INTEGRATING COMMUNITY-LEVEL DATA RESOURCES FOR PRECISION MEDICINE RESEARCH

William S. Bush, PhD, Dana C. Crawford, PhD, Farren Briggs, PhD, Darcy Freedman, PhD

Department of Population and Quantitative Health Sciences, Case Western Reserve University
Cleveland, OH, 44106, USA
Email: wsb36@case.edu; dana.crawford@case.edu; farren.briggs@case.edu; daf96@case.edu

Chantel Sloan, PhD

Department of Health Science, Brigham Young University
Provo, UT, 84602, USA
Email: chantel.sloan@byu.edu

Precision Medicine focuses on collecting and using individual-level data to improve healthcare outcomes. To date, research efforts have been motivated by molecular-scale measurements, such as incorporating genomic data into clinical use. In many cases however, environmental, social, and economic factors are much more predictive of health outcomes, yet are not systematically used in clinical practice due to the difficulties in measurement and quantification. Advances in both the availability of electronic health information, environmental exposure data, and the more systematic use of geo-coding now provide ways to systematically assess community-level indicators of health, and link these factors to electronic health records for evaluating their influence on disease outcomes. In this workshop, we discuss new electronic sources of community-level data, and provide insight into their utility and validity when compared with gold-standard data collection approaches.

1. Introduction

From the earliest efforts to identify genetic polymorphisms influencing drug response (Meyer 2004), a fundamental goal of medical practice is to tailor clinical care to the precise, individual attributes of a patient. Recent years has seen the expansion of precision medicine, with the collection of full genomic sequence data for pharmacogenomics studies (Bush et al. 2016; Rasmussen-Torvik et al. 2014). Given the dramatic and rapid expansion of knowledge in this area, it is now widely accepted that physicians cannot digest the literature fast enough to implement research findings in clinical practice (Johansen Taber and Dickinson 2014). Resources such as the Clinical Pharmacogenomics Implementation Consortium (CPIC) assist by creating clinical practice guidelines for drug prescriptions (Caudle et al. 2014). Clinical decision support systems also provide ways to implement pharmacogenomics in the clinic at point of care (Pulley et al. 2012).

While precision medicine research has focused largely on making detailed molecular measurements for each patient, efforts in the Precision Medicine Initiative will capture additional data elements including behavioral, psychosocial, and environmental factors (Collins and Varmus 2015). Part of the motivation for the expansion of scope is the increasing recognition that social, environmental, and behavioral factors are likely highly influential in disease etiology, progression, and treatment (Woolf et al. 2007). In fact, a growing body of work illustrates that considering biological and social risk factors together as a *system* may lead to better intervention strategies

(Vaughn, DeLisi, and Matto 2013). This idea of a "cells to society" model provides a framework to understand and prioritize the multitude of influential factors that comprise patient health trajectories. To date, precision medicine research has focused much more heavily on the "cells" side of this model, despite the tremendous potential of understanding and addressing the "society" side. In this workshop, we consider multiple scales on which we can derive social, environmental, and behavioral risk factors; information on the individual level, information on the community level, and information on the geographic level.

2. Information on the Individual Level

As electronic health record systems (EHRs) have increased in their adoption, information gathered from a patient encounter is now recorded digitally. Much like their paper predecessors, EHRs serve as a clearinghouse of patient-centric data gathered from clinical resources – laboratory values, vital signs, diagnoses, and procedures. Other select factors may be patient-reported, such as race/ethnicity, income, and employment, but these are rarely used to support clinical care or decision making (Community health centers leveraging the social determinants of health 2012). Despite their established importance in health, these, and other nonclinical factors, are often recorded manually and are not standardized (DeVoe et al. 2016). The National Academy of Medicine has noted the limited collection of social and behavioral factors, and has suggested new standardized data fields (based on validated instruments) for systematic adoption (Adler and Stead 2015), however these are part of a third tier of EHR meaningful use criteria, which is likely many years away.

