Journal of Biomedical Informatics 84 (2018) 75–81

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Discovering hidden knowledge through auditing clinical diagnostic knowledge bases

Matthew C. Lenert[a,*], Colin G. Walsh[b], Randolph A. Miller[b]

[a] Dept. of Biomedical Informatics, Vanderbilt University, 2525 West End Ave. Suite 1475, Nashville, TN 37203, USA
[b] Dept. of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

ARTICLE INFO

Keywords:
Data auditing
Internal medicine
Knowledge bases

ABSTRACT

Objective: Evaluate potential for data mining auditing techniques to identify hidden concepts in diagnostic knowledge bases (KB). Improving completeness enhances KB applications such as differential diagnosis and patient case simulation.

Materials and methods: Authors used unsupervised (Pearson's correlation – PC, Kendall's correlation – KC, and a heuristic algorithm – HA) methods to identify existing and discover new finding-finding interrelationships ("properties") in the INTERNIST-1/QMR KB. Authors estimated KB maintenance efficiency gains (effort reduction) of the approaches.

Results: The methods discovered new properties at 95% CI rates of [0.1%, 5.4%] (PC), [2.8%, 12.5%] (KC), and [5.6%, 18.8%] (HA). Estimated manual effort reduction for HA-assisted determination of new properties was approximately 50-fold.

Conclusion: Data mining can provide an efficient supplement to ensuring the completeness of finding-finding interdependencies in diagnostic knowledge bases. Authors' findings should be applicable to other diagnostic systems that record finding frequencies within diseases (e.g., DXplain, ISABEL).

1. Introduction

Data mining has gained popularity as a mechanism for discovery in biomedical research [1–4]. In contrast to mining patient-related data, construction and maintenance of knowledge bases (KBs) for diagnostic decision support systems remains a predominantly manual and labor-intensive process [5–7]. The authors explored the feasibility of using algorithmic methods to derive novel information relevant to general-purpose clinical diagnosis using an existing large diagnostic knowledge base.

As physicians observe pathophysiological manifestations (findings) of diseases in their patients, they require evidence-based disease-to-finding linkages to generate and confirm diagnostic hypotheses [8–10]. Diagnostic KB curators attempt to capture such background information to support clinicians' practices. Physicians combine abductive reasoning, deductive reasoning, inductive reasoning, and probabilistic reasoning with clinical knowledge to reach a diagnosis that best explains the patient's findings [8–12]. When attempting to resolve a differential diagnosis through questioning (including ordering laboratory and imaging studies), the value of information gained from seeking additional findings is not uniform [13,14]. Specifically, some findings are conditionally interdependent on other findings – a problem recognized by developers of early Bayesian systems, and more recently explored through Bayesian belief networks [15–18]. Learning that a patient with an elevated serum total bilirubin level has jaundice is not as informative as learning that the same patient's reticulocyte count is substantially elevated [19].

Knowing how each disease finding relates to other disease findings has relevant applicability in clinical systems. As noted, Bayesian systems often require conditionally independent input findings. In helping users to "work up" a case, diagnostic systems generate questions that ideally should minimize overlap in information content with already known findings [13–15]. Additionally, diagnostic patient case simulations should avoid generating too many interdependent case findings [20]. As new findings are added to a KB, the number of finding interdependencies grows at least linearly and potentially combinatorially—complicating KB maintenance tasks.

JD Myers, HE Pople Jr., and RA Miller initially developed the INTERNIST-1/QMR system at the University of Pittsburgh [21,22]. The INTERNIST-1/QMR KB has been adapted to multiple representations, including as a Bayesian Belief Network (QMR-DT) [17,18]. The latter formulation of the INTERNIST-1/QMR KB made the assumption of

* Corresponding author.
E-mail address: matthew.c.lenert@vanderbilt.edu (M.C. Lenert).

https://doi.org/10.1016/j.jbi.2018.06.014
Received 9 February 2018; Received in revised form 19 June 2018; Accepted 21 June 2018
Available online 22 June 2018
1532-0464/

conditional finding independence, which was common for diagnostic systems at the time, but had significant performance costs [23]. After more than a decade of recent dormancy, the authors and colleagues at Vanderbilt are developing new mobile applications based on a revised version of the INTERNIST-1/QMR KB. The KB contains evidence-based descriptions of approximately 630 diseases using 4600 possible findings [19]. The findings, connection weights (see below), and linked disorders associated with a disease form the disease profile. In the INTERNIST-1/QMR KB, attributes called "properties" describe unary and binary facts/relationships among findings—including interdependencies among findings [24]. Previously, relating one finding to another via properties required extensive clinical knowledge and time-consuming review effort [5]. Over the three-decade course of the INTERNIST-1/QMR project expert clinical diagnostician JD Myers manually derived 5802 such properties from 10,586,901 possible finding pairs [24]. Reliance on manual derivation presents a key challenge to the maintenance and expansion of diagnostic KB contents. The authors' new mobile applications require properties to function optimally.

