EHR-based phenome wide association study in pancreatic cancer

Tomasz Adamusiak MD PhD[1*], Mary Shimoyama PhD[1,2]
[1]Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI
[2]Department of Surgery, Medical College of Wisconsin, Milwaukee, WI
[*]tomasz@mcw.edu

Abstract

BACKGROUND. *Pancreatic cancer is one of the most common causes of cancer-related deaths in the United States, it is difficult to detect early and typically has a very poor prognosis. We present a novel method of large-scale clinical hypothesis generation based on phenome wide association study performed using Electronic Health Records (EHR) in a pancreatic cancer cohort.* ***METHODS.*** *The study population consisted of 1,154 patients diagnosed with malignant neoplasm of pancreas seen at The Froedtert & The Medical College of Wisconsin academic medical center between the years 2004 and 2013. We evaluated death of a patient as the primary clinical outcome and tested its association with the phenome, which consisted of over 2.5 million structured clinical observations extracted out of the EHR including labs, medications, phenotypes, diseases and procedures. The individual observations were encoded in the EHR using 6,617 unique ICD-9, CPT-4, LOINC, and RxNorm codes. We remapped this initial code set into UMLS concepts and then hierarchically expanded to support generalization into the final set of 10,164 clinical concepts, which formed the final phenome. We then tested all possible pairwise associations between any of the original 10,164 concepts and death as the primary outcome.* ***RESULTS.*** *After correcting for multiple testing and folding back (generalizing) child concepts were appropriate, we found 231 concepts to be significantly associated with death in the study population.* ***CONCLUSIONS.*** *With the abundance of structured EHR data, phenome wide association studies combined with knowledge engineering can be a viable method of rapid hypothesis generation.*

Introduction

The Health Information Technology for Economic and Clinical Health (HITECH) Act introduced the concept of Meaningful Use of information technology in health care. As part of this process, the legislation mandated the use of standard terminologies for electronic exchange of health information. Patient clinical records represent a largely untapped treasure trove of research information, which only recently has become more accessible thanks to the increasing adoption of Electronic Health Records and healthcare data standards. The need to integrate and exchange clinical data has long been recognized[1], but it was the HITECH Act that provided the final piece of the puzzle in terms of financial incentives.

A number of terminology standards are currently in use. **LOINC** (Logical Observation Identifiers Names and Codes) is a universal standard for identifying laboratory observations[2]. **RxNorm** is a standardized nomenclature for generic and branded drugs, as well as drug delivery devices. RxNorm provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software[3]. The Healthcare Common Procedure Coding System (HCPCS) maintained by the Centers for Medicare & Medicaid Services (CMS) is a standardized coding system for describing items and services provided in the delivery of healthcare[4]. It incorporates Current Procedural Terminology (**CPT**), a coding system maintained by the American Medical Association (AMA) to identify medical services and procedures furnished by physicians and other health care professionals[5]. International Classification of Diseases, Clinical Modification (**ICD-9-CM**) is an adaption created by the U.S. National Center for Health Statistics (NCHS) and used in assigning diagnostic and procedure codes associated with inpatient, outpatient, and physician office utilization in the United States[6]. All these terminologies are integrated within the **UMLS** (Unified Medical Language System) maintained by the National Library of Medicine (NLM)[7].

Current state of the art in extracting actionable information from EHR relies on large scale text-mining and NLP of clinical notes[8,9] or either focuses on a specific terminology within the EHR, e.g., ICD-9-CM[10,11] or looks into a handcrafted, small subset of EHR variables[12]. Our approach is novel in the sense that we analyzed the complete corpus of structured data within the EHR across all available terminology standards, as well as used an existing knowledge base (UMLS) to expand and generalize the findings.

