HemaSphere
Powered by EHA

EUROPEAN HEMATOLOGY ASSOCIATION

Perspective
OPEN ACCESS

Challenges and Solutions for Collecting and Analyzing Real World Data: The Eric CLL Database as an Illustrative Example

Anastasia Chatzidimitriou[1,2], Eva Minga[1], Thomas Chatzikonstantinou[1,3], Carol Moreno[4], Kostas Stamatopoulos[1,2], Paolo Ghia[5], on behalf of ERIC, the European Research Initiative on CLL

Correspondence: Anastasia Chatzidimitriou (e-mail: achatzidimitriou@certh.gr).

ERIC, the European research on CLL

Chronic lymphocytic leukemia (CLL) is an age-related malignancy of mature B lymphocytes.[1] While the diagnosis of CLL is relatively straightforward, the clinical course and outcome are highly heterogeneous.[2] Moreover, despite remarkable therapeutic advances achieved in recent years, the disease is mostly incurable.

ERIC, the European Research Initiative on CLL (http://www.ericll.org) is a Scientific Working Group (SWG) of the European Hematology Association (EHA) aimed at improved management of CLL through collaborative research. Thanks to the active participation of its members, now exceeding 1300 from all over Europe and beyond, ERIC engages in projects extending from basic to (mainly) translational and clinical research.

Capitalizing on such initiatives but also on our expertise in the collection, management and analysis of heterogeneous clinical and biological data,[3–5] we have developed and present here the ERIC CLL database, a registry of clinical and biological data of patients with CLL.

Challenges of gathering high-quality real-world data

Collection and analysis of real world data (RWD) can prove both effective and efficient for advancing precision medicine and improving the quality and delivery of medical care, provided these come along with data quality.[6,7] The amount of biomedical data continuously increases due to technological advances, thus raising the necessity for designing and developing standardized approaches and methodologies to be implemented in clinical practice.[8]

Data acquisition is usually a process distributed among different health professionals potentially leading to data quality problems across datasets, such as data redundancy (ie, repeated information), heterogeneity (eg, different date format) and inconsistency (eg, a date of diagnosis after the date of treatment), mainly resulting from lack of standardization and data curation processes. Such problems are particularly pertinent in the case of multi-institutional efforts, where multilevel and multi-originated data are collected. Furthermore, the rapid increase of data complexity captured during patient care, especially data produced by the application of novel methodologies (eg, next generation sequencing), poses challenges that cannot be addressed with standard computational approaches.

Thus, there is an imperative to improve real-world evidence generation by optimizing the integration of the heterogeneous information through automated and thorough quality control and curation mechanisms; and, support analysis and compatibility with established ontologies. This will provide unified and standardized access to valid, accurate and comparable datasets. Practical and feasible tools are required, capable of providing easiness in use, flexibility and simplicity, in order to facilitate the data entry procedure and encourage the registration and organization of clinically relevant data from the daily practice.[9]

Towards the development of a unified data management framework

Harmonization of heterogeneous data is a prerequisite for gathering homogenized high-quality datasets and bridging the many forms of biological and medical information.

A common approach that can be adapted to local and project-specific requirements, will inevitably facilitate biological, trans-

[1] Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece
[2] Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden
[3] Hematology Departments and Hematopoietic Cells Transplantation Unit, G. Papanikolaou Hospital, Thessaloniki, Greece
[4] Hospital de la Santa Creu i Sant Pau, Autonomous University of Barcelona, Barcelona, Spain
[5] Division of Experimental Oncology, Università Vita-Salute San Raffaele and IRCCS Ospedale San Raffaele, Milan, Italy.

The authors have indicated they have no potential conflicts of interest to disclose.

The development of ERICdb is funded partially by an unrestricted grant from AbbVie and the European Initiative on CLL (ERIC).

