Linguamatics NLP Text mining Literature Examples

STM April 2016
Susan M LeBeau, Ph.D.
Vice President, Sales

THE QUEEN'S AWARDS FOR ENTERPRISE: 2014

Linguamatics

About Linguamatics

Software	Consulting	Hosted content

- Agile, scalable, real-time NLP-based text mining
- Fact extraction and knowledge synthesis

Pharma/Biotech	Healthcare	Government
Including 27 of the top 50	Including Kaiser Permanente	Including FDA

Challenges in Unstructured Data

Different word, same meaning

cyclosporine

ciclosporin

Neoral

Sandimmune

Different expression, same meaning

Non-smoker

Does not smoke

Does not drink or smoke

Denies tobacco use

NLP

Different grammar, same meaning

5mg/kg of cyclosporine per day

5mg/kg per diem of cyclosporine

cyclosporine 5mg/kg per day

Same word, different context

Diagnosed with diabetes

Family history of diabetes

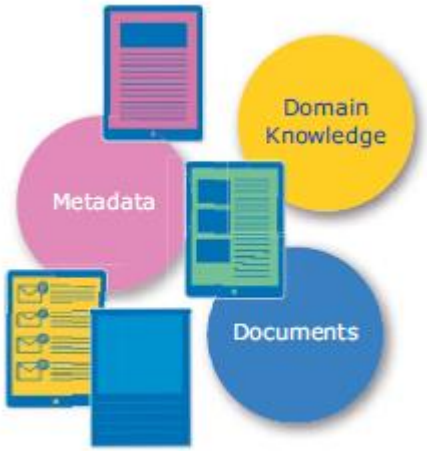
No family history of diabetes

I2E Transforms Text into Actionable Insights

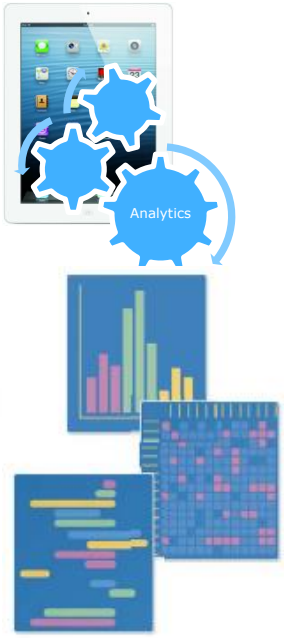
Turn text

Into structured data using sophisticated queries

To drive analytics



Doc	Dimensions	First	Units	Secc					
6280223	Dimensions	2	mm	4					9 there is a 2 x 4 mm nodule within the which is unchanged from prior examination it measured 3 mm.
6293739	Dimensions	1.9	cm						ion in the right anterior abdominal wall at the level of the umbilicus, with central fluid density and enhancement in the periphery.
6362545	Dimensions	1.7	cm	0.9	cm	5.6	cm	2	There is a focal peripherally enhancing fluid collection along the medial aspect of the distal tibia measuring 1.7 x 0.9 x 5.6 cm.
5547811	Dimensions	2.6	cm	2.6	cm	2.7	cm	1	Intracranial extension is seen with a rim enhancing mass in the anterolateral left temporal lobe, measuring approximately 2.6 cm transverse x 2.6 cm AP x 2.7 cm craniocaudally.
6317842	Dimensions	1.2	cm	1.2	cm			1	There are 2 high left parietal subcutaneous nodules that measure respectively 9 mm x 1.3 cm and 1.2 cm x 1.2 cm in largest dimensions as seen on image 11/248.
		9	mm	1.3	cm			1	There are 2 high left parietal subcutaneous nodules that measure respectively 9 mm x 1.3 cm and 1.2 cm x 1.2 cm in largest dimensions as seen on image 11/248.



