# PSB 2019 Workshop on Text Mining and Visualization for Precision Medicine

Graciela Gonzalez-Hernandez[1†], Zhiyong Lu[2†], Robert Leaman[2†], Davy Weissenbacher[1], Mary Regina Boland[1,4], Yong Chen[1], Jingcheng Du[5], Juliane Fluck[6,7,14], Casey S. Greene[8,9], John Holmes[1], Aditya Kashyap[10], Rikke Linnemann Nielsen[12], Zhengqing Ouyang[13], Sebastian Schaaf[7], Jaclyn N. Taroni[8,9], Cui Tao[5], Yuping Zhang[14], Hongfang Liu[3]

[1]*Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*

[2]*National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), 8600 Rockville Pike, Bethesda, MD 20894, USA*

[3]*Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, MN, USA*

[4]*Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, USA*

[5] *School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, USA*

[6] *ZB MED Information Centre for Life Sciences, Bonn, Germany*

[7] *Department of Bioinformatics, Fraunhofer Institute for Scientific Computing and Algorithms (SCAI), Sankt Augustin, Germany*

[8] *Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*

[9] *Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, PA, USA*

[10] *Data Science Masters Program, University of Pennsylvania, Philadelphia, PA, USA*

[11] *Department of Bio and Health Informatics, Technical University of Denmark, Lyngby, Denmark*

[11] *The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA*

[13] *Department of Statistics, University of Connecticut, Storrs, CT, USA*

[14] *Institute of Geodesy and Geoinformation, University of Bonn, Bonn, Germany*

Precision medicine, an approach for disease treatment and prevention that considers "individual variability in genes, environment, and lifestyle" [1] was endorsed by the National Institutes of Health, aided by the presidential Precision Medicine Initiative (PMI), in 2016. PMI provided funding for cancer research and for building a national cohort of one million or more U.S. participants, now known as the "All of Us" Research Program, which aims to expand its impact to all diseases. PMI was the catalyst to a widespread effort around precision medicine, as evidenced by the more than 1000 grants funded by different NIH institutes in just the last two years. The data being generated by these efforts is growing exponentially, and becomes both the greatest treasure and the greatest challenge for researchers. This workshop is a continuation of a similar session in PSB 2018, providing a forum for researchers with strong background in text mining or natural language processing (NLP) and/or machine learning (ML) who are actively collaborating with bench scientists and clinicians to tackle the challenges brought about by this explosion of data.

† Work partially supported by the National Library of Medicine of the National Institutes of Health (NIH) under grant number R01LM011176 (GGH) and its Intramural Research Program (ZL and RL). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

## 1. Introduction

According to the National Research Council, "personalized medicine" is an older term with a meaning similar to "precision medicine." However, "personalized" could be thought to imply that treatments and preventions are being developed for each individual; in contrast to what is really intended, which is identifying which approaches will be effective for which group of patients based on shared or similar genetic, environmental, and lifestyle factors. Thus, the preferred term for the presidential initiative launched in 2015 was "precision medicine" rather than "personalized medicine", heralding the switch to the later. The Precision Medicine Initiative (PMI) working group report[2] outlines the goals of precision medicine, "to redefine our understanding of disease onset and progression, treatment response, and health outcome", suggests the means to accomplish this, "more precise measurement of molecular, environmental, and behavioral factors that contribute to health and disease", and the expected outcomes "more accurate diagnoses, more rational disease prevention strategies, better treatment selection, and the development of novel therapies". However, in order to go from the means to the outcomes, one must deal with the onslaught of data that those "more precise" measurements entail.

