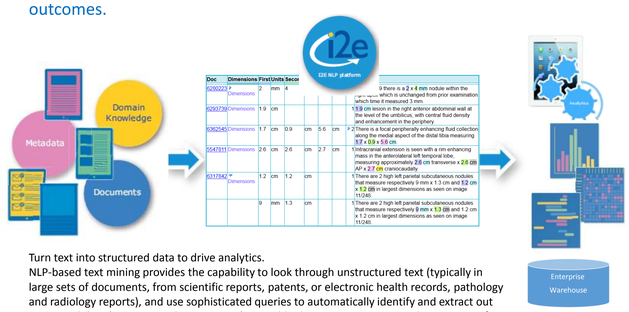
# GAINING BUSINESS INSIGHT FOR CLINICAL TRIALS - TEXT ANALYTICS FOR A DATA DRIVEN APPROACH

## Introduction

Founded 15 years ago, Linguamatics is a world-leader in deploying innovative Natural Language Processing (NLP)-based text mining for high value knowledge discovery and decision support. Our solution, I2E, helps pharma/biotech organisations (including 27 of the top 50) and healthcare industries to speed up drug discovery & development and improve patient outcomes.




Turn text into structured data to drive analytics.
NLP-based text mining provides the capability to look through unstructured text (typically in large sets of documents, from scientific reports, patents, or electronic health records, pathology and radiology reports), and use sophisticated queries to automatically identify and extract out structured data (concepts and associations) to enable the system to interpret the meaning of the text. The end result of this process is a set of structured facts, relationships or assertions, that can be used for decision support

Text mining is the process of deriving high-quality knowledge from unstructured text. Text analytics, particularly NLP, identifies relevant facts and relationships, extracts them in a structured form for review and faster analysis, and connects these facts together in new ways, to synthesise knowledge, and create actionable insights.

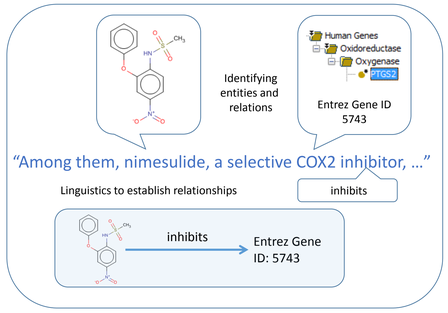
## Challenges and the power of NLP

80% of information at companies is in free text and unstructured data. Most of the answers to critical business questions can be found in free text. The challenge is to extract the key facts needed, rapidly, accurately, effectively.

| | |
|---|---|
| **Different word, same meaning**<br>cyclosporine<br>ciclosporin<br>Neoral<br>Sandimmune | **Different expression, same meaning**<br>Non-smoker<br>Does not smoke<br>Does not drink or smoke<br>Denies tobacco use |
| NLP | |
| **Different grammar, same meaning**<br>5mg/kg of cyclosporine per day<br>5mg/kg per day of cyclosporine<br>cyclosporine 5mg/kg per day | **Same word, different context**<br>Diagnosed with diabetes<br>Family history of diabetes<br>No family history of diabetes |

NLP can understand differences in grammar, expressions, meaning; and interpret sentence structures to reveal the underlying facts and relationships.

In this example sentence, the word "nimesulide" has been identified as a chemical, and I2E can assign a structure. Similarly, COX2 has been identified as a gene or protein, and again I2E can assign an ID, and a place in the Entrez Gene hierarchy. Also I2E linguistic algorithms find the compound-protein relationship, inhibits. This enables the linking, rapidly, accurately, automatically, of specific selected concepts e.g. in this example, chemistry to biology.




## Text Analytics for Clinical Trials

Clinical trials are one of the most expensive parts of the drug development process; thus one of the goals of clinical trial professionals is to increase efficiency along the process. Addressable bottlenecks include improving access to knowledge for site selection, patient populations, principal investigators, and key opinion leaders.

Text analytics and Natural Language Processing (NLP) can extract intelligence from internal and external clinical trial data, enabling clinical scientists to optimise clinical trial design and gain knowledge from legacy data.

