High-throughput Phenotyping on Electronic Health Records using Multi-Tensor Factorization[1]

Jimeng Sun (GA Tech), Joydeep Ghosh (UT Austin), Abel Kho (Northwestern), Joshua Denny & Bradley Malin (Vanderbilt)

Electronic health records (EHR) are increasingly an important source of detailed patient information. They provide insight into diagnoses [10, 14, 17, 18, 20, 21] as well as prognoses [4, 8, 16, 19], and can assist in the development of cost-effective treatment and management programs [1, 5, 8, 12, 13, 17]. However, there are formidable challenges to applying EHR data in clinical research, including:

- Diverse populations: Unlike carefully curated data from randomized control trials, EHR data cover diverse patient populations from providers who use different and incompatible EHR systems. How can we effectively gather, reconcile and analyze such large-scale, potentially diverse patient populations for retrospective and prospective analysis?
- Heterogeneous information: EHRs record a variety of inter-related aspects of patients, such as symptoms, diagnoses based on billing codes, procedures, medication orders, lab tests, physiological readings, and free-form clinical notes. How can we concisely represent such rich information and succinctly capture pertinent EHR data in terms of meaningful phenotypes?
- Noisy information: EHR data often reflect an incomplete and inaccurate representation of patients. How should we handle such noisy and missing information to extract robust signals about patients?
- Interpretation: Medical practitioners tend to be (for good reasons) conservative, such that they are reluctant to act on recommendations unless they can understand the findings and reconcile them with existing domain knowledge. For example, small improvements in accuracy to clinical decision support tools based on a "black box" may have little influence.
- Longitudinal information: EHR data captures certain patient readings as sparsely sampled event sequences with different time scales. How can we identify the temporal evolution of patients (such as disease progression and severity changes) using such data, particularly in conjunction with other resources, such as demographics and past treatments?

Clinical research requires accurate and concise clinical concepts (or phenotypes) about patients. Clinical scientists are accustomed to reasoning based on well-defined phenotypes rather than directly on high-dimensional EHR data [3, 11, 15]. The transformation from EHR data into useful phenotypes, or *phenotyping* is fundamental to EHR-based studies as they are used to obtain study cohorts.Useful phenotypes should capture multiple aspects of the patients (e.g., diagnosis, medication and lab results) and be both sensitive and specific to the target patient population. However, EHR data are typically collected in operational settings for billing and management purposes, and are not designed for clinical research. Thus, they often do not readily map to simple, let alone more sophisticated and multifaceted, phenotypes. Meanwhile, current approaches for translating EHR data into useful phenotypes are typically slow, manually intensive and limited in scope.

Given the above observations, we are working on an integrated research project on high-throughput phenotyping. This paper provides an overview of the entire project across four institutions. In this project, we try to answer the following questions:

- How can we transform such large volumes of heterogeneous, longitudinal and noisy EHR data into concise and meaningful phenotypes with minimal human intervention?
- How can we effectively involve clinical experts to refine these phenotypes in a manner that minimizes their effort?
- How can we adapt and unify transformations from different datasets so that they scale and generalize across multiple organizations?

We believe that these questions can be answered by developing a new computational learning platform based on tensors [9] because they provide a natural framework for flexible representation and analysis of

[1]This work is funded by NSF Smart Connected Health Program Award number 1418511.

multi-aspect data.

Generally speaking, tensors are high-order generalizations of matrices (2nd order tensors) and vectors (1st order tensor). Tensor factorizations are a natural generalization of data mining methodologies. Many traditional techniques for dimensionality reduction, such as principal components analysis (PCA)and topic modeling [2], are special cases of tensor factorization against second order tensors (i.e., matrices). As an example, consider that medication order information may be captured by a 3rd order tensor with 3 modes, where each mode is an aspect of a tensor: a) patient, b) medication and c) diagnosis. Techniques for tensor analysis can capture the simultaneous interaction of multiple modes; for example, to identify a small subset of diagnoses and medications that are frequently observed together in a subgroup of patient EHRs. Thus, such analysis can provide a powerful, data-driven approach to discovering phenotypes.

EHR data have many aspects that suggest they are better represented as multiple, interconnected tensors rather than a single, high-order tensor or a set of relational tables (which is the default at the present time). Fig 1 shows how some of the information about Medicare patient data, available from the Centers for Medicare and Medicaid Services (CMS), can be represented in such a form.[2] Specifically, it depicts two third order tensors and two ordinary relational tables. The three modes for tensor $\mathcal{Y}$ are patients, diagnoses and procedures. In this example, the patient mode is common to all four components, while, in general, different modes may connect different subsets of tensors or tables. It should also be noted that this data representation can be further augmented through external data sources. For instance, information from the provider referral network made public by the MedStatr project [3] can yield a provider-provider relation that does not involve a patient mode. Moreover, non-traditional health-related information, such as social networks may also be provided by patients who are part of health support groupsadding yet another tensor to Fig 1.

Once EHR data are modeled as multiple interconnected tensors, we can develop algorithms to jointly factorize them and extract latent factors that correspond to phenotypes. There are many design considerations for such algorithms that will be investigated throughout the project. And, we note this work builds on a successful pilot study using nonnegative tensor factorization on a fairly substantial dataset of over 30,000 patients [7], for which over 80% of the automatically extracted factors were deemed by a domain expert to correspond to meaningful clinical concepts. We then further extended that work to a more computationally stable algorithm [6].

