Diagnostic Free Text Analysis in Biobanks with CRIP.CodEx: Automated Matching of Classifications

Oliver Gros
Fraunhofer Institute for Cell Therapy and Immunology, Branch Bioanalytics and Bioprocesses (IZI-BB)
Am Mühlenberg 13
14476 Potsdam, Germany
oliver.gros@izi-bb.fraunhofer.de

Reinhard Thasler
Biobank under administration of HTCR, Department of General, Visceral, Vascular and Transplantation Surgery
Hospital of the University of Munich
81377 Munich, Germany
reinhard.thasler@med.uni-muenchen.de

Abstract

Biobanks represent key resources for biomedical research. To be accessible, e.g. over web-based query tools or trans-institutional metabiobanks, the stored human biospecimens have to be annotated with clinical data, transformed into harmonized and structured form, e.g. ICD codes, while currently only available from free text records.

The Biobank under Administration of Human Tissue and Cell Research Foundation HTCR at the University of Munich Medical Centre is routinely collecting remnant tissues and blood samples from treatments of patients. For diagnostic classification of the corresponding cases, a biobank specific classification was developed, but not yet matched to ICD codes.

So far done manually, we now used the automated knowledge extraction software CRIP.CodEx, not needing a training set or access to external resources, to recodify the textual description of the specialized HTCR biobank classification with ICD. We show that the information contained in the nomenclature of the individual biobank specific catalogue of diagnoses is sufficient for a mapping towards ICD-10 as well as ICD-O-3 catalogues, and deliver an automated matching of two different classification systems using CRIP.CodEx.

1 Introduction

Biobanks represent key resources for translational research and personalized medicine (Thasler et al., 2013). The Biobank under Administration of Human Tissue and Cell Research Foundation HTCR at University of Munich Medical Centre is routinely collecting remnant tissues and blood samples from treatments of patients at the Clinic for General, Visceral, Vascular and Transplantation Surgery as well as the Department for Thoracic Surgery.

However, to be a valuable resource for translational biomedical research these human biospecimens have to be annotated with clinical data, currently only available from free text records.

Access to these biospecimens and data, e.g. over web-based query tools, like the trans-institutional metabiobank CRIP (Schröder et al., 2011) or p-BioSPRE (Weiler et al., 2014), demands that information from various sources

gets integrated into harmonized and structured data to enable stratified, parameterized queries (Ambert and Cohen, 2009).

So far, analysis of free text sources and structured data entry in the data protection compliant biobank information system (Müller and Thasler, 2014) is done manually, and regarding diagnostic classification of cases, a biobank specific classification was developed. This classification however was not matched to ICD codes so far. Based on automated free text analysis this coding can now be amended with international classifications, e.g. ICD-10 or ICD-O-3, to facilitate project queries.

2 Methods

The automated knowledge extraction software CRIP.CodEx was designed to identify and extract information in free text medical records and assign corresponding codes (see Figure 1).

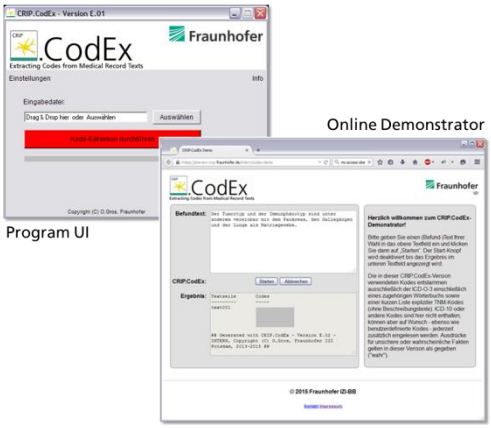

Figure 1: UI variants for CRIP.CodEx[1]

CRIP.CodEx is part of the CRIP-Toolbox[2] and will be published separately. It works automated, fast and efficient[3], identifies word relations and negation, handles extended negation scopes (Gros and Stede, 2013), but does not need access to databases or other external resources. Specifically, it does neither need a training set of pre-annotated texts, nor does it need its extraction rules input manually. The source for the self-generated extraction rules are lists of codes and their descriptions contained for example in coding guides for ICD-10 or ICD-O-3.

We used CRIP.CodEx slightly off its designed-for purpose, which is free text medical reports. Instead we used CRIP.CodEx to recodify the textual description of the specialized HTCR biobank classification with ICD (see Figure 2). The initially existing catalogue of diagnoses was based on the actual collection of cases in the biobank, reflecting indications and surgical treatments and therefore being primarily aligned along a list of organs affected by the surgical treatment. To build a catalogue of diagnoses suitable for biospecimen research, in a first step, this catalogue had to be reorganized towards distinct pathological findings across affected organs.

