RESEARCH PROTOCOL:
Creation and Validation of Dermatomyositis Algorithms through the Observational Health Data Sciences and Informatics (OHDSI) Community

Table of Contents

1. Responsible Parties 2

1.1. Investigators and Authors 2

1.2 Sponsor 2

2. Abstract 2

3. Rationale and Background 3

4. Objective 3

5. Methods 3

5.1 Data Sources 3

5.2 Study design 4

5.3 Target cohorts 4

5.4 Evaluation of algorithms 5

5.5 Phenotypes of interest 5

5.6 Characterizing Cohorts 6

5.7 Logistics of Executing a Federated Analysis 6

6. Strengths and Limitations 6

6.1 Strengths 6

6.2 Limitations 6

7. Protection of Human Subjects 7

8. Plans for Disseminating and Communicating Study Results 7

1. Responsible Parties

1.1. Investigators and Authors

Investigator/Author	Institution/Affiliation
Christopher A Mecoli, MD	Johns Hopkins University School of Medicine, Baltimore, MD, USA
Will Kelly, BS	Johns Hopkins University School of Medicine, Baltimore, MD, USA
Ben Martin, PhD	Johns Hopkins University School of Medicine, Baltimore, MD, USA
Sean Yen, MD MS	Johns Hopkins University School of Medicine, Baltimore, MD, USA
Gowtham Rao, MD	Janssen Research & Development
Azza Shoaibi, PhD	Janssen Research & Development
Joel Swerdel, PhD	Janssen Research & Development
**TBD	Columbia University, New York, NY USA
**TBD	Columbia University, New York, NY USA
**TBD	Columbia University, New York, NY USA
**TBD	Stanford University, CA, USA
**TBD	Stanford University, CA, USA
**TBD	Stanford University, CA, USA

Authorship will also include those who meaningfully contribute to study design, analysis and interpretation of results and subsequently contribute to the drafting of the work for publication, approving the final version of the study. Full guidance related to how to qualify for meaningful contribution can be found on the OHDSI website: https://www.ohdsi.org/wp-content/uploads/2021/07/OHDSI-Authorship-Guidance.pdf. The Responsible Parties involved in this protocol take accountability for the overarching protocol, package development, providing assistance to sites running the analysis and ensuring site-specific governance is adhered to in all publications generated from this protocol.

1.2 Sponsor

This study will be undertaken by Observational Health Data Science and Informatics (OHDSI), an open collaboration.

Participating data owners will be responsible for self-reporting any grants funding the conversion and maintenance of their OMOP CDM. Disclosures will be reported in accordance with publication policies of journals papers are submitted to.

2. Abstract

In this study we will evaluate several DM phenotypes (a.k.a. 'programmatic algorithms') across the OHDSI community and employ PheValuator to determine site-specific performance of each phenotype.

3. Rationale and Background

Dermatomyositis (DM) is a rare chronic autoimmune disease characterized by inflammation and weakness in the muscles and skin rash. It can affect multiple organ systems, leading to substantial morbidity and mortality. The incidence of DM varies, with estimates ranging from 1 to 10 cases per 100,000 person-years. Due to its rarity, clinical research studies on DM often have inadequate power for causal inference, particularly when conducted within single academic centers. One potential solution is to leverage real-world data such as electronic health record (EHR) data, insurance claims data, and national healthcare system registries to study DM. These data sources provide longitudinal datasets with large numbers of cases, offering great potential for studying the disease.

Historically, the evaluation of DM algorithms for identifying patients has been restricted to a single data source, typically the one the researchers are affiliated with or have access to. This focus on a single data source diminishes the generalizability of the study results. Furthermore, different studies use different DM algorithms, limiting reproducibility and comparability of results.

To address these challenges, our study has developed several DM algorithms and aims to report the performance characteristics for each across a range of different real-world data sources. We will leverage the Observational Health Data Sciences and Informatics (OHDSI) network to accomplish this. The OHDSI community has created multiple open-source tools to create and evaluate phenotypes and ultimately analytical tools to conduct network studies. OHDSI uses the common data model Observational Medical Outcomes Partnership (OMOP), which currently covers over 12% of the world's population. The OMOP Common Data Model (CDM) is a standardized data model that facilitates the harmonization of disparate healthcare data sources, enabling researchers to conduct large-scale observational studies. By using the OMOP CDM, the OHDSI community can integrate and analyze data from various sources, such as EHRs, insurance claims, and registries, to generate real-world evidence. This approach allows for more robust and generalizable findings, which are crucial for studying rare diseases like DM.