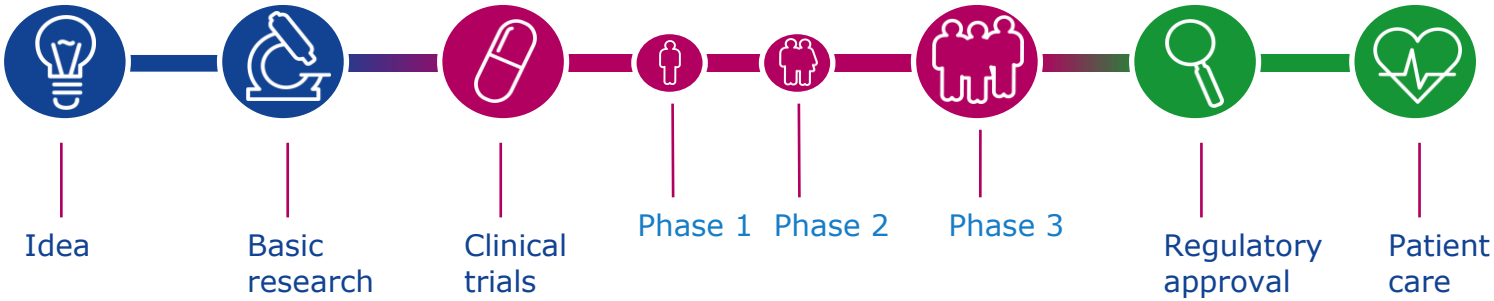
Accurate results: only retrieves relevant results

Complete results: comprehensive and systematic

Enterprise Warehouse

Linguamatics

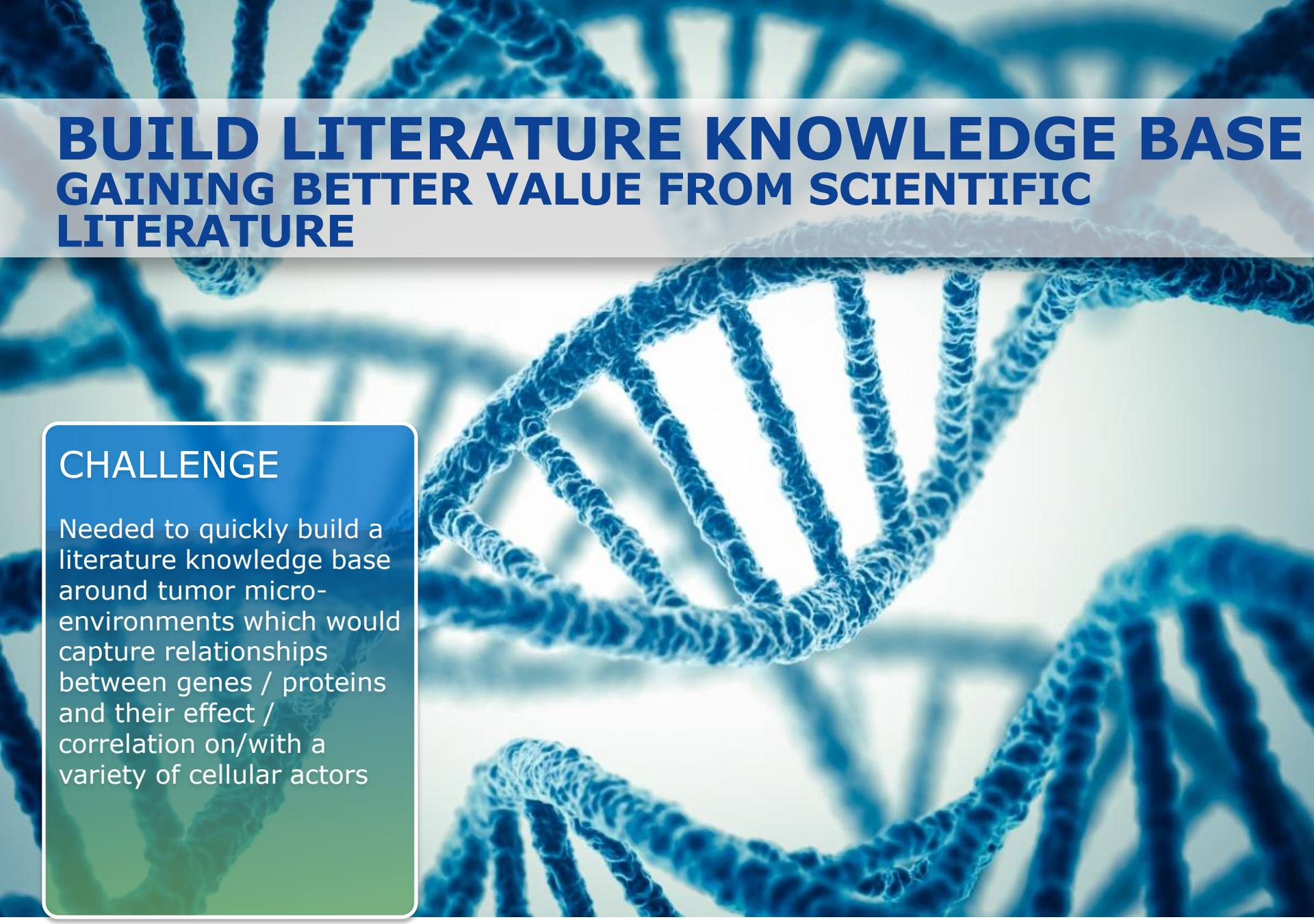
Literature Analytics – *Medline Abstracts*



