Journal of Biomedical Informatics 66 (2017) 1–10

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Single-reviewer electronic phenotyping validation in operational settings: Comparison of strategies and recommendations

Polina Kukhareva [a,*], Catherine Staes [a], Kevin W. Noonan [b], Heather L. Mueller [b], Phillip Warner [a], David E. Shields [a], Howard Weeks [b,c], Kensaku Kawamoto [a]

[a] Department of Biomedical Informatics and Knowledge Management and Mobilization, University of Utah, 421 Wakara Way, Suite #140, Salt Lake City, UT 84108, United States
[b] University of Utah Medical Group, 127 S. 500 E., Suite #660, Salt Lake City, UT 84102, United States
[c] Department of Psychiatry, University of Utah, 501 Chipeta Way, Salt Lake City, UT 84108, United States

ARTICLE INFO

Article history:
Received 23 June 2016
Revised 5 December 2016
Accepted 8 December 2016
Available online 9 December 2016

Keywords:
Electronic phenotyping
Computable phenotype
Validation
Electronic clinical quality measurement
Quality measure
Human chart review
Manual chart review

ABSTRACT

Objective: Develop evidence-based recommendations for single-reviewer validation of electronic phenotyping results in operational settings.

Material and methods: We conducted a randomized controlled study to evaluate whether electronic phenotyping results should be used to support manual chart review during single-reviewer electronic phenotyping validation (N = 3104). We evaluated the accuracy, duration and cost of manual chart review with and without the availability of electronic phenotyping results, including relevant patient-specific details. The cost of identification of an erroneous electronic phenotyping result was calculated based on the personnel time required for the initial chart review and subsequent adjudication of discrepancies between manual chart review results and electronic phenotype determinations.

Results: Providing electronic phenotyping results (vs not providing those results) was associated with improved overall accuracy of manual chart review (98.90% vs 92.46%, $p < 0.001$), decreased review duration per test case (62.43 vs 76.78 s, $p < 0.001$), and insignificantly reduced estimated marginal costs of identification of an erroneous electronic phenotyping result ($48.54 vs $63.56, $p = 0.16$). The agreement between chart review and electronic phenotyping results was higher when the phenotyping results were provided (Cohen's kappa 0.98 vs 0.88, $p < 0.001$). As a result, while accuracy improved when initial electronic phenotyping results were correct (99.74% vs 92.67%, N = 3049, $p < 0.001$), there was a trend towards decreased accuracy when initial electronic phenotyping results were erroneous (56.67% vs 80.00%, N = 55, $p = 0.07$). Electronic phenotyping results provided the greatest benefit for the accurate identification of rare exclusion criteria.

Discussion: Single-reviewer chart review of electronic phenotyping can be conducted more accurately, quickly, and at lower cost when supported by electronic phenotyping results. However, human reviewers tend to agree with electronic phenotyping results even when those results are wrong. Thus, the value of

Abbreviations: AAB, Avoidance of Antibiotic Treatment for Adults with Acute Bronchitis; AAP, Adults' Access to Preventive/Ambulatory Health Services; ABA, Adult Body Mass Index Assessment; AMM, Antidepressant Medication Management; AWC, Adolescent Well-Care Visits; BCS, Breast Cancer Screening; CAP, Children and Adolescents' Access to Primary Care Practitioners; CBP, Controlling High Blood Pressure; CCS, Cervical Cancer Screening; CDC, Comprehensive Diabetes Care; CDS, Clinical Decision Support; CHL, Chlamydia Screening in Women; CIS, Childhood Immunization Status; COL, Colorectal Cancer Screening; CWP, Appropriate Testing for Children with Pharyngitis; eCQM, electronic Clinical Quality Measurement; EHR, Electronic Health Record; eMERGE, Electronic Medical Records and Genomics; EMERSE, Electronic Medical Record Search Engine; FN, False Negative; FP, False Positive; FPC, Frequency of Ongoing Prenatal Care; HEDIS, Healthcare Effectiveness Data and Information Set; HPV, Human Papillomavirus Vaccine for Female Adolescents; IMA, Immunizations for Adolescents; LBP, Use of Imaging Studies for Low Back Pain; LSC, Lead Screening in Children; MPM, Annual Monitoring for Patients on Persistent Medications; NCS, Non-Recommended Cervical Cancer Screening in Adolescent Females; PBH, Persistence of Beta-Blocker Treatment After a Heart Attack; PPC, Prenatal and Postpartum Care; PPV, Positive predictive value; SHARPn, Strategic Health IT Advanced Research Project area four; TN, True Negative; TP, True Positive; URI, Appropriate Treatment for Children with Upper Respiratory Infection; UUHC, University of Utah Health Care; W15, Well-Child Visits in the First 15 Months of Life; W34, Well-Child Visits in the First 36 Years of Life.

* Corresponding author at: 421 Wakara Way, Ste 140, Salt Lake City, UT 84108-3514, United States.

E-mail addresses: polina.kukhareva@utah.edu (P. Kukhareva), catherine.staes@hsc.utah.edu (C. Staes), kevin.noonan@hsc.utah.edu (K.W. Noonan), heather.mueller@hsc.utah.edu (H.L. Mueller), phillip.warner@utah.edu (P. Warner), david.shields@utah.edu (D.E. Shields), howard.weeks@hsc.utah.edu (H. Weeks), kensaku.kawamoto@utah.edu (K. Kawamoto).

http://dx.doi.org/10.1016/j.jbi.2016.12.004
1532-0464/

providing electronic phenotyping results depends on the accuracy of the underlying electronic phenotyping algorithm.

