Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240412

Data Integration for a Global Deep-Phenotyping Registry for Pulmonary Hypertension – Lessons Learned

Meike T. FUENDERICH[a,b,1], Philipp KRIEB[a,b], Werner SEEGER[a,b] and Raphael W. MAJEED[a,b,c]

[a] *Department of Internal Medicine, Universities of Giessen and Marburg Lung Center (UGMLC), Member of the German Center for Lung Research (DZL), Germany*
[b] *Institute for Lung Health, Cardio-Pulmonary Institute, Giessen, Germany*
[c] *Institute of Medical Informatics, University Hospital RWTH Aachen, Aachen, Germany*

Abstract. The integration of data from various healthcare centers into disease registries is pivotal for facilitating collaborative research and enhancing clinical insights. In this study, we investigate the integration process of existing registries into the PVRI GoDeep meta-registry, focusing on the complexities and challenges encountered. We detail the integration process, including data transformation, mapping updates, and feedback mechanisms. Our findings underscore the importance of standardized processes and proactive communication in addressing data quality issues, ultimately enhancing the reliability and trustworthiness of meta-registry data. Through careful harmonization of the data and transparent documentation of data processing, we pave the way for leveraging registry data to drive advancements in pulmonary hypertension research and patient care.

Keywords. Meta-registry, data integration; data mapping; data quality.

1. Introduction

Integrating data from multiple healthcare centers into registries presents challenges but offers vast potential for advancing medical research and improving patient care [1-3]. As data volumes grow and diversify, the need to consolidate this information becomes increasingly relevant. However, this task is complicated by the inherent heterogeneity in healthcare data and the varied methods of data management across diseases, hospitals, and countries [4]. To address these challenges, systematic approaches are required to identify and resolve issues throughout the integration process.

Pulmonary hypertension (PH) is a chronic disease characterized by elevated blood pressure within the arteries of the lungs [5-7]. Given the rarity of specific PH subtypes, future research requires multinational data collections [7]. The Pulmonary Vascular Research Institute (PVRI) GoDeep meta-registry integrates clinical data from existing registries worldwide [8]. Our aim is to systematically analyze the challenges of large-scale multinational data integration within the context of the PVRI GoDeep meta-registry,

[1] Corresponding Author: Meike Fuenderich; E-mail: meike.fuenderich@innere.med.uni-giessen.de.

enhancing its reliability and utility [8]. Collaborating with clinicians, researchers, and data managers, the goal is to accelerate PH research and improve healthcare delivery, despite the hurdles posed by its rarity and challenges in gathering real-world data, thereby paving the way for others taking the challenge of building meta-registries.

2. Methods

This section outlines the process of integrating data from existing PH registries into the PVRI GoDeep meta-registry, as shown in Figure 1, and the identification and resolution of issues encountered. Our objective is to establish a comprehensive repository of PH patient data to facilitate robust research and improve clinical care. Registries seeking inclusion must meet specific criteria, including maintaining a local or regional PH registry with at least 100 patients diagnosed via right heart catheterization in accordance with international guidelines [5], ensuring high-quality, standardized data. Before data transfer, an assessment determines whether approval from the ethics committee or Institutional Review Board (IRB) is required; if needed, approval must be obtained beforehand. Since data is anonymized before transfer, existing consent typically suffices, eliminating the need for additional specialized patient consent.

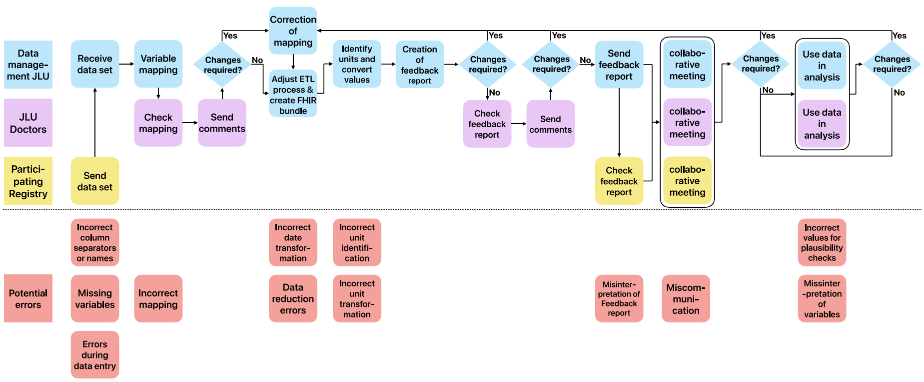

Figure 1. Flowchart illustrating the data integration process, divided into data management, doctors, and the participating registry, with potential errors listed at each step.
JLU = Justus Liebig University Giessen.