Methods

Extract, Load and Transform (ELT)

A *Limited Data Set*, as defined under the Health Insurance Portability and Accountability Act (HIPAA), was obtained from the Medical College of Wisconsin Clinical Research Data Warehouse for this analysis. The data extract was in the form of standard Epic Clarity tables for a subset of patients that had an encounter or a problem list code in the *Malignant neoplasm of pancreas* (ICD9:157) code subset:

157 Malignant neoplasm of pancreas

- **157.0** Malignant neoplasm of head of pancreas
- **157.1** Malignant neoplasm of body of pancreas
- **157.2** Malignant neoplasm of tail of pancreas
- **157.3** Malignant neoplasm of pancreatic duct
- **157.4** Malignant neoplasm of islets of langerhans
- **157.8** Malignant neoplasm of other specified sites of pancreas
- **157.9** Malignant neoplasm of pancreas, part unspecified

Data was loaded into our in-house clinical analytics portal (ClinMiner), which was used to dynamically translate between any of the underlying clinical terminologies, and provided a consolidated view of the underlying patient data in a single UMLS perspective[13]. Drug information in EHR was encoded using MediSpan terminology, one of the RxNorm sources, which facilitated its automatic translation into UMLS. Labs were encoded as orders using CPT-4 codes or using a fixed category from the *CLARITY_COMPONENT* lookup table. We have manually mapped 130 tests from *CLARITY_COMPONENT* to LOINC, which provided coverage for over 97% of all lab observations (1 493 101 observations in total). Remaining ~3% lab observations were left unmapped and excluded from further analysis.

The source annotation space covered 6 617 unique ICD-9, CPT-4, LOINC, and RxNorm codes. This code set was then remapped into UMLS to facilitate further analysis, which resulted in 6 741 distinct UMLS CUIs (Concept Unique Identifiers). This code set was then expanded across a limited set of *is_a* and selected other relationships (e.g., *has_ingredient* for RxNorm drugs) as an extension of the method previously proposed in a method similar to that of parent child analysis described by Grossmann et al.[14].

but not beyond the original set of UMLS Metathesaurus semantic types of the expanded concepts to exclude functional concepts from the analysis and to keep the general meaning of the originating concept in the expansion. Additionally, the UMLS traversal was limited to either the UMLS Metathesaurus itself, or any of the following terminologies specific to Meaningful Use: RxNorm, NDF-RT, LOINC, SNOMED CT, HCPCS, and ICD-9-CM. This resulted in 18 038 concepts. Finally, we discarded 7 874 concepts that did not increase information content (i.e., were redundant in terms of partitioning of the underlying data) to reach the final ‘phenome’ of 10 162 concepts.

Statistical analysis

A chi-squared test was used to (χ^2) to test the significance of the associations. To correct for multiple testing we used a Bonferroni correction and tested at a level of $p < 4.9 \times 10^{-6}(0.05/10162)$. Odds Ratio (OR) and Relative Risk (RR) were used to assess the effect size of associations found to be significant.

Results

713 concepts were found to be significantly associated with death in the study population. Where both parent and its child concepts were found to be significant, child concepts were removed to further generalize the results and final

result set was thus reduced to 231 terms. A breakdown of all concepts by category and number of observations is shown in Figure!1.

Most of the terms were positively correlated with death and only the following 9 concepts were found to be associated with lower relative risk of death in the study population:

- Immunoassay for tumor antigen, quantitative; CA 125
- Vitamin D; 25 hydroxy, includes fraction(s), if performed
- Prealbumin measurement
- Racial group
- Benzoic acid or derivative
- Iodine AND/OR iodine compound
- Ionic iodinated contrast media
- Triiodobenzoic Acids
- sevoflurane Inhalant Solution

For practical reasons, only the top ten (five from each side) significant associations are shown in Table 1. The complete result set encompassing all 231 significant associations is available as supplementary materials at http://dx.doi.org/10.6084/m9.figshare.816958.

Discussion

As with any retrospective observation the primary limitation is a lack of a prospective control group, which means the results can be biased due to an imbalanced design. It is also worth noting, that correlation does not imply causation. For example, while cytopathology was found to be associated with an increased risk of death, it is more likely due to selection bias. Patients with more advanced disease more frequently underwent the procedure as part of their diagnostic process. There are also limitation due to data incompleteness. For example, here we looked at known deaths from the EHR only and did not include data from outside sources such the National Death Index. On the other hand, retrospective designs have the advantage of observing real clinical practice.

