# Saturday, November 3

## 8:30 am – 4:30 pm W12: Natural Language Processing Working Group Pre-Symposium: Graduate Student Consortium, Year-in-Review, and Community Shared Tasks

**Hongfang Liu, Ozlem Uzuner, Sivaram Arabandi, Dina Demner-Fushman,** [**Stephane Meystre,**](https://amia2016.zerista.com/profile/member/903038) **Jon Patrick,** [**Guergana**](https://amia2016.zerista.com/profile/member/922988)[**Savova,**](https://amia2016.zerista.com/profile/member/922988) **Kavishwar Wagholikar,** [**Chunhua Weng,**](https://amia2016.zerista.com/profile/member/923296) **Hua Xu, Rong Xu, Meliha Yetisgen, 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 constrained 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 resources are needed to facilitate the progress observed in the general NLP domain. To address 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 biomedical NLP. The event will consist of three sections: 1) a graduate student consortium, where students can present their work and get feedback from experienced researchers in the field; 2) a year-in-review session, where significant NLP articles in the biomedical domain will be presented followed by a panel discussion; and 3) a NLP community challenge session. This session will invite organizers and participants of NLP challenges in the biomedical domain to present the design, implementation, and results of the challenges.

**8:30 am – 4:30 pm W14: Analyzing Large Drug Prescription Datasets - Principles, Tools and Examples (sponsored by Pharmacoinformatics Working Group)**

**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 manufacturer and 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. developed by the National Library of Medicine, can facilitate the mapping of NDCs to RxNorm concepts for clinical drugs. RxNorm also supports aggregation by linking clinical drug products to their ingredients, and to drug classes from ATC, MED-RT and DailyMed. The RxNorm and RxClass application programming interfaces (APIs) and companion browsers facilitate the use of RxNorm for aggregation purposes. Additionally, features have recently been added to the drug APIs to facilitate the interpretation of obsolete drug identifiers often found in clinical data warehouses.

The first part of this tutorial presents basic information about drug datasets and resources for analyzing them, with emphasis on RxNorm. The audience will be invited to participate (active exploration of RxNorm through RxNav and RxClass; follow-along activities with RxMix).

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 (Observational Medical Outcomes Partnership) clinical data model, and we address issues in integrating drugs from different countries.

# Sunday, November 4

## 8:30 am – 12:00 pm W20: AMIA CRI-WG Pre-Symposium: Community Strengthening and Knowledge Sharing Towards Systematic and Scalable Clinical Data Quality Assessment

## Chunhua Weng, Abu Mosa, Monika Ahuja, Ana Szarfman, Anthony Solomonides, Vojtech Huser, Sigfried Gold, Meredith Zozus

**Abstract:** Given the increasing need for a rapid learning health system based on clinical data and the emerging culture of data-driven clinical research, unreliable clinical data quality can undermine the evidence base for biomedicine. To address this need, the AMIA CRI working group pre-symposium aims to provide a unique opportunity for community building for clinical data quality assessment and for facilitating close interactions among trainees, scholars, stakeholders, and industry professionals interested in scalable and standards-based approaches to clinical data quality improvement. The event will consist of four sections: 1) a panel presenting the state-of-art clinical data quality checking methods, tools and best practices from experts; 2) a journal club-style highlight session, where significant data quality articles will be presented followed by a panel discussion; 3) a tutorial introducing endorsed data quality assessment resources followed by Q&A with the participants; and 4) a moderated brainstorming session aiming to develop a research agenda for clinical data quality improvement with identified knowledge gaps and prioritized research tasks.

**8:30 am – 12:00 pm W21: Forms and Flowsheets for FHIR Medical Records - and a Large Medical Record Test Set for Experimenting**

**Clement McDonald, Joel Buchanan, Paul Lynch, Ye Wang, Shennon Lu**

**Abstract:** This workshop will present a NLM/LHCNBC set of tools for inputting and displaying FHIR medical record data. We will introduce a thumbnail overview of FHIR, and also present a large medical record test set for use in experimenting with novel applications built on FHIR medical record. We learned many lessons about efficient ways to load medical record data into FHIR while loading the 54 million records of our test set, and will share those lessons.

