
AMIA 2016 ANNUAL SYMPOSIUM

NOVEMBER 12-16, 2016

NLM STAFF PAPERS AND PRESENTATIONS

Saturday, November 12

8:30 am – 4:30 pm – WG03: Natural Language Processing Working Group Pre-Symposium: Graduate Student Consortium and ‘Hackathon’ (Salon A-4)

NLP WG pre-symposium 2016: Graduate Student Consortium and ‘Hackathon’

Sivaram Arabandi, Dina Demner-Fushman, Hongfang Liu, Stephane Meystre, Jon Patrick, Guergana Savova, Ozlem Uzuner, Kavishwar Waghlikar, Chunhua Weng, Hua Xu, Rong Xu, Pierre Zweigenbaum

Abstract: The application of Natural Language Processing (NLP) methods and resources to clinical and biomedical text has received growing attention over the past years, but progress has been limited by difficulties to access shared tools and resources, partially caused by patient privacy and data confidentiality constraints. Efforts to increase sharing and interoperability of the few existing re- sources are needed to facilitate the progress observed in the general NLP domain. To answer this need, the AMIA NLP working group pre-symposium continues the tradition since its inception in 2012 to provide a unique platform for close interactions among students, scholars, and industry professionals who are interested in clinical NLP. The event will consist of two sections: 1) a graduate student consortium, where students can present their work and get feedback from experienced researchers in the field; and 2) a ‘hackathon’ of NLP tools, where developers of these tools will present them to users and help these users implement their tools to work on practical NLP tasks in groups.

8:30 am – 4:30 pm – WG05: Data Mining for Medical Informatics (DMMI) - Learning Health (Williford A)

Desiderata for Drug Classification Systems for their Use in Analyzing Large Drug Prescription Datasets (9:50 am – 10:05 am)

Fabrcio Kury*, Olivier Bodenreider

Background: Information about the billions of prescriptions written for patients each year is collected in large datasets. Drug classification systems (DCSs) are key to analyzing these datasets. However, their ability to support such analyses has not been studied.

Methods: We identified six desirable features for drug classification systems (DCSs) from the perspective of analyzing large drug prescription datasets. In addition to offering operational definitions for these desiderata, we also applied them to clinically significant drugs in RxNorm for six DCSs, and used them to assess the impact of these DCSs on the analysis of a large prescription dataset from Medicare Part D claims.

Results: Based on these desiderata, we could determine that ATC, VAC and EPC seem to be better suited for the analysis of large drug prescription datasets, because of their coverage and granularity, and because ATC and VAC support aggregation.

1:00 pm – 4:30 pm – T10: Resources for Analyzing Drug Prescription Datasets (Continental B)

Resources for Analyzing Drug Prescription Datasets

Olivier Bodenreider, Vojtech Huser, Christian Reich

Abstract: Large prescription datasets have become increasingly available to researchers (e.g., claims data from Medicare and private insurance companies, pharmacy data from clinical institutions, feeds from health information networks, such as Surescripts). Prescription data are generally recorded at a level that is very detailed (e.g., with National Drug Codes (NDCs) that include packaging information), and often need to be aggregated for meaningful clinical analysis (e.g., at the level of the ingredient or drug class).

Resources such as RxNorm, the standard terminology for drugs in the U.S., can help map NDCs to RxNorm concepts for clinical drugs. RxNorm also supports aggregation by linking clinical drugs to their ingredients, and to drug classes from ATC, MeSH, NDF-RT and DailyMed. The RxNorm and RxClass APIs facilitates the use of RxNorm for aggregation purposes.

The first part of this tutorial presents basic information about prescription datasets and resources for analyzing them, with emphasis on RxNorm. In the second part, we demonstrate an application of these resources to common use cases, including the comparison of prescribed vs. defined daily doses for drugs and the identification of potentially inappropriate medications (e.g., during pregnancy, for the elderly). Finally, we present the experience of the OHDSI (Observational Health Data Sciences and Informatics) community in integrating various kinds of drug data in a large clinical data warehouse compliant with the OMOP clinical data model.

