Nursing Informatics 2016
W. Sermeus et al. (Eds.)

doi:10.3233/978-1-61499-658-3-387

Integration and Analysis of Heterogeneous Colorectal Cancer Data for Translational Research

Jitendra JONNAGADDALA [a], Joanne L CROUCHER [a], Toni Rose JUE [a], Nicola S MEAGHER [a], Lena CARUSO [a], Robyn WARD [b], Nicholas J HAWKINS [c,1]
[a] *Prince of Wales Clinical School, UNSW Australia,* [b] *Office of DVC-Research, University of Queensland,* [c] *School of Medicine, University of Queensland, St Lucia, Australia*

Abstract. Cancer is the number one cause of death in Australia with colorectal cancer being the second most common cancer type. The translation of cancer research into clinical practice is hindered by the lack of integration of heterogeneous and autonomous data from various data sources. Integration of heterogeneous data can offer researchers a comprehensive source for biospecimen identification, hypothesis formulation, hypothesis validation, cohort discovery and biomarker discovery. Alongside the increasing prominence of big data, various translational research tools such as tranSMART have emerged that can converge and analyse different types of data. In this study, we show the integration of different data types from a significant Australian colorectal cancer cohort. Additionally, colorectal cancer datasets from The Cancer Genome Atlas were also integrated for comparison. These integrated data are accessible via http://www.tcrn.unsw.edu.au/transmart. The use of translational research tools for data integration can provide a cost-effective and rapid approach to translational cancer research.

Keywords. Colorectal cancer, data integration, data analysis, cohort discovery, biomarker discovery

1. Introduction

Cancer is one of the leading causes of mortality in Australia with estimated deaths of 45,700 per year. Colorectal cancer (CRC) is the second most common type of cancer in Australia. It is also one of the major burdens to health expenditure [1]. The basic biology of CRC development is well studied with major discoveries made in the last two decades. However, the translation of new discoveries into practice often takes many years. The analysis of vast amounts of heterogeneous data collected and generated over long periods is a key issue hindering the implementation of translational research. Often this valuable data is derived from varied data sources, while its storage using different formats and standards makes the data difficult to integrate and analyse. Clinical data, experimental data, biospecimen data and imaging data are the major heterogeneous data types observed in colorectal cancer translational research. From a personalized medicine point of view, integrating heterogeneous data is needed to provide a unified view for analysis [2, 3]. It

[1] Corresponding author: Nicholas J Hawkins, School of Medicine, University of Queensland, St Lucia, QLD 4072, Australia; E-Mail: n.hawkins@uq.edu.au

could provide translational researchers a comprehensive source for hypothesis formulation and validation, as well as cohort and biomarker discovery. At the same time, it also enables reuse of valuable existing data, thereby minimising cost and increasing research effectiveness.

The aims of this study are to i) integrate heterogeneous CRC data from both public and private sources ii) increase and provide easy global access to an Australian CRC study data and associated biospecimens, thereby fostering global collaborations and data reuse and iii) analyse and generate hypothesis using the integrated data.

2. Methods

The Molecular and Cellular Oncology (MCO) study is a major CRC study in Australia, recruiting over 1,500 participants between 1993 and 2010 [4]. Biospecimens were collected along with clinical data. During this time, large amounts of experimental and imaging data were also generated. Imaging data mainly included whole slide images. All patients were assessed for key biomarkers such as microsatellite instability (MSI), CIMP status, KRAS mutation and BRAF mutation. Additionally, public data was sourced from The Cancer Genome Atlas (TCGA), specifically the Colon adenocarcinoma (COAD) and Rectum adenocarcinoma (READ) studies [5]. The TCGA COAD and READ open access tier data was used for this project which included de-identified clinical and biospecimen data.

2.1. Distributed Data Sources

The MCO study clinical data was originally stored in a Microsoft Access™ database. As part of this study, the clinical data was extracted, clinically coded using SNOMED-CT terminology and migrated into the clinical data management system OpenClinica [2] to improve management of clinical data. The biospecimens collected as part of the MCO study are physically stored in the UNSW Biorepository. The associated biospecimen data, including specimen type, morphological abnormality and tissue site are stored in the biobanking laboratory information system (LIMS) OpenSpecimen (previously known as caTissue) [6, 7]. Over 1,700 whole slide images together with their metadata are stored in the web-based application Aperio Spectrum[3] (now part of Leica Biosystems). However, due to the size of the whole slides images are large only metadata of these images was used for data integration. The experimental data, mainly biomarker-related data was stored in Microsoft Excel. Figure 1 illustrates the distributed data sources and flow of data at a high level.

