Journal of Biomedical Informatics 51 (2014) 280–286

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Design patterns for the development of electronic health record-driven phenotype extraction algorithms

Luke V. Rasmussen [a,*], Will K. Thompson [a,g], Jennifer A. Pacheco [a], Abel N. Kho [a], David S. Carrell [b], Jyotishman Pathak [c], Peggy L. Peissig [d], Gerard Tromp [e], Joshua C. Denny [f], Justin B. Starren [a]

[a] Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[b] Group Health Research Institute, Seattle, WA, United States
[c] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States
[d] Marshfield Clinic Research Foundation, Marshfield, WI, United States
[e] Sigfried and Janet Weis Center for Research, Geisinger Health System, Danville, PA, United States
[f] Departments of Biomedical Informatics and Medicine, Vanderbilt University, Nashville, TN, United States
[g] Center for Biomedical Research Informatics, NorthShore University HealthSystem, Evanston, IL, United States

ARTICLE INFO

Article history:
Received 23 December 2013
Accepted 16 June 2014
Available online 21 June 2014

Keywords:
Electronic health record
Phenotype
Algorithms
Software design
Design patterns

ABSTRACT

Background: Design patterns, in the context of software development and ontologies, provide generalized approaches and guidance to solving commonly occurring problems, or addressing common situations typically informed by intuition, heuristics and experience. While the biomedical literature contains broad coverage of specific phenotype algorithm implementations, no work to date has attempted to generalize common approaches into design patterns, which may then be distributed to the informatics community to efficiently develop more accurate phenotype algorithms.

Methods: Using phenotyping algorithms stored in the Phenotype KnowledgeBase (PheKB), we conducted an independent iterative review to identify recurrent elements within the algorithm definitions. We extracted and generalized recurrent elements in these algorithms into candidate patterns. The authors then assessed the candidate patterns for validity by group consensus, and annotated them with attributes.

Results: A total of 24 electronic Medical Records and Genomics (eMERGE) phenotypes available in PheKB as of 1/25/2013 were downloaded and reviewed. From these, a total of 21 phenotyping patterns were identified, which are available as an online data supplement.

Conclusions: Repeatable patterns within phenotyping algorithms exist, and when codified and cataloged may help to educate both experienced and novice algorithm developers. The dissemination and application of these patterns has the potential to decrease the time to develop algorithms, while improving portability and accuracy.

1. Introduction

Electronic health records (EHRs) have been shown to be a valuable source of information for biomedical research, including the definition and identification of clinical phenotypes [1–5]. The increasing use of EHRs [6,7] has resulted in large quantities of data available for secondary purposes such as research. In order to better handle this growing source of data, we need to improve methods and approaches to phenotype more efficiently.

The electronic Medical Records and Genomics (eMERGE) network has been a leader in the development of phenotype algorithms based on EHR data. In addition to the work done through eMERGE for genome-wide association studies (GWAS) [8–15], there are additional examples of electronic algorithms to mine EHRs for identifying diseases for biomedical research and clinical care [16–20] disease surveillance [21], pharmacovigilance [22], as well as for decision support [23]. These studies have provided some guidance on dealing with the challenges of using EHR and claims data [2,20,24–26]. This guidance has often been in the context of a single algorithm, although more recent work has begun to address the broader challenges of using EHR data for phenotyping [1,27]. Additionally, research is being conducted to identify how electronic phenotype algorithms may be represented and made more portable across disparate EHRs [28,29], which has the

* Corresponding author. Address: Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University, Feinberg School of Medicine, Rubloff Building, 11th Floor, 750 North Lake Shore Drive, Chicago, IL 60611, United States. Tel.: +1 (312) 503 2823.
E-mail address: luke.rasmussen@northwestern.edu (L.V. Rasmussen).

http://dx.doi.org/10.1016/j.jbi.2014.06.007
1532-0464/

potential to automate approaches to handle the complexities and nuances of EHR data.

A major goal of the current phase of eMERGE is to improve the ease and speed of developing new phenotype definitions. No known work to date, however, has attempted to broadly classify challenges and solutions to using EHR data for the development of electronic phenotype algorithms, or demonstrated an approach to widely disseminate the findings. This knowledge could potentially reduce the time to develop phenotype algorithms, improve portability to other sites and even accuracy by describing experiences developing other algorithms. The primary goal of this paper is to apply lessons from prior work in software design patterns to the problem of defining and disseminating EHR-based phenotype algorithms.

