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

doi:10.3233/SHTI240613

Cohort Builder: A Software Pipeline for Generating Patient Cohorts with Predetermined Baseline Characteristics from Medical Records and Raw Ophthalmic Imaging Data

Sepehr MOUSAVI[a,b,*], Ali GARJANI[a,b,*], Adham ELWAKIL[a,b,*], Laurent Pierre BROCK[a,b], Alexandre DHERSE[a,b], Edwige FORESTIER[a], Marine PALAZ[a], Emilien SEILER[a,b], Alexia DURIEZ[a,b], Thibaud MARTIN[a,b], Thomas WOLFENSBERGER[a], Reinier SCHLINGEMANN[a,c], Ilenia MELONI[a,b,†], Ciara BERGIN[a,†] and Mattia TOMASONI[1,a,b,†]

[a] *Department of Ophthalmology, University of Lausanne, Fondation Asile des Aveugles, Jules Gonin Eye Hospital, Lausanne, Switzerland*

[b] *Platform for Research in Ocular Imaging, Fondation Asile des Aveugles, Jules Gonin Eye Hospital, Lausanne, Switzerland*

[c] *Department of Ophthalmology, Amsterdam University Medical Centres, Amsterdam, The Netherlands*

Mattia TOMASONI https://orcid.org/0000-0001-8775-2384

Abstract. In clinical research, the analysis of patient cohorts is a widely employed method for investigating relevant healthcare questions. The ability to automatically extract large-scale patient cohorts from hospital systems is vital in order to unlock the potential of real-world clinical data, and answer pivotal medical questions through retrospective research studies. However, existing medical data is often dispersed across various systems and databases, preventing a systematic approach to access and interoperability. Even when the data are readily accessible, clinical researchers need to sift through Electronic Medical Records, confirm ethical approval, verify status of patient consent, check the availability of imaging data, and filter the data based on disease-specific image biomarkers. We present Cohort Builder, a software pipeline designed to facilitate the creation of patient cohorts with predefined baseline characteristics from real-world ophthalmic imaging data and electronic medical records. The applicability of our approach extends beyond ophthalmology to other medical domains with similar requirements such as neurology, cardiology and orthopedics.

Keywords. Clinical Research, Data Pipeline, Biomedical Information Retrieval, Ophthalmology, Real-World Data.

---

[1] Corresponding Author: M Tomasoni; E-mail: mattia.tomasoni@fa2.ch.

1. Introduction

The advent of artificial intelligence (AI) and machine learning (ML) technologies heralds a new era in healthcare, offering unprecedented opportunities for advancements in diagnostics and patient care [1–3]. In particular, specialties that utilize image-based diagnostics, such as ophthalmology, have seen significant benefits from the integration of AI for disease detection, medical imaging analysis, and predictive health outcomes [4–12]. The capabilities of AI to support early disease detection, enhance the precision of medical image interpretations, disease prediction and evolution have been widely recognized [13]. However, the practical application of AI in clinical practice is contingent upon the availability of extensive, well-organized datasets [14,15]. Existing literature acknowledges the arduous but critical steps required to prepare medical imaging data for AI analysis, emphasizing the need for ethical approvals, data anonymization, quality assurance, and structured data storage to support AI training effectively [16,17]. Nevertheless, there exists a discernible gap in research regarding the methodologies for consolidating disparate data sources for medical imaging AI applications.

2. Methods

Cohort Builder is a software pipeline designed to facilitate the creation of patient cohorts from real-world ophthalmic imaging data. Image Management System: We used the Discovery® software by RetinAI as an Image Management System (IMS) and Image Viewer. It can automatically segment and extract biomarkers from medical image acquisitions using AI [18,19]. It also serves as a tool to perform automatic medical image segmentation, which allows monitoring of disease progression.

3. Results

Cohort Builder is composed of three main modules: Cohort Planner, Cohort Extractor and Cohort Labeler (as shown in Figure 1). Integration of these subcomponents enables clinical researchers to efficiently extract and label patient data for research purposes. An instance of Cohort Builder has been deployed on the servers of the Swiss Ophthalmic Imaging Network [20] and it is available to researchers and clinicians at partner institutions.

3.1. The Cohort Planner Module

This module allows researchers with a specific clinical question to estimate the number of potential patients available for analysis based on specified baseline characteristics (such as age, gender, or disease). Patient consent for data usage is verified by querying the patient consent database, ensuring compliance with privacy regulations and legal provisions on research involving human subjects. By estimating sample size and providing support for power calculations, it assists in planning the subsequent cohort extraction phase.

3.2. The Cohort Extractor Module

The main module, called Cohort Extractor, streamlines the process of cohort assembly. It automatically "uploads" raw images to an Image Management System (IMS). It then performs an AI-assisted extraction of retinal biomarkers and "downloads" the data.

