Linguamatics

Power of text mining for precision medicine research

Introduction

Precision, personalized, genomic, and stratified medicine are all terms that relate to the ability to better tailor treatments to the most appropriate groups of patients. This can be at the clinical level, or within drug discovery and development.

Within the clinical arena, in order to understand the best treatment pathway for a particular patient or group of patients, it is important to be able to access and analyze information about many different aspects of patients' lives beyond just their medical history. The decades of discussion on nature vs. nurture have demonstrated the importance of genetics, lifestyle choices, and environmental influences.

In the pharma industry, there has been a move over past decades toward more adoption of pharmacogenomics strategies and the pursuit of targeted therapies. Precision medicine offers the prospect of lower variability of drug response, improved safety, and increased treatment efficacy. The annotation of high-throughput biological screens, such as next generation sequencing (NGS), microarray data, microRNA assays, knockout mouse phenotype data, and proteomic screens, can provide information for pharmacogenomic-related tailored drug development, from biomarker discovery and target evaluation, through to patient stratification and clinical profiling.

One key aspect in precision medicine is the ability to access the most in-depth information for gene-disease and genotype-phenotype associations. However, many of the sources needed for understanding genotype-phenotype relationships comprise unstructured text, which is not easily analyzed. I2E text analytics can unlock the value from sources such as electronic health records (EHRs), scientific literature, conference abstracts, or internal reports.

Challenges

Clear and comprehensive information on a particular disease phenotype, related genes and gene variants, and the nuances of potential causal relationships is highly valuable. These data can inform all phases of drug development, and diagnosis, treatment pathways, and drug use in healthcare settings. High-throughput genetic screens (e.g. gene panels, whole exome screens) are providing large volumes of gene variant data. However, biological interpretation of the output is highly time-consuming. Many sources of potential information on disease associations or patient data are in an unstructured text format (e.g. scientific literature, EHRs). Manually searching literature, reviewing EHRs, and annotating genetic results is slow, and generally not comprehensive. For the full picture, researchers and clinicians need to standardize and integrate both structured and unstructured data, and map output to standards (e.g. medical codes, ontologies, formats) to enable better analysis and visualization.

Solution

Linguamatics' natural language processing (NLP) text-mining platform, I2E, can provide a solution to these challenges—extracting key facts from unstructured documents, using relevant ontologies and focused search strategies or queries. Such queries can be written to extract a more comprehensive view of a patient's medical, social, and environmental history from EHRs, such as medical narratives, pathology reports, clinician notes, radiology reports, and more. Mapping symptoms to an ontology, such as the Human Phenotype Ontology, creates a clear understanding of each individual patient, as well as the ability to make comparisons across patient populations more easily, enabling more focused treatment and precise monitoring of the patient journey.

On the genotype side, I2E can help scientists, researchers, and clinicians to rapidly annotate gene-variant associations with disease from the latest literature, which can assist with diagnosis, treatment targeting, or therapy development. I2E enables researchers to collate a comprehensive gene profile, with key biological annotation from a combination of sources (e.g. MEDLINE, full-text literature, OMIM, NIH Grants), accessing the most up-to-date information.

Use of extensive ontologies and I2E's advanced linguistic analytics, along with pattern- and rule-based approaches for mutation and genetic information (e.g. SNPs, CNV, indels, or nucleotide substitutions), means that I2E can extract the most up-to-date published knowledge for genotype-phenotype associations. Of particular importance is the ability to identify and normalize gene mutation or gene variant terms from the unstructured text, which frequently have multiple formats for the same gene (e.g. Val 158 Met, Val by Met at codon 158, V158M, 158V>M, 158V/M, V to M mutation at position 158).

Use of NLP text mining is important when there is little or no available structured annotation about particular genetic variants. This is especially vital for rare diseases that may not have coverage in standard databases such as the Human Gene Mutation Database (HGMD) or ClinVar.

Benefits

As can be seen in the case studies below, pharmaceutical and healthcare customers are using the Linguamatics I2E text-mining solution to provide knowledge that assists with delivery of precision medicine. I2E's NLP-based text mining can filter and extract clinically relevant information from unstructured patient notes or scientific literature, and do so in a timely manner without the need for manual review. The text-mined results provide structured, standardized output for rapid analysis, visualization, or integration into research or clinical databases. According to customers, I2E gets to actionable results 10 to 1000 times faster than a traditional keyword search.

Precision medicine case studies

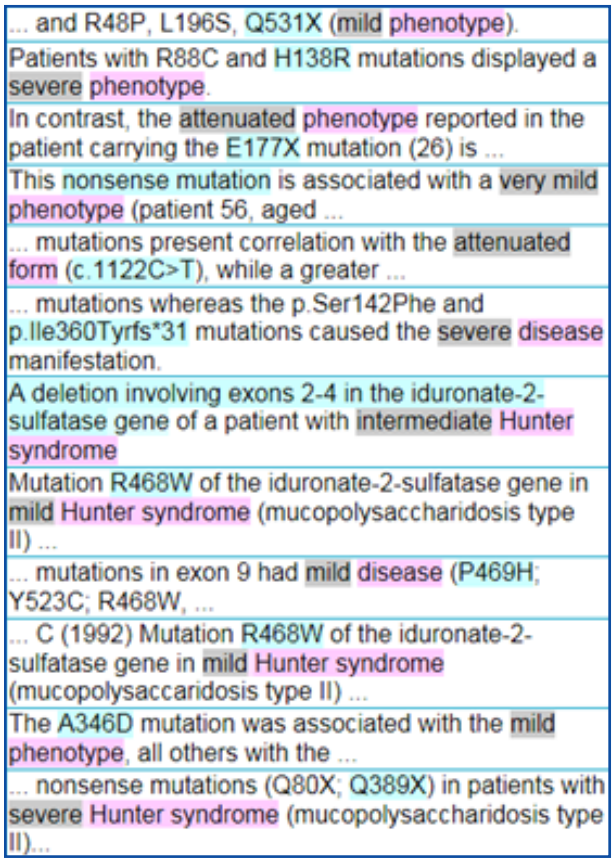
Shire's use of NLP to uncover genotype-phenotype associations in rare disease