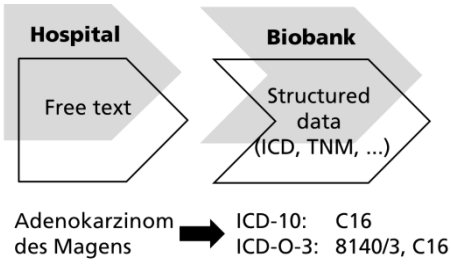

Figure 2: Knowledge Extraction with CRIP.CodEx – text example

The resulting categories therefore were composed of two separate elements, a diagnosis from pathological report and the affected organ. These categories, 65 in total, were then manually referred to the ICD-10 catalogue resulting in 82 codes.

2.1 Automated Matching

After ensuring that the final categories were sufficiently selective, we analyzed them in a first run using CRIP.CodEx for ICD-O-3 as well as ICD-10 classifications. The result was checked manually and discussed with a Pathologist.

However, the discussion of the first run showed that the problems identified relied not only to CRIP.CodEx, but also to the diagnostic catalogue, mainly the categories' composition as separate elements of diagnosis and organs affected by the surgical treatment. Both sources could be tackled – and therefore both, the software configuration and also the categories have been worked on to reach improved results in the second run.

[1] Online demonstrator available at https://preview-crip.fraunhofer.de/intern/codex-demo/

[2] www.crip.fraunhofer.de/en/toolbox

[3] Time per text less than one up to a few seconds (10^3 words)

2.2 Optimization

The configuration of CRIP.CodEx was optimized including these aspects: adding variants of local wording/synonyms, e.g. 'Carzinom' instead of 'Karzinom'/'carcinoma', into the internal dictionary, increasing the number of allowed multiple matches when detecting combined and hyphenated words with synonyms, as well as restricting ICD-10 neoplasm code assignment depending on detected ICD-O-3 classification.

The categories of the biobank specific classification (see Table 1) have been amended mainly by integrating the pathological diagnosis and the primarily diseased organ in to one descriptive text. Further amendments include specifying the affected organs and information on primary and secondary tumors more consistently, as well as adding tumor types in some categories while shortening the description of others and dividing into subcategories, accompanied by also amending the ICD-O-3 code assignments with an expert from the regional tumor registry towards the WHO "Blue Book" (Bosman et al., 2010) as international standard.

After both, optimizing the configuration of CRIP.CodEx, as well as further amending of the categories, we performed a second run for analyzing the categories.

ICD-10	**Diagnosis category**	**ICD-O-3 M**
C15	Plattenepithelkarzinom des Ösophagus	8070/3
C16	Adenokarzinom des Magens	8140/3
C16, C17, C25, C74	Sarkom (Bauchraum: Leber, Magen, Pankreas, Dünndarm, Dickdarm, Niere)	8800/3
C16, C22.9	Hepatozelluläres Karzinom (HCC)	8010/3, 8170/3
C16.9, C16, C22.9	Hepatozelluläres Karzinom (HCC), Subtyp: fibrolamelläres Leberzellkarzinom	8010/3, 8170/3
C17	Karzinome des Dünndarmes	8010/3
C17, C18.9, C20	Kolorektales Karzinom (CRC), auch Adenokarzinome	8010/3, 8140/3
C17, C22.9	Leiomyosarkom (Magen, Abdomen, Peritoneum, Retroperitoneum, Bindegewebe)	8890/3
C17, C22.9, C25	Cholangiokarzinom, extrahepatisch, auch Klatskin-Tumoren	8160/3
C22.1, C22.9	Cholangiokarzinom (CCC), intrahepatisch	8010/3, 8160/3
C22.3	Angiosarkom der Leber	9120/3
C23	Gallenblasenkarzinom	8010/3
C25	Adenokarzinom des Pankreas	8140/3
C45, C45.0, C45.9	Mesotheliom (Pleura, Peritoneum)	9050-55/0-3
C73	Schilddrüsenkarzinom (papilläres, follikuläres, medulläres, anaplastisches)	8010/3, 8021/3, 8050/3, 8260/3, 8330/3, 8510/3
C74	Nebennierenkarzinom	8010/3
C78.7	[Lokalisation, z.B. Leber]metastase nach [diverse Primärkarzinome]	8000/6, 8010/6
C80	Lebermetastase bei unbekanntem Primärtumor (CUP-Syndrom)	8000/6
D13.4, D13.6, D35.0, D37.2, D44.1	Adenom (einschl. Zystadenom) des/der Leber, Pankreas, Dünndarm, Dickdarm, Nebenniere	8140/0, 8440/0
D13.4, D18.0	Hämangiom der Leber	9120/0
D30.0, D35.0, D41.0	Phäochromozytom	8700/0
E66	Adipositas per magna (BMI >40)	
K51	Colitis ulcerosa	
K57.32, K57.33	Divertikulitis	

Table 1: Final catalogue of diagnoses (excerpt) and their classification in ICD-10 and ICD-O-3

3 Results

3.1 First run

The reorganized catalogue of diagnoses contained 64 categories and has been automatedly matched to classifications by CRIP.CodEx. For each of the categories CRIP.CodEx assigned matching classifications in ICD-10 or ICD-O-3. All together there have been 237 correct code assignments (true positives), while we identified 6 wrong assignments (false positives) and 12 missing codes (false negatives).