HemaSphere (2020) 4:5(e425). http://dx.doi.org/10.1097/HS9.0000000000000425.
Received: 27 April 2020 / Accepted: 26 May 2020

lational and clinical research, enabling multi-center projects on, for example, clinical association studies and translational medicine at large.[10]

Standardization

Agreement on common policies along with a user-friendly integration of ontologies, terminologies and standards, paves the way for standardized registration of RWD; thus, the collected data can be seamlessly combined into a data integration framework, achieving semantic interoperability.

Retrospective data integration

Development and use of reliable "Extract Transform Load" (ETL) software enables and facilitates the massive import of datasets from external, diverse sources into a centralized repository.[11]

Quality assessment

The development of a semantic framework with data cleaning processes[12] capable of identifying quality problems in the collected data, is a prerequisite in order to minimize barriers to data sharing, availability and reusability for research purposes.

A standard-based cleaning, organization and integration approach can provide relevance and data accuracy while guaranteeing long-term usability of the collected high-quality data.[13]

The definition of rules for syntactical and semantic errors, out of range values, missing data, unique value and functional dependency violations, is a well-recognized objective of any strategy aiming at improving quality control.

Building an integrated research infrastructure

Organizing clinical and translational RWD in a standardized and centralized data repository allows unified access for research purposes, improving research efficiency and quality on multiple levels. That said, such integrated approaches must ensure data security and privacy, in particular when aiming for multi-institutional, transnational efforts. In this context, the preferred systems for data collection and retrieval are web-based, with remote data entry, consisting of custom project-related electronic case report forms with simple, user-friendly interfaces to facilitate efficient data registration.

Moreover, data collection, management and sharing must be conducted with standardized procedures ensuring compliance and adherence to ethical, regulatory and legal standards and preventing unauthorized access and unintended disclosure. Personal data protection and conforming to EU regulations, at least for member states of the European Union (EU),[14] and the principles of beneficence, non-maleficence, respect for autonomy and confidentiality must be guaranteed through the application of anonymization methods in the captured data[15] and the development of access control and activity monitoring mechanisms.

The ERIC CLL database

The ERIC CLL database is a data management system that supports research and medical knowledge discovery in CLL implementing the aforementioned methodologies. The database is designed in expandable modules that allow the rapid introduction of additional categories and values when and if needed, based on the type of project run at any single moment in time, making it flexible and adjustable. All new information eventually ends up in the central dataset as stable asset of the database. Currently, the ERIC CLL database includes data from 9147 cases coming from 19 centres in 10 different countries.

The main objectives of the ERIC CLL database are to (1) collect and transform clinically relevant RWD into evidence and correlate with biological data in order to provide accurate information about the state-of-the-art diagnosis; (2) generate hypotheses regarding important disease characteristics, laboratory studies and therapies; and, (3) define relevant parameters influencing CLL impact on health systems.

The data categories composing all the relevant information to form an accurate and complete representation of the disease course include basic demographic data, disease-related information, treatment options and response, laboratory results and relevant outcome data. The data model has been designed in order to meet the requirements for an accurate description of diagnosis, prognostic assessment and management of patients with CLL.

A relational database developed in PostgreSQL has been designed in a way to provide data integrity and promote data correlations and statistical analysis. To the benefit of the scientific community at large, open-source tools have been used for the development of the ERIC CLL database. A standards-based approach is used to determine efficacy of data registration and integration, providing useful, accurate and valid information, increasing the availability of clinically relevant structured data and fulfilling the data quality assurance requirements.

A web-based user interface has been developed for prospective data collection as part of patient's routine care, designed to ensure data protection, security and availability. The interface allows for controlled database login, real-time registration with data validation mechanisms, data retrieval and management.

Moreover, a retrospective data registration and import tool has been developed in order to efficiently and effectively load into the database retrospective patient data, collected in purpose-specific template registration spreadsheets. The tool deploys data cleaning processes based on certain rules ensuring content validation and detecting data inconsistency and redundancy errors. A mapping mechanim is then applied using transformation rules to convert data to predefined types and import them into the database in the appropriate form. Accordingly, the tool can be configured and applied to transform exports of data coming from different sources, regardless the diversity of the software currently utilized in each single institution (ie, different databases), thus enabling interoperability.