Big data in health is both a blessing and a curse. It is enabling, promising, but has been the largest roadblock to true progress in precision medicine, as much of key information remains hidden in descriptive text or in patterns that are only obvious after cleverly feeding massive amounts of the right data to machine learning algorithms. Selecting, integrating, and analyzing the right data from medical records (EMRs), standardized clinical data (such as what is required by Medicare), administrative data –from hospitals, insurance companies, and pharmacies-, patient surveys and self-reports in social media or health forums, or via wearable sensors, the published literature, clinical trials, and research data deposited in public collections such GenBank or the Gene Expression Omnibus (GEO) database, and many curated databases of interactions and pathways, to name just a few, is one of the major challenges to precision medicine.

Big data and the advance of machine learning, especially deep learning, has led to an explosion of the application of machine learning techniques in precision medicine. For example, deep learning algorithms have been able to diagnose of pneumonia on chest x-ray images[3], apply for personalized risk stratification based on clinical data[4], and detect spread of breast cancer into lymph node tissue on microscopic specimen images[5]. However, there is no silver bullet. The majority of such studies have not been conducted with scientific rigor regarding data reproducibility and model validity/portability in real-world scenario, and are thus limited to the framework and data used for the study itself.

We have also seen significant advances in NLP methods that have enabled unstructured data to be used for decision support systems and predictive algorithms, given that such data was found exclusively in unstructured form, as recent studies comparing text-mining results with curated databases showed[6-8]. Barriers to progress include ambiguity in the data itself, as variant names in the papers are written irregularly and hard to be grounded and even recognized[9,10], as well as lack of trust and standard validation approaches. For example, whereas there is almost universal acceptance of ICD based cohort selection, NLP does not enjoy the same level of trust, and inclusion

of a patient record in a study based solely on NLP based selection will be frowned upon unless it is followed by manual annotation.

This workshop highlights original research and invited presentations on novel text mining, natural language processing (NLP), and visual analytics approaches at the intersection of lifestyle, environment, and genetics that enable further understanding of disease processes and effective treatment for individuals and cohorts that share specific characteristics.

## 2. Workshop Summary

The workshop includes a keynote talks by Christopher Chute, plus 6 oral presentations by authors of abstracts submitted for competitive review and selected for presentation based on their innovation and significance. In addition, the workshop closes with presentations by a panel of experts, focusing on 'Current Challenges in Incorporating Genomic, Clinical, Published, and User-generated Data for Precision Medicine', which gives attendees a view of state of the art approaches and roadblocks to the advancement of text mining and machine learning methods that will enable the next big breakthrough in this area.

### 2.1 *Keynote: Comparability and Consistency of NLP for Biomedical Discovery and Translation*

The keynote talk is given by Dr. Christopher Chute, the Bloomberg Distinguished Professor of Health Informatics, Professor of Medicine, Public Health, and Nursing at Johns Hopkins University, and Chief Research Information Officer for Johns Hopkins Medicine. He received his undergraduate and medical training at Brown University, internal medicine residency at Dartmouth, and doctoral training in Epidemiology and Biostatistics at Harvard. He is Board Certified in Internal Medicine and Clinical Informatics, and an elected Fellow of the American College of Physicians, the American College of Epidemiology, HL7, and the American College of Medical Informatics (ACMI), as well as a Founding Fellow of International Academy of Health Sciences Informatics; he is currently president of ACMI through 2018.

Dr Chute's career has focused on how we can represent clinical information to support analyses and inferencing, including comparative effectiveness analyses, decision support, best evidence discovery, and translational research. He has had a deep interest in semantic consistency, harmonized information models, and ontology. His current research focuses on translating basic science information to clinical practice, and how we classify dysfunctional phenotypes (disease). He became founding Chair of Biomedical Informatics at Mayo Clinic in 1988, retiring from Mayo in 2014, where he remains an emeritus Professor of Biomedical Informatics. He is presently PI on a spectrum of high-profile informatics grants from NIH spanning translational science. He has been active on many HIT standards efforts and chaired ISO Technical Committee 215 on Health Informatics and the World Health Organization (WHO) International Classification of Disease Revision (ICD-11).