## Clinical Trial Optimisation

Drivers
- Deliver clinical trials faster, cheaper, more effectively
- Improve site selection, study design, cohort selection
- Better patient care

Typical questions
- Understand current trial protocols
- Identify key opinion leaders and investigators
- Patient identification and selection

I2E can integrate and interrogate key data from different sources (clinical trials, scientific literature, EHRs) to address these drivers and provide answers to key questions.

## Easy Access to Powerful Text Analytics

Linguamatics provides access to a range of content options, all accessible on the cloud via I2E OnDemand or via our Connected Data Technology for those with I2E Enterprise edition.
ClinicalTrials.gov is included in this knowledge cloud, and novice users can get immediate benefit from this valuable source using I2E's Clinical Trial Query Bundle.

This includes four queries, for stand-alone data extraction or more powerful combination use:
- Indications in Clinical Trials: Finds the diseases covered by the trial.
- Trial Overview: Finds data about clinical trials, including tracking, recruitment and administration.
- Study Design: Finds trial design information, for example the allocation, masking, and intervention model.
- Study Arm: Finds details about the number and type of study arms and their labels.

## Case Study: Clinical Trial Site Selection

Merck's Experimental Medicine division needed to locate a clinical trial site that would be able to conduct gastric bypass trials, with the ability to measure gut peptides before and after surgery. The ideal trial site needed to fit many different criteria (see below). Searching for study sites that fulfil all these criteria would be very labour-intensive to do manually.

- Experience with gastric bypass trials
- Experience with randomised trials
- Not be a children's hospital
- Ability to measured blood or urine biomarkers, gut peptides or other Diabetes parameters
- Experience other than just comparing two methods of bypass
- Can handle a surgical trial rather than a drug trial.
- Experience in methods that are not staples
- Has the ability to handle a large study
- Has a proven track record in producing quality results in a timely manner
- Use a central Institutional Review Board
- Experienced trial coordinator with a proven track record
- Is in one of a specified list of countries

Using I2E to index ClinicalTrials.gov enabled Merck to run specific queries, to extract "gastric bypass" terms from intervention fields, as well as biomarker, gut peptide and other key metrics from unstructured clinical trials record text. The output was a structured summary table which enabled systematic and rapid analysis and ranking of the possible sites.
The search yielded three ideal sites for this trial, one of which was previously unknown to the Merck Experimental Medicine group. I2E integrated searches across ClinicalTrials.gov and MEDLINE, reduced time burden for clinical trial site planning, hence decreasing overall cost.

| | | | | | |
|---|---|---|---|---|---|
| MEDLINE | ClinicalTrials.gov | Patents | NIH Grants | OMIM | Content sources available via I2E OnDemand |
| TR Cortellis | CCC Full Text | FAERS | PMC Open Subset | FDA Drug Labels | |

## Conclusions

Clinical trial records contain valuable data, and text analytics can transform these data assets into structured information that can be used to answer high-value questions in real time. Use cases for clinical trial data include:
- Trial design, protocol modification, site selection
- In-licensing & out-sourcing intelligence
- Competitive intelligence across the industry
- Key opinion leader identification

Linguamatics' agile natural language processing (NLP) text mining software, I2E, enables users to rapidly identify, extract, synthesise and analyse relevant information such as clinical trial site, selection criteria, study characteristics, patient numbers and characteristics that would not be possible using other approaches.

Text analytics for business insights over clinical trial data leads to:
- Enhanced revenues
- Reduced costs
- Reduced risk
- Improved care

## Case Study: Extracting Summary Statistics from Clinical Trial Databases

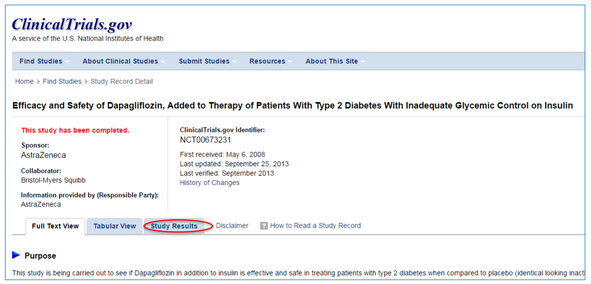
Eli Lilly needed a solution for expediting the extraction, analysis and synthesis of specific statistics from on-going and legacy clinical trial records, for a specific set of therapeutic indications. Doing so would enable the organisation to understand the competitive landscape and focus ongoing research and trial planning.