Some information on social determinants of health is reported in ancillary EHR data through routine clinical communications (i.e. intake questions). Indicators of socio-economic status, such as measures of occupational prestige, unemployment, education, and homelessness (Hollister et al. 2016), along with country-of-origin (Farber-Eger et al. 2017) have been extracted from clinical free-text. Smoking status and alcohol use have also been extracted using natural language processing (Chen and Garcia-Webb 2014; Savova et al.). While text extraction has limitations, the use of structured elements (billing codes, etc.) may also be subject to reporting bias as well (Men 2015). Furthermore, data extracted from clinical text ultimately relies on patient self-report, which is also subject to reporting bias and differences in health perception (Campos-Castillo and Anthony 2015; Sen 2002). For many of measures like alcohol use (Bradley et al. 2011), data suggest that patients may be reluctant to divulge sensitive information associated with social stigma, also known as social desirability bias (Althubaiti 2016).

3. Information on the Community Level

Much like pharmacogenomics, the concept of using community data to inform clinical care in the US dates back to the 1960s (Adashi, Geiger, and Fine 2010), however early basic concepts, such as organizing patient charts by family and neighborhood (Froom 1977) were never widely adopted and have not transitioned to modern EHR systems. Now, due to the digitization and public availability of data, there are unprecedented opportunities to gather community-level data for precision medicine studies.

The Accelerating Data Value Across a National Community Health Center Network (ADVANCE) pilot study has begun conducting electronic assessments of the built environment,

environmental exposures, and neighborhood economic conditions to synthesize a "community vital sign" (Bazemore et al. 2016). This work is similar to other assessments, like the SocioEconomic Status Index (Roblin 2013) and the Neighborhood Deprivation Index (Messer et al. 2006). Another critical component of the built environment is access to healthy food; multiple studies have examined the impact of perceived availability of fast food and healthy food options (Barnes et al. 2016, 2017). While some information on community status can also be obtained through self-reporting, perceptual biases may be especially problematic when defining community-level characteristics – how do you self-report the status of your community relative to others? External, objective sources of community-level data may provide better estimates of their health impact.

4. Information on the Geographic Level

Geographic Information Systems (GIS) have become an extremely useful tool for analyzing publicly available geospatial information (Steiniger and Bocher 2009), such as census and environmental exposure data. These data resources have been used to derive community-level metrics integrated with EHR data, including characteristics of the walkable built environment in Pennsylvania (Nau et al. 2015), Massachusetts (D. T. Duncan et al. 2014), and Ohio (Roth et al. 2014). Similar approaches have been applied to characterize the food environment (Fiechtner et al. 2015, 2016). There are also multiple sources of atmospheric pollution in the US (B. Duncan 2014), which can provide insights into pulmonary conditions (C. D. Sloan and Johnston 2016), along with other transient, spatio-temporal factors like weather (C. Sloan et al. 2017).

Geographic data has also been combined with electronic health record information to assess the distribution of preventable emergency department visits (Fishman 2015), to identify geographic risk factors for sexually transmitted infections (Comer et al. 2011), and to assess asthma risk (Xie et al. 2017). Importantly, studies that have performed geospatial mapping of EHR data find high concordance to traditionally collected studies like the Centers for Disease Control 500 Cities Project (Birkhead 2017; CDC 2017).

5. Closing

It is clear that effective precision medicine will require a complete understanding of the patient's current health status, including risk factors from cells to society, to forecast disease development and implement treatment response. Standards for data collection are common for clinical data derived from physical examinations, and even the collection of genetic data must now adhere to Clinical Laboratory Improvement Amendments (CLIA)-certified processes, but many EHRs are missing complete and uniform documentation of environmental, social, and behavioral contexts despite their strong influence in disease processes and treatment outcomes. Despite this, the realization of a cells to society vision of precision medicine is within reach. EHRs continue to evolve, along with the infrastructure to collect, store, and integrate these community-level data into EHRs and research databases to enable the *precision* of precision medicine on multiple scales.

6. References

Adashi, Eli Y., H. Jack Geiger, and Michael D. Fine. 2010. "Health Care Reform and Primary Care — The Growing Importance of the Community Health Center." *New England Journal of Medicine* 362(22): 2047–50.
Adler, Nancy E., and William W. Stead. 2015. "Patients in Context — EHR Capture of Social and Behavioral

Determinants of Health.” *New England Journal of Medicine* 372(8): 698–701.