2.2. Development of Use Cases

The integration of heterogeneous data for translational research is challenging and the expected outcome of data integration widely varies from one translational researcher to another. In order to overcome this issue, we developed use cases by seeking feedback from translational researchers with diverse backgrounds, including pathologists, cancer epidemiologists, molecular biologists, bioinformaticians and medical oncologists. The

[2] https://openclinica.com/
[3] http://www.leicabiosystems.com/pathology-imaging/aperio-digital-pathology/

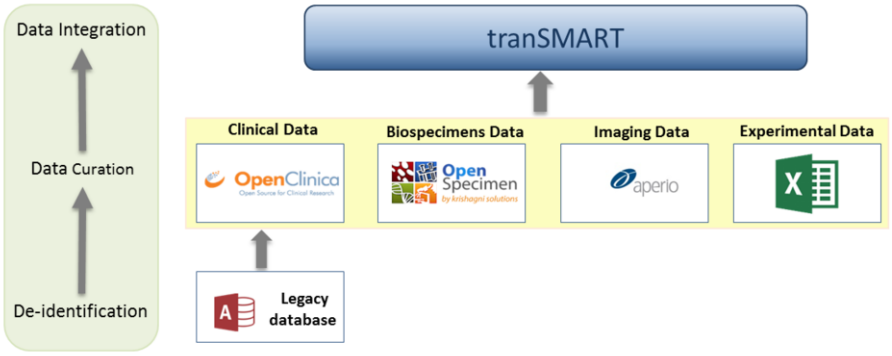

Figure 1. Distributed data sources and flow of data.

main use cases for this study included capabilities like biospecimen identification, cohort discovery, survival analysis, hypothesis generation and comparisons across other public datasets. The development of use cases helped us to identify the need to collect new data or summarise existing data. These developed use cases were later validated with the newly integrated data.

2.3. Hierarchical Representation of Heterogeneous Data

Flexible and sustainable representation of data in a hierarchical format is vital for effective downstream analysis of the integrated data in tools such as tranSMART. Thus, we employed an iterative design process with constant feedback from translation researchers to develop a model suitable for our developed use cases. The model was designed by employing a three step approach: i) evaluate and assess the suitability of existing models, ii) assess the standards and terminologies used in conjunction with data sources and iii) design and extend the model beyond CRC. The final model heavily relied on SNOMED-CT and ICD-10 terminologies for clinical data. The data from distributed data sources was curated to comply with the designed model and finally integrated using the tranSMART tool. Based on privacy and confidentiality requirements, potentially identifiable data were removed or replaced. Figure 2 below illustrates the high level structure of the hierarchical data model.

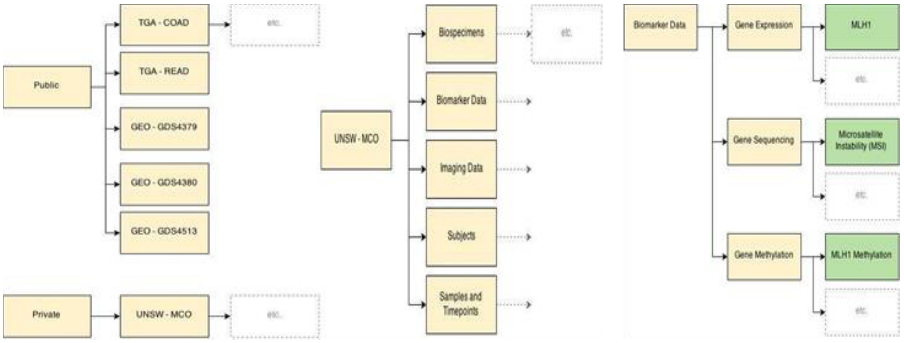

Figure 2. Snapshot of the data hierarchy at the study and data type level.

3. Results

Different types of CRC data, has been successfully integrated and analysed by adopting the tranSMART tool. The study took a little over 12 months to achieve its aims. More than 600 data variables were curated and integrated, representing more than 1,500 patients and thousands of associated biospecimens and whole slide images. TCGA data representing more than 600 patients was also integrated. Over twenty key clinical and biospecimen data variables were included for each of the TCGA COAD and READ datasets. Different use cases were identified and were verified using the integrated CRC data via the tranSMART tool. Complex queries can be constructed without any SQL programming experience. Biospecimen and cohort discovery has been made accessible with the added advantage of being able to request access to the biospecimens of patients from the subsequently identified cohort. Biospecimens may be browsed based on data variables such as anatomical site, specimen type, morphological abnormality and pathological status. Physical access to identified biospecimens (from the MCO cohort) is subject to standard governance and request procedures set forth by the UNSW Biorepository [8]. Cross study comparisons was also made available by integrating the Australian MCO and the American TCGA CRC cohort. Hypothesis generation was another important use case identified in this study. The integrated CRC data can assist researchers in developing new hypotheses or in validating an existing hypothesis using the tranSMART tool. In the absence of access to integrated CRC data via tools like tranSMART, similar analyses would take from weeks to months.