The forms and flowsheet tools we will present in this workshop are bookends to a FHIR medical record. In the interactive component of the workshop, participants will learn to use these tools. One tool provides machinery for entering data into FHIR EMR via input forms, and the other tool provides machinery for displaying that content in variety of flowsheet styles. All of the LHC tools are open source and based on standards. Participants will have access to the tools after the workshop.

The input side (https://lhncbc.nlm.nih.gov/project/lhc-forms) of this pair of bookends includes tools for defining input forms, generating forms from these definitions, entering EMR data through them, validating and converting units of measure. These LHC-Forms include skip logic, calculated values, survey instrument scoring, repeating groups of questions, extensive error checks, responsive design and many other capabilities. The attendees will learn to use and modify such forms. All LOINC panels also serve as LHC-Forms. There are more than 2000 such LHC-Forms to play with.

The output side is a flowsheet. Users can choose sets (panels) of terms to include in the flowsheet, show each panel member in a timeline tailored to the panel, and alternatively, one can insert all results in one common timeline. Each variable shows a spark graph at the beginning of its timeline. Yet another display presents the flowsheet in a problem-oriented fashion, including tests, drugs, and imaging studies.

# Sunday, November 4

**3:30 pm – 5:00pm S05: Oral Presentations - Big Data Analytics and Machine Learning**

**Online evaluation of PubMed's new relevance search algorithm**

**Nicolas Fiorini\*, Zhiyong Lu (3:30 PM–3:52 PM)**

**Abstract:** With the rapid growth of biomedical literature in PubMed – about two articles every minute – finding and retrieving the most relevant papers for a given query is increasingly challenging. We demonstrate that Best Match, a recently introduced relevance search in PubMed, provides state-of-the-art retrieval performance in online experiments. Particularly, we find that this positive algorithmic change translates into increased click-through rate and improved user experience in real-world circumstances. Since the new algorithm was fully deployed in June 2017, we have also observed a steady increase (over 30%) in Best Match usage by PubMed users: assisting millions of searches in PubMed on a weekly basis.

# Monday, November 5

## 10:30am – 12:00pm S26: Panel - Data Science in Biomedical Informatics Education: Critical Problems and Innovative Solutions

## Harry Hochheiser, Javed Mostafa, Nils Gehlenborg, Shannon McWeeney, Valerie Florance

## Abstract: The growth of interest in data science presents educators with opportunities in preparing researchers and professionals for future changes in biomedical informatics research and training. The NIH Strategic Plan for Data Science, released as a draft in March 2018, identifies workforce development for biomedical data science as a key goal. In 2017, NLM Training Programs in Biomedical Informatics and Data Science launched an initiative aimed at developing new data science educational materials, curricula, and instructional delivery modalities. The initiative now has reached approximately its halfway mark and recently a meeting was held for 14 training programs in which updates on critical barriers, lessons learned, and new strategies for data science pedagogy and training were collected from all the program leaders. A key goal of the panel is to provide a summative update to a wider audience, based on the cumulative knowledge from all the participating NLM training programs, on core skills and competencies associated with data science, proposed approaches for addressing data science, and the role that data science plays in biomedical informatics education, research, and practice. Panelists will share perspectives on data science education efforts at their institutions and discuss the relationship between data science and biomedical informatics.