THE QUEEN'S AWARDS FOR ENTERPRISE: 2014

Linguamatics

BUILD LITERATURE KNOWLEDGE BASE
GAINING BETTER VALUE FROM SCIENTIFIC LITERATURE



CHALLENGE

Needed to quickly build a literature knowledge base around tumor micro-environments which would capture relationships between genes / proteins and their effect / correlation on/with a variety of cellular actors

Challenges: the Customer Viewpoint

- Define the different concepts
  - E.g. 30,000 human genes, their aliases, manage term disambiguation * morphological variations
- Analyse the semantic relationships between the objects including negation
  - Capture the meaning and structure the facts
- Harmonise the vocabulary
  - Ontologies, preferred terms….
  - Flexibility to use customised thesauri, ontologies
- Applicable to 30 million abstract records
  - Queries efficiently executed, remotely, with results retrieved within seconds or minutes
- Complex queries
  - Requires an efficient and user friendly interface to test and tune
- Export in convenient formats for post-processing

BUILD LITERATURE KNOWLEDGE BASE
GAINING BETTER VALUE FROM SCIENTIFIC LITERATURE

CHALLENGE

Needed to quickly build a literature knowledge base on tumor micro-environments which would capture relationships between genes / proteins and their effect / correlation on/with a variety of cellular actors

SOLUTION

Linguamatics I2E provided the ability to run a single query across the entire set of MEDLINE abstracts to extract genes, effects, cell types, phenotypes, and obtain comprehensive results for analysis.

Structured results retrieved within seconds/minutes

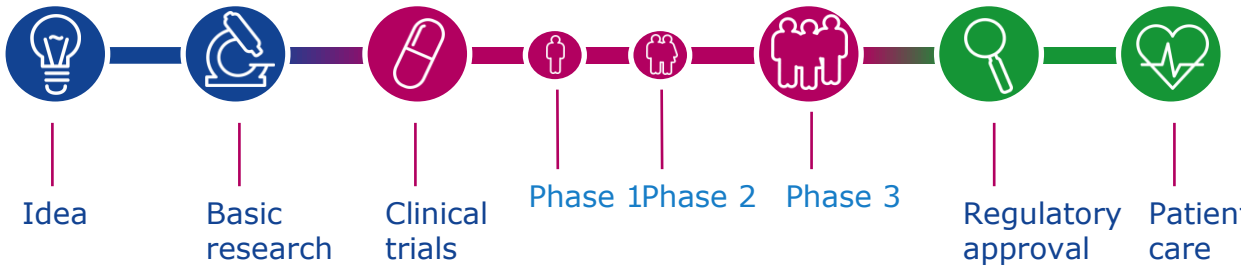
BENEFIT

This equates to ~20 billion unique keyword searches

Rapidly added new knowledge to internal translational science database for direct use in projects

