Tracking Health Disparities Through Natural-Language Processing

Mark L. Wieland, MD, MPH, Stephen T. Wu, PhD, Vinod C. Kaggal, BS, and Barbara P. Yawn, MD, Msc

Health disparities and solutions are heterogeneous within and among racial and ethnic groups, yet existing administrative databases lack the granularity to reflect important sociocultural distinctions. We measured the efficacy of a natural-language–processing algorithm to identify a specific immigrant group. The algorithm demonstrated accuracy and precision in identifying Somali patients from the electronic medical records at a single institution. This technology holds promise to identify and track immigrants and refugees in the United States in local health care settings. (*Am J Public Health.* 2013; 103:448–449. doi:10.2105/AJPJ.2012.300943)

Characterizing and closing the gap of racial and ethnic health disparities is a national priority,[1(p4)] but disparities and solutions are heterogeneous for different groups.[2] For example, a specific health-related assessment and intervention may take very different forms when applied to a Somali American community than to an ancestral African American community. This example reveals an important limitation of health disparities research: existing regional and national databases lack the granularity to reflect this sociocultural heterogeneity. Therefore, assessment of disease prevalence and intervention impact is compromised by the labeling of both communities in our example as African American in existing databases. Adding this texture to administrative databases has been recommended, but implementation is costly and many years away.[3]

Natural-language processing (NLP) holds the potential to bypass these limitations. NLP is an informatics discipline that allows computers to process and understand human languages. Application of NLP to the health care arena is an active area of research with escalating opportunity for impact in the context of a national mandate to expand electronic medical record (EMR) infrastructure. A recent demonstration project showed that NLP review of a health care system EMR outperformed administrative databases in documenting postoperative complications.[4]

We tested the hypothesis that application of NLP to EMRs can identify a subset racial/ethnic group for the purposes of eventually documenting and tracking health disparities. Persons from Somalia compose the largest African refugee population in the United States, with a particular concentration in Minnesota. Furthermore, data support the existence of health care disparities among this population.[5,6] Therefore, we designed our NLP tool to identify this population.

METHODS

We conducted our study at a large academic medical center in the midwestern United States that serves a relatively large regional Somali population, Mayo Clinic in Rochester, Minnesota. For Somali cohort identification, we used a tool with proven effectiveness in finding specific, customized clinical terms for high-throughput phenotyping.[7] This is a rule-based NLP algorithm in which it is possible to encode a customized dictionary and search the unstructured text of EMRs for inclusion and exclusion criteria. We reconfigured the tool for cohort identification so that descriptive terms such as "Somali" and "refugee" (and their variants) ruled in patients, and other terms ruled them out. We constructed the algorithm for local demographic context as follows:

> Somali OR Somalian OR Somalia OR Immigrant OR Refugee NOT Spanish NOT Hispanic NOT Latino NOT Latina NOT Mexican NOT Mexico NOT Cambodian NOT Cambodia NOT Vietnamese NOT Vietnam NOT Sudanese NOT Sudan.

We applied the algorithm to a set of all patients aged 18 years and older who were seen in the outpatient primary care clinics in the divisions of Primary Care Internal Medicine and Family Medicine during a 15-day period in March 2011. A single clinician with experience caring for Somali patients manually reviewed all charts to identify patients as Somali or not Somali. First and last names were used to identify possible Somali patients; EMRs of these patients were then reviewed for direct or indirect documentation of Somali ancestry.

We used the results of this manual chart review as a gold standard for evaluating the efficacy of the algorithm for identifying Somali patients. We calculated sensitivity, specificity, positive predictive value, and negative predictive value for the algorithm.

RESULTS

We identified 5782 patients during the study interval; the NLP algorithm identified 122 of these patients as Somali. Compared with manual identification, the algorithm demonstrated sensitivity of 92.2%, specificity of 99.9%, positive predictive value of 97.5%, and negative predictive value of 99.8%.

Error analysis showed that the EMR for each of the 10 false negatives contained the term Somali, but the algorithm ruled the patient out because of other factors.

DISCUSSION

In this demonstration project, an NLP algorithm showed accuracy and precision in identifying patients from a subset immigrant group. This was a single-center study, with resultant implications for generalizability.

Future research should work to develop and disseminate a generalized NLP cohort identification tool for use in identifying patients from other vulnerable populations. It will be important that these tools have the ability to interact with users to create customizable cohorts that incorporate regional knowledge of populations. For example, exclusionary terms used in our algorithm were informed by local demographic data. This technology holds promise to identify and track immigrants and refugees in the United States at a local health care level, paving the way for improved patient care and reduction of health disparities. ■

About the Authors

Mark L. Wieland is with the Division of Primary Care Internal Medicine and Stephen T. Wu and Vinod C. Kaggal are with the Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN. Barbara P. Yawn is with the Department of Research, Olmsted Medical Center, Rochester, MN.

Correspondence should be sent to Mark L. Wieland, 200 First St SW, Rochester, MN 55904 (e-mail: wieland.mark@mayo.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This article was accepted June 11, 2012.