3.3. The Cohort Labeler Module

Designed to facilitate expert labelling of the extracted patient cohort, this software module enables clinical experts to systematically view the extracted data and assign a label via an interactive Graphical User Interface. The output can be used to train AI algorithms and answer the original clinical question. Cohort Labeler enables the visualization of selected scans in the patient history, and displays the segmentation of relevant ocular structures. It shows the distribution of pathological retinal fluids via histogram overlays and it allows the recording of annotator's notes.

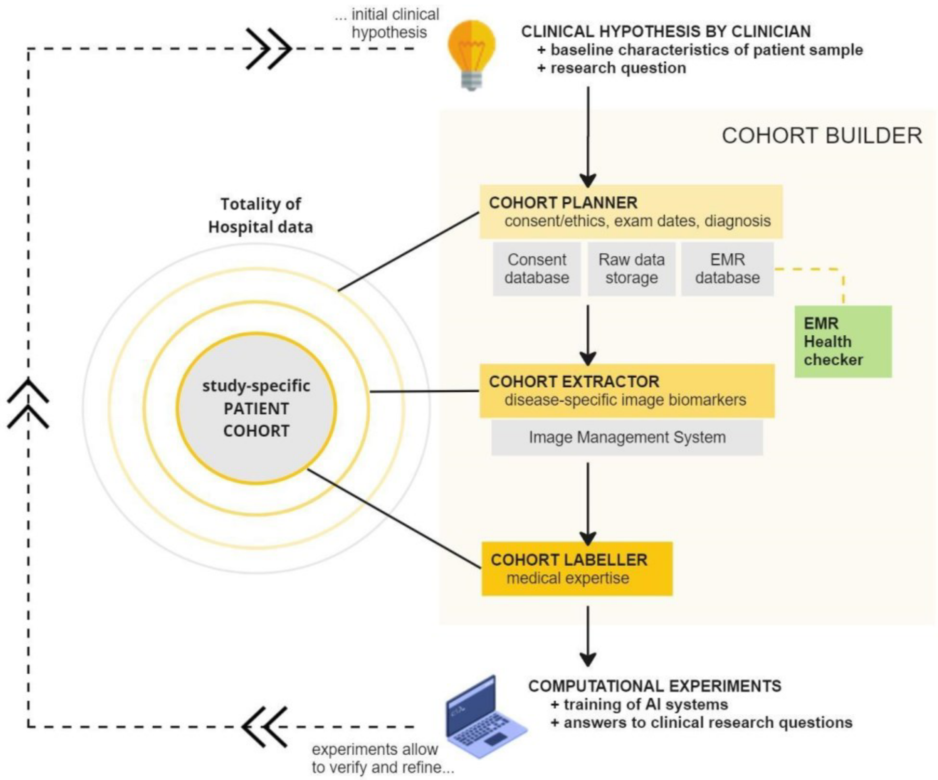

Figure 1. Overview of the Cohort Builder software pipeline. The pipeline comprises three main modules: Cohort Planner, Cohort Builder, and Cohort Labeler. Cohort Planner assists clinicians in estimating potential patient numbers and planning data extraction. Cohort Builder automates the extraction of retinal biomarkers, streamlining cohort assembly, while Cohort Labeler facilitates expert labelling of patient datasets for AI algorithm training and potential future studies.

3.4. The EMR Health Checker Module

This software module, separated from the rest of the pipeline, analyses the Electronic Medical Record (EMR) database to provide indicators of the completeness of certain fields, such as the diagnosis status. EMR Health Checker offers an estimation of EMR hygiene, necessary for the correct identification of subjects with certain baseline characteristics, which is a prerogative of Cohort Builder.

3.5. Use Cases and Code availability

The Cohort Builder pipeline has played a crucial role in the creation of patient cohorts with specific baseline characteristics for a range of ophthalmology projects, ranging from grading of ocular inflammation [22] to ocular genomics [23–25]. More detailed information is available on the website of the Swiss Ophthalmic Imaging Network (SOIN): https://sphn.ch/network/projects/completed-projects_tiles/project-page_soin. The Cohort Builder pipeline software, to which access is granted upon request, is available at https://github.com/JulesGoninRIO/cohortbuilder.

4. Discussion

Our methodology and our open-source pipeline hold the potential to serve as a strong foundational implementation for other institutions, impacting clinical research on large-scale retrospective studies. Furthermore, the adoption of the innovative infrastructure developed in this project holds promise for addressing prevalent challenges across various healthcare settings, beyond ophthalmology.

5. Conclusions

Acknowledging the importance of patient cohorts for addressing clinical research questions in everyday medical practice, we propose a modular software solution to address the fragmentation of patient data across disparate systems and the lack of a systematic approach to data access and interoperability. Cohort Builder is a software pipeline that ensures effective utilization of real-world data. It is designed to streamline the creation of patient cohorts integrating information on consent, diagnoses from EMRs, and AI-based disease-critical biomarkers. Jointly, Cohort Planner, Cohort Extractor, and Cohort Labeler automate data preparation and processing and improves the efficiency and accuracy of patient cohort creation in clinical research settings. Our approach holds the potential to serve as a strong foundational implementation for other institutions, impacting clinical research on large-scale beyond ophthalmology.