We have observed that the use of contrast media and medical gases used to induce anesthesia lowered the risk of death in the study population. This confirms an already known association between hospital resource utilization and patient mortality[15,16].

Cimetidine, an H2 receptor antagonist, has a known off-label use as an anticancer drug[17–19]. Paradoxically, its use was associated with an increased risk of death in our study population. However, this subpopulation was also older than the rest of the cohort, and likely increased mortality was due to a more advanced disease process. Without access to clinical notes, we can only speculate that perhaps this was a part of an experimental treatment.

We see the potential to use this approach to automatically generate groupers or value sets of closely related concepts. This could be used either in the EHR to alert the physician to other possibly relevant features of patient presentation as well as on the research side to make more informative patient cohort selections.

A major critique would be that they we only looked at association of the concept and not the value. Presence of an observation on a patient-level also discards the temporal and frequency information. On the other hand, this would also increase dimensionality of the analysis (cf. *curse of dimensionality*) and would require not only a more sophisticated statistical approach but could also suffer from lower statistical power. These are some of the challenges that we hope to address in future work.

Label	***CUI***	***Semantic Type***	**Exposed Deceased**	**Exposed Alive**	**Not Exposed Deceased**	**Not Exposed Alive**	**p**	**OR**	**RR**
Increased Risk (RR > 1)									
Cytopathology, fluids, washings or brushings, except cervical or vaginal; smears with interpretation	*C0374051*	*Laboratory Procedure*	28	8	808	310	8.31×10^{-11}	9.12	**2.80**
Cimetidine	*C0008783*	*Pharmacologic Substance*	17	6	810	321	2.03×10^{-6}	7.14	**2.60**
Hyposmolality and/or hyponatremia	*C0020645*	*Finding*	22	10	806	316	6.54×10^{-7}	5.61	**2.44**
Osmolality; blood	*C0373690*	*Laboratory Procedure*	43	25	791	295	2.29×10^{-10}	4.61	**2.32**
Haptoglobin; quantitative	*C0373631*	*Laboratory Procedure*	25	14	802	313	1.17×10^{-6}	4.57	**2.28**
Decreased Risk (RR < 1)									
sevoflurane Inhalant Solution	*C1253873*	*Clinical Drug*	7	85	731	331	1.90×10^{-6}	0.18	**0.24**
Triiodobenzoic Acids	*C0041013*	*Organic Chemical*	56	332	484	282	3.00×10^{-15}	0.28	**0.39**
Ionic iodinated contrast media	*C0361904*	*Indicator, Reagent, or Diagnostic Aid*	57	332	484	281	6.66×10^{-15}	0.29	**0.39**
Iodine AND/OR iodine compound	*C0303013*	*Inorganic Chemical Pharmacologic Substance*	60	338	478	278	1.38×10^{-14}	0.30	**0.40**
Benzoic acid or derivative	*C0578497*	*Organic Chemical*	58	328	488	280	4.41×10^{-14}	0.30	**0.41**

Table 1: Top ten events by effect size significantly associated with death in the study cohort. Abbreviations: CUI – Concept Unique Identifier; OR – Odds Ratio; RR – Relative Risk.