AMIA 2016 ANNUAL SYMPOSIUM

NOVEMBER 12-16, 2016

NLM STAFF PAPERS AND PRESENTATIONS

Sunday, November 13

3:30 pm – 5:00 pm – S06: Papers - Analytics, Costs, Utilization, & Data Quality (Williford A/B)

Analysis of Healthcare Cost and Utilization in the First Two Years of the Medicare Shared Savings Program Using Big Data from the CMS Enclave (3:52 pm – 4:14 pm)

Fabrício Kury*, Seo Baik, Clement McDonald

Abstract: The Medicare Shared Savings Program (MSSP) is the larger of the first two Accountable Care Organization (ACO) programs by the Centers for Medicare and Medicaid Services (CMS). In this study we assessed healthcare cost and utilization of 1.71 million Medicare beneficiaries assigned to the 333 MSSP ACOs in the calendar years of 2013 and 2014, in comparison to years 2010 and 2011, using the official CMS data. We employed doubly robust estimation (propensity score weighting followed by generalized linear regression) to adjust the analyses to beneficiary personal traits, history of chronic conditions, previous healthcare utilization, ACO administrative region, and ZIP code socioeconomic factors. In comparison to the care delivered to the control cohort of 17.7 million non-ACO beneficiaries, we found that the care patterns for ACO beneficiaries shifted away from some costly types of care, but at the expense of increased utilization of other types, increased imaging and testing expenditures, and increased medication use, with overall net greater increase in cost instead of smaller increase.

3:30 pm – 5:00 pm – S11: ACMI Featured Presentation - Enduring Challenges in Biomedical and Health Informatics - Celebrating the 40th Anniversary of SCAMC (Williford C)

Featured Presentation

Michael Ackerman, Marion Ball, Joshua Denny, Robert Grennes, Casimir Kulikowski, Jacqueline Merrill, Anne Moen, Judy Ozbolt, Edward Shortliffe

Abstract: The first SCAMC in 1976 began a long and successful series of conferences that helped define and coalesce biomedical and health informatics as a professional field in the US, while also encouraging considerable international participation. In 1988, SCAMC was one of the three organizations that merged to create AMIA. Coverage of the very wide breadth of biomedical informatics disciplines and depth of scientific and clinical involvement has been carried on successfully by AMIA through its meetings to the present day. The panelists will briefly compare and contrast the challenges and opportunities for the field that have evolved but also endure since the first SCAMC, after which the audience will be invited to share their experiences on this topic.

3:30 pm – 5:00 pm – S07: Podium Presentations - Terminologies: Nursing, Oncology, and Dentistry (Salon A-4)

Development of an Oncology Subset of SNOMED CT Based on Patient Notes (4:30 pm – 4:45 pm)

Sina Madani*, Jerry Henderson, Kin Wah Fung

Abstract: MD Anderson Cancer Center (MDA) is one of the world's largest institutions involved exclusively in cancer care, research, and prevention. More than 120,000 clinical transcribed documents are added to the MDA EMR system on a monthly basis. We used natural language processing methods to generate a subset of SNOMED CT concepts that are frequently documented as cancer diagnoses in patient notes.

AMIA 2016 ANNUAL SYMPOSIUM

NOVEMBER 12-16, 2016

NLM STAFF PAPERS AND PRESENTATIONS

Monday, November 14

10:30 am – 12:00 pm – S29: Papers - Making it SNOMED (Salon A-4)

Leveraging Lexical Matching and Ontological Alignment to Map SNOMED CT Surgical Procedures to ICD-10-PCS (10:52 am – 11:14 am)

Kin Wah Fung*, Julia Xu, Filip Ameye, Arturo Romero, Arabella D'Havé

Abstract: In 2015, ICD-10-PCS replaced ICD-9-CM for coding medical procedures in the U.S. We explored two methods to automatically map SNOMED CT surgical procedures to ICD-10-PCS. First, we used MetaMap to lexically map ICD-10-PCS index terms to SNOMED CT. Second, we made use of the axial structure of ICD-10-PCS and aligned them to defining attributes in SNOMED CT. Lexical mapping produced 45% of correct maps and 44% of broader maps. Ontological mappings were 40% correct and 5% broader. Both correct and broader maps will be useful in assisting mappers to create the map. When the two mapping methods agreed, the accuracy increased to 93%. Reviewing the MetaMap generated body part mappings and using additional information in the SNOMED CT names and definitions can lead to better results for the ontological map.