2.1. Integration Process

After resolving legal matters, a sample data export is created to verify anonymization and key column, including age at diagnosis, survival status and diagnosis details. Discrepancies are addressed, and a comprehensive data export is transmitted. Variable annotation is maximized prior to data conversion into a standardized format within the ETL (Extract, Transform, Load) process, focusing on time specifications, special characters handling and further necessary processing. Due to the varying data storage methods, the received data may be in various formats: a single Excel file (sometimes with multiple sheets), an Excel file for each patient, multiple files (TSV, Excel, etc.), each representing a different category. During the mapping and ETL process, the file(s) are reordered by category, and column names are renamed to match the registries standard. Next, standardization to international standard terminologies Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) and Logical Observation Identifiers Names and Codes (LOINC) is conducted, followed by transforming the data into a Fast Healthcare Interoperability Resources (FHIR) bundle for interoperability [8].

After identifying variable units, the data points are converted to the registry's standard unit as needed for data comparability. Subsequently, a detailed feedback report is compiled, documenting variables, their units, and any queries or observations arising during the integration process. If errors in the data transformation process are identified within this report, e.g. wrong units and missing data, causes are investigated and clarified before a new report is created. This comprehensive report is shared with the respective centers for transparent communication and feedback. An interactive online meeting is then scheduled to discuss the report and address remaining queries, fostering mutual understanding and alignment among stakeholders for a seamless integration process.

2.2. Identification of Possible Erroneous Data Alterations

Various issues can affect data quality from initial entry in a local registry to integration into the PVRI GoDeep registry. Steps to find errors include verifying patient counts pre- and post-transformation, checking if all mapped variables are present, and the handling of missing values. Temporal consistency checks ensure a logical sequence of events by eliminating inconsistencies and identifying any errors that may have arisen from date conversion. Boxplot analyses are used to compare the data of each variable of the new center with the data of that variable of the entire registry and between centers, highlighting parameters not yet converted to the standard unit and potential misclassifications. The feedback report additional provides the information of the units a variable is reported in and the data completeness per variable. This report, along with any arising questions and issues, is also discussed with the respective center. Plausibility checks are conducted before every analysis, including scrutiny for biologically implausible values and ensuring cross-variable consistency. Additionally, irregularities in the expected value distribution during imputation may suggest unit variations.

3. Results

The integration of data from multiple registries into the PVRI GoDeep meta-registry unveiled diverse challenges impacting data quality and reliability.

Category A. Data entry and export: Problems including column shifts, where data points were moved to neighboring columns due to empty fields, transfer difficulties that led to data loss in certain rows and incorrect column labels introducing ambiguity. Inaccurate date entries, including birth and death dates, as well as input errors, collectively affected the precision and reliability of the datasets. In one occasion, poorly managed delimiters distorted a dataset and needed to be retransferred.

Category B. ETL process: Particularly challenging was the different anonymization of dates, discrepancies in birth and death date calculation, and double mapping, requiring meticulous attention for accurate data transformation. Unique anonymization possibilities for dates often prevented reuse of existing solutions, posing errors, as did special character conversion, e.g. for laboratory variables.

Category C. Data mapping: A significant issue arose regarding the precision of column names, stemming from the ambiguity caused by various linguistic context abbreviations or instances where a single abbreviation represents multiple variables. Utilizing boxplots to compare values of previously verified data for that variable assisted in addressing this challenge. Typographical errors in codes or units hinder accurate data assignment, while dataset updates may overhaul factor levels, as well as altering the label for missing values or separator usage, introducing discrepancies.

Category D. Language and communications: Differences in handling special characters, umlauts and syntax across diverse operating systems caused compatibility issues during data processing. Communication challenges with registries, such as delays or language barriers, complicated information exchange. Unclear column designations led to misunderstandings, especially when registries lacked precise data knowledge. Despite pre-analysis error correction, various errors persisted, including impossible values, inaccuracies, and missing data. Plausibility checks specific to PH and thorough data examination are crucial before every statistical analysis. Notably, imputation control revealed instances where standardizing values to the same unit for a variable was overlooked, highlighting the importance of careful data management.

4. Discussion

Integrating data from various PH registries into the PVRI GoDeep meta-registry presents complex challenges, spanning from data export to final analysis for publication. These challenges have been grouped into four main categories: problems beyond our direct control, issues within the ETL process, data mapping challenges, and other irregularities.