### *2.2 Oral Presentations*

In *Development and Validation of the PEPPER Framework (Prenatal Exposure PubMed ParsER) with Applications to Food Additives,* **Mary Regina Boland, Aditya Kashyap, Jiadi Xiong, John**

**Holmes, and Scott Lorch,** note that although environmental factors contribute to 36% of child deaths worldwide, no comprehensive list of all prenatal environmental exposures exists. They present a method called PEPPER: Prenatal Exposure Pubmed ParsER that utilizes all full-text research articles from Pubmed Central to learn the 'state-of-the-field'. They found that of 31,764 prenatal exposure studies, only 53.0% were methodology studies. When PEPPER is coupled with the FDA's food additive database (called EAFUS), PEPPER is able to capture 56.4% of the studied exposures. Prenatal exposure effects of food additives were studied for 176 compounds out of 3,968 (4.4%) compounds contained in EAFUS. Of 16,832 prenatal exposure methodology studies, only 1,886 (11.2%) investigate food additive effects. In total, 3,117 studies investigated prenatal exposure to food additives. The majority of these were methodology studies (60.5%), followed by non-methodology studies (27.2%), PDF only (8.9%) and systematic reviews (3.4%). Prenatal exposure to commonly used food additives (EAFUS category ASP) are rarely studied with a rate of only 0.24% of methodology studies. Surprisingly, there is also a paucity of research on the effects of banned food additives on prenatal development. Of 2,105 research articles investigating banned food additives, only four (0.19%) investigate effects during the prenatal period and only three (0.14%) were methodology studies.

**Jingcheng Du, Yang Xiang, Jing Huang, Xinyuan Zhang, Rui Duan, Jiayi Tong, Jiang Bian, Sahiti Myneni, Yong Chen, and Cui Tao,** in *Mining HPV Vaccination Health Beliefs from Twitter Using Deep Learning: A Longitudinal Analysis of Four-Year Data (2014 - 2017),* focus on understanding the public perceptions of vaccines as it is the first step towards developing effective vaccine promotion strategies to fight against the increase of vaccine refusal and delay observed in the last two decades. Traditional surveying methods suffer significant limitations on accessing large-scale public perceptions. The popularity of social media opens a new dimension. However, most of the studies were focusing on analyzing the frequency rather than contents of social media postings. The accurate understanding of the contents in the perspective of grounded behavior change theories is fundamental for the design of precise and targeted vaccination promotion strategies. According to the authors, their study is the first effort to map Twitter vaccination discussion to the grounded behavior change theory - Health Belief Model. They propose and evaluate a deep learning model and apply the model to automatically and accurately extract vaccination health belief from large-scale Twitter data. The deep learning model shows superiority over machine learning baseline model. They identify manifestation of health belief constructs in Twitter corpus of vaccine discussions in a four-year Twitter dataset.

In *Data integration for prediction of time to insulin in type 2 diabetes patients,* the subject of **Rikke Linnemann Nielsen, Louise Donnelly, Agnes Martine Nielsen, Kaixin Zhou, Bjarne Ersboll, Ewan Pearson, and Ramneek Gupta** present Type an approach to predicting risk of a fast or slow disease progression, which varies between individuals. This variation is captured in electronic medical records of T2D patients and identification of biomarkers that are predictive of diabetes progression can possibly reveal relevant patient subgroups characteristics that may assist clinical decisions in T2D treatment management. In their study they analyze electronic medical records from a cohort-based population in Tayside, UK registered from 1994 to 2010 using machine learning approaches. They investigate if integration of life-style data, anthropometry, biochemical data, drug-prescription data and genetic variants could predict slow and fast progression based on

classification of time to insulin (TTI) in T2D patients using random forest and artificial neural network models. TTI is defined as the first day of insulin treatment or as the clinical need for insulin (HbA1c >8.5% treated with two or more non-insulin diabetes therapies) since the day diagnosis was confirmed by HbA1c. Prediction targets is TTI within year 1, 3 or 5 since time of diagnosis. The best performing ANN models with all data except genetics most accurately predicts T2D patients with fast progression. The authors also discuss inclusion of genetic variants in the machine learning models as well as further longitudinal work with the phenotype.