1:45 pm – 3:15pm – S44: Late Breaking Session – Ignite Presentations -1 (Salon A-2)

FDA Adverse Reaction Extraction Challenge

Dina Demner-Fushman

5:00 pm – 6:30 pm – Poster Session 1 (Stevens Salon D)

PubMed 'Early Alerts': Towards Better Precision of Literature Searching for Pharmacovigilance Information based on an Assessment of Relevance Feedback

Anna Ripple*, Alfred Sorbello, Shahrukh Haider, Olivier Bodenreider

Abstract: The PubMed 'Early Alerts' provide FDA regulatory reviewers with weekly topical searches of the most recently submitted citations to PubMed/MEDLINE to support prospective detection of emerging adverse drug events for specific drugs. We seek to increase precision based on a relevance assessment for a subset of retrieved citations. We identified several candidate text words to potentially increase the precision of our literature search strategy for pharmacovigilance information, but further validation is required.

Exploring the Use of ClinicalTrials.gov Trial Results Data for Pharmacovigilance

Vojtech Huser*, Olivier Bodenreider

Abstract: In recent years, data from clinical trials are increasingly being shared using trial result data- bases. ClinicalTrials.gov is the largest repository of trial summary results, and it grew at a pace of 3711 deposited trial results per year (based on a trend from the last 3 years). We explore the feasibility of using trial registration data (interventions) and trial basic summary results data (adverse events) to provide drug-event pairs data for pharmacovigilance platforms. We analyze the current structured data and estimate in what proportion of trial results semi- automated pharmacovigilance analysis is feasible. We also explore conversion of free-text drug intervention entries into coded RxNorm concepts and, similarly, coded terminology issues for adverse events reporting.

NDC Properties in RxNorm

Lee Peters*, Olivier Bodenreider

Abstract: The National Drug Code (NDC) is a universal product identifier for human drugs in the United States. The most used query from users in the RxNorm Application Program Interface (API) is to map NDCs to the corresponding RxNorm concepts. However, until recently, the RxNorm API did not expose the properties associated with NDC (extracted from SPLs). The new API function now provides this information.

AMIA 2016 ANNUAL SYMPOSIUM

NOVEMBER 12-16, 2016

NLM STAFF PAPERS AND PRESENTATIONS

Tuesday, November 15

9:00 am – 10:00 am - Keynote Presentation: Patricia F. Brennan, PhD, RN (International Ballroom)

10:30am – 12:00pm S58: Papers – Making sense through NLP (Continental B)

Differentiating Sense through Semantic Interaction Data (10:30 am – 10:52am)

Terri Workman*, Charlene Weir, Thomas Rindfleisch

Abstract: Words which have different representations but are semantically related, such as dementia and delirium, can pose difficult issues in understanding text. We explore the use of interaction frequency data between semantic elements as a means to differentiate concept pairs, using semantic predications extracted from the biomedical literature. We applied datasets of features drawn from semantic predications for semantically related pairs to two Expectation Maximization clustering processes (without, and with concept labels), then used all data to train and evaluate several concept classifying algorithms. For the unlabeled datasets, 80% displayed expected cluster count and similar or matching proportions; all labeled data exhibited similar or matching proportions when restricting cluster count to unique labels. The highest performing classifier achieved 89% accuracy, with F1 scores for individual concept classification ranging from 0.69 to 1. We conclude with a discussion on how these findings may be applied to natural language processing of clinical text.

1:45pm – 3:15pm – S69: Papers - NLP for Patient Satisfaction and Questions (Salon A-4)

Resource Classification for Medical Questions (2:07 pm – 2:29 pm)

Kirk Roberts*, Laritza Rodriguez, Sonya Shooshan, Dina Demner-Fushman

Abstract: We present an approach for manually and automatically classifying the resource type of medical questions. Three types of resources are considered: patient-specific, general knowledge, and research. Using this approach, an automatic question answering system could select the best type of resource from which to consider answers. We first describe our methodology for manually annotating resource type on four different question corpora totaling over 5,000 questions. We then describe our approach for automatically identifying the appropriate type of resource. A supervised machine learning approach is used with lexical, syntactic, semantic, and topic-based feature types. This approach is able to achieve accuracies in the range of 80.9% to 92.8% across four datasets. Finally, we discuss the difficulties encountered in both manual and automatic classification of this challenging task.

