A Collaborative Analytic Enclave for the Metabolic Dysregulation and Obesity Cancer Risk Program (MeDOC) Consortium: Extensions of the OMOP Common Data Model for Translational Research

Authors: Madhan Subramanian[1], Nisha Grover[1], Maddie Wheeler[1] Marinella Temprosa[1]

Affiliation: Biostatistics Center and Dept of Biostatistics and Bioinformatics, George Washington University [1]

Background

The MeDOC (Metabolic Dysregulation and Obesity Cancer Risk Program) Consortium, established in 2022, aims to advance our understanding of the underlying mechanisms that connect obesity, metabolic dysregulation, and cancer risk through individual and collaborative projects. The six-member consortium, sponsored by the National Cancer Institute (NCI), collects data for both human and pre-clinical translational research. This research focuses on the hallmarks of metabolic dysregulation encompassing hormones, microbiome, inflammation, immunity, glycemia, insulinemia, adipokines and lipids. In order to integrate data across all projects and provide an infrastructure to conduct collaborative analysis, we adopted the OMOP CDM[1] and developed extensions to support the diverse data and biospecimens from both published and stored human and animal studies. Advances in biomedical research are galvanized by data-driven discoveries for which data mapping, harmonization, and documentation of study results, using FAIR principles[2], are essential components. Through these initiatives, MeDOC aims to enhance the standardization and interoperability of biomedical data, facilitating more efficient and effective research into the links between obesity, metabolic dysregulation, and cancer risk by creating tools for the open-source community.

Methods

The MeDOC Coordinating Center (GW-CC) is developing an analytic enclave that incorporates the OMOP CDM with three extensions (Figure 1): MeDOC-BioRep, virtually catalogues biospecimens and research data collected from human and animal studies across the Consortium; MeDOC-KB, a knowledge base for documenting associations between metabolic dysregulation and obesity cancer/phenotype diseases; and MeDOC-Miner uses natural language processing (NLP) and large language models (LLM) to conduct bibliometric analyses of internal and external PubMed publications for hypothesis generation and synthesis.

These three extensions will work in conjunction with OMOP's CDM and OHDSI tools allowing MeDOC investigators research metabolically driven cancers through cross-species data harmonization. With OMOP's CDM structure, our extensions provide a unified framework to align human and animal data, which will enhance data sharing and foster novel biomedical insights. As an example, we will describe two collaborative projects (CP) within MeDOC that can leverage this infrastructure. The MICR CP investigates incretin associated weight loss within mice with breast cancer. The METABO-GUARD CP is researching the effects of bariatric surgery within human colorectal cancer patients. Both collaborative projects will be collecting and storing blood and stool samples, tracked via MeDOC-BioRep, and store additional samples for future cross-consortium

analyses. The results from these studies will then be available to the consortium through MeDOC-KB. In turn MeDOC-Miner will work with MeDOC-KB to inform our future cross-Consortium analyses.

MeDOC-BioRep

MeDOC-BioRep serves as a virtual warehouse of human and animal biospecimens collected across the Consortium. Using OMOP's CDM[1] as a backbone, will help align the vocabularies from the different projects. This extension of the CDM was developed to standardize data collected from animal studies (e.g., animal strain, housing type, intervention, diet, etc.). Data collection templates will be used to facilitate the harmonization of the biospecimen vocabulary. Once harmonized, the data will be displayed visually in static and interactive plots using a web application developed with the Posit Shiny package.

MeDOC-KB

MeDOC-KB will use preexisting, publicly available data to document associations in the underlying mechanisms that link obesity, metabolic dysregulation, and increased cancer risk, employing knowledge graphs, with the ultimate goal of creating an infrastructure to catalog emerging data from MeDOC studies. These will be used in conjunction with the CDM to develop a framework for the documentation, exploration, and validation of potential targets for investigation. Initially, we captured information from multiple sources and existing cancer atlases: (1) Nightingale for lipidomics (2) Olink proteomics and (3) PubMed central, then developed user friendly visualization tools in Tableau and Shiny.

MeDOC-Miner

MeDOC-Miner utilizes topic modelling to identify common themes and groupings within the corpus comprised of internal and external PubMed publications using R[3]. CDM Source identifies the source of the publication (PubMed or MeDOC project(s)). All publications were collected in a web-based reference manager folder in RefWorks identified by PubMed ID. Data preprocessing was done by creating a "clean text" function to tokenize words, remove special characters and stopwords, perform stemming/lemmatization, and convert to lowercase. We represented the corpus using the Latent Dirichlet Allocation (LDA) model, which is a probabilistic model that assumes documents are a mixture of topics and topics are a mixture of words. The number of topics for each corpus was found using the ldatuning package. Our LDA model was created with Gibbs Sampling and visualized using the tm, topicmodels, and LDAvis packages[3,4]. The LDA model derives two measures: the rows representing the distribution of topics over documents within a corpus (Theta) and rows representing distribution of words over topics (Phi). Summaries of the topics were derived using LLM on the publications identified to contribute to the topic[5]. MeDOC-Miner includes an R-shiny app to display results and allow for exploration of topics and articles. Using references from the funded projects, we applied MeDOC-Miner to identify themes across the Consortium.