Shire develops and provides healthcare therapies in the areas of behavioral health, gastrointestinal conditions, rare diseases, and regenerative medicine. The company uses text mining for systematic examination of gene-disease associations, exemplified by its research to uncover genotype-phenotype

associations for diseases such as Hunter syndrome (also known as mucopolysaccharidosis II), which is a rare disease caused by an X-linked deficiency in the iduronate 2-sulfatase (IDS) gene.

Shire provide an enzyme replacement therapy for Hunter syndrome, but to address the central nervous system aspects of the disease, this needs to be delivered directly to the cerebrospinal fluid via an intrathecal device. This is a potentially life-changing intervention, yet is invasive and unpleasant for young patients. Shire wanted to find a reliable way to identify those patients who had the greatest potential to benefit from this involved procedure. A text-mining project used I2E to extract all the associations reported in full-text literature between the relevant gene, any mutation or variant, and phenotype descriptions for neurocognitive impairment. The result was a set of prognostic genetic markers that enabled clinicians to make informed decisions on which pediatric patients would benefit from this enzyme replacement therapy.

Figure 1: Text analytics for rare disease genotype-phenotype annotations. Mucopolysaccharidosis II or Hunter syndrome is an X-linked deficiency in iduronate 2-sulfatase. Onset of the severe form usually occurs at 2–4 years of age, and the disease presents with symptoms including bone deformities, hearing loss, frequent respiratory infections, cardiomyopathy, hepatosplenomegaly, and often some level of neurocognitive impairment. The figure shows results from I2E text mining to identify associations between mentions of genetic variants (highlighted in blue), the disease (in pink), and level of severity (in grey).



> *Text mining was remarkably successful. Results were significantly better than any genetic database of reported genotypes available.*
>
> Madhu Natarajan, Director, Systems Pharmacology, Shire

Sanofi's use of NLP for genotype-phenotype associations in multiple sclerosis

Sanofi is a long-term user of I2E. One example use case involved using I2E to annotate the output of NGS for a multiple sclerosis (MS) biomarker project. Sanofi researchers needed a comprehensive catalogue of disease annotations for human leukocyte antigen (HLA) alleles, and turned to Linguamatics I2E to text mine the scientific literature. The HLA region is the most polymorphic region of the human genome. HLA alleles have been associated with more than 40 different autoimmune diseases, various types of cancer, infectious disease, and drug adverse events. However, there are no known public databases or resources that systematically annotate the association of HLA alleles and diseases. For the Sanofi MS project, a workflow was established for HLA typing and analysis using whole exome sequencing. This identified more than 400 HLA alleles. The Linguamatics I2E platform was used to search the literature, to annotate

the association of the HLA alleles with diseases and drug hypersensitivity. Vocabularies around the HLA alleles were developed, and the Linguamatics Disease ontology enabled a broad set of disease subclasses and synonyms under "autoimmune diseases" to be searched. This project resulted in more than double the previous disease associations, and the curated annotations were fed into a knowledge base for wider use within the Sanofi team.

> *We use [I2E] for gene disease mapping, target ID, toxicity and gene mutation analyses, biomarker discovery, drug repurposing, and patent analysis […] our results demonstrate that text mining can be applied to identify accurate associations between HLA alleles, haplotypes, and disease, as well as drug hypersensitivity.*
>
> Dongyu Liu, Associate Director, Translational Sciences, Sanofi

Leading Academic Medical Center (AMC)

With the rapid growth in precision medicine approaches, a leading AMC is using genomic techniques to support its treatment selection. One technique is chromosomal microarray (CMA) testing, a genomic scale clinical test that detects sub-microscopic deletions and duplications of genomic material in the DNA of patients suspected of having a genetic disorder. Using current standard of care practices, up to 40% of patients will receive a test result of "variant of unclear clinical significance" (VUS) based on Copy Number Variants (CNVs). This level of uncertainty can be very unsettling to patients and their families, and necessitates further clinical testing to elucidate a genetic diagnosis.

Using standard methodology, a healthcare professional is required to manually acquire additional phenotypic information from the EHR to facilitate interpretation of VUS lesions. It takes a highly skilled, certified cytogenetic technologist approximately 40–60 minutes per patient to manually extract this information from the EHR. Interpretation is then based on a manual, qualitative association of gene content and function with phenotypic relevance in the patient.

This AMC is now using Linguamatics I2E to understand how NLP can reduce the cost of phenotypic information extraction from patient EHRs. The patient charts are searched via I2E and annotated with the Human Phenotype Ontology (HPO), to improve the diagnostic yield of the CMA testing.

Why wait?

I2E is a world-leading, agile, scalable, real-time NLP-based text-mining solution. Since 2001, I2E has been used by top pharmaceutical companies and healthcare providers to speed effective drug discovery, development, and delivery of healthcare therapeutics. Linguamatics I2E is the "bench to bedside and back" NLP solution.

To understand more about using I2E for precision medicine research, contact us at: **enquiries@linguamatics.com**