

Automatic extraction of adverse drug reaction terms from medical free text

Ola Caster

Uppsala Monitoring Centre

WHO Collaborating Centre for International Drug Monitoring

ISCB 2009, Prague, Czech Republic

26th August

Joint work with Johanna Strandell, Andrew Bate, and I. Ralph Edwards



Safety monitoring of drugs

- Marketed drugs are continuously monitored for adverse drug reactions (ADRs) not detected during clinical testing
- So called individual case safety reports (ICSRs) are a key component of this monitoring
 - Primarily spontaneously reported incidents from clinical practice
- The Uppsala Monitoring Centre (UMC) collects ICSRs from more than 90 countries
 - Vigibase: pooled database of almost 5 million reports



ADR terminologies and free text sources

- Reactions are classified according to standard, hierarchical, ADR terminologies
 - MedDRA
 - WHO-ART, which is maintained by UMC
- Statistical analyses depend on this classification
 - For example data mining for pairs of drugs and ADR terms occurring more often than expected
- However, lots of relevant information is only available in free text
 - For example the Summary of Product Characteristics



ADR terminologies and free text sources

- Terms from ADR terminologies do occur in medical free text sources...
- ...however, exact matching is likely to exhibit poor sensitivity
- This precludes efficient comparison between results from statistical analyses and published information
 - Of particular interest: Is a drug-ADR pair highlighted by data mining of ICSRs already listed in (free text) standard literature sources?
 - Purpose is to detect as yet unknown adverse reactions



Objective

- To develop an algorithm that can extract ADR terms from medical free text at higher sensitivity than exact matching, yet with high precision



Data sources

- XML version of *Stockley's Drug Interactions* [1]
 - Well-renowned source of information on drug-drug interactions
 - Each interaction, including its potential adverse effects, is described in free text
 - 40259 database entries in used version
- The WHO-ART terminology for ADRs
 - 5770 terms (lowest level of hierarchy)



The algorithm exemplified

	ADR term	Free text
	QT prolonged	Some quinolones (e.g. gatifloxacin) can prolong the QT interval and [...] when used with quinidine.

- A real example from Stockley's database
- Exact matching does not work
- Important aspects of the algorithm will be explained



Algorithmic step	ADR term	Free text
None	QT prolonged	Some quinolones (e.g. gatifloxacin) can prolong the QT interval and [...] when used with quinidine.
Remove non-alphanumerics (+ lower case)	qt prolonged	some quinolones e g gatifloxacin can prolong the qt interval and [...] when used with quinidine

- Problem 1: There is a **the** in the free text but not in the ADR term
 - Solution 1: Remove stopwords



Stopwords

- Words that carry none or little semantical meaning
 - ... if, of, any, for, at, are, also ...
- Quite common to remove stopwords in algorithms like this
- The list used here comes from the Snowball project [2]



Algorithmic step	ADR term	Free text
None	QT prolonged	Some quinolones (e.g. gatifloxacin) can prolong the QT interval and [...] when used with quinidine.
Remove non-alphanumerics (+ lower case)	qt prolonged	some quinolones e g gatifloxacin can prolong the qt interval and [...] when used with quinidine
Remove stopwords	qt prolonged	quinolones e g gatifloxacin can prolong qt interval [...] used quinidine

- Problem 2: The words are in different grammatical forms (**prolonged** vs **prolong**)
 - Solution 2: Use stemming



Stemming

- Each word is replaced by its stem
- Increases the chance to match words with the same semantic meaning, but with different grammatical functions
 - 'interact', 'interacts', 'interacted', 'interaction', 'interacting' would all be mapped to 'interact'
- The stemming algorithm used here was the so called Porter algorithm [3]



