AMIA 2015 Annual Symposium
November 14 - 18, 2015

NLM Staff Papers and Presentations

Sunday, November 15

3:30 pm – 5:00 pm – S05: Papers - Methods and Tools for Deep Phenotyping (Continental 7/8/9)

LORD: a phenotype-genotype semantically integrated biomedical data tool to support rare disease diagnosis coding in health information systems
R. Choquet*; M. Maaroufi; Y. Fonjallaz; A. De Carrara; P. Vandenbussche; F. Dhombres; P. Landais

Abstract: Characterizing a rare disease diagnosis for a given patient is often made through expert's networks. It is a complex task that could evolve over time depending on the natural history of the disease and the evolution of the scientific knowledge. Most rare diseases have genetic causes and recent improvements of sequencing techniques contribute to the discovery of many new diseases every year. Diagnosis coding in the rare disease field requires data from multiple knowledge bases to be aggregated in order to offer the clinician a global information space from possible diagnosis to clinical signs (phenotypes) and known genetic mutations (genotype). Nowadays, the major barrier to the coding activity is the lack of consolidation of such information scattered in different thesaurus such as Orphanet, OMIM or HPO. The Linking Open data for Rare Diseases (LORD) web portal we developed stands as the first attempt to fill this gap by offering an integrated view of 8,400 rare diseases linked to more than 14,500 signs and 3,270 genes. The application provides a browsing feature to navigate through the relationships between diseases, signs and genes, and some Application Programming Interfaces to help its integration in health information systems in routine.

3:30 pm – 5:00 pm – S06: Papers - Labs and Tests (Plaza B)

Identifying the Clinical Laboratory Tests from Unspecified "Other Lab Test" Data for Secondary Use
X. Pan*; J. J. Cimino

Abstract: Clinical laboratory results are stored in electronic health records (EHRs) as structured data coded with local or standard terms. However, laboratory tests that are performed at outside laboratories are often simply labeled "outside test" or something similar, with the actual test name in a free-text result or comment field. After being aggregated into clinical data repositories, these ambiguous labels impede the retrieval of specific test results. We present a general multi-step solution that can facilitate the identification, standardization, reconciliation, and transformation of such test results. We applied our approach to data in the NIH Biomedical Translational Research Information System (BTRIS) to identify laboratory tests, map comment values to the LOINC codes that will be incorporated into our Research Entities Dictionary (RED), and develop a reference table that can be used in the EHR data extract-transform-load (ETL) process.

AMIA 2015 Annual Symposium
November 14 - 18, 2015

NLM Staff Papers and Presentations

Monday, November 16

8:30 am – 10:00 am – S17: Papers - Taking a Risk (and Addressing It) (Continental 6)

Automatic Classification of Structured Product Labels for Pregnancy Risk Drug Categories, a Machine Learning Approach
L. M. Rodriguez*; D. Demner-Fushman

Abstract: With regular expressions and manual review, 18,342 FDA-approved drug product labels were processed to determine if the five standard pregnancy drug risk categories were mentioned in the label. After excluding 81 drugs with multiple-risk categories, 83% of the labels had a risk category within the text and 17% labels did not. We trained a Sequential Minimal Optimization algorithm on the labels containing pregnancy risk information segmented into standard document sections. For the evaluation of the classifier on the testing set, we used the Micromedex drug risk categories. The precautions section had the best performance for assigning drug risk categories, achieving Accuracy 0.79, Precision 0.66, Recall 0.64 and F1 measure 0.65. Missing pregnancy risk categories could be suggested using machine learning algorithms trained on the existing publicly available pregnancy risk information.

8:30 am – 10:00 am – S18: Papers/Podium Presentations - Clinical Studies (Imperial A)

Reproducing a Prospective Clinical Study as a Retrospective Study in MIMIC-II
F. S. Kury*; V. Huser; J. J. Cimino

Abstract: In this paper we sought to reproduce, as a computational retrospective study in an EMR database (MIMIC-II), a recent large prospective clinical study: the 2013 publication, by the Japanese Association for Acute Medicine (JAAM), about disseminated intravascular coagulation, in journal Critical Care (PMID: 23787004). We designed in SQL and Java a set of electronic phenotypes that reproduced the study's data sampling, and used R to perform the same statistical inference procedures. All source code we produced is available online at https://github.com/fabkury/paamia2015. Our program identified 2,257 eligible patients in MIMIC-II, and the results remarkably agreed with the prospective study. A minority of the needed data elements was not found in MIMIC-II, and statistically significant inferences were possible in the majority of the cases.

