Master's Internship on

Ontology-Driven Phenotyping from Electronic Health Records

Basic Information

Duration: 6 months, ideally between February - August 2022

Supervisors: Vianney Jouhet (Bordeaux University Hospital - IAM unit, Bordeaux), Fleur Mougin (Inserm U1219 - Erias unit, Bordeaux)

Host lab: Bordeaux University Hospital (Clinical data warehouse)

Context

With the increasing adoption of electronic health records (EHRs), the amount of data produced at the patient bedside is rapidly increasing. These data provide new perspectives to: create and disseminate new knowledge; consider the implementation of personalized medicine; offer to patients the opportunity to be involved in the management of their own medical data [1]. Indeed, the secondary use of biomedical data produced throughout the patient's care is an essential issue [2] and has been the subject of numerous studies for several years [1, 2, 3, 4, 5, 6, 7]. Although centers such as Harvard and Vanderbilt have enabled researchers to develop platforms for secondary use of EHRs [8, 9], its adoption and implementation within healthcare remain slow due to its complexity [10]. Secondary use of EHRs is not an easy task.

For Hripcsak et al., phenotyping `*transforms the raw EHR data into clinically relevant features'* [11]. Although clinical care now provides a large amount of data, using these raw EHRs for phenotyping purposes remains a challenge. Indeed, when reusing EHRs, multiple issues must be addressed: completeness, complexity, and biases that limit the feasibility [11]. These issues have to be taken into account, and the data generated from EHRs need to be evaluated [12].

The internship can lead to a funded PhD position within the INTENDED project.

Research Topic

Querying EHRs in order to identify patients corresponding to a specific phenotype (such as patients with metastatic lung cancer treated with tyrosine kinase inhibitors) often results in an incomplete and noisy set of patients due to incomplete, erroneous, and inconsistent data. For instance, the data may not contain a precise diagnostic because patients may have been diagnosed outside the hospital, so this information must be extracted from free-text documents or inferred from other types of available information (treatments, procedures...). Therefore, it is necessary to identify *possible* patients who meet these criteria and classify them based on the consistency and completeness of their data.

The main goal of this internship is to investigate the ability of a high-level domain ontology (including only broad domain concepts) to drive: (i) phenotype definition, (ii) patient retrieval, (iii) identification and visualization of inconsistent / incomplete EHRs regarding the defined phenotype. The developed methods will be evaluated on real-life research projects (metastatic lung cancer, Vexas syndrome).

Candidate Profile

This internship is best suited to candidates who have prior experience with knowledge representation (ontologies, description logic) and medical informatics.

Biological / medical knowledge would be an asset.

How to Apply

Candidates for the Master's internship should contact the two supervisors by email:

- Vianney Jouhet (vianney.jouhet@chu-bordeaux.fr)
- Fleur Mougin (fleur.mougin@u-bordeaux.fr)

The email should include a CV, course transcripts (last two years), and a short description of how the internship topic relates to their prior experience and research interests.

The position will remain open until a suitable candidate is found. However, for full consideration, applicants should get in touch by **January 3, 2022**.

Bibliography

[1] Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. JAMA: the journal of the American Medical Association, 309(13):1351–1352, April 2013.

[2] H U Prokosch and T Ganslandt. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods of information in medicine, 48(1):38–44, 2009.

[3] Charles Safran, Meryl Bloomrosen, W Edward Hammond, Steven Labkoff, Suzanne Markel-Fox, Paul C Tang, Don E Detmer, and Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association: JAMIA, 14(1):1–9, February 2007.

[4] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. AMIA Summits on Translational Science Proceedings, 2010:1–5, March 2010.

[5] Cynthia Barton, Crystal Kallem, Patricia Van Dyke, Donald Mon, and Rachel Richesson. Demonstrating “Collect once, Use Many” – Assimilating Public Health Secondary Data Use Requirements into an Existing Domain Analysis Model. AMIA Annual Symposium Proceedings, 2011:98–107, 2011.

[6] AbdenNaji El Fadly, Bastien Rance, No¨el Lucas, Charles Mead, Gilles Chatellier, Pierre-Yves Lastic, Marie-Christine Jaulent, and Christel Daniel. Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. Journal of biomedical

informatics, 44 Suppl 1:S94–102, December 2011.

[7] Georges De Moor, Mats Sundgren, Dipak Kalra, Andreas Schmidt, Martin Dugas, Brecht Claerhout, Töresin Karakoyun, Christian Ohmann, Pierre-Yves Lastic, Nadir Ammour, Rebecca Kush, Danielle Dupont, Marc Cuggia, Christel Daniel, Geert Thienpont, and Pascal Coorevits. Using electronic health records for clinical research: The case of the EHR4CR project. Journal of Biomedical Informatics, 53:162–173, 2015.

[8] Ioana Danciu, James D. Cowan, Melissa Basford, Xiaoming Wang, Alexander Saip, Susan Osgood, Jana Shirey-Rice, Jacqueline Kirby, and Paul A. Harris. Secondary use of clinical data: the Vanderbilt approach. Journal of Biomedical Informatics, 52:28–35, December 2014.

[9] Ruth Nalichowski, Diane Keogh, Henry C. Chueh, and Shawn N. Murphy. Calculating the benefits of a Research Patient Data Repository. AMIA Annual Symposium proceedings, page 1044, 2006.

[10] Bastien Rance, Vincent Canuel, Hector Countouris, Pierre Laurent-Puig, and Anita Burgun. Integrating Heterogeneous Biomedical Data for Cancer Research: the CARPEM infrastructure. Applied Clinical Informatics, 7(2):260–274, 2016.

[11] George Hripcsak and David J. Albers. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association: JAMIA, 20(1):117–121, January 2013.

[12] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association: JAMIA, 20(1):144–151, January 2013.