Combining Open-domain and Biomedical Knowledge for Topic Recognition in Consumer Health Questions (2:29 pm – 2:51 pm)

Yassine Mrabet*, Halil Kilicoglu, Kirk Roberts, Dina Demner-Fushman

Abstract: Determining the main topics in consumer health questions is a crucial step in their processing as it allows narrowing the search space to a specific semantic context. In this paper we propose a topic recognition approach based on biomedical and open-domain knowledge bases. We first recognize named entities in consumer health questions using an unsupervised method that relies on a biomedical knowledge base, UMLS, and an open-domain knowledge base, DBpedia. We cast topic recognition as the binary classification problem of deciding whether a detected named entity is the question topic or not. We evaluated our approach on a dataset from the National Library of Medicine (NLM), introduced in this paper, and another from the Genetic and Rare Disease Information Center (GARD). The combination of knowledge bases outperformed the results obtained by individual knowledge bases by up to 16.5% F1 score on the NLM dataset. We also achieved state-of-the-art performance on the GARD dataset. Our results demonstrate that combining open-domain knowledge bases with biomedical knowledge bases can lead to a substantial improvement in understanding user-generated health content.

Recognizing Question Entailment for Medical Question Answering (2:51 pm – 3:13pm)

Asma Ben Abacha*, Dina Demner-Fushman

Abstract: With the increasing heterogeneity and specialization of medical texts, automated question answering is becoming more and more challenging. In this context, answering a given medical question by retrieving similar questions that are already answered by human experts seems to be a promising solution. In this paper, we propose a new approach for the detection of similar questions based on the Recognition of Question Entailment (RQE). In particular, we consider Frequently Asked Question (FAQs) as a valuable and widespread source of information. Our final goal is to automatically provide an existing answer if FAQ similar to a consumer health question exists. We evaluate our approach using consumer health questions received by the National Library of Medicine and FAQs collected from NIH websites. Our first results are promising and suggest the feasibility of our approach as a valuable complement to classic question answering approaches.

1:45pm – 3:15pm – S73: Podium Presentations - Making Care Safe (Continental A)

National Patterns of Potentially Inappropriate Drug Use in the US Elderly: An Evaluation of the CMS Enclave as a Scalable Population Assessment Tool (2:30 pm – 2:45 pm)

Mallika Mundkur*, Fabricio Kury, Ferdinand Dhombres, Vojtech Huser, Olivier Bodenreider

Abstract:

Background: Elderly patients are at high risk for adverse drug events and better assessment tools are needed to monitor use patterns.

Methods: We evaluated use of 128 high-risk medications utilizing Medicare drug claims within the CMS enclave.

Results: Out of 1 million patients in our cohort, 47% received at least one high-risk medication. **Conclusions:** Elderly patients are frequently exposed to high-risk medications. While the enclave enables large-scale assessments; efficiency remains a limitation.

Assessing the potential risk in drug prescriptions during pregnancy (2:45 pm – 3:00 pm)

Ferdinand Dhombres*, Vojtech Huser, Laritza Rodriguez, Olivier Bodenreider

Abstract: This investigation demonstrates the feasibility of assessing the potential risk in drug prescriptions during pregnancy from a large claims dataset (14.7M prescriptions; 3,7M pregnant women) using RxNorm and the Briggs reference (new FDA regulation of June 2015), with finer-grained recommendations compared to the old FDA categories, as well as stronger evidence. In our cohort, there is human data evidence for 87.8% of the prescriptions for drugs with potential risk.

