criteria, population health managers can improve the risk profile of their enrolled populations, reducing cost and improving outcomes.

This was illustrated in a recent study by Vanderbilt University and Mount Sinai Medical Center.

Surgical complications comprise more than half of the healthcare adverse events and most are preventable injuries that occur to patients as a result of their medical management and not their underlying disease process. The monitoring of surgical complications starts with data collection. But self-reporting by hospitals grossly underreports safety issues, manual chart reviews are expensive and require an enormous commitment of healthcare resources and administrative data using codes designed for billing are of limited clinical value.

The goal of the study was to develop a Post-Operative Event Monitor (POEM) to detect surgical complications using EHR data. The researchers analyzed 8,186 surgical procedures on 7,743 patients from 1999 to 2006, targeting nine adverse events, including sepsis, deep vein thrombosis, cardiac arrest and others.

They analyzed structured data, such as demographic, vital signs, labs and pharmacy and nearly 300,000 unstructured reports, such as progress notes, operative reports and discharge summaries. Using a tool that performed NLP and then matched the parsed data to SNOMED CT, they constructed a database of the unstructured content. They then used a series of SNOMED queries to build algorithms that filter the data and present those patients most likely to develop complications.

As this was retrospective analysis, they then compared the results of the data using CNLP to the actual data.

The researchers measured sensitivity (true positive rate or recall rate) and specificity (true negative rate) in the test set. The results of the study are shown in Figure 4 below.

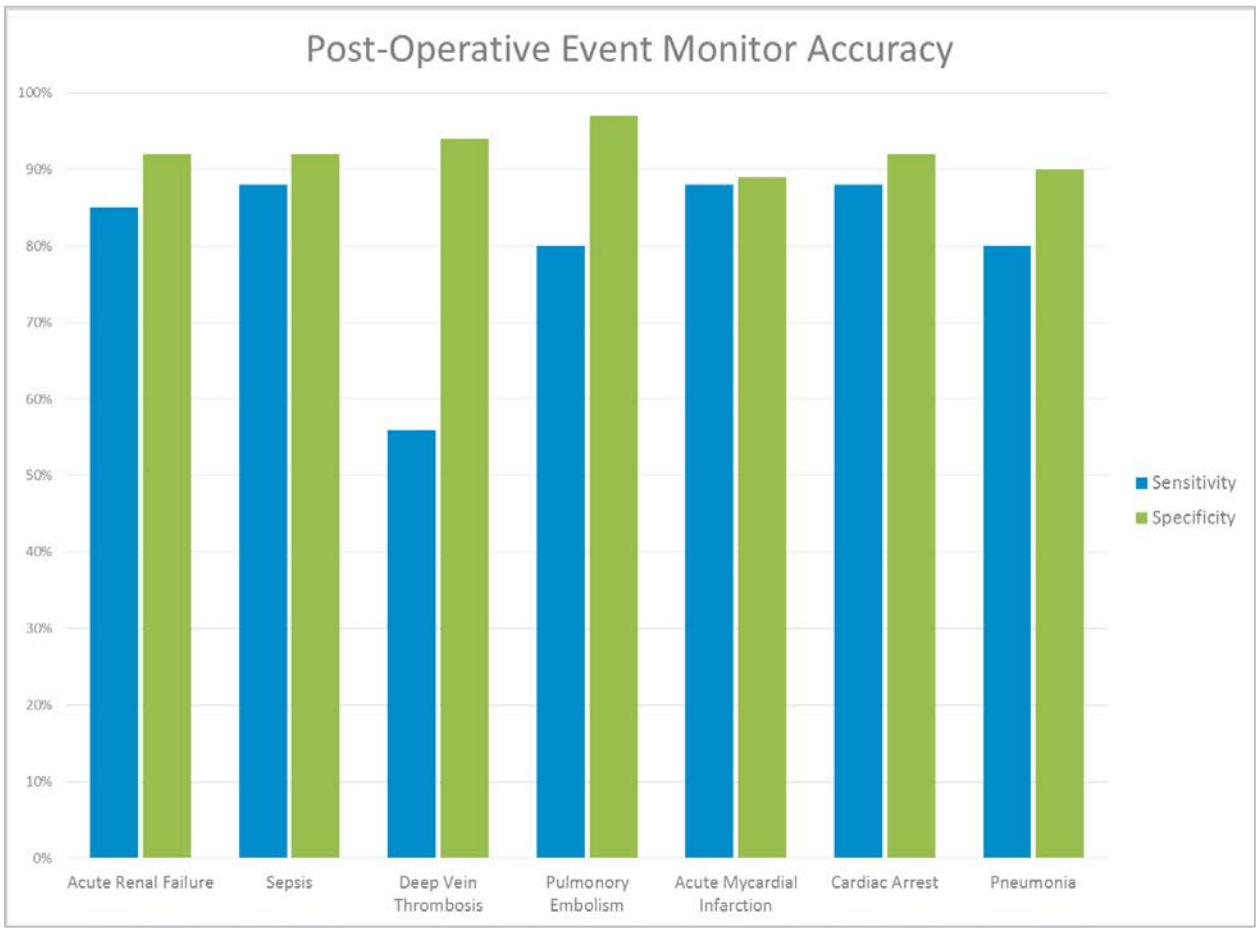

Figure 4

The poor sensitivity score for deep vein thrombosis was the result of a coding error by the clinician, not a fault in the CNLP system.