This would have taken weeks or not be possible at all

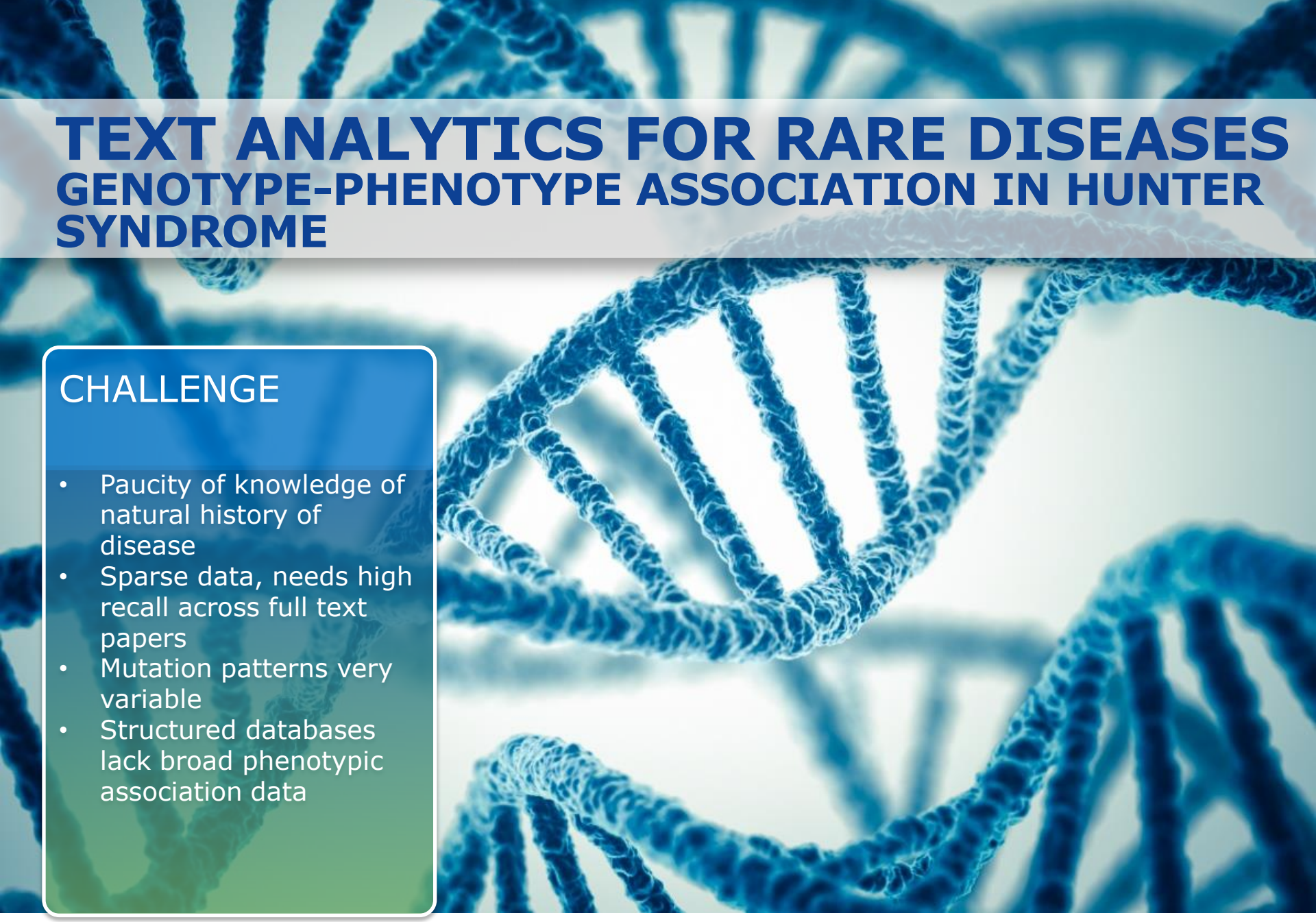
Genotype-Phenotype analytics
Full Text PubMed Central