Dr. Eric Su, Principal Research Scientist at Lilly, utilised I2E to develop automated ways to extract summary statistics on various endpoints from clinical trial databases. Eric indexed two clinical trial databases, Citeline's Trialtrove and ClinicalTrials.gov, and developed queries to extract endpoints for two disease areas (oncology and diabetes).

These queries are now deployed through I2E Smart Query and Web Portal interfaces, so that other researchers and clinical scientists, including non-I2E users, can edit query parameters and generate user-defined output in html or Excel-compatible format.

"I2E provides data that would take 10s or 100s times longer with tedious manual work. It enables downstream calculations to provide insight. Some work would not have been done or not done comprehensively without I2E."

Lilly


ClinicalTrials.gov
A service of the U.S. National Institutes of Health

Find Studies | About Clinical Studies | Submit Studies | Resources | About This Site

Home > Find Studies > Study Record Detail

**Efficacy and Safety of Dapagliflozin, Added to Therapy of Patients With Type 2 Diabetes With Inadequate Glycemic Control on Insulin**

This study has been completed.

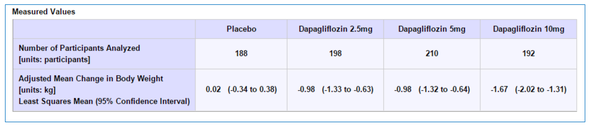
Sponsor: AstraZeneca
Collaborator: Bristol-Myers Squibb
Information provided by (Responsible Party): AstraZeneca

ClinicalTrials.gov Identifier: NCT00673231
First received: May 6, 2008
Last updated: September 25, 2013
Last verified: September 2013
History of Changes

Full Text View | Tabular View | Study Results | Disclaimer | How to Read a Study Record

**Purpose**

This study is being carried out to see if Dapagliflozin in addition to insulin is effective and safe in treating patients with type 2 diabetes when compared to placebo (identical looking inact

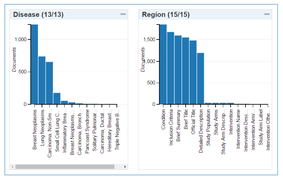

**Measured Values**

| | Placebo | Dapagliflozin 2.5mg | Dapagliflozin 5mg | Dapagliflozin 10mg |
|---|---|---|---|---|
| Number of Participants Analyzed [units: participants] | 188 | 198 | 210 | 192 |
| Adjusted Mean Change in Body Weight [units: kg] Least Squares Mean (95% Confidence Interval) | 0.02 (-0.34 to 0.38) | -0.98 (-1.33 to -0.63) | -0.98 (-1.32 to -0.64) | -1.67 (-2.02 to -1.31) |

| Doc | n | Measure Title | Least Squares Mean | Units | Weeks | Actual Treatment | LowerLimit | UpperLimit |
|---|---|---|---|---|---|---|---|---|
| NCT00673231 | 188 | Adjusted Mean Change in Body Weight | 0.02 | kg | 24 | Placebo | -0.34 | 0.38 |
| | 192 | Adjusted Mean Change in Body Weight | -1.67 | kg | 24 | Dapagliflozin tablet oral 10 mg total daily dose once daily 24 weeks | -2.02 | -1.31 |
| | 198 | Adjusted Mean Change in Body Weight | -0.98 | kg | 24 | Dapagliflozin tablet oral 2.5 mg total daily dose once daily 24 weeks | -1.33 | -0.63 |
| | 210 | Adjusted Mean Change in Body Weight | -0.98 | kg | 24 | Dapagliflozin tablet oral 5 mg total daily dose once daily 24 weeks | -1.32 | -0.64 |
| NCT01294423 | 86 | Adjusted Mean Change in Body Weight | -2.13 | kg | 24 | Dapagliflozin : Dapagliflozin 5mg/matching placebo for Dapagliflozin 10mg oral dose | -2.65 | -1.60 |
| | 87 | Adjusted Mean Change in Body Weight | -0.84 | kg | 24 | Placebo : Matching placebo for Dapagliflozin 5mg/10mg oral dose | -1.36 | -0.32 |
| | 88 | Adjusted Mean Change in Body Weight | -2.22 | kg | 24 | Dapagliflozin : Dapagliflozin 10mg/matching placebo for Dapagliflozin 5mg oral dose | -2.73 | -1.71 |
| | 97 | Adjusted Mean Change in Body | -1.74 | kg | 24 | Dapagliflozin tablet 5 mg once daily | -2.34 | -1.15 |