Conclusion: We recommend using a mix of phenotyping validation strategies, with the balance of strategies based on the anticipated electronic phenotyping error rate, the tolerance for missed electronic phenotyping errors, as well as the expertise, cost, and availability of personnel involved in chart review and discrepancy adjudication.

1. Introduction

Electronic phenotyping (i.e., automated identification of patients satisfying specified conditions) is essential for a variety of biomedical informatics domains including electronic clinical quality measurement (eCQM), clinical decision support (CDS), predictive analytics, risk adjustment, clinical registries, public health reporting, and cohort identification for clinical trials and research [1–3]. Moreover, the need for electronic phenotyping continues to increase due to the ongoing digitalization of health care [4]. For instance, many regulatory bodies are starting to require quality metrics to be calculated electronically instead of through manual chart audits. Given this increased need for electronic phenotyping, a number of projects have emerged for developing electronic phenotype definitions such as the Electronic Medical Records and Genomics (eMERGE), Electronic Medical Record Search Engine (EMERSE), mini-Sentinel, and Strategic Health IT Advanced Research area four (SHARPn) projects [5–7].

An electronic phenotype definition includes a set of inclusion and exclusion criteria that allow for the algorithmic selection of sets of individuals based on stored clinical data [1]. For example, a CDS system might recommend medications to improve glycemic control for individuals with diabetes who have hemoglobin A1c (HbA1c) levels of 9% or greater [8]. Similarly, an eCQM might require identifying the same set of individuals and determining whether those individuals received recommended care or achieved care goals. For a quality measure, denominator inclusion criteria refer to criteria specifying the set of all individuals for whom the measure is applicable (e.g., diagnosis of diabetes) [9]. Denominator exclusion criteria identify individuals who should be removed from the measure population before determining if numerator criteria are met (e.g., diagnosis of ischemic vascular disease). The numerator criteria are the processes or outcomes expected for each individual identified in the denominator (e.g., HbA1c < 7%). If denominator and numerator criteria are calculated using electronic algorithms, the determination of the patient status is referred to as an electronic phenotyping result [3].

Given the increasing importance of electronic phenotyping, it is essential that such phenotyping is as accurate as possible. For the purposes of this paper, we define an erroneous electronic phenotyping result as misclassification of a patient according to the phenotyping definition. For example, misclassifying an individual without diabetes as meeting denominator criteria for a comprehensive diabetes care measure is an inaccurate electronic phenotyping result. Such inaccurate phenotyping in the context of CDS can lead to alert fatigue and potentially patient harm [10], and inaccurate phenotyping in the context of eCQM can lead to misleading characterization of practice performance, and/or limit the ability of performance feedback to catalyze improvements in care quality [11,12]. Sources of error in electronic phenotyping include an incorrect or incomplete phenotyping definition, inaccurate interpretation of the phenotyping definition, erroneous identification or use of source clinical data, and insufficient data quality and consistency [13]. Many phenotyping algorithm implementations have been found to have problems with accuracy [14,15]. Therefore, electronic phenotyping results must be appropriately validated before they can be used with confidence. Electronic phenotyping validation is the process of establishing the accuracy of electronic phenotypes.

Electronic phenotyping validation requires comparing electronic phenotyping results to a reference standard. Such a reference standard is usually developed using one of the following three manual chart review methods: 'gold standard' (i.e., double manual chart review with at least two independent reviewers and adjudication, performed to resolve inter-reviewer discrepancies), 'trained standard' (i.e., one expert reviewer with validity of review checked), and 'regular practice' (single human reviewer) [16]. Among 113 studies of automated clinical coding and classification systems described by Stanfill et al., the majority (51%) used the single-reviewer 'regular practice' approach to create a reference standard [16]. While many phenotyping validation efforts use the 'gold standard' approach for iterative validation of phenotype definitions [1,5,17–19], such studies are usually performed in resource-rich research settings [6,13,19,20]. For example, the eMERGE network has conducted extensive research in phenotype definition validation such as identifying individuals with cataracts, type 2 diabetes, or dementia [13]. Unfortunately, 'gold standard' double manual chart review is often not feasible in operational settings due to resource constraints. Indeed, in operational settings, the reference standard for validation is often single manual chart review coupled with expert adjudication of any discrepancies with electronic phenotyping results.

Despite the importance of the single human reviewer approach in operational settings, this approach to electronic phenotyping validation is not well described in the literature. Thus, we have a limited understanding of the strengths and limitations of different single-reviewer strategies. In particular, there is limited guidance available in the literature on how a maximum number of phenotyping errors can be identified with minimal person-hours, thereby optimizing the quality of electronic phenotyping results given available resources.

At our academic medical center, we were faced with the need to efficiently validate electronic phenotypes being implemented for enterprise clinical quality measurement and physician compensation using a single-reviewer approach. It has been previously shown that providing electronic phenotyping results to humans can improve the efficiency of manual phenotyping tasks such as diagnosis coding and quality measurement [21,22]. However, no prior literature was available in the context of single-reviewer electronic phenotyping validation. Thus, we hypothesized that supporting a single-reviewer chart review process with electronic phenotyping results would make the review faster and more precise by reducing the validator's cognitive load. For example, providing the date of the last retinal eye exam for an individual with diabetes would allow the reviewer to more efficiently confirm that the patient received the recommended care. However, we were concerned that the reviewer may be influenced by the provided results and may over-agree with erroneous electronic phenotyping