THE QUEEN'S AWARDS FOR ENTERPRISE: 2014

Linguamatics

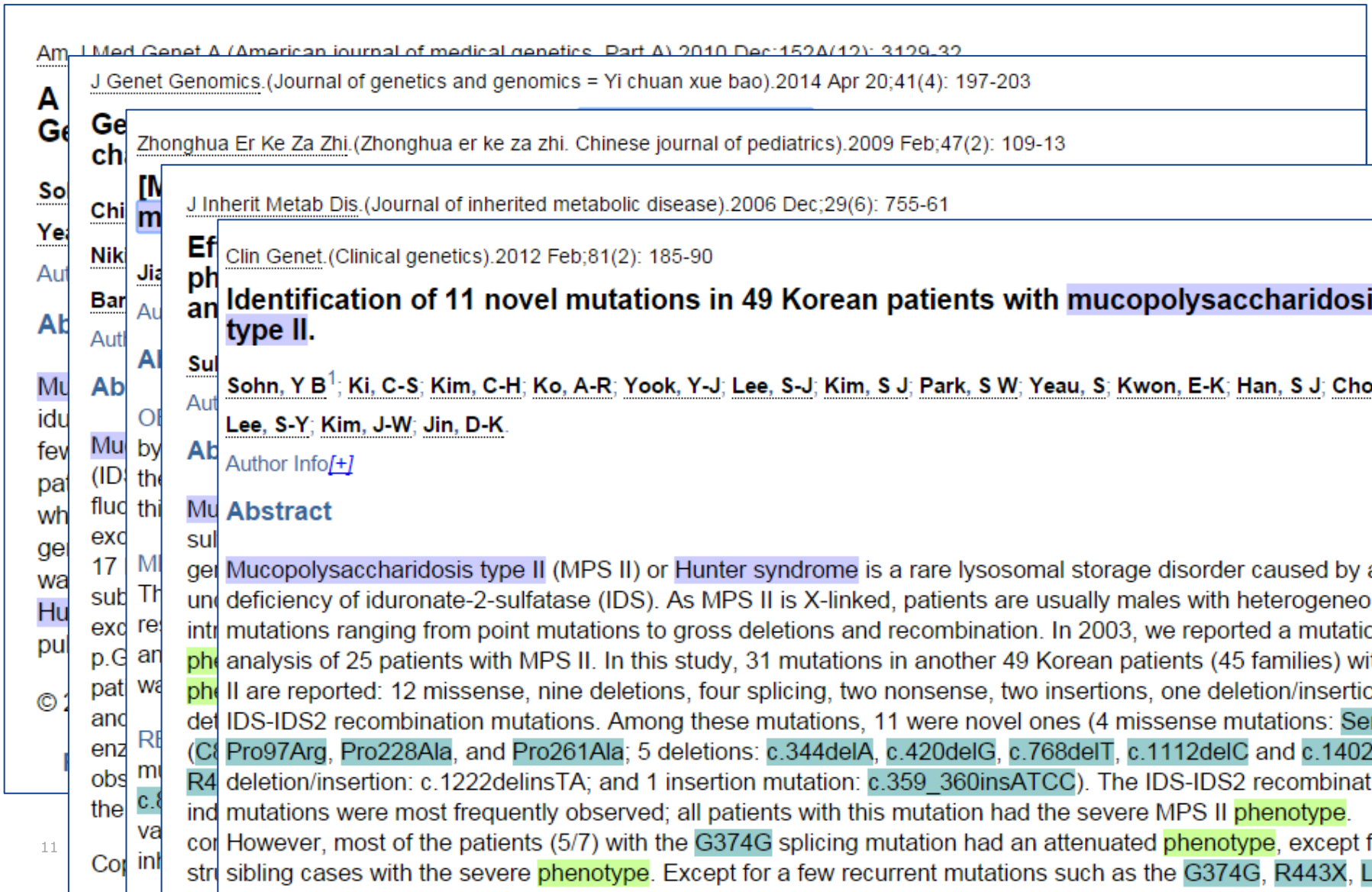
TEXT ANALYTICS FOR RARE DISEASES
GENOTYPE-PHENOTYPE ASSOCIATION IN HUNTER SYNDROME