In neurodegeneration, knowledge on etiologies and underlying mechanisms is still sparse, resulting in late diagnosis and a lack of effective therapies. Until longitudinal studies deliver sufficient data, mining and integrating complementary clinical routine data appears promising. In *Longitudinal visualization of heterogeneous data from neurodegenerative patients for clinical hypothesis generation,* **Sebastian Schaaf, Mischa Uebachs, Vyara Tonkova, Kilian Krockauer, Lisa Langnickel, Philipp Koppen and Juliane Fluck** identify a variety of data sources and create an extraction strategy involving text mining, collecting diagnoses, cognitive test scores, biomarker lab measurements as well as medications. The integration into their longitudinal clinical data model allows a semantic access to normalized data from both routine and study contexts, using standards like FHIR, OMOP and adequate public terminologies. Besides programmatic access, they set up an interactive visualization interface, providing views on aggregated data for exploratory settings, but also a custom longitudinal patient viewer, depicting events and measurements for individuals on a timeline. Beyond supporting principal data exchange and review, they regard the recent developments to be crucial for efficient hypotheses generation, stratification and recruitment.

In *MultiPLIER: a transfer learning framework reveals systemic features of rare autoimmune disease,* **Jaclyn Taroni, Peter Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter Merkel, and Casey Greene** present a feature-representation-transfer approach, MultiPLIER, which consists of training Pathway Level Information ExtractoR (PLIER) models on large compendia comprised of multiple experiments, tissues, and biological conditions and transferring this information to small rare disease datasets. They demonstrate that MultiPLIER better describes biological processes related to more active or severe disease in a rare autoimmune disorder than models trained on individual datasets.

**Yuping Zhang, Zhengqing Ouyang, and Hongyu Zhao** in *A statistical framework for data integration through graphical models with application to cancer genomics,* building on a previous study[11], present the problem of discovering regulatory relationships among heterogeneous genomic variables from biological conditions with potentially shared regulatory mechanisms. The genomic variables can be genetic variants, epigenetic states, and gene expression profiles, etc. The heterogeneous genomic variable types may be binary, categorical, or continuous. The biological conditions can be different tissue types or disease types, etc. They may have both shared and tissue- or disease-specific regulations. The authors develop a new general network estimation framework, named DIG, to jointly learn conditional independence among a set of heterogeneous types of variables across a set of distinct but related conditions. They illustrate the method by integrating mutations and copy number variations, and apply it to COAD and BRCA using TCGA data. Their study identify both common and distinct network modules in COAD and BRCA, which shows that the modules are biologically meaningful.

## References


1. Collins FS, Varmus H. A New Initiative on Precision Medicine. *N Engl J Med*. 2015;372(9):793-795. doi:10.1056/NEJMp1500523.
2. https://www.nih.gov/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf Accessed October 1, 2018.
3. Rajpurkar P, Irvin J, Zhu K, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:171105225 2017.18
4. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine 2018;1:18.
5. Babak Ehteshami Bejnordi, MS1; Mitko Veta, PhD2; Paul Johannes van Diest, MD, PhD3; et al, Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer, JAMA. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585
6. Allot et al. LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC *Nucleic Acids Research*, 2018
7. Singhal et al. Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. *PLoS Comput Biol*, 2016
8. Lee et al. Scaling up data curation using deep learning: An application to literature triage in genomic variation resources *PLoS Comp Biol*, 2018
9. Wei et al. tmVar 2.0: Integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine *Bioinformatics*, 2017.
10 Wei et al tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, 2013.
11. Zhang, Y., Ouyang, Z. and Zhao, H., 2017. A statistical framework for data integration through graphical models with application to cancer genomics. *The Annals of Applied Statistics*, 11(1), pp.161-184.