Althubaiti, Alaa. 2016. “Information Bias in Health Research: Definition, Pitfalls, and Adjustment Methods.” *Journal of multidisciplinary healthcare* 9: 211–17. https://www.dovepress.com/information-bias-in-health-research-definition-pitfalls-and-adjustment-peer-reviewed-article-JMDH (October 2, 2017).

Barnes, Timothy L et al. 2016. “Geographic Measures of Retail Food Outlets and Perceived Availability of Healthy Foods in Neighbourhoods.” *Public Health Nutrition* 19(8): 1368–74.

Barnes, Timothy L. et al. 2017. “Do GIS-Derived Measures of Fast Food Retailers Convey Perceived Fast Food Opportunities? Implications for Food Environment Assessment.” *Annals of Epidemiology* 27(1): 27–34.

Bazemore, Andrew W et al. 2016. “*‘Community Vital Signs’*: Incorporating Geocoded Social Determinants into Electronic Records to Promote Patient and Population Health.” *Journal of the American Medical Informatics Association* 23(2): 407–12.

Birkhead, Guthrie S. 2017. “Successes and Continued Challenges of Electronic Health Records for Chronic Disease Surveillance.” *American journal of public health* 107(9): 1365–67.

Bradley, Katharine A et al. 2011. “Quality Concerns with Routine Alcohol Screening in VA Clinical Settings.” *Journal of general internal medicine* 26(3): 299–306.

Bush, William S et al. 2016. “Genetic Variation among 82 Pharmacogenes: The PGRN-Seq Data from the eMERGE Network.” *Clinical pharmacology and therapeutics*.

Campos-Castillo, Celeste, and Denise L Anthony. 2015. “The Double-Edged Sword of Electronic Health Records: Implications for Patient Disclosure.” *Journal of the American Medical Informatics Association* 22(e1): e130–40.

Caudle, Kelly E et al. 2014. “Incorporation of Pharmacogenomics into Routine Clinical Practice: The Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline Development Process.” *Current drug metabolism* 15(2): 209–17.

CDC. 2017. “500 Cities | About the Project.” *8/28/2017*. https://www.cdc.gov/500cities/about.htm (October 2, 2017).

Chen, Es, and M Garcia-Webb. 2014. “An Analysis of Free-Text Alcohol Use Documentation in the Electronic Health Record: Early Findings and Implications.” *Applied clinical informatics* 5(2): 402–15.

Collins, Francis S., and Harold Varmus. 2015. “A New Initiative on Precision Medicine.” *New England Journal of Medicine* 372(9): 793–95.

Comer, Karen Frederickson et al. 2011. “Incorporating Geospatial Capacity within Clinical Data Systems to Address Social Determinants of Health.” *Public health reports (Washington, D.C. : 1974)* 126 Suppl 3(Suppl 3): 54–61.

Community Health Centers Leveraging the Social Determinants of Health. 2012. http://www.altfutures.org/pubs/leveragingSDH/IAF-CHCsLeveragingSDH.pdf.

DeVoe, Jennifer E et al. 2016. “Perspectives in Primary Care: A Conceptual Framework and Path for Integrating Social Determinants of Health Into Primary Care Practice.” *Annals of family medicine* 14(2): 104–8.

Duncan, Brian. 2014. “Satellite Data of Atmospheric Pollution for U.S. Air Quality Applications: Examples of Applications, Summary of Data End-User Resources, Answers to FAQs, and Common Mistakes to Avoid.” *Atmospheric Environment* 94: 647–62.

Duncan, Dustin T et al. 2014. “Characteristics of Walkable Built Environments and BMI Z-Scores in Children: Evidence from a Large Electronic Health Record Database.” *Environmental health perspectives* 122(12): 1359–65.

Farber-Eger, Eric et al. 2017. “Extracting Country-of-Origin from Electronic Health Records for Gene- Environment Studies as Part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Study.” *AMIA Joint Summits on Translational Science* 2017: 50–57.

Fiechtner, Lauren et al. 2015. “Food Environments and Childhood Weight Status: Effects of Neighborhood Median Income.” *Childhood Obesity* 11(3): 260–68.

———. 2016. “Effects of Proximity to Supermarkets on a Randomized Trial Studying Interventions for Obesity.” *American journal of public health* 106(3): 557–62.