The authors recently observed that interdependencies between finding pairs occur in two forms: etiological and pragmatic. Etiological interdependencies reflect a common cause (elevated bilirubin → jaundice) or alternatively a broader than-narrower than relationship (hepatomegaly → moderate liver enlargement). Pragmatic relationships represent common sense knowledge about mutually exclusive clinical circumstances (males do not develop pregnancy-related findings or diseases). By definition (i.e., due to a common cause), findings related by etiological properties probabilistically co-occur within diseases more often than by random chance; conversely, pragmatic properties identify findings unlikely to co-occur in the same disorder. Table 1 provides examples of INTERNIST-1/QMR properties and their purposes.

Not all etiologically interdependent findings merit representation as properties. For example, patients experiencing hepatocellular injury will typically have elevated serum aspartate aminotransferase (AST) and alanine aminotransferase (ALT). Nevertheless, one would not want to create a property that states if AST is elevated, so is ALT, because in some instances – such as muscle injury – this does not occur. The authors have used the term "facet" to describe sets of findings in a disease that share a common etiology [25]. For example, in inflammatory arthritis, the findings morning stiffness, severe joint pain, and decreased joint range of motion all share a facet relationship with one another. Facet relationships do not qualify as pairwise property relationships (e.g., AST/ALT) [25]. Superimposing facets on the INTERNIST-1/QMR disease profiles in the past-involved intense manual efforts [26]. Facets are useful for patient case simulation and for causal deductive reasoning in diagnosis [25].

The current study used the disease profiles curated in the INTERNIST-1/QMR KB to determine the extent to which existing and new properties could be discovered via mining techniques. Semi-automated discovery of "missing" properties and derivation of new facet relationships would both improve upon human imperfections during arduous review of a large dataset. Each disease profile in the INTERNIST-1/QMR KB contains an average of 85 findings [24]. For each disease-finding pair, the KB includes an estimated positive predictive value (evoking strength) on a 0–5 scale and a literature-derived sensitivity (frequency) of the finding in the disease (i.e., how often patients with the disease present with the finding, on a 1–5 scale) [19]. A sensitivity of 1 means the finding is rarely seen with the disease; 3 means the finding occurs in approximately half of all cases, and 5 means the finding is present in essentially all cases [14].

Table 1
Property examples.

Index finding	Relationship	Affected finding	Type
Sputum production	Implies	Cough	Etiological
History of prostatitis	Rules out	Sex female	Pragmatic
Heart murmur systolic apical	Implies	Heart murmur present	Etiological

2. Materials and methods

The authors hypothesized that two pathophysiologically interrelated findings would likely co-occur together in parallel across multiple disease profiles in a diagnostic KB. Additionally, if the findings' co-occurrences were etiologically determined, their relative frequencies in the diseases (as represented by sensitivity scores across the disease profiles) would vary conjointly in parallel. Based on the foregoing hypotheses, the authors first developed a heuristic algorithm to discover new properties by searching for finding pairs meeting the given criteria. Authors then evaluated the heuristic algorithm against statistical discovery methods to determine which approach might be superior. Pragmatic properties (of form "A" contradicts "B" – e.g., females cannot have a history of prostatitis) are not discoverable in this fashion, because contradictory findings rarely co-occur within disease profiles. Correlating the presence of a specific finding with the absence of a different specific finding across the 630 disease profiles would be difficult, given that, on average, only 85 of the 4601 potential findings appear in a disease profile. The authors chose the INTERNIST-1/QMR KB for this study because of its breadth. Other formulations of the KB, such as QMR-DT, have fewer disease profiles and/or lacked the concept of finding interdependencies. The authors considered a finding's frequency and evoking strength, whether on an ordinal or probability scale, of less value for determining etiological relationships, than the number of overlapping disease profiles. Furthermore, the authors believed the ordinal scale to be more robust, because an ordinal ranking is less sensitive to the curator's interpretations of probabilities in the literature.

The authors also investigated the use of a supervised machine learning method for property auditing. The study employed Python's Scikit Learn random forest model trained on a labeled set of 56,000 ("property" or "no relationship") pairwise finding combinations, to discover new properties with the remaining 10,530,901possible finding pairs [27]. Authors selected a random forest model, because those models tend to be resistant to over fitting. Over fitting was a concern, because the training set was significantly smaller than the search space [28]. The model used the sum of disease frequencies between two findings as its feature space. This meant the model had 630 numeric features, ranging from 0 to 10 due to the summation. The authors chose to combine frequency scores with a sum to preserve information of disease finding frequency, as compared to a multiplicative effect. The authors used frequency scores to parody the type of data used by unsupervised methods. The authors explored several different ratios of properties to non-properties in the training set; some training sets involved an acceptable level of noisy negative labels [29].

2.1. Unsupervised knowledge auditing methods

The project analyzed all co-occurrences of 4601 findings across all 630 disease profiles, totaling 10,586,901 unique pairwise combinations. This large search space imposed pragmatic computational considerations, as the authors needed to evaluate 630 disease profiles for each of the nearly 11 million finding combinations. The statistical and heuristic methods designated a score of zero for finding pairs that co-occurred in fewer than two disease profiles, and higher scores based on how well a pair of findings tracked together across multiple disease profiles. Findings that frequently co-occur in multiple disease profiles together are more likely to share a common pathophysiologic cause, and therefore merit a property relationship to prevent diagnostic systems from seeking redundant or uninformative findings [30]. Since etiological dependencies between findings imply a common