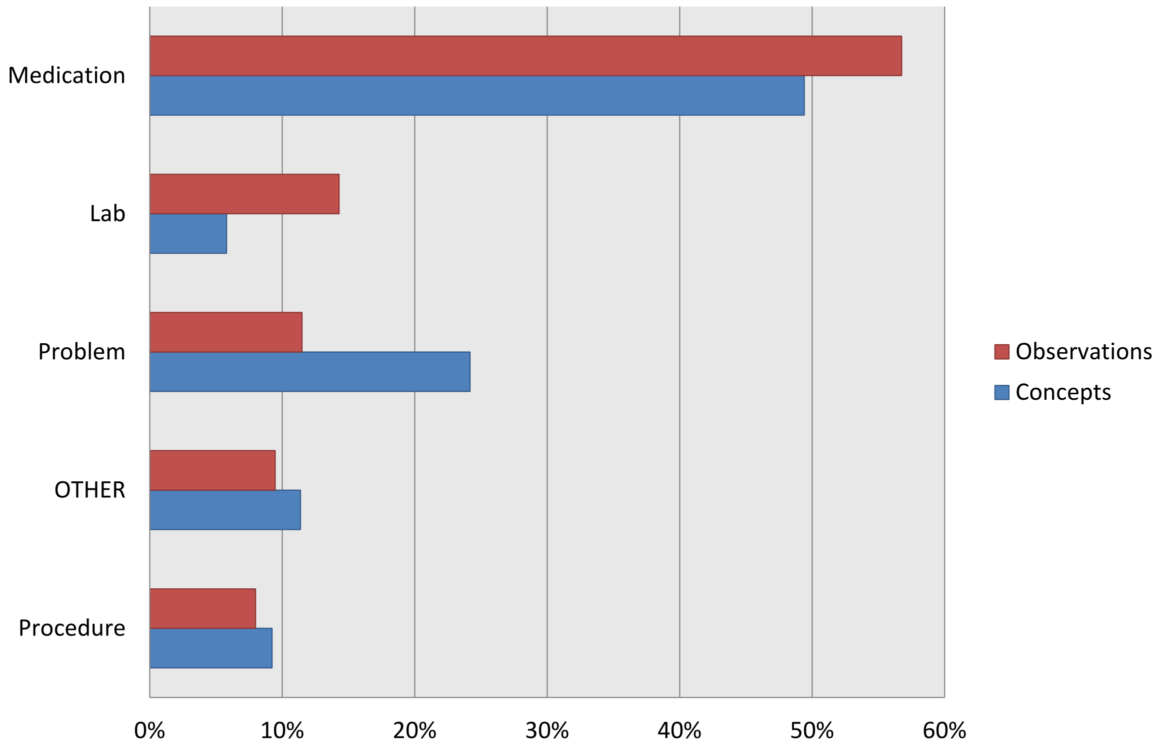

Figure 1: Breakdown of all 10 164 concepts by semantic type category and proportion of observations annotated with a particular concept. Categories are defined as follows. **Lab** is any UMLS concept in the semantic type tree: *A2.3.1. Clinical Attribute*, *A2.2.1 Laboratory or Test Result*, or *B1.3.1.1 Laboratory Procedure*. **Procedure** is a UMLS concept that is in the *B1.3.1 Health Activity* branch, but is not a *1.3.1.1 Laboratory Procedure*. **Problem** is a concept with a semantic type under *B2.2.1.2 Pathologic function* or *A2.2.2 Sign or symptom*. **Medication** groups all concepts classified by the UMLS Semantic Network either under semantic type *A1.4 Substance* or under *A1.3.3 Clinical Drug*. Finally, **OTHER** groups all other semantic types.

Conclusions

Information contained in EHR combined with knowledge engineering could be used a a viable method of rapid hypothesis generation, but requires comprehensive validation.

Acknowledgments

This project was funded in part by the Advancing a Healthier Wisconsin endowment at the Medical College of Wisconsin and the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through grant UL1 RR031973. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

We thank Stacy Zacher, Glenn Bushee, and Bradley Taylor for their help.

References

[1] J.J. Cimino and E.H. Shortliffe, editors. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine (Health Informatics)*. Springer-Verlag New York, Inc., Secaucus, NJ, 2006.

[2] C. J. McDonald. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. *Clinical Chemistry*, 49(4):624–633, April 2003. ISSN 0009-9147. doi: 10.1373/49.4.624.

[3] Fola Parrish, Nhan Do, Omar Bouhaddou, and Pradnya Warnekar. Implementation of RxNorm as a terminology mediation standard for exchanging pharmacy medication between federal agencies. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, page 1057, January 2006. ISSN 1942-597X.

[4] R Finnegan. HCPCS–outpatient procedures. *Journal (American Medical Record Association)*, 58(10):20–2, October 1987. ISSN 0273-9976.

[5] Current procedural terminology (CPT). *JAMA : the journal of the American Medical Association*, 212(5):873–4, May 1970. ISSN 0098-7484.

[6] R Finnegan. ICD-9-CM. *Journal (American Medical Record Association)*, 57(7):34–5, July 1986. ISSN 0273-9976.