Results

As previously discussed, the MICR CP and METABOGUARD CP are examples within the consortium that would highly benefit from adopting and extending OMOP's CDM. Since both studies will be collecting and storing blood and stool samples, of both mice and human data, the data will be stored in our extension of the CDM, MeDOC-Biorep, to ensure the data can still work with

OMOP's infrastructure (Figure 2). Then the investigator can use our tools (MeDOC-KB and MeDOC-Miner) to look at associations and generate hypotheses. Investigators can also utilize OHDSI's tools such as ATLAS to perform real time analysis on data or use the HADES R packages to perform analysis on observational data.

MeDOC-BioRep

MeDOC-BioRep will be hosted on the MeDOC study website providing investigators with convenient access to browse available biospecimens at the local project-sites. An interactive catalogue allows users to view biospecimens by type or project. Summary data will be presented visually, showcasing trends across time and projects.

MeDOC-KB

Data obtained from existing cancer atlases published by Nightingale and Olink include associations for a lipidomics panel, (249 measures of lipoprotein subclasses, fatty acids, and small molecules), across 13 MeDOC relevant cancers, and a proteomics panel, (1,471 protein assays for 1,462 unique proteins), for 2 MeDOC relevant cancers[6, 7]. The associations harvested were implemented in the CDM extension and visualized using Tableau (Figure 3), which will be available to the Consortium as an R-shiny app, and analyzed using knowledge graph analytics.

MeDOC-Miner

Utilizing LDA, we found the relevant topics that that were studied at each MeDOC project site. For example, the REMEDY project has 10 topics within their corpus. The LDAvis function which shows each topic within a principal component grid and the top 30 terms based on frequency within each topic. The Theta and Phi values were also calculated. A summary describing each topic was produced via an LLM (Figure 3).

Conclusion

To ensure standardization and interoperability within MeDOC, we extended the OMOP CDM vocabulary to include animal studies (MeDOC-BioRep) and disease-target associations (MeDOC-KB). For future development, we will use MeDOC-Miner in conjunction with the MeDOC-KB to mine PubMed Texts and use LLM for summarization. Successful utilization of Topic Modelling not only identifies current research topics within MeDOC but also guides future research direction, enhancing the impact of MeDOC. The MeDOC-BioRep is expected to serve as a valuable resource by facilitating collaborative projects and harmonizing prospective data collections. These initiatives aim to establish a collaborative analytic enclave, guiding MeDOC in addressing unmet scientific priorities related to metabolic dysregulation and obesity cancer risk.

Figures

Figure 1. Three initiatives to enhance the standardization and interoperability of MeDOC data: 1) MeDOC-BioRep will add an Animal Table underneath the Standardized Clinical Data portion of the CDM; 2) MeDOC-KB adds an Association Table, that is mapped from the Standardized Vocabularies Table via the Concept_Relationship table, an Omics_Experiment Table, and Target Table which both map to the Association Table; 3) MeDOC-Miner will extend the Standardized Meta-data by adding a Citation Table that is mapped from the CDM_Source Table and maps to the Omics_Experiment table.

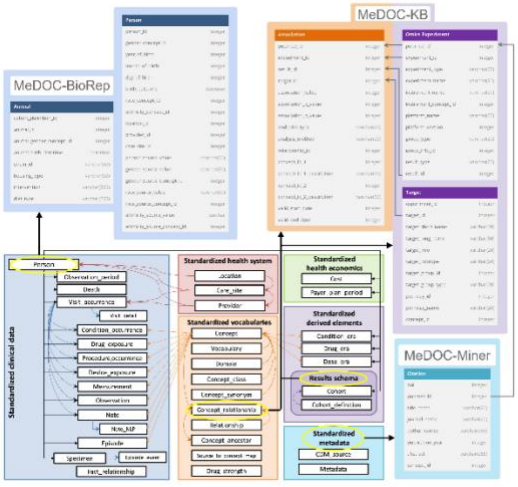


Figure 2. This figure shows the flow of data within our extension of OMOP's CDM. This set up will allow MeDOC investigators to use the tools we are developing to understand emerging data and for hypothesis generation, however, they will also be able to access OHDSI tools such as ATLAS or HADES for further analysis of their data.