Contributors

M. L. Wieland and S. T. Wu conceptualized the study, oversaw the analysis, and led the writing of the article. V. C. Kaggal performed the analysis and contributed to writing the article. B. P. Yawn conceptualized and supervised the study and contributed to writing the article. All authors edited and approved the article.

Acknowledgements

This work was funded in part by Mayo Clinic's Center for Individualized Medicine and the SHARPn (Strategic Health IT Advanced Research Projects) Area 4: Secondary Use of EHR Data (US Department of Health and Human Services grant 90TR000201).

Human Participant Protection

The study procedures were approved by the institutional review board at Mayo Clinic.

References

1. Institute of Medicine. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care.* Washington, DC: National Academies Press; 2003.

2. Sue S, Dhindsa MK. Ethnic and health disparities research: issues and problems. *Health Educ Behav.* 2006;33(4):459–469.

3. Desai J. State-based diabetes surveillance among minority populations. *Prev Chronic Dis.* 2004;1(2):A03.

4. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA.* 2011;306(8):848–855.

5. Wieland ML, Morrison TB, Cha SS, Rahman AS, Chaudhry R. Diabetes care among Somali immigrants and refugees. *J Community Health.* 2012;37(3):680–684.

6. Morrison TB, Wieland ML, Cha SS, Rahman AS, Chaudhry R. Disparities in preventive health services among Somali immigrants and refugees. *J Immigr Minor Health.* 2012;14(6):968–974.

7. Savova GK, Fan J, Ye Z, et al. Discovering peripheral arterial disease cases from radiology notes using natural language processing. In: *AMIA Annual Symposium Proceedings.* Bethesda, MD: American Medical Informatics Association; 2010:722–726.

Young Adult Smoking Cessation: Predictors of Quit Attempts and Abstinence

Lori M. Diemert, MSc, Susan J. Bondy, PhD, K. Stephen Brown, PhD, and Steve Manske, PhD

We examined young adult smoking cessation behaviors, coding cessation behavior as no attempt, quit attempt (< 30 days), or abstinence (≥ 30 days) during follow-up from July 2005 through December 2008, observed in 592 young adult smokers from the Ontario Tobacco Survey. One in 4 young adults made an attempt; 14% obtained 30-day abstinence. Cessation resources, prior attempts, and intention predicted quit attempts, whereas high self-efficacy, using resources, having support, and low addiction predicted abstinence, indicating that young adult smokers require effective and appropriate cessation resources. (*Am J Public Health.* 2013;103:449–453. doi:10.2105/AJPJ.2012.300878)

Young adults have had the highest smoking prevalence among all age groups.[1,2] Over the past decade, the prevalence of quit attempts increased among Americans aged 45 to 64 years; however, it remained stable among young adults.[3] A recent review concluded that the determinants of young adult cessation are not well understood.[4] Previous longitudinal studies in this population have long follow-up intervals—3 to 7 years[5–12]—which means that certain measures (e.g., self-efficacy) may have changed across time and are no longer relevant to predict the later behavior. We examined proximate predictors of young adult smoking cessation behaviors in a prospective study with a 6-month follow-up.

METHODS

We compiled data from 592 young adult smokers (aged 18–29 years) with a 6-month follow-up from July 2005 through December 2008 from the Ontario Tobacco Survey, a population-representative cohort of smokers in Ontario, Canada.[13,14] We classified smoking cessation behavior as no quit attempt, attempt to quit (lasting < 30 days), and 30-day abstinence during follow-up. Guided by social cognitive theory,[15,16] we chose the following covariates: sociodemographic characteristics, smoking addiction,[17,18] quitting history, intentions, beliefs, and social and environmental factors (Table 1).

Using multivariable logistic regression models with covariates associated with the outcome ($P<.2$), we examined predictors of quit attempts (vs no attempt) and abstinence (vs attempt and no attempt). We conducted analyses using SAS version 9.2,[19] accounting for the complex survey design and weighted to the population.

RESULTS

Sixty percent of young adults made no attempt to quit smoking; 25% made an attempt, and 14% were abstinent for 30 days or longer during follow-up. Education, level of addiction, using resources, having support, prior attempts, quit intention, and perceived addiction were significantly associated with young adult cessation behaviors ($P<.05$; Table 1).

Four factors predicted quit attempts among young adults in the multivariate models: using resources, 2 or more prior quit attempts, quit intention, and knowledge of stop smoking medication benefits (Table 2). Abstinence for 30 days or longer was greater among those who were confident in their ability to quit smoking, had used cessation resources, had support, or had lower levels of addiction (Table 2).

DISCUSSION

We assessed prospective predictors of cessation among young adult smokers. We identified different predictors for quit attempts and abstinence; only the use of cessation resources—known to increase cessation[20,21]—contributed to both. We combined all forms of cessation resources; however, resources used for making an attempt may differ from those used to maintain abstinence. There is limited evidence of effective interventions for young adult smokers[22–24]; thus, the resources used for