10:30 am – 12:00 pm – S28: Papers - Analysis of Scientific Literature (Continental 1/2/3)

Classification of Clinically Useful Sentences in MEDLINE
M. A. Morid*; S. R. Jonnalagadda; M. Fiszman; K. Raja; G. Del Fiol

Abstract: **Objective**: In a previous study, we investigated a sentence classification model that uses semantic features to extract clinically useful sentences from UpToDate, a synthesized clinical evidence resource. In the present study, we assess the generalizability of the sentence classifier to Medline abstracts. **Methods**: We applied the classification model to an independent gold standard of high quality clinical studies from Medline. Then, the classifier trained on UpToDate sentences

was optimized by re-retraining the classifier with Medline abstracts and adding a sentence location feature. **Results**: The previous classifier yielded an F-measure of 58% on Medline versus 67% on UpToDate. Re-training the classifier on Medline improved F-measure to 68%; and to 76% ($p<0.01$) after adding the sentence location feature. **Conclusions**: The classifier's model and input features generalized to Medline abstracts, but the classifier needed to be retrained on Medline to achieve equivalent performance. Sentence location provided additional contribution to the overall classification performance.

10:30 am – 12:00 pm – S32: Systems Demonstrations - Interoperability in the Clinical Space (Plaza A)

Navigating between Drug Classes and RxNorm Drugs with RxClass

O. Bodenreider*; L. Peters; T. Nguyen

Abstract: *RxClass* is a web-based, interactive browser and companion API to explore the relationships between RxNorm drugs and drug classes from several sources including ATC, MeSH and NDF-RT. Like *RxNav*, *RxClass* is publicly available at: http://rxnav.nlm.nih.gov

1:45 pm – 3:15 pm – S39: Papers - Trials and Tribulations (Imperial A)

Extracting Characteristics of the Study Subjects from Full-Text Articles

D. Demner-Fushman*; J. G. Mork

Abstract: Characteristics of the subjects of biomedical research are important in determining if a publication describing the research is relevant to a search. To facilitate finding relevant publications, MEDLINE citations provide Medical Subject Headings that describe the subjects' characteristics, such as their species, gender, and age. We seek to improve the recommendation of these headings by the Medical Text Indexer (MTI) that supports manual indexing of MEDLINE. To that end, we explore the potential of the full text of the publications. Using simple recall-oriented rule-based methods we determined that adding sentences extracted from the methods sections and captions to the abstracts prior to MTI processing significantly improved recall and F1 score with only a slight drop in precision. Improvements were also achieved in directly assigning several headings extracted from the full text. These results indicate the need for further development of automated methods capable of leveraging the full text for indexing.

3:30 pm – 5:00 pm – S48: Didactic Panel - ClinicalTrials.gov: Adding Value through Informatics (Imperial B)

ClinicalTrials.gov: Adding Value through Informatics

N. R. Smalheiser*; V. Huser; A. McCray; A. Tasneem; C. Weng

Abstract: ClinicalTrials.gov is a repository of registered clinical trials maintained by NLM containing detailed descriptions of trial sponsorship, design, and results (when available). ClinicalTrials.gov plays an increasingly important, pivotal role in evidence-based medicine, and serves a diverse audience ranging from clinical researchers, who are designing and conducting new trials and recruiting patients; systematic reviewers, who are summarizing the best available evidence regarding safety and efficacy; bio-entrepreneurs, who are looking for drug repurposing or new therapeutic opportunities; and patients, who may be looking for a suitable clinical trial that might accept them. This panel will present an overview of ClinicalTrials.gov and discuss several ongoing lines of informatics research that are adding value -- for example, using text mining to improve the computability of eligibility criteria, design attributes and outcome results and connect these with patient EHR data; linking a registered trial with the publications arising from that trial; and performing aggregate analyses across trials to extract reusable design knowledge, understand design patterns and trends, and uncover systematic biases. The panelists will also discuss challenges and opportunities for further evolution of ClinicalTrials.gov, particularly in light of

emerging trends such as patient-centered clinical trials, or the use of unpublished trial data in meta-analyses.