1:45pm – 3:15pm – S77: Systems Demonstrations - Capturing Data and Connecting to Public Health (Salon A-3)

LHC-Forms and Related Widgets for Capturing and Tuning Health Data (1:45 pm – 2:30 pm)

Ye Wang, Paul Lynch, Ajay Kanduru, John Hook, Lee Mericle, Christophe Ludet, Daniel Vreeman, Clement McDonald*

Abstract: NLM's Lister Hill National Center for Biomedical Communications (LHC) developed four inter-related open source, web-based, JavaScript tools for gathering and processing clinical data:

1) LHC-Forms is a data capture widget designed in partnership with Regenstrief Institute to produce a browser executable form from a stored form description. LHC-Forms supports HL7 data types, repeating groups of questions, survey scoring, validation checks, answer lists, and default values that can be derived from answers given to preceding questions. LHC-Forms can execute any LOINC panel as a form, and generate an HL7 v2 message of the entered data. Try it:

<https://lhc-forms.lhc.nlm.nih.gov>.

2) Clinical Table Search Service is an auto-completion tool that provides an autocomplete menu to fields that take answers from large tables. It is accessed via URLs whose parameters control what table to search, which fields to return to the choice menu grid, and which fields of the selected item to store as hidden content in the input fields. Our implementation provides preconfigured access to many clinical tables: LOINC, RxTerms, ICD-10-CM, many NCBI genomics tables, COSMIC and others. Try it: <https://clin-table-search.lhc.nlm.nih.gov>.

3) An LHC-Form building tool. Try it: <https://lhc-formbuilder.lhc.nlm.nih.gov>.

4) A JavaScript validator and converter for UCUM units of measure. Try it: <https://ucum-validator.lhc.nlm.nih.gov>. The four modules can be used together or separately in web applications.

3:30pm – 5:00pm – S82: Panel - Women in Informatics Leadership Forum (Waldorf)

Interactive Panel

Suzanne Bakken, Wendy Chapman, Valerie Florance, Rebecca Jacobson, Jessica Tenenbaum

Abstract: Abstract Biomedical Informatics is a diverse field encompassing many different areas of study and providing many different potential career paths. Our field has benefited greatly by the intersection of different disciplines which enhance the diversity of Biomedical Informatics as well as the demographics of its contributors. While women have always represented a significant fraction of the discipline, there have been far fewer women in leadership positions in biomedical informatics, particularly in traditional academic roles and professional organizations. Anecdotally, we observe that there are an increasing number of women entering training programs and early academic positions in bio- medical informatics. What can we do now to increase the potential for a larger cohort of women leaders in Biomedical Informatics within the next decade? This AMIA Panel will consist of two parts intended to form a single cohesive experience for participants, although participants may choose to participate in either one or both of the related events.

5:00 pm – 6:30 pm – Poster Session 2 (Stevens Salon D)

Characterizing the semantic composition of the UMLS Metathesaurus over time

Olivier Bodenreider*, Lee Peters

Abstract: In this investigation, we leverage the semantic groups to characterize the semantic composition of the UMLS Metathesaurus over time. Semantic types are grouped into fifteen semantic groups, which represent broad subdomains of biomedicine, such as Anatomy, Chemicals and Drugs, and Disorders. The UMLS semantic groups have been used to create semantic profiles for source vocabularies, but can also be applied to the Metathesaurus as a whole.

Deriving the “Number Needed to Treat” from PubMed Structured Abstracts

Paul Fontelo*, Fang Liu

Abstract: The Number Needed to Treat (NNT) indicates the number of patients needed be treated to prevent one unfavorable outcome. In some IMRAD-formatted abstracts, the Results section may contain adequate data of interventions to estimate the NNT. We developed a Web interface that automatically calculates the NNT after the table is manually populated with clinical trial results. A link is provided to the fulltext article if the abstract is insufficient.

A Feasibility Study of Answering Clinical Questions Using askMEDLINE at the Point of Care

Kyungsook Gartrell*, Gwenyth Wallen, Caitlin Brennan, Cheryl Fisher, Paul Fontelo

Abstract: Wireless networks, mobile devices, and Internet resources allow clinicians to potentially access evidence from the literature at the point of care. This feasibility study was intended to demonstrate that clinical questions that arose during rounds could be answered using a mobile application and PubMed resources. Our results show that real-time search and retrieval of reliable and potential useful information for clinical decision-making is feasible at the point of care.