CHALLENGE

- Paucity of knowledge of natural history of disease
- Sparse data, needs high recall across full text papers
- Mutation patterns very variable
- Structured databases lack broad phenotypic association data

Data buried in scientific literature



Am J Med Genet A.(American journal of medical genetics. Part A).2010 Dec;152A(12): 3129-32

J Genet Genomics.(Journal of genetics and genomics = Yi chuan xue bao).2014 Apr 20;41(4): 197-203

Zhonghua Er Ke Za Zhi.(Zhonghua er ke za zhi. Chinese journal of pediatrics).2009 Feb;47(2): 109-13

J Inherit Metab Dis.(Journal of inherited metabolic disease).2006 Dec;29(6): 755-61

Clin Genet.(Clinical genetics).2012 Feb;81(2): 185-90

Identification of 11 novel mutations in 49 Korean patients with mucopolysaccharidosis type II.

Sohn, Y B[1]; **Ki, C-S**; **Kim, C-H**; **Ko, A-R**; **Yook, Y-J**; **Lee, S-J**; **Kim, S J**; **Park, S W**; **Yeau, S**; **Kwon, E-K**; **Han, S J**; **Choi,**

Lee, S-Y; **Kim, J-W**; **Jin, D-K**.

Author Info[+]

Abstract

Mucopolysaccharidosis type II (MPS II) or Hunter syndrome is a rare lysosomal storage disorder caused by a deficiency of iduronate-2-sulfatase (IDS). As MPS II is X-linked, patients are usually males with heterogeneous mutations ranging from point mutations to gross deletions and recombination. In 2003, we reported a mutation analysis of 25 patients with MPS II. In this study, 31 mutations in another 49 Korean patients (45 families) with II are reported: 12 missense, nine deletions, four splicing, two nonsense, two insertions, one deletion/insertion, IDS-IDS2 recombination mutations. Among these mutations, 11 were novel ones (4 missense mutations: Ser6 Pro97Arg, Pro228Ala, and Pro261Ala; 5 deletions: c.344delA, c.420delG, c.768delT, c.1112delC and c.1402de deletion/insertion: c.1222delinsTA; and 1 insertion mutation: c.359_360insATCC). The IDS-IDS2 recombinatio mutations were most frequently observed; all patients with this mutation had the severe MPS II phenotype. However, most of the patients (5/7) with the G374G splicing mutation had an attenuated phenotype, except for sibling cases with the severe phenotype. Except for a few recurrent mutations such as the G374G, R443X, L5

Extracted, Structured with I2E

Linguamatics

Found **95** assertions from **1000** hits (user limit reached) in **27** docs.
Examined 23747489 (92%) of 25757954 docs.
Took 15.1144 secs (CPU 6.36).
[more details]

HTML as in
Docs/assertion: All Hits/doc/assertion: 10
Cross product Zip archive: None Page Results

