Attributes for Causal Inference in Electronic Healthcare Databases

Jenna Reps, Jonathan M. Garibaldi,
Uwe Aickelin and Daniele Soria
IMA, School of Computer Science
The University of Nottingham
{jzr,jmg,uxa,dqs}@cs.nott.ac.uk

Jack E. Gibson and Richard B. Hubbard
Clinical Sciences Building
Nottingham City Hospital
{jack.gibson,
richard.hubbard}@nottingham.ac.uk

Abstract

Side effects of prescription drugs present a serious issue. Existing algorithms that detect side effects generally require further analysis to confirm causality. In this paper we investigate attributes based on the Bradford-Hill causality criteria that could be used by a classifying algorithm to definitively identify side effects directly. We found that it would be advantageous to use attributes based on the association strength, temporality and specificity criteria.

1. Introduction

The aim of medication is to improve patients' standard of living, but medication can lead to side effects, also known as adverse drug reactions (ADRs). Existing ADR signalling algorithms have a high false positive rate. This reduces their efficiency as the signals they generate need to be confirmed with more rigorous analysis.

A novel approach for signalling ADRs is to develop a causality classifier with suitable input attributes. Such an algorithm would be more efficient at signalling ADRs as it would not require additional analysis. The Bradford Hill causality criteria (BHCC) [1] is an excellent starting point for developing suitable attributes as it is often considered when determining causal relationships. In this paper we investigate attributes based on the BHCC to aid future ADR classifying algorithms. In the continuation of this paper we summarise the existing algorithms, the BHCC and the feature selection applied in the next section, followed by the results and finish with the conclusion.

2 Background & Methodology

Spontaneous Reporting System (SRS) databases and Electronic Healthcare Databases (EHDs) are the databases generally used for post marketing drug surveillance. The SRS databases rely of voluntary reports of suspected ADRs whereas the EHD databases are often extracted directly from medical practitioners records. Existing algorithms measure association rather than determining causality directly. The BHCC were developed to distinguish between association and causation. The nine factors of interest (in the context as ADR signalling) are:

- Association Strength - how strong the association is.
- Temporality - the direction of the association.
- Specificity - how specific the relationship is.
- Experimentation - does the medical event stop and start in sync with the drug?
- Dosage - correlation between dosage and medical event occurrence?
- Analogy - do similar drugs have similar side effects?
- Coherence - does the association make sense?
- Plausibility - is the association possible?
- Consistency - association found in different databases?

The SRS and EHD algorithms calculate a measure of association strength and also cover temporality, as the EHDs apply filters to removed medical events that cause the drug and people submitting reports to SRS algorithms will only report medical events that occur after the drug. Furthermore, people will only report a suspected ADR if it is plausible, so the SRS algorithms indirectly cover plausibility.

The attributes detained in Table 1 were derived using The Health Improvement Network database (www.thin-uk.com). Feature selection was applied using a multivariate filter, the Correlation-based Feature Selection (CFS) algorithm [2], as this algorithm is not dependent on a specific classifier.

3. Results & Discussion

The attributes chosen by the CFS algorithm were LEOPARD, RD_{13BNF}, ABratio Lv3, Gender Ratio and Read Code Level. The reason that the majority of attributes were not selected by the CFS algorithm is because they had a high correlation with either LEOPARD or the RD_{13BNF}. The

Table 1. Attribute Summary Table

Feature	Criterion	Description
RR, RD, OR	Strength	The Risk Ratio, Risk Difference and Odds Ratio [5] for all prescriptions.
RR_{13d},RD_{13d},OR_{13d}	Strength	The Risk Ratio, Risk Difference and Odds Ratio for drugs prescribed for the first time in 13 months.
RR_{13BNF},RD_{13BNF} ,OR_{13BNF}	Strength	The Risk Ratio, Risk Difference and Odds Ratio for drugs corresponding to a bnf that has not been prescribed in the last 13 months.
IC_{Δ}	Strength	The Information Component as calculated in [3]
lowerIC_{Δ}	Strength	The lower 95% interval of the Information Component as calculated in [3]
Age STDEV	Specificity	Standard deviation of patient's age who experience medical event after drug divided by standard deviation of the ages for all the patients.
Gender Ratio	Specificity	Male proportion of patients experiencing the medical event within 30 days of the drug divided by male proportion of patients prescribed the drug.
RR drug / RR bnf	Specificity	The RR of the drug divided by the RR for all the drugs in the same family.
Read Code Level	Specificity	The specificity level of the medical event: general (level 1)- specific (level 5).
ABratio Level 2	Temporality	How often the level 2 version of the medical event is recorded after the prescription compared to before.
ABratio Level 3	Temporality	How often the level 3 version of the medical event is recorded after the prescription compared to before.
LEOPARD [4]	Temporality	1 if the drug is prescribed significantly more after the medical event than before, 0 otherwise.
OE_{filt1} [3]	Temporality	1 if the IC_{Δ} is greater the month before the drug than the month after, 0 otherwise.
OE_{filt2} [3]	Temporality	1 if the IC_{Δ} is greater on the day of prescription compared to the month after, 0 otherwise.
Dosage Ratio	Dosage	Average dosage of patients experiencing the medical event within 30 days of the drug divided by average dosage of patients prescribed the drug.
High Low Ratio	Dosage	Proportion of patients given the highest dosage that experience the medical event (within 30 days) divided by the proportion of patients given the lowest dosage that experience the medical event (within 30 days).
Spearman's rank	Dosage	The Spearman's rank correlation coefficient between the patient dosage and $\{0, 1\}$ indicating if the patient experienced the medical event within 30 days.
Pearson product-moment	Dosage	The Pearson product-moment correlation coefficient between the patient dosage and $\{0, 1\}$ indicating if the patient experienced the medical event within 30 days.
$Repeat_1$	Experiment	Number of patients that have medical event in at least two distinct hazard periods and not in their non-hazard periods divided by the number of patients that have at least two distinct hazard periods and have medical event in one hazard period.
$Repeat_2$	Experiment	Number of patients that have medical event in two distinct hazard periods and not in their non-hazard periods divided by the occurrence in the non-hazard periods.

results show that the specificity attributes Gender Ratio and Read Code level can complement the temporal and strength attributes for ADR signalling. The experiment and dosage attributes investigated in this paper did not offer sufficient additional information than what could be gained from the RD_{13BNF} or the LEOPARD attributes.

4. Conclusion

In this paper we investigated novel attributes based on the Bradford Hill causality criteria. We found that the specificity attributes offer additional information for ADR signalling and it would be advantageous to include them in future algorithms. Future work could involve investigating other BHCC based attributes.

References

[1] A. Bradford-Hill. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58:295–300, 1965.

[2] M. A. Hall. Correlation-based feature selection for machine learning. Technical report, 1999.

[3] G. N. Noren, J. Hopstadius, A. Bate, and et al. Temporal pattern discovery in longitudinal electronic patients records. *Data Min Knowl Disc*, 20:361–387, 2010.

[4] M. J. Schuemie. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. *Pharmacoepidemiol Drug Saf*, 20(3):292–299, 2011.

[5] C. L. Sistrom and C. W. Garvan. Proportions,odds and risk. *Radiologyt*, 230:12–19, 2004.