Acknowledgements

This work was supported by the Swiss Personalized Health Network (2018DRI13 to Thomas J. Wolfensberger). This work was also supported by the Claire et Selma Kattenburg Foundation.

References

[1] Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69S: S36–S40.
[2] Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak. 2021;21: 125.
[3] Bohr A, Memarzadeh K. Chapter 2 - The rise of artificial intelligence in healthcare applications. In: Artificial Intelligence in Healthcare. Academic Press; 2020. pp. 25–60.
[4] Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019;103: 167–175.
[5] Badar M, Haris M, Fatima A. Application of deep learning for retinal image analysis: A review. Computer Science Review. 2020;35: 100203.
[6] Dai L, Wu L, Li H, Cai C, Wu Q, Kong H, et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat Commun. 2021;12: 3242.
[7] Liu X, Ali TK, Singh P, Shah A, McKinney SM, Ruamviboonsuk P, et al. Deep Learning to Detect OCT-derived Diabetic Macular Edema from Color Retinal Photographs: A Multicenter Validation Study. Ophthalmol Retina. 2022;6: 398–410.
[8] Potapenko I, Thiesson B, Kristensen M, Hajari JN, Ilginis T, Fuchs J, et al. Automated artificial intelligence-based system for clinical follow-up of patients with age-related macular degeneration. Acta Ophthalmol. 2022;100: 927–936.
[9] Ran AR, Cheung CY, Wang X, Chen H, Luo L-Y, Chan PP, et al. Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep-learning analysis. Lancet Digit Health. 2019;1: e172–e182.
[10] Xiong J, et al. Multimodal Machine Learning Using Visual Fields and Peripapillary Circular OCT Scans in Detection of Glaucomatous Optic Neuropathy. Ophthalmology. 2022;129: 171–180.
[11] Yellapragada B, Hornauer S, Snyder K, Yu S, Yiu G. Self-Supervised Feature Learning and Phenotyping for Assessing Age-Related Macular Degeneration Using Retinal Fundus Images. Ophthalmol Retina. 2022;6: 116–129.
[12] Yim J, Chopra R, Spitz T, Winkens J, Obika A, Kelly C, et al. Predicting conversion to wet age-related macular degeneration using deep learning. Nat Med. 2020;26: 892–899.
[13] Al Kuwaiti A, Nazer K, Al-Reedy A, Al-Shehri S, Al-Muhanna A, Subbarayalu AV, et al. A Review of the Role of Artificial Intelligence in Healthcare. J Pers Med. 2023;13. doi:10.3390/jpm13060951
[14] Kahn CE Jr, Carrino JA, Flynn MJ, Peck DJ, Horii SC. DICOM and radiology: past, present, and future. J Am Coll Radiol. 2007;4: 652–657.
[15] Strickland NH. PACS (picture archiving and communication systems): filmless radiology. Arch Dis Child. 2000;83: 82–86.
[16] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing Medical Imaging Data for Machine Learning. Radiology. 2020;295: 4–15.
[17] Diaz O, Kushibar K, Osuala R, Linardos A et al. Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. Phys Med. 2021;83: 25–37.
[18] Bogunovic H, et al. RETOUCH: The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge. IEEE Trans Med Imaging. 2019;38: 1858–1874.
[19] De Zanet S, Ciller C, Wolf S, Sznitman R. Pathological OCT Retinal Layer Segmentation Using Branch Residual U-Shape Networks. Medical Image Computing and Computer Assisted Intervention − MICCAI 2017. 2017; 294–301.
[20] Bergin et al. SOIN / MI Data Lab: Personalized ophthalmology through collaborative data collection and dynamic patient consent. proceedings of MIE2022. 2022.
[21] Lawrence AK, Selter L, Frey U. SPHN - The Swiss Personalized Health Network Initiative. Stud Health Technol Inform. 2020;270: 1156–1160.
[22] Amiot V, Jimenez-Del-Toro O, Eyraud P, Guex-Crosier Y, Bergin C, Anjos A, et al. Fully automatic grading of retinal vasculitis on fluorescein angiography time-lapse from real-world data in clinical settings. 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2023. doi:10.1109/cbms58004.2023.00301
[23] Bergmann S, Vela S, Beyeler M, Trofimova O, Tomasoni M, Iuliani I, et al. Phenotypic and Genetic Characteristics of Retinal Vascular Parameters and their Association with Diseases. 2023 [preprint]. doi:10.21203/rs.3.rs-3413660/v1
[24] Tomasoni et al. Genome-wide Association Studies of Retinal Vessel Tortuosity Identify Numerous Novel Loci Revealing Genes and Pathways Associated With Ocular and Cardiometabolic Diseases. Ophthalmology Science. 2023;3: 100288.
[25] Milloz A, Molas G, Paychère Y, Bouillon A, Amiot V, Gurtler L, et al. Estimating Quality of OCT Cubes using Phase-Level Unified Metric Evaluation (PLUME-OCT). 2024 [preprint]. doi:10.21203/rs.3.rs-4171462/v1