Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220487

Causal Associations Among Diseases and Imaging Findings in Radiology Reports

Ronnie SEBRO[a] and Charles E. KAHN, Jr.[b,1]
[a] *Mayo Clinic, Jacksonville, Florida, USA*
[b] *University of Pennsylvania, Philadelphia, Pennsylvania, USA*

Abstract. This study explored the ability to identify causal relationships between diseases and imaging findings from their co-occurrences in radiology reports. A natural language processing (NLP) system with negative-expression filtering detected positive mentions of 16,912 disorders, interventions, and imaging findings in 1,702,462 consecutive radiology reports; the 55,564 causal relations defined by the Radiology Gamuts Ontology (RGO) served as reference standard. Conditions were considered to co-occur if they were present in reports from the same patient. The φ and κ statistics both achieved AUC≥0.70, P<0.001 in identifying causal relationships from pairwise co-occurrence data. Analysis of radiology reports can identify a large proportion of known causal associations among diseases and imaging findings. Automated approaches hold promise to identify causal relationships among diseases and imaging findings from their co-occurrence in text-based radiology reports.

Keywords. Knowledge discovery, Ontologies, Health data science, Big data, Natural language processing, Radiology, Reporting

1. Introduction

Narrative-text radiology reports provide a rich source of information that can be extracted using natural language processing (NLP) to identify the presence of diseases and imaging findings. This study applied an ontology of disorders, interventions, and imaging observations to detect co-occurrence of diseases and imaging findings in a large corpus of narrative-text radiology reports. The study evaluated the ability of two statistical tests to identify causal relationships based on the co-occurrence of terms in a large cohort of patients**.**

2. Materials and Methods

1,702,462 consecutive radiology reports of 1,396,293 patients were analyzed in this IRB-approved study. The Radiology Gamuts Ontology (RGO) defined 16,912 imaging findings, diseases, and interventions, and 55,564 causal relationships between them [1]. An occurrence was defined as positive mention of an RGO entity; reports were

[1] Corresponding Author, Charles E. Kahn, Jr, MD, MS, Department of Radiology and Institute for Biomedical Informatics, University of Pennsylvania, 3400 Spruce Street, Philadelphia PA 19104, USA; E-mail: ckahn@upenn.edu.

aggregated by patient. Analysis included pairs of entities that co-occurred in at least 25 patients. The absolute value of the phi coefficient (φ) and Cohen's kappa statistic (κ) were analyzed: receiver operator characteristic (ROC) curves were compared using DeLong's test. Area under the ROC curve (AUC) was computed.

3. Results

Analysis was limited to 95,620 pairs where both RGO entities occurred in 25 or more patients, of which 161 pairs (0.17%) had causal associations in RGO. Both φ and κ achieved AUC ≥ 0.70; their AUCs showed no statistically significant difference (DeLong's test p=0.288).

4. Conclusion

The φ and κ statistics showed moderate performance in identifying known causal relationships based on pairwise co-occurrence of terms in radiology reports; there was no statistically significant difference between the two metrics. Both metrics' performance can be considered remarkably strong given that "positive" pairs of causally related entities constituted only 0.17% of all of pairs of entities evaluated. This finding suggests that automated approaches may be able to help identify causal relationships among diseases and imaging findings from text-based radiology reports. Our findings agree with reports of pairwise disease-finding associations in clinical data and have mined observational data to identify causal pathways [2, 3]. In most settings, most EHR information exists as narrative text, including radiology reports.

Our larger goal is to discover new causal relationships among disorders and imaging findings directly from their patterns of co-occurrence. Discovery of causal relationships in observational data has been grounded in Bayesian network learning, a computationally intensive (and NP-complete) problem; various approaches have sought to reduce the computational burden to make the approach more tractable on large datasets [4, 5]. Given our large cohort of observational data, our current work includes induction of a Bayesian network model over the entire set of co-occurring terms. Preliminary analysis has shown significantly stronger performance than pairwise co-occurrence statistics and has yielded plausible new causal relations.

References

[1] Budovec JJ, Lam CA, Kahn CE, Jr. Radiology Gamuts Ontology: differential diagnosis for the Semantic Web. RadioGraphics. 2014;34(1):254-64.

[2] Cao H, Hripcsak G, Markatou M. A statistical methodology for analyzing co-occurrence data from a large sample. J Biomed Inform. 2007;40(3):343-52.

[3] Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. AMIA Annu Symp Proc. 2005:106-10.

[4] Jin Z, Li J, Liu L, Le TD, Sun B, Wang R, editors. Discovery of causal rules using partial association. IEEE 12th International Conference on Data Mining; 2012 10-13 Dec. 2012.

[5] Ramanan N, Natarajan S. Causal learning from predictive modeling for observational data. Front Big Data. 2020;3:535976.