Problems beyond our direct control, such as data entry and data export issues, can only be detected and minimized by careful analysis of the data and attempts to identify irregularities, collect them, and report them back to the registries.

The harmonization of the ETL process effectively reduced careless errors; yet, there is still potential for errors, particularly during date transformation. When using functions from programming language packages, it is crucial to critically assess their impact on other functions while ensuring the desired result. In this phase, checklist-based verification and validation has also proven to be effective to minimize unwanted changes. Initially, unit transformation was included in the ETL process. By converting the values to their standard units after merging all centers, the risk of incorrect conversions was reduced. However, this shift now requires careful attention to unit spelling accuracy.

Separating the mapping from the ETL pipeline and employing an easily updatable CSV mapping table streamlined the process, fostering a faster and easier feedback loop between data managers and registries. As a result, it created a more uniform structure for files post-ETL pipeline and enabled automated creation of the data source template.

While language discrepancies posed minor challenges, the use of varying abbreviations employed for identical variables, along with the utilization of identical abbreviations for multiple variables, will continue to complicate mapping in some cases. Proactive communication and collaboration with registries remain pivotal.

Most future registries will have unique characteristics needing tailored solutions in variable mapping or the ETL process. Despite this, the implemented processes alleviate these challenges and will continue to do so. Regularly checking the raw data, ETL process, and mapping, ensures data integrity, enabling reliable analysis. While this may be time-consuming, it reduces the risk of late-stage errors, which can be resource-intensive and have cumulative effects.

Moving forward, developing a detailed provenance logging framework would enhance data traceability and accountability by carefully tracking and recording changes to each data point, improving transparency and reliability, and reducing the time and effort required to identify issues during analysis. Despite these advancements, data entry errors are almost inevitable, thus, the application of basic plausibility checks remains indispensable before starting any data analysis. Furthermore, regular meetings with clinicians and data managers are essential for maintaining good data quality.

5. Conclusions

In conclusion, integrating data from multiple centers into the PVRI GoDeep meta-registry presents numerous challenges, from data entry and export to complexities within the ETL process and mapping. The systematic approach to identify and resolve data quality issues enables effective navigating through the integration process, potentially benefiting other meta-registries. Communication, collaboration, and the implementation of mitigation strategies have enhanced data reliability. Optimizing processes and standards has minimized errors, resulting in greater integration accuracy.

Central to the success of any registry is tracking and documenting data changes for faster error detection, ensuring transparency and accountability in data management. Despite the challenges and time involved in creating an international registry of data from different countries and centers, it serves as indispensable repository that provides access to expansive and diverse datasets covering a spectrum of patient demographics and clinical scenarios [3]. This diversity empowers researchers to explore multifaceted factors influencing disease progression and treatment efficacy, facilitating tailored interventions to meet individual patient needs.

References

[1] Torbicki A, Bacchi M, Delcroix M, Farber HW, Ghofrani H-A, Hennessy B, et al. Integrating data from randomized controlled trials and observational studies to assess survival in rare diseases. Circ Cardiovasc Qual Outcomes. 2019;12(5):e005095.

[2] Gagne JJ, Thompson L, O'Keefe K, Kesselheim AS. Innovative research methods for studying treatments for rare diseases: methodological review. BMJ. 2014;349:g6802.

[3] Richesson R, Vehik, K. Patient registries: utility, validity and inference. Rare diseases epidemiology (2010): 87-104. DOI 10.1007/978-90-481-9485-8_6

[4] Gohar A, AbdelGaber S, Salah M. A patient-centric healthcare framework reference architecture for better semantic interoperability based on blockchain, cloud, and iot. Ieee Access 2022;10:92137-92157. https://doi.org/10.1109/access.2022.3202902

[5] Simonneau G, Montani D, Celermajer DS, Denton CP, Gatzoulis MA, Krowka M, et al. Haemodynamic definitions and updated clinical classification of pulmonary hypertension. European respiratory journal 53.1 (2019).

[6] Naeije R, Richter MJ, Rubin LJ. The physiologic basis of pulmonary arterial hypertension. Eur Respir J. 2022;59(6):2102334

[7] Weatherald J, Reis A, Sitbon O, Humbert M. Pulmonary arterial hypertension registries: past, present and into the future. European Respiratory Review 28.154 (2019).

[8] Majeed R., Wilkins M, Howard L, Hassoun P, Anthi A, Cajigas H, et al. Pulmonary Vascular Research Institute GoDeep: A meta-registry merging deep phenotyping data from international PH reference centers. Pulmonary Circulation 2022;12(3). https://doi.org/10.1002/pul2.12123