New Tool To Overcome Missing Race/Ethnicity In Our Electronic Record

March 25, 2019

On behalf of our co-authors (Evan Sholle [first author], Prakash Adekanattu, Marcos Davila, Stephen Johnson, Jyotishman Pathak, Sanjai Sinha, Cassidie Li, Stasi Lubansky, and Tom Campion) and the Cornell Center for Health Equity, we are pleased to announce a new tool to correctly classify patients with missing race/ethnicity data in our electronic health record (EHR).

As you may know, our EHR has as much as 40% missing data on our patients’ race and/or ethnicity. This makes it challenging to use the EHR to recruit minorities, who are historically under studied, for our research studies. It also makes it difficult to track our own performance on the quality of the care we provide our patients. National studies show differences in the quality of care provided to minority patients compared with majority whites, and it is a high priority for both Weill Cornell Medical College and NewYork/Presbyterian Health System to provide the same high quality of care to all of our patients. This project provides the tools to make it possible to examine our own performance, and target recruitment of minorities with confidence.

As an initiative of the Cornell Center for Health Equity, we developed and validated a rule-based natural language processing (NLP) algorithm using outpatient physician notes in Weill Cornell’s EHR to identify patients who are Black/African American or Hispanic/Latino. After applying this NLP algorithm, we were able to increase the number of patients identified as Black by 26%, and Hispanic by 20%. Details of the NLP approach and validation are described in a manuscript entitled "Underserved Populations with Missing Race/Ethnicity Data Differ Significantly From Those with Structured Race/Ethnicity Documentation" that will be published this summer in a special focus issue on health informatics and health equity in the Journal of the American Medical Informatics Association.

Tom Campion has kindly agreed to make this algorithm available to you via the Architecture for Research Computing in Health (ARCH) program. Reports describing patient demographics, including NLP-extracted race and ethnicity data, are available for both quality improvement efforts and IRB-approved studies. In the coming months, Weill Cornell investigators will also be able to use i2b2 to query the NLP-extracted race data alongside diagnoses, procedures, medications, and other clinical concepts without IRB approval to support activities preparatory to their research. Anyone interested should contact arch-support@med.cornell.edu to learn more about this opportunity and to initiate a request.

Monika Safford, MD and Laura Pinheiro, PhD
Division of General Internal Medicine

Centerforhealthequity.cornell.edu

Cornell Center for Health Equity