Due to the traceability of CRIP.CodEx's each individual code assignment, it was possible to identify the causes for all false positives and negatives. All but two causes could be resolved by the outlined optimization in the CRIP.CodEx configuration and the read in coding guide respectively.

3.2 Second run

The configuration of CRIP.CodEx as well as the original reorganized biobank specific catalogue of diagnoses has been optimized and amended as outlined above. Then again the local catalogue has been automatedly matched to ICD classifications by CRIP.CodEx. The final catalogue of diagnoses contains 73 categories, for each of them the software assigned matching classifications in ICD-10 or ICD-O-3 in the final second run. All together there have been 442 correct code assignments (true positives) by CRIP.CodEx, while we identified zero wrong assignments (false positives) and 14 missing codes (false negatives).

4 Conclusion & Outlook

Since diagnostic information contained in medical free text is extracted and codified by the automated CRIP.CodEx software, we also showed that the information contained in the nomenclature of the individual biobank specific catalogue of diagnoses is sufficient for a mapping towards ICD-10 and basically also ICD-O catalogues. As a remaining issue however, for categories such as e.g. "Carcinoma of the gallbladder", which summarize a wide range of different morphologies, ICD-O code extraction from nomenclature is too general, and amending the nomenclature with a listing of these morphologies is also dissatisfying.

For the implementation of the restructured and amended biobank-specific catalogue in the Biobanks Database or the HTCR Web Application however, the categories have been now structured as a table by key features that can be maintained:

- "Primary tumor vs. secondary tumor vs. no tumor"
- "affected Organ"
- "originating organ in case of secondary tumor"
- "included morphologies"

So as a next step, by tapping complementary data sources, mainly extracting diagnostic information from pathology reports by CRIP.CodEx, in addition to coding of specialized local classification, we deliver an automated matching of two different classification systems.

Even if all automated classifications have to be checked thoroughly according to an individual request from a research project, before providing samples and data, this initial, very effective automated classification to great extent facilitates case related database research as well as data export for display of the collection in biobank registries and even metabiobanks. Thereby we have not only enhanced the biobank's availability for translational research but also proposed a general protocol for matching internal codes with international classifications and standards.

Acknowledgements

The authors specially thank Dr. Jens Neumann from Institute of Pathology, Ludwig-Maximilians-University Munich and Dr. Gabriele Schubert-Fritschle of Tumor Registry Munich for discussing the results of CodEx code assignments and the amended categories and manual code assignments.

References

Ambert KH and Cohen AM (2009): A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. Journal of the American Medical Informatics Association 16, 590-595.

Bosman FT, Carneiro F, Hruban RH, Theise ND (Editors): WHO classification of tumours of the digestive system – 4th edition, IARC: Lyon, 2010.

Gros O and Stede M (2013): Determining Negation Scope in German and English medical diagnoses, in Taboada M and Trnavac R (Eds.): Nonveridicality and Evaluation – Theoretical, Computational and Corpus Approaches, Studies in Pragmatics 11, BRILL. ISBN: 9789004258167.

Müller TH, Thasler R (2014): Separation of personal data in a biobank information system. Stud Health Technol Inform. 014; 205:388-92.

Schröder C, Heidtke KR, Zacherl N, Zatloukal K, & Taupitz J (2011): Safeguarding donors' personal rights and biobank autonomy in biobank networks: the CRIP privacy regime. Cell and tissue banking, 12(3), 233-240.

Thasler WE, Thasler RM, Schelcher C, Jauch KW (2013): Biobanking for research in surgery: are surgeons in charge for advancing translational research or mere assistants in biomaterial and data preservation? Langenbecks Arch Surg. 2013 Apr; 398(4):487-99.

Weiler G, Schröder C, Schera F, Dobkowicz M, Kiefer S, Heidtke KR, Hänold S, Nwaknwo I, Forgó N, Stanulla M, Eckert C, and Graf N (2014): p-BioSPRE – an information and communication technology framework for transnational biomaterial sharing and access. ecancer 2014, 8:401-419.