PMID	Source	Mutation Genes/Prote..	Genes/Proteins	Severity	Phenotype		Doc		Hit
8111411	PDF	Q531X		mild	general	2	Hopwood_gene_8111411	1	... and R48P, L196S, Q531X (mild phenotype).
15614569	PDF	H138R		severe	general	2	Chang Ex_II_15614569	1	Patients with R88C and H138R mutations displayed a severe phenotype.
17391447	PDF	E177X		attenuated	general	2	Froissar_ppl_17391447	1	In contrast, the attenuated phenotype reported in the patient carrying the E177X mutation (26) is ...
9660053	PDF	nonsense mutation		very mild	general	2	Froissar_enet_9660053	1	This nonsense mutation is associated with a very mild phenotype (patient 56, aged ...
24125893	PDF	c.1122C>T		attenuated	general	1	Mucopoly_nts_24125893	1	... mutations present correlation with the attenuated form (c.1122C>T), while a greater ...
24780617	Abstract	p.Ile360Tyrfs*31		severe	general	1	24780617	2	... mutations whereas the p.Ser142Phe and p.Ile360Tyrfs*31 mutations caused the severe disease manifestation.
9712538	PDF	A deletion involving exons 2-4 in the iduronate-2-sulfatase gene	IDS	intermediate	disease	1	Bonuccel_enet_9712538	2	A deletion involving exons 2-4 in the iduronate-2-sulfatase gene of a patient with intermediate Hunter syndrome
1284597	Abstract	R468W		mild	disease	1	1284597	1	Mutation R468W of the iduronate-2-sulfatase gene in mild Hunter syndrome (mucopolysaccharidosis type II) ...
7887413	PDF	P469H		mild	general	1	Jonsson_enet_7887413	1	... mutations in exon 9 had mild disease (P469H; Y523C; R468W, ...
7981716	PDF	R468W		mild	disease	2	Mutation_S II_7981716	1	... C (1992) Mutation R468W of the iduronate-2-sulfatase gene in mild Hunter syndrome (mucopolysaccaridosis type II) ...
8566953	Abstract	A346D		mild	general	1	8566953	1	The A346D mutation was associated with the mild phenotype, all others with the ...
9501270	PDF	Q389X		severe	disease	1	Isogai_1_bDis_9501270	1	... nonsense mutations (Q80X; Q389X) in patients with severe Hunter syndrome (mucopolysaccharidosis type II)...

TEXT ANALYTICS FOR RARE DISEASES
GENOTYPE-PHENOTYPE ASSOCIATION IN HUNTER SYNDROME

CHALLENGE

- Paucity of knowledge of natural history of disease
- Sparse data, needs high recall across full text papers
- Mutation patterns very variable
- Structured databases lack broad phenotypic association data

SOLUTION

- Abstracts identified in MEDLINE using broad vocabularies.
- Full text PDFs processed for text analytics.
- I2E mutation ontology and bespoke severity vocabs enabled extraction of genotype-phenotype associations.

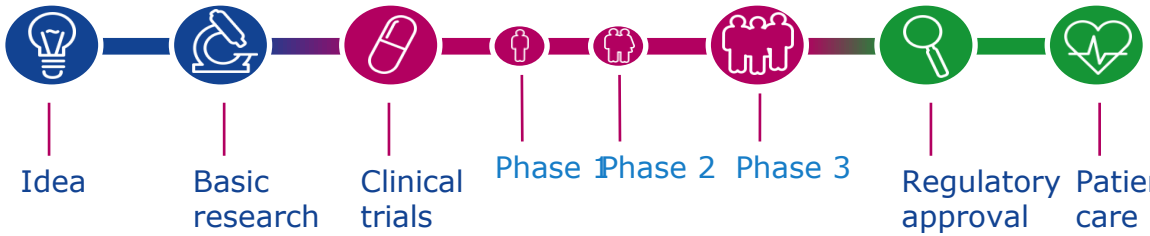
BENEFIT

- Extraction of patient mutations matched or bettered genetic databases
- Increased understanding of IDS mutational spectrum for provider diagnostics and patient awareness
- Enabled rational approach to immune response classification