## 10:30am – 12:00pm S31: Oral Presentations - Terminology Standards and Mapping

## Re-purposing the ICD-9-CM Procedures Index for Coding in ICD-10-PCS and SNOMED CT

## Kin Wah Fung\*, Julia Xu, Filip Ameye, Arturo Romero-Gutiérrez, Ariel Busquets

## (11:24 AM–11:42 AM)

## Abstract: Compared to the ICD-10-PCS index, the ICD-9-CM Procedure (ICD9V3) Index is richer and contains more clinician-friendly terms including abbreviations and eponyms. We re-purposed the ICD9V3 index by mapping the index terms to SNOMED CT and the ICD-9-CM codes to ICD-10-PCS codes through the General Equivalent Mappings. The re-purposed index outperformed the ICD-10-PCS index in the retrieval of ICD-10-PCS codes using a list of commonly used procedure names, with significantly higher recall, precision and F-score. We also derived a SNOMED CT to ICD-10-PCS map from the re-purposed ICD-9-CM index, which had a higher coverage of SNOMED CT concepts and comparable accuracy compared to a map derived from the ICD-10-PCS index. The re-purposed index will be a useful resource for ICD-10-PCS coders and for mapping between SNOMED CT and ICD-10-PCS.

# Monday, November 5

## 1:45pm – 3:15pm S38: Oral Presentation - Health Information Systems and Design

## Analyzing Real-World Use of Research Common Data Elements

## Vojtech Huser\*, Liz Amos (2:51 PM–3:13 PM)

## Abstract: Common Data Elements (CDEs) are defined as "data elements that are common to multiple data sets across different studies" and provide structured, standardized definitions so that data may be collected and used across different datasets. CDE collections are traditionally developed prospectively by subject-matter and domain experts. However, there has been little systematic research and evidence to demonstrate how CDEs are used in real-world datasets and the subsequent impact on data discoverability. Our study builds upon previous mapping work to investigate the number of CDEs that could be identified using a varying level of commonness threshold in a real-world data repository, the Database of Phenotypes and Genotypes (dbGaP). In an analyzed collection of mapped variables from 426 dbGaP studies, only 1,414 PhenX variables are observed (out of all 24,938 defined PhenX variables). Results include CDEs that are identified with varying levels of commonness thresholds. After the semantic grouping of 68 PhenX variables collected in at least 15 studies (n=15), we observed 32 truly "common" common data elements. We discuss benefits of post-hoc mapping of study data to a CDE framework for purposes of findability and reuse, as well as the informatics challenges of pre-populating clinical research case report forms with data from Electronic Health Record that are typically coded in terminologies aimed at routine healthcare needs.

## 1:45pm – 3:15pm S39: Oral Presentation - Information Extraction and Classification

## Adverse Reactions and Drug-Drug Interaction Extraction tracks at the Text Analysis Conference (TAC)

## Dina Demner-Fushman\*, Joseph Tonning, Kin Wah Fung, Phong Do, Richard Boyce, Kirk Roberts (2:03pm–2:21pm)

**Abstract:** The National Library of Medicine is collaborating with FDA on developing automated approaches to extraction of information from the Structured Product Labels and facilitating research in this area. To that end, NLM and FDA co-organized an Adverse Drug Reaction extraction evaluation in 2017 and a Drug-Drug Interaction extraction evaluation in 2018. We describe the benchmark datasets developed for the evaluations in the shared tasks, the tasks, and the results of the tasks.

## 5:00 pm – 6:30 pm Poster Session 1

**Biomarkers in Prostate and Breast Cancers: Leading Causes of Malignancies in Men and Women**

**Paul Fontelo, Fang Liu\***

**Abstract:** Biomarkers are now routinely used in determining the clinical outcome, diagnosis, prognosis, staging and treatment of human cancers. In May 2017, the FDA for the first time in its history, approved the checkpoint blockade drug Pembrolizumab, solely for biomarker indications instead of tumor histopathology or location. We developed a search tool to discover important biomarkers in human malignancies. We use breast and prostate cancer to demonstrate this new search tool.