[7] D A Lindberg, B L Humphreys, and A T McCray. The Unified Medical Language System. *Methods of information in medicine*, 32(4):281–91, August 1993. ISSN 0026-1270.

[8] Nicholas J Leeper, Anna Bauer-Mehren, Srinivasan V Iyer, Paea Lependu, Cliff Olson, and Nigam H Shah. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. *PloS one*, 8(5): e63499, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0063499.

[9] Svetlana Lyalina, Bethany Percha, Paea Lependu, Srinivasan V Iyer, Russ B Altman, and Nigam H Shah. Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. *Journal of the American Medical Informatics Association : JAMIA*, August 2013. ISSN 1527-974X. doi: 10.1136/amiajnl-2013-001933.

[10] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics (Oxford, England)*, 26(9):1205–10, May 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq126.

[11] Jeremy L Warner, Amin Zollanvari, Quan Ding, Peijin Zhang, Graham M Snyder, and Gil Alterovitz. Temporal phenome analysis of a large electronic health record cohort enables identification of hospital-acquired complications. *Journal of the American Medical Informatics Association : JAMIA*, August 2013. ISSN 1527-974X. doi: 10.1136/amiajnl-2013-001861.

[12] George Hripcsak and David J Albers. Correlating electronic health record concepts with healthcare process events. *Journal of the American Medical Informatics Association : JAMIA*, August 2013. ISSN 1527-974X. doi: 10.1136/amiajnl-2013-001922.

[13] Tomasz Adamusiak, Shimoyama Naoki, Tutaj Marek, and Shimoyama Mary. Next Generation Ontology Browser. In *Proceedings International Conference on Biomedical Ontology 2013*, pages 131–132, 2013.

[14] Steffen Grossmann, Sebastian Bauer, Peter N Robinson, and Martin Vingron. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics (Oxford, England)*, 23(22): 3024–31, November 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm440.

[15] Corrado Cecchetti, Riccardo Lubrano, Sebastian Cristaldi, Francesca Stoppa, Maria Antonietta Barbieri, Marco Elli, Raffaele Masciangelo, Daniela Perrotta, Elisabetta Travasso, Claudia Raggi, Marco Marano, and Nicola Pirozzi. Relationship between global end-diastolic volume and cardiac output in critically ill infants and children. *Critical care medicine*, 36(3):928–32, March 2008. ISSN 1530-0293. doi: 10.1097/CCM.0B013E31816536F7.

[16] Marko Kavcic, Brian T Fisher, Yimei Li, Alix E Seif, Kari Torp, Dana M Walker, Yuan-Shung Huang, Grace E Lee, Sarah K Tasian, Marijana Vujkovic, Rochelle Bagatell, and Richard Aplenc. Induction mortality and resource utilization in children treated for acute myeloid leukemia at free-standing pediatric hospitals in the United States. *Cancer*, 119(10):1916–23, May 2013. ISSN 1097-0142. doi: 10.1002/cncr.27957.

[17] O Sürücü, M Middeke, I Höschele, J Kalder, S Hennig, C Dietz, and I Celik. Tumour growth inhibition of human pancreatic cancer xenografts in SCID mice by cimetidine. *Inflammation research : official journal of the European Histamine Research Society ... [et al.]*, 53 Suppl 1:S39–40, March 2004. ISSN 1023-3830. doi: 10.1007/s00011-003-0318-1.

[18] Yisheng Zheng, Meng Xu, Xiao Li, Jinpeng Jia, Kexing Fan, and Guoxiang Lai. Cimetidine suppresses lung tumor growth in mice through proapoptosis of myeloid-derived suppressor cells. *Molecular immunology*, 54(1): 74–83, May 2013. ISSN 1872-9142. doi: 10.1016/j.molimm.2012.10.035.

[19] Martina Kubecova, Katarina Kolostova, Daniela Pinterova, Grzegorz Kacprzak, and Vladimir Bobek. Cimetidine: an anticancer drug? *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences*, 42(5):439–44, April 2011. ISSN 1879-0720. doi: 10.1016/j.ejps.2011.02.004.