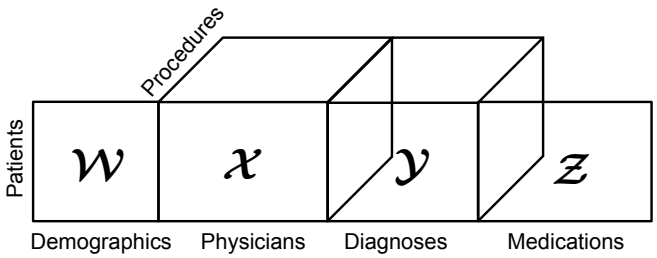

Figure 1: Multi-relational representation of CMS data.

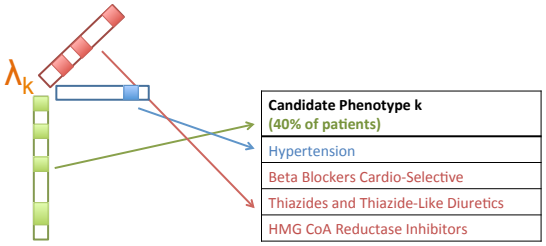

Figure 2: Hypertension phenotype example.

Overall, this project aims at developing a comprehensive computational framework that will substantially increase our ability to simultaneously cater to multiple aspects of large-scale EHR data, leading to a suite of tools for personalized health and well-being. These tools can generalize to data across multiple institutions and will help identify existing, as well as novel, phenotypes; thus making them much more likely to be readily accepted and used in clinical research and practice. The proposed work will result in a variety of phenotypes that are derived from real EHR data and vetted by domain experts. The accompanying suite of algorithms and methods will provide relatively automated ways of (high-throughput) phenotype generation, refinement, adaptation, and application, in a broad range of health informatics settings. This includes the

[2]This version of tensors is derived from CMS data from the Research Data Assistance Center (ResDac, `http://www.resdac.org`) through Claims and Claims Line Feed (CCLF) data files.

[3]`http://www.medstartr.com/projects/93-phase-ii-next-level-doctor-social-graph`

analysis of billing data for Medicare patients for proactive patient management and the prediction of progress (or deterioration) of patients already admitted to ICUs. This suite will reflect algorithmic innovations, as well as adaptation to specific characteristics of EHR data and clinical practice.

References

[1] Bennett, C., Doub, T., and Selove, R. EHRs connect research and practice: Where predictive modeling, artificial intelligence, and clinical decision support intersect. *Health Policy and Technology 1*, 2 (June 2012), 105–114.

[2] Blei, D., Ng, A., and Jordan, M. Latent Dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.

[3] Cimino, J. J. From data to knowledge through concept-oriented terminologies: Experience with the Medical Entities Dictionary. *Journal of the American Medical Informatics Association 7*, 3 (May 2000), 288–297.

[4] Ebadollahi, S., Sun, J., Gotz, D., Hu, J., Sow, D., and Neti, C. Predicting patient's trajectory of physiological data using temporal trends in similar patients: A system for near-term prognostics. *AMIA Annual Symposium Proceedings 2010* (2010), 192–196. PMID: 21346967.

[5] Greengard, S. A new model for healthcare. *Communications of the ACM 56*, 2 (2013), 17–19.

[6] Ho, J., Ghosh, J., and Sun, J. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *KDD* (2014).

[7] Ho, J. C., Ghosh, J., Steinhubl, S., Stewart, W., Denny, J. C., Malin, B. A., and Sun, J. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Submitted for review*.

[8] Jensen, P. B., Jensen, L. J., and Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews: Genetics 13*, 6 (June 2012), 395–405.

[9] Kolda, T. G., and Bader, B. W. Tensor decompositions and applications. *SIAM Review 51*, 3 (2009), 455–500.

[10] Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine 23*, 1 (Aug. 2001), 89–109.

[11] Ledley, R. S., and Lusted, L. B. Reasoning foundations of medical diagnosis. *M.D. computing : computers in medical practice 8*, 5 (Sept. 1991), 300–315.

[12] Ramakrishnan, N., Hanauer, D., and Keller, B. Mining electronic health records. *Computer 43*, 10 (2010), 77–81.

[13] Romano, M. J., and Stafford, R. S. Electronic health records and clinical decision support systems: impact on national ambulatory care quality. *Archives of internal medicine 171*, 10 (May 2011), 897–903.

[14] Roque, F. S., Jensen, P. B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T., Jensen, L. J., and Brunak, S. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Computational Biology 7*, 8 (Aug. 2011), e1002141.

[15] Rose, J. S. *Medicine and the information age*. American College of Physician Executives, Tampa, FL, 1998.

[16] Saria, S., Rajani, A. K., Gould, J., Koller, D., and Penn, A. A. Integration of early physiological responses predicts later illness severity in preterm infants. *Science Translational Medicine 2*, 48 (Sept. 2010), 48ra65–48ra65.

[17] Savage, N. Better medicine through machine learning. *Communications of the ACM 55*, 1 (Jan. 2012), 17–19.

[18] Sun, J., Hu, J., Luo, D., Markatou, M., Wang, F., Edabollahi, S., Steinhubl, S. E., Daar, Z., and Stewart, W. F. Combining knowledge and data driven insights for identifying risk factors using electronic health records. *Proceedings of the American Medical Informatics Association Annual Symposium* (2012), 901–910.

[19] Sun, J., Sow, D. M., Hu, J., and Ebadollahi, S. A system for mining temporal physiological data streams for advanced prognostic decision support. In *ICDM* (2010), pp. 1061–1066.

[20] Visweswaran, S., Angus, D. C., Hsieh, M., Weissfeld, L., Yealy, D., and Cooper, G. F. Learning patient-specific predictive models from clinical data. *Journal of Biomedical Informatics 43*, 5 (Oct. 2010).

[21] Wu, J., Roy, J., and Stewart, W. F. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical Care 48*, 6 Suppl (June 2010), S106–13.