# Monday, November 5

**Real World Database for Validation of Units for Clinical Laboratory Tests**

**Vojtech Huser\*, Doyeop Kim, Ajit Londhe**

**Abstract:** Sufficient data quality in large Electronic Health Record (EHR) databases is important to producing valid analytical findings. Existing data quality assessment tools currently lack comprehensive rules for assessing laboratory results recorded in EHR databases. Our study and resulting database (called ThemisUnits) addressed this gap using real world data using multiple international EHR datasets. ThemisUnits database was developed within the Observational Data Science and Informatics (OHDSI) consortium as part of the Themis initiative which aims to arrive at stricter data model specifications that would promote higher semantic interoperability within the OHDSI Common Data Model (CDM; www.ohdsi.org/data-standardization). We obtained data from 13 OMOP datasets. The database contains 437 distinct measurements and 513 measurement-unit pairs. It can serve as a knowledge base for data quality assessment of laboratory data.

**Improving Spelling Correction with Consumer Health Terminology**

**Chris Lu, Dina Demner-Fushman\***

**Abstract:** NLM launched the Consumer Health Information and Question Answering project to provide consumers with reliable health information. A consumer spelling tool (CSpell) was developed for spelling error correction in the NLP pipeline due to the high spelling error rate in consumer questions. A systematic approach was developed to retrieve consumer health terminology (CHT) from the UMLS Metathesaurus/MEDLINE. After adding CHT to the default dictionary in CSpell, we observed a 6% improvement in F1.

**Enhancing Identification of Relation Arguments in SemRep**

**Graciela Rosemblat, Dongwook Shin, Halil Kilicoglu\***

**Abstract**: We implemented and evaluated two argument identification enhancements to SemRep, a biomedical semantic relation extraction tool. A precision-focused rule uses lexical information and prepositional phrase attachment heuristics to address prepositionally-triggered relations, and a recall-focused rule restricts argument scope for relations triggered by hyphenated adjectives. Both rules improved precision, while the first led to some recall loss. Their overall effect for downstream applications is considered positive.

**Searching for Health Information Quality Indicators While Seeking Cures**

**Kira Zhovnirovskii, David Kaufman, Catherine Smith, Anita Murcko, Alla Keselman\***

# Abstract: This pilot case study compares author’s credentials and commercial interests as criteria for evaluating websites that describe natural remedies for type 2 diabetes. The top 25 relevant results of a search for “diabetes reversal natural” were analyzed with respect to content and presentation. Findings suggest that physicians were not less likely to criticize the healthcare establishment or promise a complete cure, but sites that did not sell any products were.

# Tuesday, November 6

## 8:30am – 10:00am S60: Oral Presentation - NLP and Machine Learning for Patient Safety

## Finding medication doses in the literature

## Dina Demner-Fushman\*, James Mork, Willie Rogers, Sonya Shooshan, Laritza Rodriguez, Alan Aronson (8:48 AM–9:06 AM)

## Abstract: Medication doses, one of the determining factors in medication safety and effectiveness, are present in the literature, but only in free-text form. We set out to determine if the systems developed for extracting drug prescription information from clinical text would yield comparable results on scientific literature and if sequence-to-sequence learning with neural networks could improve over the current state-of-the-art. We developed a collection of 694 PubMed Central documents annotated with drug dose information using the i2b2 schema. We found that less than half of the drug doses are present in the MEDLINE/PubMed abstracts, and full-text is needed to identify the other half. We identified the differences in the scope and formatting of drug dose information in the literature and clinical text, which require developing new dose extraction approaches. Finally, we achieved 83.9% recall, 87.2% precision and 85.5% F1score in extracting complete drug prescription information from the literature.

**8:30am – 10:00am** **S64: System Demonstrations - Ontology Driven Health Information Systems Architectures**

**Using RxNav for drug analytics – How to interpret obsolete drug identifiers?**

## Olivier Bodenreider\* (9:30 AM–10:00 AM)

**Abstract:** Analytics was not among the use cases RxNorm was initially designed to support. One specific issue here is that many drug identifiers recorded in clinical data warehouses may no longer be valid in the current release of RxNorm and detailed information about the corresponding drugs may be missing. To address this issue, we have developed specific API functions to support a history mechanism for NDC and RxNorm identifiers.