3:30 pm – 5:00 pm – S54: Systems Demonstrations - From Patients to Research (Plaza A)

OHDSI: An Open-Source Platform for Observational Data Analytics and Collaborative Research
J. Duke*; F. DeFalco; C. Knoll; V. Huser; R. D. Boyce; P. B. Ryan

Abstract: Observational Health Data Sciences and Informatics (OHDSI, pronounced 'Odyssey') is an international consortium focused on large-scale analysis of observational data. OHDSI has developed an open-source platform for data analytics, visualization, and collaborative research based on a widely used common data model (CDM). In this demonstration, we will present the OHDSI suite of software including tools for loading data into the CDM, assessing data quality, defining and characterizing clinical cohorts, and conducting observational research. We will also present an API for retrieving adverse drug event evidence from a wide range of sources.

5:00 pm – 6:30 pm – Poster Session 1 – Grand Ballroom

Automated searches for personalized evidence to prevent hospital acquired infection
A. Cahan*; S. E. Shooshan; L. M. Rodriguez; D. Demner-Fushman

Abstract: We aimed to facilitate knowledge retrieval at point of care by using automated search. We constructed a "Risk for Infection" PubMed® search filter and evaluated it using a dataset of de-identified clinical notes. The precision and inferred average precision rates of the filter were significantly higher than an unfiltered PubMed search but lower than a proprietary search engine.

An Easy-to-Use Clinical Text De-identification Tool for Clinical Scientists: NLM Scrubber
M. Kayaalp*; A. Browne; Z. Dodd; P. Sagan; C. J. McDonald

Abstract: Health Insurance Portability and Accountability Act (HIPAA) requires that clinical documents be stripped of personally identifying information prior to their secondary use for clinical research. We have been studying clinical text de-identification for more than a decade and developing NLM Scrubber—it is a tool for every clinical scientist who conducts retrospective research using clinical reports. Although we continuously improve and add new functionalities to it, it is very simple to install and use.

Generating the MEDLINE N-Gram Set
C. J. Lu*; D. L. Tormey; L. McCreedy; A. C. Browne

Abstract: The MEDLINE n-gram set is a very useful resource in Natural Language Processing (NLP) and Medical Language Processing (MLP). Currently, there is no MEDLINE n-gram set available in the public domain. Due to the large scale of data, it is a challenge to generate MEDLINE n-grams to fit into a research schedule with limited computer resources. The Lexical System Group (LSG) developed an algorithm to generate the MEDLINE n-gram set for adding multiwords into the SPECIALIST Lexicon. We believe the NLP community can benefit from access to this big data. We processed 2.6 billion single words from 22.4 million MEDLINE documents (titles and abstracts) to generate MEDLINE n-grams (n = 1 to 5) with terms appearing at least 30 times and having less than 50 characters for the 2014 release.

ArticlesAboutMe.org: Disseminating Clinical Trials Results to Patients
V. Huser*; A. Yaman; C. Weng; J. J. Cimino

Abstract: We created a service (available at ArticlesAboutMe.org) that enables clinical trial participants to register and receive an email every time an article that reports the results of the trial is published. Since January 2015, the service has been used for monitoring of over twelve trials. Existing clinical research informatics resources (ClinicalTrials.gov, PubMed) enable relatively simple implementation. Keeping participants informed about study results may provide additional motivation to enroll in a clinical study.

An analysis of PubMed4Hh App User Distribution

F. Liu*; P. Fontelo

Abstract: PubMed for Handhelds (PubMed4Hh) is an app for finding relevant health information from the National Library of Medicine on mobile devices. Apple's iOS app developer tool provides daily downloads data and regional distribution data. Comparison between PubMed4Hh download distribution and the regions of PubMed citations shows a consistent match between the number of users and the number of PubMed indexed publications of a region.

AMIA 2015 Annual Symposium
November 14 - 18, 2015

NLM Staff Papers and Presentations

Tuesday, November 17

10:30 am – 12:00 pm – S65: Papers - Consumer Text and Ontologies (Yosemite C)

Automatic Extraction and Post-coordination of Spatial Relations in Consumer Language
K. Roberts*; L. M. Rodriguez; S. E. Shooshan; D. Demner-Fushman

Abstract: To incorporate ontological concepts in natural language processing (NLP) it is often necessary to combine simple concepts into complex concepts (post-coordination). This is especially true in consumer language, where a more limited vocabulary forces consumers to utilize highly productive language that is almost impossible to pre-coordinate in an ontology. Our work focuses on recognizing an important case for post-coordination in natural language: spatial relations between disorders and anatomical structures. Consumers typically utilize such spatial relations when describing symptoms. We describe an annotated corpus of 2,000 sentences with 1,300 spatial relations, and a second corpus of 500 of these relations manually normalized to UMLS concepts. We use machine learning techniques to recognize these relations, obtaining good performance. Further, we experiment with methods to normalize the relations to an existing ontology. This two-step process is analogous to the combination of concept recognition and normalization, and achieves comparable results.