Illustration of data for clinical trial summary statistics. Top: title page in ClinicalTrials.gov for NCT00673231, a trial of Dapagliflozin for type 2 diabetes. Middle: a snapshot of some of the study results tables from ClinicalTrials.gov. Bottom: structured results table from I2E, showing the adjusted mean change in body weight data extracted from NCT00673231 table, with number of patients, treatment arm, outcome statistic, numeric metrics extracted automatically. Left: structured results extracted from Citeline's TrialTrove database using I2E.

| Pharmacologic Substa.. | Phase | ClinicalTrial Title | Median Survival | Number | Time Unit | Study Arm | | Doc | Hit |
|---|---|---|---|---|---|---|---|---|---|
| Sorafenib | II | A Phase II Multicenter Uncontrolled Trial of BAY 43-9006 in Patients With Advanced Hepatocellular Carcinoma | Median OS | 3.1 | months | responders and non-responders | 1 | 5705 | 1 Median survival (from landmark analysis) was 3.1 months (4.8 months and 3.1 months in responders and non-responders, respectively) |
| | | Sorafenib versus capecitabine in the Management of Advanced Hepatocellular Carcinoma | Median OS | 5.07 | months | the capecitabine group | 1 | 1752991 | Results: Median overall survival was 7.05 months in the sorafenib group and 5.07 months in the capecitabine group (hazard ratio in the capecitabine group 2.36; 95 % confidence interval 1.174-4.74; P < 0.016). |




Results from the I2E Indications query, for clinical trials relevant to either breast cancers or lung cancers. The bar charts (top) enable the user to review and filter the results; the results table (below) provides more detail, with drill-down to the underlying evidence.

| Disease | Region | Doc | #Hits | Hit# | Hit |
|---|---|---|---|---|---|
| Lung Neoplasms | Condition | NCT00268785 | 1 | 1 | Lung Cancer |
| Lung Neoplasms | Official Title | NCT00119613 | 1 | 1 | A Randomized, Double Blind, Placebo-Controlled Study of Subjects With Previously Untreated Extensive-Stage Small-Cell Lung Cancer (SCLC) Treated With Platinum Plus Etoposide Chemotherapy With or Without Darbepoetin Alfa |
| Lung Neoplasms | Condition | NCT00119470 | 1 | 1 | Lung Cancer |
| Lung Neoplasms | Condition | NCT00120939 | 1 | 1 | Lung Neoplasms |
| Lung Neoplasms | Brief Summary | NCT00122239 | 1 | 1 | This study will examine, for the first time, the independent contribution of a patient's own genetic makeup to the development of post-radiation complications, permitting the future development of predictive tests to avoid radiation injury..... prostate, brain and lung cancer survivors who have received conformal ... |
| Lung Neoplasms | Condition | NCT00085813 | 1 | 1 | Non-Small-Cell Lung Cancer |
| Lung Neoplasms | Condition | NCT00002947 | 1 | 1 | Lung Cancer |
| Lung Neoplasms | Brief Summary | NCT00086268 | 1 | 1 | This study will evaluate the effects of an investigational drug in combination with chemotherapy in patients with stage IIIB/IV non-small cell lung cancer ... the combination on progression of lung |

Dr. Jane Reed is the head of life science strategy at Linguamatics. She is responsible for developing the strategic vision for Linguamatics' growing product portfolio and business development in the life science domain. Jane has extensive experience in life sciences informatics. She has worked for more than 15 years in vendor companies supplying data products, data integration and analysis and consultancy to pharma and biotech - with roles at Instem, BioWisdom, Incyte, and Hexagen.

THE QUEEN'S AWARDS FOR ENTERPRISE: 2014

Linguamatics


324 Cambridge Science Park, Milton Rd, Cambridge, CB4 0WG, UK
1900 West Park Drive Suite 280 Westborough MA 01581 USA
www.linguamatics.com jane.reed@Linguamatics.com