Personal identifying data are not requested or stored in the ERIC CLL database. Concerning the collection of retrospective data, anonymization of data takes place during the registration and validation processes; anonymized datasets are then saved and imported into the ERIC CLL database, conforming to EU regulations. Moreover, a user management system has been developed and configured for authentication and authorization of users to ensure data confidentiality and privacy, providing efficient and secure handling and exchange of information. Center-based, lab-based and role-based privileges are defined to restrain access, control, monitor and facilitate data management procedures, according to general and local requirements. Data is stored in a secure dedicated database server, controlled by an ISMS which is ISO27001-2013 certified and abides to GDPR.

The infrastructure includes system failover mechanisms, backup processes to prevent data loss and history management mechanisms providing information about data modifications.

Concluding remarks

The uniqueness of CLL in terms of clinico-biological heterogeneity and rapidly evolving therapeutic paradigms underlines the need for large-scale collaboration and multidisciplinarity aimed towards the realization of precision medicine in CLL. This essentially requires refined understanding of CLL at the fundamental, pathophysiological level as well as integration of multiple layers and sources of biological data with information about disease trajectories and outcomes. The ERIC CLL database is a concrete step in this direction, tailored to user needs and aspiring to contribute to improved management of patients with CLL.

References

1. Kipps TJ, Stevenson FK, Wu CJ, et al. Chronic lymphocytic leukaemia. *Nat Rev Dis Primers.* 2017;3:17008.
2. Hallek M, Cheson BD, Catovsky D, et al. iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. *Blood.* 2018;131:2745–2760.
3. Baliakas P, Hadzidimitriou A, Sutton LA, et al. Clinical effect of stereotyped B-cell receptor immunoglobulins in chronic lymphocytic leukaemia: a retrospective multicentre study. *Lancet Haematol.* 2014;1: e74–e84.
4. Baliakas P, Hadzidimitriou A, Sutton LA, et al. Recurrent mutations refine prognosis in chronic lymphocytic leukemia. *Leukemia.* 2015;29:329–336.
5. Herishanu Y, Shaulov A, Fineman R, et al. Frontline treatment with the combination obinutuzumab +/- chlorambucil for chronic lymphocytic leukemia outside clinical trials: Results of a multinational, multicenter study by ERIC and the Israeli CLL study group. *Am J Hematol.* 2020;95:604–611.
6. US Food and Drug Administration. Framework for FDA's real-world evidence program. 2018. https://www.fda.gov/media/120060/download. Accessed September 19, 2020.
7. Personalized Medicine Coalition. Personalized Medicine at FDA: The Scope & Significance of Progress in 2019. http://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/PM_at_FDA_The_Scope_and_Significance_of_Progress_in_2019.pdf. Accessed September 22, 2020.
8. Khozin S, Blumenthal GM, Pazdur R. Real-world data for clinical evidence generation in oncology. *J Natl Cancer Inst.* 2017;109: 10.1093/jnci/djx187.
9. Meystre SM, Lovis C, Burkle T, et al. Clinical data reuse or secondary use: current status and potential future progress. *Yearb Med Inform.* 2017;26:38–52.
10. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018.
11. Meyer J, Ostrzinski S, Fredrich D, et al. Efficient data management in a large-scale epidemiology research project. *Comput Methods Programs Biomed.* 2012;107:425–435.
12. Rahm EDH. Data cleaning: Problems and current approaches. *IEEE Data Eng Bull.* 2000;23:3–13.
13. Noy NF. Semantic integration: a survey of ontology-based approaches. *ACM Sigmod Record.* 2004;33:65.
14. Rumbold JM, Pierscionek B. The effect of the general data protection regulation on medical research. *J Med Internet Res.* 2017;19:e47.
15. Kushida CA, Nichols DA, Jadrnicek R, et al. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med Care.* 2012;50 Suppl:S82–S101.