I2E for Clinical Decision Support in Hospital Rounds:

Real-time access to medical knowledge for on the spot patient care

Medline Abstracts and Science Direct

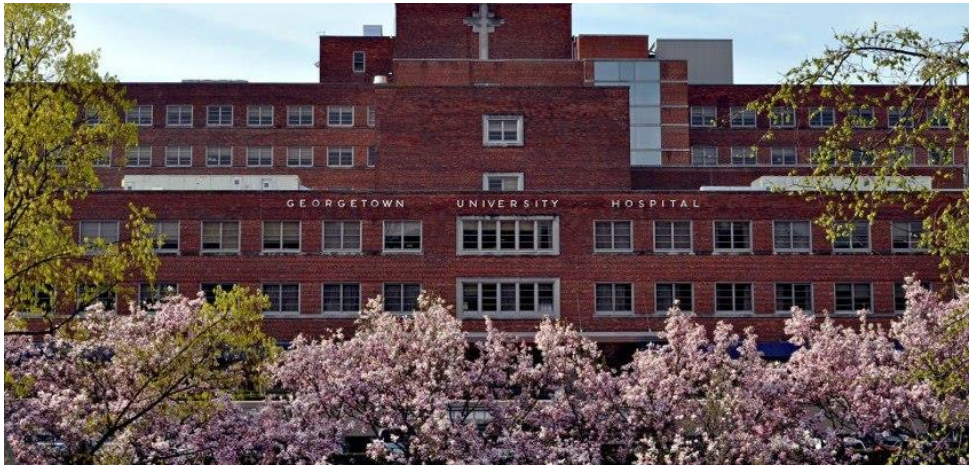


THE QUEEN'S AWARDS FOR ENTERPRISE: 2014

Linguamatics

Georgetown University Medical Center

- Internationally recognized academic medical center
- Dahlgren Memorial Library serves GUMC
- Jonathan Hartmann is Senior Clinical Informationist at DML and provides services to MedStar Georgetown University Hospital



GU Medical Center Requirements

- Informationist accompanies clinical teams on daily rounds
  - General Pediatrics
  - Pediatric and Neo Natal Intensive Care
  - Internal Medicine
- Clinical staff ask Informationist questions
  - Normal saline vs lactated ringers for pancreatitis patients?
  - Causes of pseudomembrane other than C. difficile infection?
- Tablets can be conveniently carried around during rounds
- Informationist can retrieve most needed information on rounds, but in some cases has to go back to office to find out more and provide to clinical staff later

Why use I2E?

- In house research for building database
  - MEDLINE
- Access to published research during rounds
  - MEDLINE
  - Full Text Articles
  - Eliminate the need to go back to desk, retrieve information and provide it to clinical staff at a later stage
  - On the spot answers help clinical staff to make decisions more promptly and improve patient care
- ***Information retrieved at the point of care allows physicians to make critical decisions in a shorter timeframe***

Summary

- Unstructured text in literature is growing across bench-to-bedside continuum
- Application of analytics and NLP is key to future drug discovery, development and delivery of better healthcare
- Linguamatics I2E provides agile NLP text mining:
  - Interactive and scalable search
  - Workflow can be automated
  - Precise, structured results in the format you need

Thank You!

For more information…
Visit: www.linguamatics.com

Contact: Susan LeBeau, VP Sales
Email: susan.lebeau@linguamatics.com
Phone: +1 (774) 571-1117
Email: enquiries@linguamatics.com

Meet our experts at upcoming events:
Visit http://www.linguamatics.com/welcome/events/conferences.html

THE QUEEN'S AWARDS FOR ENTERPRISE: 2014

Linguamatics