Text Processing of Clinical Research Protocols and Informed Consents to Facilitate Tracking of Re- search Procedures

Alok Sagar Panny, Vojtech Huser*

Abstract: Computable representation of events in a clinical study is an ongoing Clinical Research Informatics (CRI) challenge. At our institution (NIH Clinical Center), which conducts more than 1600 clinical studies at any given time, we piloted an approach where readily available clinical study documents (informed consent and study protocols) are parsed (using MetaMap natural language processing tool) for presence of procedure terms that can be used to construct a computable representation of the study. Part of our project also aims to suggest extensions to the existing CRI standards (such as CDISC ODM) to advance computable representation of clinical protocols.

A Comparison of ESpell and GSpell Spell-Checker Tools

Fang Liu*, Paul Fontelo

Abstract: Accurate spelling is crucial for obtaining optimal results with PubMed search tools. With increased usage of mobile devices and smartphones, we have noticed higher spelling errors. Auto spell-checkers, like, ESpell and GSpell, provide spelling suggestions and improve data retrieval. Using actual 500 search terms, we compared the use of these two utilities in 'PubMed for Handhelds'. Both utilities are useful but we found that ESpell has higher precision and recall.

Multiword Frequency Analysis Based on the MEDLINE N-gram Set

Chris Lu*, Destinee Tormey, Lynn McCreedy, Allen Browne

Abstract: Multiwords are vital to better precision and recall in NLP applications. The Lexical Systems Group (LSG) developed an effective approach to add multiwords to the SPECIALIST Lexicon from the MEDLINE n-gram set. This paper describes a frequency analysis on LexMultiwords (LMWs) and acronym expansions (e.g. blood pressure for BP) based on the word count (WC) in MEDLINE. Results show most LMWs locate in the low WC range with better precision and F1 score.

Resolving Hierarchical Ambiguity in Indexing Recommendations

James Mork, Dina Demner-Fushman*

Abstract: Mapping key points extracted from text to the MeSH hierarchy is an essential information extraction task requiring decisions on the best level of MeSH specificity. We developed a corpus-based approach to analyze methods for selecting the appropriate specificity level given both a general and a specific MeSH term from the same MeSH tree. A rule derived using our analysis eliminated over 50% of the erroneous results with very little loss to recall.

Towards automatic discovery of Genes related to Human Placenta

Laritza Rodriguez*, Stephanie Morrison, Kathleen Greenberg, Dina Demner-Fushman

Abstract: Discovery of gene pathways and biochemical mechanisms explaining disease causality can be facilitated by automated extraction of information from the literature. The first step in understanding complex pathways is extraction of disease-specific genes. Our goals are to extract genes and related biomarkers, to create a specialized human placenta gene repository, and to identify target genes for pregnancy-related diseases. Here we present the first phase of the study: extraction of gene mentions from text.

AMIA 2016 ANNUAL SYMPOSIUM
NOVEMBER 12-16, 2016

NLM STAFF PAPERS AND PRESENTATIONS

Wednesday, November 16

10:30 am – 12:00 pm – S109: Panel - Standardizing Research Common Data Elements: Initiatives, Exchange Formats and Current Use by Patient-level Trial Results Databases (Salon A-1)

Didactic Panel

Dikla Shmueli-Blumberg, Rachel Hess, Vojtech Huser, Murat Sincan

Abstract: Sharing of de-identified patient level data from clinical trials is increasingly becoming a norm and this trend is increasing the importance of research Common Data Elements. CDEs have the potential to increase the value of trial data by making it easier to integrate data across multiple trials. CDEs can reduce study start-up costs, improve the quality of collected data, and facilitate cross study comparisons, data aggregation and meta-analyses. This panel will (1) describe current CDE initiatives, highlight several trial data sharing platforms (including their use of CDEs), and describe CDE informatics standards; (2) describe a case study a site using REDCap Electronic Data Capture system and their CDE experience; (3) describe a case study of a trial data sharing platform of the National Institute on Drug Abuse; and (4) showcase a multi-site observational study that first selected suitable CDEs and implemented collection of trial data directly within an Epic EHR.