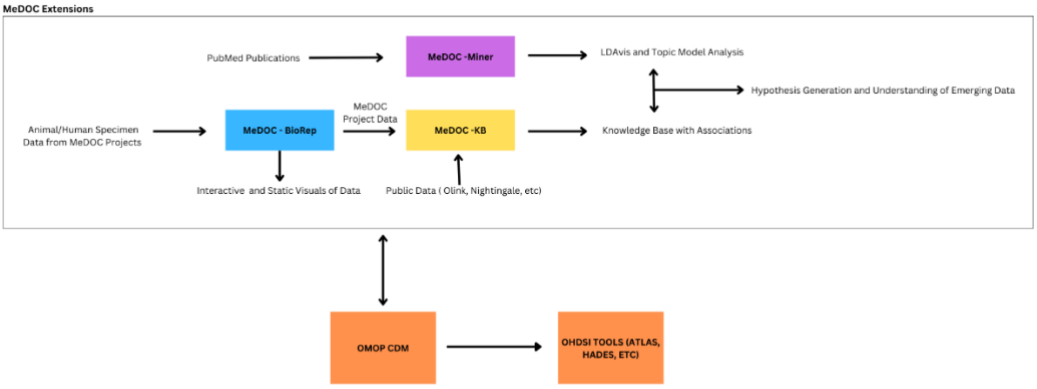
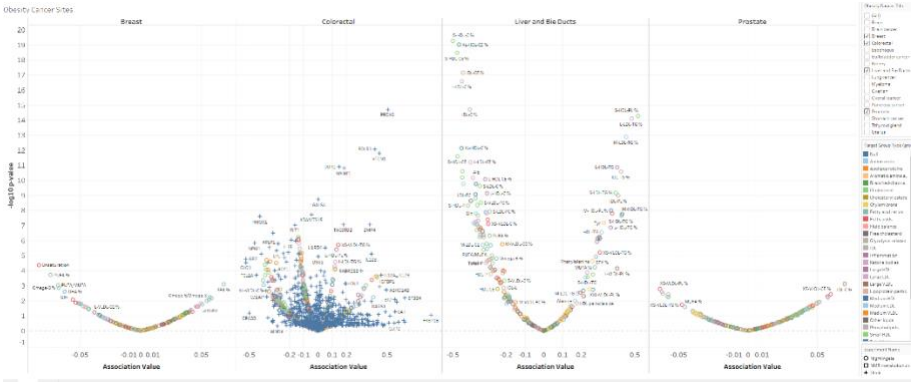
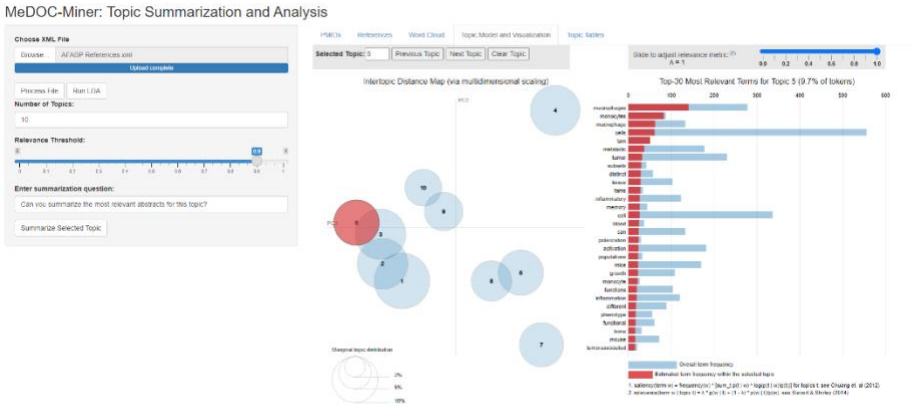


Figure 3. This Figure plotted association values of Breast, Colorectal, Liver, and Prostate cancers that were collected from publicly available datasets such as Nightingale and Olink. We also see the MeDOC- Miner Shiny Application: The image depicts how the user can upload their XML file with Publication information, run the LDA analysis based on their Corpus, and receive and output of a LDA visualization, summary of the topic via Clinfo.AI/OpenAI, and the top documents that contribute to the Specific Topic.





References

1. OHDSI. The Book of OHDSI: Observational Health Data Sciences and Informatics. OHDSI; 2019. 458 p. ISBN: 1088855199, 9781088855195.
2. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016 Mar 15;3(1):1-9.
3. Silge J, Robinson D. Text mining with R: A tidy approach. " O'Reilly Media, Inc."; 2017 Jun 12.
4. Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. InProceedings of the workshop on interactive language learning, visualization, and interfaces 2014 Jun (pp. 63-70).
5. Cintron DW, Montrosse-Moorhead B. Integrating big data into evaluation: R code for topic identification and modeling. American Journal of Evaluation. 2022 Sep;43(3):412-36.
6. Julkunen H, Cichońska A, Tiainen M, et al. Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK biobank. Nature communications. 2023;14(1):604. https://www.ncbi.nlm.nih.gov/pubmed/36737450. doi: 10.1038/s41467-023-36231-7. Updated 2024. Accessed March 5, 2024.
7. Álvez MB, Edfors F, Von Feilitzen K, et al. Next generation pan-cancer blood proteome profiling using proximity extension assay. Nat Commun. 2023;14(1). doi: 10.1038/s41467-023-39765-y.