These results compare to results of a prior study in which a computer assisted algorithm relied only on structured data to identify surgical site infection which had a sensitivity of 38%.[7]

It was their conclusion that algorithms based on structured and unstructured data extracted from the EHR produced respectable sensitivity and specificity across a large sample of patients seen in six different medical centers. This study demonstrates the utility of a solution, such as CLiX ENRICH, that combines CNLP with structured data for mining the information contained within the EHR.

[7] Price CS, Savitz LA. Final Report (Prepared by Denver Health and its partners under Contract No. 290-2006-00-20). AHRQ Publication No. 12-0046-EF. Rockville, MD: Agency for Healthcare Research and Quality; Mar, 2012. Improving the measurement of surgical site infection risk stratification/outcome detection; http://www.ahrq.gov/research/findings/final-reports/ssi/

Clinical Quality Measures

Measuring adherence to clinical quality measures (CQM) has a crucial need for healthcare organizations globally. Adherence to these measures is the keystone of many new healthcare initiatives and are required to determine if changes to the system are truly generating better care.

In one study, US Veterans' Administration, University of Utah and Stanford University wanted to determine how effective a CNLP solution would be to measure ejection fraction values found in the unstructured narrative. Left ventricular ejection fraction (EF) is a key component of heart failure quality measures used within the Department of Veteran Affairs (VA).

A reference standard needed to compare the test results was produced by human annotators using an annotation schema and guidelines, which were also validated by two domain experts. The researchers then determined the document-level classification of EF greater than or less than 40% in two steps. First they used CNLP to extract concepts from the report text and then by using this information to determine the binary classification of whether or not the document contained an EF of <40%.

Researchers looked at 490 records and found that by using CNLP, document-level classification of EF of <40% had a sensitivity (recall) of 98.41%, a specificity of 100%, a positive predictive value (precision) of 100%.

Figure 5

It was their conclusion that an EF value of <40% can be accurately identified in VA echocardiogram reports. An automated information extraction system can be used to accurately extract EF for quality measurement.[8]

Biosurveillance

Biosurveillance typically refers to the ability to provide early detection of and situational awareness to potentially catastrophic biological events. In many cases, agencies such as the CDC (Centers for Disease Control and Prevention) and ECDC (European Centre for Disease Prevention and Control) use both structured data, such as diagnostics codes, as well as unstructured text (chief complaint, progress notes, and discharge summaries) to monitor possible disease outbreaks. All of this data is generated from hospitals and physicians caring for patients and is submitted to these agencies for analysis. Improving accuracy is essential to their mission of early detection and prevention.

[8] Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure, J Am Med Inform Assoc 2012; http://jamia.bmj.com/content/19/5/859.full.pdf+html

Pneumonia is a critically important syndromic condition to monitor, as it can be an indication of other disease processes such as influenza or even anthrax inhalation. In one study, researchers from the Mayo Clinic, Vanderbilt University and the CDC evaluated whether CNLP could be used to identify pneumonia patients by analyzing radiology reports.

They used a CNLP system that encoded the reports to SNOMED CT. They then used a set of SNOMED based queries to specify the clinical concepts in which they were interested. Evaluated first were 400 radiology reports and these were then encoded by physicians as the "gold standard" to be used to evaluate how well CNLP did. Those same 400 reports where then encoded using their CNLP engine.

The accuracy of the CNLP system in the identification of pneumonias was high with a sensitivity of 100%, a specificity of 98% and a positive predictive value (precision) of 97%. They concluded that this SNOMED CT based CNLP system using SNOMED queries was accurate for the automated biosurveillance of pneumonias from radiological reports.[9]

Conclusion

The economics of healthcare are unsustainable and are causing fundamental changes in how healthcare is managed and delivered across the globe. These seismic shifts are placing greater focus on improving care at reduced costs. Consequently, the adoption of technology in all phases of healthcare is essential to survival and success. Understanding the enormous amount of data that is being generated is a cornerstone of the new healthcare ecosystem. Thus, clinical analytics has become essential for governments, healthcare organizations, payers, researchers and others involved in the management and delivery of healthcare.

Historically the focus has been on using structured data. However, 80% of clinical information is contained in the unstructured data, such as progress notes, discharge summaries, operative reports, etc. CNLP is a proven technology that unlocks this unstructured data for use in a variety of different use cases.

Clinithink's CLiX ENRICH platform is purpose built for use in healthcare to solve these complex problems. It leverages Clinithink's CNLP engine along with advanced query, cross mapping and data persistence capabilities to provide the data necessary for analytics, population health management, clinical research and many other uses.

[9] NLP-based Identification of Pneumonia Cases from Free-Text Radiological Reports, AMIA 2008 Symposium Proceedings; http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656026/pdf/amia-0172-s2008.pdf

Atlanta

555 Northpoint Center East, 4th Fl.
Alpharetta, GA 30022, USA

London

2-8 Scrutton Street, 3rd Fl.
London, EC2A 4RT, UK

Wales

4 Derwen Road, 1st Fl.
Bridgend, CF31 1LH, UK

US (+1) 978 296 5275
UK (+44) 292 125 0190
info@clinithink.com

www.clinithink.com