**10:30am-12:00pm S66: Panel - The Future Ain't What it Used to Be: Negotiating the Changing Digital Health Ecosystem**

**David Kaufman, Catherine Smith, Alla Keselman, Anita Murcko**

**Abstract:** The questions of how to judge the quality of online health information and how to support the public in the difficult task of information evaluation has been a central issue in informatics for many years. However, with the emergence of new Internet technologies, including the spread of multimedia channels and the ubiquity of social media, the need for new approaches to critical evaluation has taken on a new dimension and urgency. Fake news, which names a troubling phenomenon of deceptive communication, has become ubiquitous across media platforms. This raises important concerns for public health and consumer health informatics. The panelists are collaborators in a research program focusing on the assessment of quality of online health information from “unconventional sources,” including those that make claims unsupported by the medical establishment. At the conclusion of this panel, audience members will be able to: (1) Characterize the dimensions of the problem confronting health consumers and patients; (2) Describe evolving criteria to render judgments about the soundness of information resources; (3) Discuss the role that education can play in empowering lay health information seekers; (4) recognize the challenges that physicians are confronted with in helping patients negotiate the morass of online health information.

# Tuesday, November 6

## 1:45pm – 3:15pm S83: Oral Presentation - Data Visualization

## Experimenting new search and user interface in PubMed Labs

## Zhiyong Lu\*, Nicolas Fiorini (2:39 PM–2:57 PM)

## Abstract: PubMed is a freely accessible system for searching the biomedical literature, with approximately 2.5 million users worldwide on an average workday. We have recently developed PubMed Labs (www.pubmed.gov/labs), an experimental platform for users to test new features/tools and provide feedback, which enables us to make more informed decisions about potential changes to improve the search quality and overall usability of PubMed. In doing so, we hope to better meet our user needs in an era of information overload (each day, thousands of new articles are added to PubMed with more than 28 million in total now). Another novel aspect of PubMed Labs lies in its mobile-first and responsive layout, which offers better support for accessing PubMed on the increasingly popular use of mobile and small-screen devices. Currently, PubMed Labs only includes a core subset of PubMed functionalities, e.g. search, facets. We encourage users to test PubMed Labs and share their experience with us, based on which we expect to continuously improve PubMed Labs with more advanced features and better user experience.

## 3:30pm – 5:00pm S88: Panel - Methods and Tools to Enhance Rigor and Reproducibility of Biomedical Research

## Halil Kilicoglu, Aurélie Névéol, Timothy Clark, Hua Xu, Neil Smalheiser

## Abstract: Rigor and reproducibility of biomedical research has been the topic of much debate in recent years. Cases of replication failures, increasing number of retractions, and pervasiveness of questionable research practices lead to a lack of confidence in published findings, indicating that a portion of the biomedical research investment is wasted. Some ongoing efforts aim to address the issues in research conduct and dissemination by focusing mainly on standardization. Guidelines and principles pertaining to data, code, and publications have been proposed. The goal of this didactic panel is to engage the medical informatics community in a discussion about strategies to complement such efforts using informatics methods, tools, and resources. In the panel, first, we will provide a brief overview of standardization initiatives. Next, the panelists will present their informatics-based approaches toward improving rigor and reproducibility of biomedical research, focusing on such areas as information retrieval, natural language processing/text mining, and semantic modeling. Finally, with audience participation, we will discuss challenges facing informatics research aiming to address these problems and seek to identify some potentially fruitful research directions.