10:30 am – 12:00 pm – S66: Papers - EHRs for Hospital Teams (Plaza A)

The State and Trends of Barcode, RFID, Biometric and Pharmacy Automation Technologies in US Hospitals
R. Y. Uy*; F. S. Kury; P. Fontelo

Abstract: The standard of safe medication practice requires strict observance of the five rights of medication administration: the right patient, drug, time, dose, and route. Despite adherence to these guidelines, medication errors remain a public health concern that has generated health policies and hospital processes that leverage automation and computerization to reduce these errors. Bar code, RFID, biometrics and pharmacy automation technologies have been demonstrated in literature to decrease the incidence of medication errors by minimizing human factors involved in the process. Although evidence suggests the effectivity of these technologies, adoption rates and trends vary across hospital systems. The objective of study is to examine the state and adoption trends of automatic identification and data capture (AIDC) methods and pharmacy automation technologies in U.S. hospitals. A retrospective descriptive analysis of survey data from the HIMSS Analytics® Database was done, demonstrating an optimistic growth in the adoption of these patient safety solutions.

10:30 am – 12:00 pm – S67: Podium Presentations - Advanced Data Analytics (Plaza B)

Process Mining of Growing Adoption of Genomic Precision Medicine Testing Using Commercial Claims and Encounters Database
V. Huser*

Abstract: A new set of molecular pathology (MoPath) codes in Current Procedural Terminology, that covers many genomic precision medicine tests, went into effect in 2013. We analyzed 324 thousand genetic testing instances of 146 thousand patients in MarketScan Commercial Claims and Encounters dataset showing an increasing adoption of genomic testing and analyzing cost and testing context trends. This work is part of a larger effort to characterize a genomic patient in claims and EHR databases.

3:30 pm – 5:00 pm – S83: Didactic Panel - State of the Art of Clinical Narrative Report De-Identification and Its Future (Continental 6)

State of the Art of Clinical Narrative Report De-Identification and Its Future
M. Kayaalp*; J. Aberdeen; S. Meystre; P. Szolovits

Abstract: While automatic de-identification systems exist, release of de-identified data usually requires significant multi-round expensive effort for validation. To overcome this barrier, we need a consensus on the parameters of successful automatic de-identification. Although we can establish such parameters relative to error rates of human annotators, it is ultimately a policy question whose answer needs to be vetted by the public. When personal identifiers are substituted with surrogates or pseudonyms, it could be very difficult to spot the residual identifiers missed by the de-identifier, but in absolute terms, it is difficult to ensure that de-identified clinical text contains no references that might indirectly identify the patient; hence, de-identified clinical text is usually shared through a data use agreement. When such an agreement is in place, dates and some address parts can be left identified in a limited data set. If we reframe the de-identification problem by focusing on the pertinent identifiers, we may smooth the path to data sharing. Installing and running a clinical de-identification system may require substantial expertise, which small institutions and clinical scientists may lack. Although we would like to develop capable systems with sophisticated functionalities, we also should strive for simplicity for routine de-identification tasks.

3:30 pm – 5:00 pm – S87: Papers - NLP Miscellaneous Applications (Plaza A)

An Ensemble Method for Spelling Correction in Consumer Health Questions
H. Kilicoglu*; M. Fiszman; K. Roberts; D. Demner-Fushman

Abstract: Orthographic and grammatical errors are a common feature of informal texts written by lay people. Health-related questions asked by consumers are a case in point. Automatic interpretation of consumer health questions is hampered by such errors. In this paper, we propose a method that combines techniques based on edit distance and frequency counts with a contextual similarity-based method for detecting and correcting orthographic errors, including misspellings, word breaks, and punctuation errors. We evaluate our method on a set of spell-corrected questions extracted from the NLM collection of consumer health questions. Our method achieves a F1 score of 0.61, compared to an informed baseline of 0.29, achieved using ESpell, a spelling correction system developed for biomedical queries. Our results show that orthographic similarity is most relevant in spelling error correction in consumer health questions and that frequency and contextual information are complementary to orthographic features.