Fishman, Jamie Andrew. 2015. “Geospatial Analysis of Preventable Emergency Department Visits in Chicago, IL.” https://www.ideals.illinois.edu/handle/2142/78400 (October 2, 2017).

Froom, J. 1977. “An Integrated Medical Record and Data System for Primary Care. Part 1: The Age-Sex Register: Definition of the Patient Population.” *The Journal of family practice* 4(5): 951–53.

Hollister, Brittany M et al. 2016. “DEVELOPMENT AND PERFORMANCE OF TEXT-MINING ALGORITHMS TO EXTRACT SOCIOECONOMIC STATUS FROM DE-IDENTIFIED ELECTRONIC HEALTH

RECORDS." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 22: 230–41.

Johansen Taber, Katherine, and Barry Dickinson. 2014. "Pharmacogenomic Knowledge Gaps and Educational Resource Needs among Physicians in Selected Specialties." *Pharmacogenomics and Personalized Medicine* 7: 145.

Men, Jessica. 2015. "Obesity and Smoking, Alcohol Use Underreported in National Database." *AJMC*. http://www.ajmc.com/newsroom/obesity-and-smoking-alcohol-use-underreported-in-national-database-

Messer, Lynne C. et al. 2006. "The Development of a Standardized Neighborhood Deprivation Index." *Journal of Urban Health* 83(6): 1041–62.

Meyer, Urs A. 2004. "Pharmacogenetics - Five Decades of Therapeutic Lessons from Genetic Diversity." *Nature reviews. Genetics* 5(9): 669–76.

Nau, Claudia et al. 2015. "Community Socioeconomic Deprivation and Obesity Trajectories in Children Using Electronic Health Records." *Obesity* 23(1): 207–12.

Pulley, J M et al. 2012. "Operational Implementation of Prospective Genotyping for Personalized Medicine: The Design of the Vanderbilt PREDICT Project." *Clinical pharmacology and therapeutics* 92(1): 87–95.

Rasmussen-Torvik, L J et al. 2014. "Design and Anticipated Outcomes of the eMERGE-PGx Project: A Multicenter Pilot for Preemptive Pharmacogenomics in Electronic Health Record Systems." *Clinical pharmacology and therapeutics* 96(4): 482–89.

Roblin, Douglas W. 2013. "Validation of a Neighborhood SES Index in a Managed Care Organization." *Medical Care* 51(1): e1–8.

Roth, Caryn, Randi E Foraker, Philip R O Payne, and Peter J Embi. 2014. "Community-Level Determinants of Obesity: Harnessing the Power of Electronic Health Records for Retrospective Data Analysis." *BMC medical informatics and decision making* 14(1): 36.

Savova, Guergana K et al. "Mayo Clinic NLP System for Patient Smoking Status Identification." *Journal of the American Medical Informatics Association : JAMIA* 15(1): 25–28.

Sen, Amartya. 2002. "Health: Perception versus Observation." *BMJ (Clinical research ed.)* 324(7342): 860–61.

Sloan, Chantel et al. 2017. "The Impact of Temperature and Relative Humidity on Spatiotemporal Patterns of Infant Bronchiolitis Epidemics in the Contiguous United States." *Health & Place* 45: 46–54.

Sloan, Chantel D., and James D. Johnston. 2016. "Can Extreme Air Pollution Events Provide a Window into Incident Asthma?" *American Journal of Respiratory and Critical Care Medicine* 194(12): 1440–41.

Steiniger, Stefan, and Erwan Bocher. 2009. "An Overview on Current Free and Open Source Desktop GIS Developments." *International Journal of Geographical Information Science* 23(10): 1345–70.

Vaughn, Michael G., Matt. DeLisi, and Holly C. Matto. 2013. *Human Behavior : A Cell to Society Approach*.

Woolf, Steven H, Robert E Johnson, Robert L Phillips, and Maike Philipsen. 2007. "Giving Everyone the Health of the Educated: An Examination of Whether Social Change Would Save More Lives than Medical Advances." *American journal of public health* 97(4): 679–83..

Xie, Sherrie, Rebecca Greenblatt, Michael Z Levy, and Blanca E Himes. 2017. "Enhancing Electronic Health Record Data with Geospatial Information." *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2017: 123–32.