MIE 2011

Checking coding completeness by mining discharge summaries

Stefan Schulz, Thorsten Seddig, Susanne Hanser, Albrecht Zaiß, Philipp Daumke

Undercoding of in-patient treatment episodes is a common problem in hospital information systems

- Incompleteness of disease encoding (ICD10) in hospitals
  - main diagnosis coded
  - comorbidities (secondary diseases) often not coded
  - typical: multimorbid patient admitted for surgical intervention (hip replacement, lens implant, prostatectomy…)
- Investigation: Does undercoding affect reimbursement given a DRG (diagnosis related group) - related reimbursement system ?
- Setting:
  - University Hospital of Freiburg (Germany)
  - Only very severe comorbidities have impact on DRG grouping in German DRG system

Undocumented diagnoses can be detected by mining the EPR

- Hypothesis: drug (ingredient) names in the EHR for which there is no justifying ICD code point to undocumented diseases
- Most trustworthy source of drug prescriptions: Discharge summary (in many departments no structured documentation of drug administration)
- Focus on three diseases, known to be readily omitted
  1. **Diabetes mellitus**,
  2. **Parkinson's disease**,
  3. **Bronchial asthma and chronic obstructive pulmonary disease (COPD)**.

Rule base checks medical texts annotated by ICD codes for completeness

- For each of 34,865 treatment episodes:
  - discharge summary
  - one or more ICD codes
  - 17,000 used for training, 17,865 for testing
- Rule base for diabetes, Parkinson's and COPD:
  - Drug indications (in ICD) manually extracted from two databases (Rote Liste, MMI) and enriched by off-label use from the training corpus
  - Including brand names and ingredient names
  - Each rule encoded as a triple $R = (D, P, N)$ with
    D = string characterizing a drug
    P = "positive list" of ICD codes for the diseases under scrutiny
    N = "negative list" of ICD codes for other indications
- Exact string match

Documents with unjustified drug mentions are filtered

Filter algorithm retrieves documents (cases) for which no justification for a drug name (ICD code in the HIS) is found:

```
For each d = {diabetes, Parkinson's, COPD}
    For each document:
      For each drug name specific to d:
        If drug name matches text token in document:
           If no match between any discharge ICD
           code and any code in the negative
           or positive list for d:
           Return document (candidate for undercoding)
```

Estimation of Precision and Recall

- Precision: text samples (n = 3 * 50) of the retrieved texts were analyzed by a domain expert
- Recall: roughly estimation by set of documents already annotated with a ICD code of interest.

```
For each d = {diabetes, Parkinson's, COPD}
  For each document:
    If annotated with ICD code from positive list:
      For each drug name specific to d:
        If drug name matches text token in document:
           If no match between any discharge ICD
           code and any code in the negative list for d:
           Return document
```

- Recall estimator:
  1 – (# docs returned / # docs with ICD code from pos. list)

High rate of false positives for Parkinson's and COPD

		Diabetes	Parkinson	Asthma / COPD
Summaries with relevant drug names		984	232	875
Summaries without justifying ICD annotations		201	65	172
Sample for expert rating		50	50	50
Code missing? (expert rating)	Yes	39	7	27
	No	11	43	23
Precision		79%	14%	45%
Estimated number of undercoded episodes		158	9	77

Candidates for missing codes as returned by algorithm 1 and estimated precision after rating of 50 treatment episodes per disease.

Most false positives due to other indications not coded or not in rule base

Diabetes drugs	Parkinson drugs	Asthma / COPD drugs
Multi organ failure.	Restless legs syndrome not coded	Foreign body in lung.
Hyperglycemia as side effect of severe respiratory infection	Essential tremor not coded	Pulmonary atresia combined with rhinitis and varicella.
Patient participates in a clinical trial	Huntington's chorea	Acute myleoid leukaemia and fever.
Unique insulin dose given to mitigate steroid side effects	Acute seizure	Lymph node tuberculosis. Oxis to be taken on demand.
Patient with a glucose tolerance test.	Richardson Olszewski syndrome	Pneumonia after stem cell transplantation.
Lab result for serum insulin mesurement.	Paranoid schizophrenia	Salbutamol to decrease the potassium level
18 months-old infant is resuscitated	Hypokinetic rigid syndrome	Lung cancer

Analysis of false positives

Recall low for Diabetes and COPD

	Retrieved cases using filter			
	+	-	Total	Recall
Diabetes	783	1031	1814	783/1814 = 43%
Parkinson	106	45	151	106/151 = 70%
Asthma / COPD	99	173	272	99/272 = 36%

Recall estimation based on correctly coded diagnoses (algorithm 2).

Most false negatives due to diseases not treated by drugs

		Diabetes	Parkinson	Asthma / COPD	Rate
Disease treatment without drug administration	Specific drug administration not mentioned in summary	11	14	11	72%
	Specific drug administration mentioned in summary	1			2%
Disease treatment with drug administration	Drug not listed in rule base	5		3	16%
	Drug name typing variant not matched with rule base		2		4%
	Drug not correctly recognized		1	2	6%

Analysis of false negatives

Undercoding significant, but not relevant for hospital revenue; methods can be optimized

- Of all treatment episodes under scrutiny, 2% were undercoded re diabetes mellitus, Parkinson's or COPD
- Diseases deemed secondary or unrelated to the actual clinical problem tend to be omitted, given that that they have no impact for DRG grouping
- Very severe comorbidities (with relevance for DRG grouping) are normally coded; no single case of DRG-relevant undercoding
- Improvement of the method: context sensitivity, spelling correction, automation of rule base construction, searching for other text elements