In software engineering, the use of design patterns are frequently used to generate solutions to common problems or scenarios [30]. These patterns are free from any technical implementation details, such as programming language or database platform. Design patterns are not applicable only in the domain of software development. They have roots in architecture [31], and have recently been applied to the development of ontologies [32,33] and health information technology (HIT) solutions [34,35]. Even though design patterns are used in multiple domains, they share similar constructs that form a basis of overall pattern languages [36]. Generally, design patterns provide: (1) a description of a scenario or problem that exists and that the pattern may address; (2) a template for a solution; and, (3) considerations for when to apply the pattern, or what its implications may be [30,31,36]. Design patterns are not intended to capture every possible pattern that may occur in the target domain; rather, they represent best practices and common approaches to solving a problem. In practice, they may be derived from intuition, heuristics and experience.

In order to more widely disseminate solutions to common problems and scenarios found in the development of electronic phenotype algorithms, we propose the creation of "EHR-driven phenotype extraction design patterns"—logical patterns recurring frequently in phenotyping algorithms that are EHR and technology agnostic. This paper presents an initial catalog of such patterns from experiences within the eMERGE network.

2. Methods

The steps used to define, develop and review phenotype design patterns are shown in Fig. 1, and are explained in more detail below.

2.1. Setting

The eMERGE network [37] is a National Human Genome Research Institute (NHGRI)-sponsored initiative that has demonstrated the feasibility of EHR-derived phenotypes in order to conduct genome-wide association studies (GWAS). Within the network, sites develop and locally validate an EHR-based phenotype algorithm, which are then implemented and validated at one or more additional network sites. While phenotype algorithms themselves are largely recorded as text documents [38] and have to be re-implemented in a format that can be executed at each site, the transfer of phenotypes from one site to other sites with different EHR systems demonstrates the broader application of EHR-derived phenotyping.

2.2. Phenotype selection

Phenotypes created by the eMERGE network are publicly available on the Phenotype KnowledgeBase website (PheKB, http://www.phekb.org), and are classified by the group or consortium under which the algorithm was created, as well as a status to indicate how mature the algorithm is in its development process. For this study, all phenotype algorithms associated with the eMERGE network that were marked with a "Final" or "Validated" status were downloaded on January 25, 2013. The algorithm set consists of both case/control studies (i.e. Cataracts, Resistant Hypertension) as well as quantitative measures (i.e. Red Blood Cell Indices, White Blood Cell Indices). The algorithms were developed, implemented, and validated by chart review by at least one other eMERGE site. Phenotypes that were available in PheKB but had not been validated were considered too preliminary for study, and were excluded. In addition, the selected algorithms within the eMERGE network were not developed independently (sites collaborated on and built new algorithms after having reviewed others), which allowed evaluation of shared experiences as algorithms were developed over time.

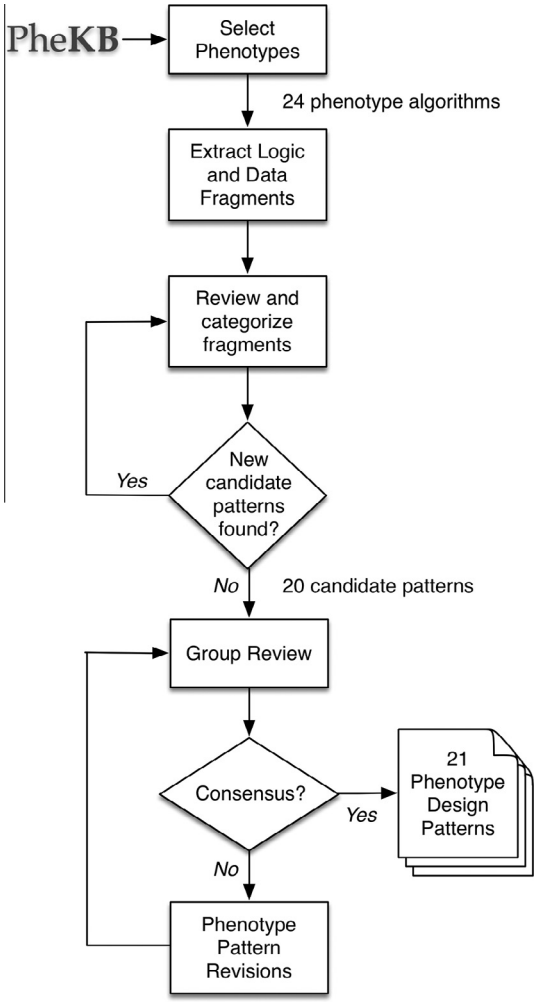

Fig. 1. Methods for developing and reviewing phenotype design patterns.

2.3. Phenotype algorithm review

One of the authors (LVR) reviewed each of the phenotype algorithms, and identified unique, discrete fragments in the text definitions that represented the inputs, logic, and constraints within the algorithms. As multiple artifacts can exist for each phenotype algorithm (i.e. chart abstraction forms for validation, data dictionary definitions), only documentation containing a textual description of the algorithm was reviewed.