5:00 pm – 6:30 pm – Poster Session 2 – Grand Ballroom

Finding Similar Drug Classes using RxClass
L. Peters*; T. Nguyen; O. Bodenreider

Abstract: *RxClass*, a web-based browser for drug classes, supports navigation between RxNorm drugs and drug classes from several sources (ATC, MeSH, DailyMed and NDF-RT), and allows users to explore similar classes.

PubMed 'Early Alerts': A Pilot Study to Support Prospective Detection of Emerging Adverse Drug Events

A. Sorbello*; A. Ripple; O. Bodenreider

Abstract: The FDA traditionally monitors the safety of marketed drugs by analyzing reports submitted by manufacturers and consumers to the FDA Adverse Event Reporting System (FAERS). In order to enhance prospective detection of emerging adverse drug events (ADE), we investigated leveraging existing PubMed "MyNCBI" functionalities and searching resources to survey the biomedical literature for the latest published safety information in the use case of the new oral hepatitis C drugs.

Bridging the MedlinePlus Cloud to askMEDLINE

P. Fontelo*; F. Liu

Abstract: We developed a Web interface to show the top searched terms from MedlinePlus Cloud and direct the top searched terms to askMEDLINE's query database. The bridge from MedlinePlus, a patient and family health information resource to askMEDLINE provides recent evidence from PubMed for patients and health care professionals.

AMIA 2015 Annual Symposium
November 14 - 18, 2015

NLM Staff Papers and Presentations

Wednesday, November 19

10:30 am – 12:00 pm – S108: Papers - Your Ontologies on Drugs (Yosemite B)

Characterization of the Context of Drug Concepts in Research Protocols: An Empiric Study to Guide Ontology Development
J. J. Cimino*; V. Huser

Abstract: We examined a large body of research study documents (protocols) to identify mentions of drug concepts and established base concepts and roles needed to characterize the semantics of these instances. We found these concepts in three general situations: background knowledge about the drug, study procedures involving the drug, and other roles of the drug in the study. We identified 18 more specific contexts (e.g., adverse event information, administration and dosing of the drug, and interactions between the study drug and other drugs). The ontology was validated against a test set of protocol documents from NIH and ClinicalTrial.gov. The goal is to support the automated extraction of drug information from protocol documents to support functions such as study retrieval, determination of subject eligibility, generation of order sets, and creation of logic for decision support alerts and reminders. Further work is needed to formally extend existing ontologies of clinical research.

10:30 am – 12:00 pm – S108: Papers - Your Ontologies on Drugs (Yosemite B)

Approaches to Supporting the Analysis of Historical Medication Datasets with RxNorm
O. Bodenreider*; L. Peters

Abstract: **Objective**: To investigate approaches to supporting the analysis of historical medication datasets with RxNorm. **Methods**: We created two sets of National Drug Codes (NDCs). One is based on historical NDCs harvested from versions of RxNorm from 2007 to present. The other comprises all sources of NDCs in the current release of RxNorm, including proprietary sources. We evaluated these two resources against four sets of NDCs obtained from various sources. **Results**: In two historical medication datasets, 14-19% of the NDCs were obsolete, but 91-96% of these obsolete NDCs could be recovered and mapped to active drug concepts. **Conclusion**: Adding historical data significantly increases NDC mapping to active RxNorm drugs. A service for mapping historical NDC datasets leveraging RxNorm was added to the RxNorm API and is available at https://rxnav.nlm.nih.gov/.

10:30 am – 12:00 pm – S109: Papers - The User Perspective on Informatics Tools (Yosemite C)

Challenges and Insights in Using HIPAA Privacy Rule for Clinical Text Annotation
M. Kayaalp*; A. Browne; P. Sagan; T. McGee; C. J. McDonald

Abstract: The Privacy Rule of Health Insurance Portability and Accountability Act (HIPAA) requires that clinical documents be stripped of personally identifying information before they can be released to researchers and others. We have been manually annotating clinical text since 2008 in order to test and evaluate an algorithmic clinical text de-identification tool, NLM Scrubber, which

we have been developing in parallel. Although HIPAA provides some guidance about what must be de-identified, translating those guidelines into practice is not as straightforward, especially when one deals with free text. As a result we have changed our manual annotation labels and methods six times. This paper explains why we have made those annotation choices, which have been evolved throughout seven years of practice on this field. The aim of this paper is to start a community discussion towards developing standards for clinical text annotation with the end goal of studying and comparing clinical text de-identification systems more accurately.