Algorithmic step	ADR term	Free text
None	QT prolonged	Some quinolones (e.g. gatifloxacin) can prolong the QT interval and [...] when used with quinidine.
Remove non-alphanumerics (+ lower case)	qt prolonged	some quinolones e g gatifloxacin can prolong the qt interval and [...] when used with quinidine
Remove stopwords	qt prolonged	quinolones e g gatifloxacin can prolong qt interval [...] used quinidine
Use stemming	qt prolong	quinolon e g gatifloxacin can prolong qt interv [...] use quinidin

- Problem 3: The order is wrong
 - Solution 3: Permute the ADR term



Permutation of ADR terms

- Many ADR terms that include an adjective are constructed in the opposite order of what would be normal in free text
 - Terminology: *scoliosis congenital*
 - Free text: *congenital scoliosis*
- Use every possible order of the individual words of the ADR term



Algorithmic step	ADR term	Free text
None	QT prolonged	Some quinolones (e.g. gatifloxacin) can prolong the QT interval and [...] when used with quinidine.
Remove non-alphanumerics (+ lower case)	qt prolonged	some quinolones e g gatifloxacin can prolong the qt interval and [...] when used with quinidine
Remove stopwords	qt prolonged	quinolones e g gatifloxacin can prolong qt interval [...] used quinidine
Use stemming	qt prolong	quinolon e g gatifloxacin can prolong qt interv [...] use quinidin
Use permutation	prolong qt	quinolon e g gatifloxacin can prolong qt interv [...] use quinidin



Outline of algorithm

1. For each ADR term and free text entry
 - Remove all non-alphanumerics and set to lower case
 - Remove all stopwords
 - Exchange each individual word by its stem
2. For each free text entry as resulting from step 1
 - Look for occurrences of each ADR term as resulting from step 1, as well as all permutations of its individual words
3. Study frequently non-matched words that are relevant, and, if possible, add synonyms to ADR terms in order to match these words (*not mandatory*)
4. Reiterate steps 1 and 2 (*not mandatory*)



Synonyms

- A few examples from this study:

Original word	Synonym
haemorrhage	bleeding
oestrogen	estrogen
cardiac	heart



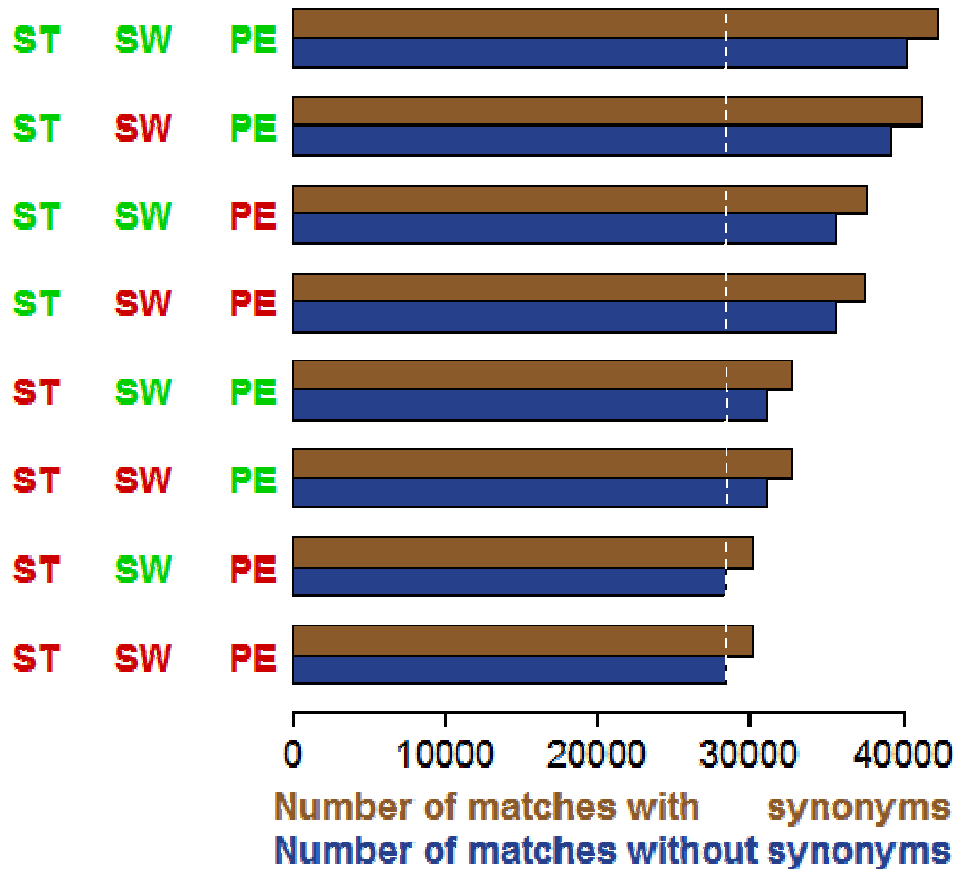
Overall result

- The algorithm extracted 42330 ADR terms from 40259 database entries



Results disseminated

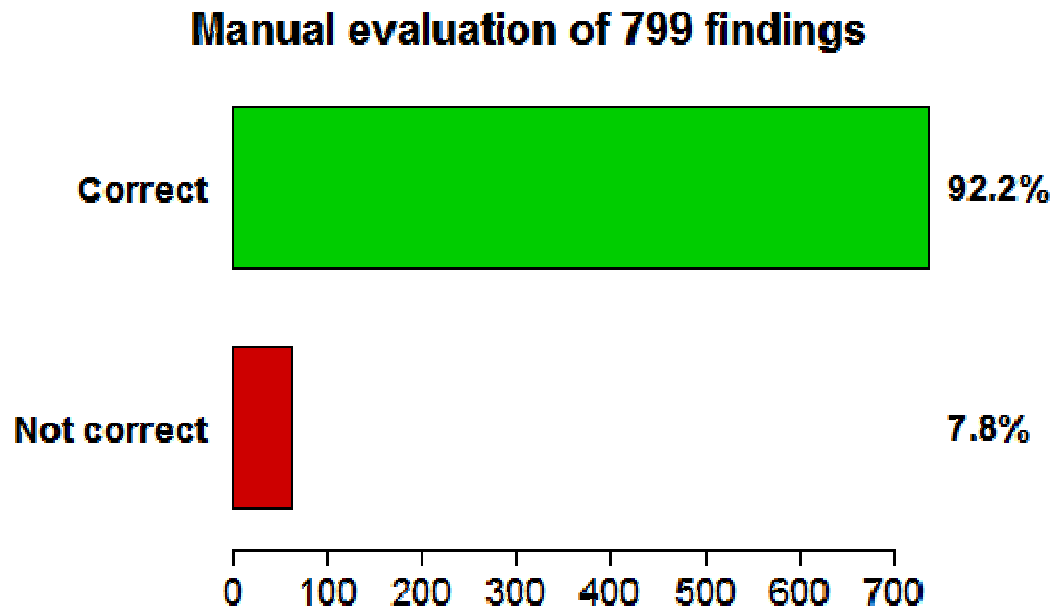
Effect of using stemming (ST), removing stopwords (SW), and using permutation (PE)



- Removal of stopwords was less important than using permutation, which was less important than stemming
- Combined effect (including synonyms): **49% increase** in number of matches
 - Suggests increased sensitivity



What about specificity?



- Relevant to examine whether the increased sensitivity comes at a loss of specificity
- A subset of findings were manually reviewed
- The rate of false positives is acceptable



Conclusion

- The algorithm, which combines generic principles with addition of source-specific synonyms, is able to increase the hit rate, and thus the sensitivity, compared to exact matching, while maintaining a high specificity.



References

- [1] Baxter K (ed.). *Stockley's Drug Interactions (XML)*. London: Pharmaceutical Press, 2008.
- [2] <http://snowball.tartarus.org/>
- [3] Porter MF. *An algorithm for suffix stripping*. Program 1980; **14**(3):130–137.