4. Discussion

TranSMART an open source translational research tool was implemented in Australia for the first time with access to MCO & TCGA CRC data. The custom features developed for the tranSMART tool as part of this project are available[4] under open source license.

There are some limitations in this study. Firstly, although high dimensional data is available for MCO and TCGA studies, integration of this data was outside the scope of this initial project and remains a planned future development. In addition, we observed that there is no support for temporal data in the tranSMART tool causing an overlap between some queries [9]. For example, a simple query was constructed to identify patients with and without MLH1 biomarker expression. From the descriptive statistics generated, we observed that 181 patients had loss of MLH1 staining, 1,311 with normal MLH1 staining and 17 with both loss and normal MLH1 staining. Given the query, it would be logical to observe absolute patient count for one or the other but not for both. In addition, for the calculation of survival time, the date of the initial surgical resection is considered as entry point due to the lack of diagnosis date data. Similar types of assumptions have been made with other data variables either because of missing data or noisy data [10]. Therefore, it is important for researchers to understand the underlying assumptions made during data integration and the limitations of translational tools during data analysis.

Data integration depends on numerous factors including heterogeneity, temporality and granularity of data; number of data sources; amount of missing and noisy data; common data models used and specific research objectives [11, 12]. While the

[4] https://github.com/TCRNBioinformatics/UNSWTransMart

development of use cases and data hierarchies was a useful process, some important technical limitations were identified during data integration, testing and analysis. A rigorous requirements analysis, technical solutions review and biocuration framework is recommended. The framework should take into account factors like quality of data, governance, validation, privacy, compliance and security requirements. Furthermore, with the rise of institutional biobanking, biospecimens are now associated with rich and routinely collected clinical data [13], prompting the need for sophisticated translational tools capable of integrating data in real time.

In summary, this study has used a large Australian colorectal cancer study to demonstrate the feasibility of using tranSMART as a tool for supporting translational cancer research. In future, we would like to explore the possibilities of integrating high dimensional data, and other CRC datasets from Australia and other countries. In addition, we also would like to develop a conceptual manual-biocuration framework for translational research based on our experiences from this study.

Acknowledgements

The MCO tranSMART project is supported by the Australian National Data Service (ANDS) through the National Collaborative Research Infrastructure Strategy Program, as well as through the Cancer Institute NSW and UNSW Australia. We acknowledge the MCO Study Group, the TCGA Research Network and Jack London, Vivek Ratnaparkhi, Santosh Maskar and Manish Kumar for their technical assistance in adoption and customization of tranSMART tool.

References

[1] AIHW, *Cancer in Australia: an overview*, Australian Institute of Health and Welfare, Canberra, Australia, 2014.
[2] A. Halevy, A. Rajaraman & J. Ordille, *Data integration: the teenage years*. Paper presented at the Proceedings of the 32nd international conference on Very large data bases, Seoul, Korea, 2006.
[3] M. Lenzerini, *Data integration: a theoretical perspective*. Paper presented at the Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Madison, Wisconsin, 2002.
[4] R. Ward, & N. Hawkins, *Molecular and Cellular Oncology (MCO) Study Data*. Retrieved from: https://researchdata.ands.org.au/mco-study-tumour-collection/17113, 2014.
[5] TCGA Research Network. TCGA. Retrieved from http://cancergenome.nih.gov/, 2015.
[6] J. Jonnagaddala, J. Li & P. Ray, *Evaluation of caBIG® caTissue Software.* Paper presented at the World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China, 2013.
[7] L.D. McIntosh, et al., Catissue Suite to Openspecimen: Developing An Extensible, Open Source, Web-Based Biobanking Management System, *Journal of Biomedical Informatics* **57** (2015): 456-464.
[8] UNSW Australia. Lowy Biorepository. Retrieved from http://biorepository.unsw.edu.au/, 2015.
[9] V. Canuel, B. Rance, P. Avillach, P. Degoulet & A. Burgun, Translational research platforms integrating clinical and omics data: a review of publicly available solutions, *Brief Bioinformatics* **16** (2014), 280-290.
[10] L. Ohno-Machado, Modeling medical prognosis: survival analysis techniques, *Journal of Biomedical Informatics* **34(6)** (2001), 428-439.
[11] C. Goble, R. Stevens, D. Hull, K. Wolstencroft & Lopez, R., Data curation + process curation = data integration + science, *Brief Bioinformatics,* **9** (2008), 506-517.
[12] J. Jonnagaddala, T.R. Jue, P. Ray & A. Talaei-Khoei, Data Sharing Challenges and Recommendations for Human Biorepositories: A Systematic Literature Review, *The International Technology Management Review* **4** (2014), 68 - 77.
[13] L. Wyld, S. Smith, N.J. Hawkins, J. Long, & R.L. Ward, Introducing Research Initiatives into Healthcare: What Do Doctors Think? *Biopreservation and Biobanking,* **12** (2014), 91-98.