# Tuesday, November 6

## 3:30pm – 5:00pm S98: Late Breaking Session - National Health IT Priorities to Advance Research

## Teresa Zayas-Cabán, Jonathan Wald, P. Jon White, Patricia Brennan

**Abstract:** The pace of research and evidence generation is anticipated to accelerate dramatically due to significant increases in the routine use of health information technology (IT) and the introduction of novel sources of electronic data. It is critical for the health IT community to anticipate health IT infrastructure changes required to realize future advances. The Office of the National Coordinator for Health Information Technology (ONC) Chief Scientist Division (CSD), charged with developing and evaluating ONC’s overall scientific efforts and activities, is leading an effort to understand and define the work needed to advance the nation’s health IT infrastructure over the next 3 to 5 years in support of innovative biomedical and health services research. This late-breaking session will describe ongoing work to identify health IT infrastructure gaps and the actions needed to address them. Each presenter will share their perspective briefly, allowing time for participant interaction during the session. The session will address a review of relevant efforts and gaps that emerged from that review, findings from a recent workshop that discussed possible actions needed, and national priorities. An important objective of the session is to inform the informatics community and gather participant input that will inform and strengthen the work being undertaken by ONC.

## 5:00 pm – 6:30 pm Poster Session 2

## The Impact of Inferring Treatments on Information Retrieval for Precision Medicine

## Travis Goodwin\*, Sanda Harabagiu

## Abstract: In 2017, the Precision Medicine track (TREC-PM) of the Text REtrieval Conference (TREC) aimed to facilitate the design of automatic systems capable of providing useful precision-medicine related information to clinicians treating patients with cancer. Given a query describing a patient's tumor, genetic variants, and demographics, systems were evaluated on their ability to retrieve and rank (1) scientific articles from MEDLINE describing targeted treatments, or (2) pertinent clinical trials for which the patient may be eligible. In this poster, we explore whether automatically inferring targeted medications and treatments improved performance in the 2017 TREC-PM evaluation.

## Complete Patient Privacy via Comprehensive Privacy for All

**Mehmet Kayaalp\***

**Abstract:** Despite strong regulations and protections in place, complete patient privacy remains elusive. The main problem lies neither in healthcare nor in tools developed to protect patient privacy. The culprit is the lack of protection of privacy outside of the healthcare sector. We study how other nations attempt to solve the problem, how we can approach the problem in the US, and how informatics would help us reach the solution.

# Tuesday, November 6

**Increasing UMLS Coverage and Reducing Ambiguity via Automated Creation of Synonymous Terms**

**Francois-Michel Lang, James Mork, Dina Demner-Fushman\*, Alan Aronson**

**Abstract:** Extensive synonymy is one of the greatest strengths of the UMLS Metathesaurus. We propose a methodology to further expand this strength by using the existing synonymy to semi-automatically identify new synonymous strings not already in the UMLS and add them as new synonyms within existing UMLS concepts and reduce ambiguity of UMLS mappings.

**Discovering Biomarker Interactions in Cancer Publications**

**Fang Liu\*, Paul Fontelo**

**Abstract:** We propose an approach to discovering prevailing N-gram phrases with biomarkers or common terms from PubMed publications for human cancers. The method was performed on skin cancer publications from 1992 to 2016. It may provide useful information on top biomarkers and related terms on skin cancer. This method may also be applied to other research areas related to biomarkers and cancer research.

**RxNav-in-a-Box – A locally-installable version of RxNav and related APIs**

**Lee Peters, Richard Rice, Olivier Bodenreider\***

**Abstract:** RxNav-in-a-Box provides users with a locally-installable version of the RxNorm APIs, allowing their applications to access the RxNorm data without dependency on NLM servers. Until recently, NLM has not provided a locally-installable version of RxNav and the drug APIs, because of the complexity of the underlying data and tooling. With the recent development of container-based technology and virtualization software, a platform-independent solution in now available.

**Linking Section Labels with Biomedical Relations in SemMedDB**

**Dongwook Shin\*, Halil Kilicoglu**

**Abstract:** We extended SemMedDB, a repository of biomedical semantic relations, by incorporating structured abstract section information for the relations. NLM’s Structured Abstract Label List and Category Mappings were used for this purpose. We found that 39.7% of the relations (from 40.9% of the sentences) in SemMedDB could be associated with explicit section information. This information can serve as a relation filtering mechanism by applications that rely on SemMedDB.