4. Objective

The primary objective of this study is to evaluate and validate different DM phenotypes across a range of OMOP databases. Secondary objectives include (a) Raise awareness of OHDSI and OMOP as a CDM in the clinical rheumatology community and (b) Provide proof of concept of potential to perform large-scale network studies in rare diseases.

5. Methods

5.1 Data Sources

This study is a multinational cohort study evaluating several phenotypes for adult dermatomyositis (DM).

We intend to solicit participation from a variety of healthcare settings in multiple geographies. Should more data partners wish to participate, this analysis could extend to any additional databases that are formatted to the Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM).

The study will be conducted using data from real world data sources that have been mapped to the OMOP-CDM in collaboration with the Observational Health Data Sciences and Informatics (OHDSI) and European Health Data and Evidence Network (EHDEN) initiatives. The OMOP-CDM (https://github.com/OHDSI/CommonDataModel/wiki) includes a standard representation of health care experiences (such as information related to drug utilization and condition occurrence), as well as common vocabularies for coding clinical concepts, and enables consistent application of analyses across multiple disparate data sources.

To be included as a data source, participating Data Partners must meet the following inclusion criteria:

1. Have a database that complies with specifications for the OMOP CDM, version 5.4.
2. Have a minimum population of >= 300 of participants with a condition occurrence of dermatomyositis.
3. Can execute the study package in their local environment.
4. Have appropriate approval from any applicable governance organizations such as an IRB or Ethics Commitee to provide summary results from the HADES DatabaseDiagnostics and PheValuator packages to the study team for analysis or ability obtain such approval by the data submission deadline.

5.2 Study design

The study will be a retrospective observational cohort study based on routinely collected health care data which has been mapped to the OMOP-CDM. Cohorts of individuals with DM will be identified.

5.3 Target cohorts

The phenotype we desired to create is adult patients who have a clinical diagnosis of dermatomyositis, with the goal of studying this patient population in large network studies using real-world data (electronic health records, claims, registries). We began by conducting a literature review on prior studies examining the performance characteristics of algorithms for dermatomyositis. Drawing from the literature, we created 8 DM algorithms using OHDSI methodology through the open-source tool ATLAS – a web-browser interface that allows the transparent creation of disease phenotypes.

5.4 Evaluation of algorithms

The evaluation of each algorithm will be performed in two ways. First, we will perform the gold-standard manual chart review of our entire Johns Hopkins Myositis Center Cohort. Our cohort from 2016 onward has previously been converted to the OMOP common data model, totaling approximately 1500 patients, where systematic and detailed chart reviews have been conducted to confirm DM diagnosis based on validated classification criteria for dermatomyositis [ACR/EULAR 2017 Dermatomyositis Classification Criteria]. Furthermore, both symptom-onset date and diagnosis date are recorded systematically. Thus, our Hopkins Myositis Cohort provides an ideal environment to assess the performance of each algorithm including sensitivity, specificity, positive predictive value, and negative predictive value, as well as quantify index date misclassification.

Second, to estimate the performance of each algorithm across multiple databases where we did not have access to detailed data, we will employ a probabilistic approach using an open-source tool called PheValuator. PheValuator utilizes LASSO regression to develop a diagnostic predictive model. The process involves the following steps:

1. Identifying a cohort of patients with a very high probability of having the phenotype of interest, in this case, dermatomyositis (DM).
2. Identifying a cohort of patients with a very high probability of not having the phenotype of interest.
3. Identifying a cohort of patients to estimate the prevalence of the phenotype of interest.

With these three cohorts identified, PheValuator can develop a predictive model for DM and generate a probabilistic gold standard in any data source used. The generated DM phenotypes can then be tested against this 'probabilistic gold standard' derived from PheValuator. This approach allows researchers to understand how the phenotypes perform in any given data source without needing access to individual patient charts.

5.5 Phenotypes of Interest

Phenotype Algorithm Cohort Number in ATLAS OHDSI Demo	Phenotype Algorithm	Index Date	Observation Period Prior to Index Date	Additional Inclusion Criteria
1781804	1	1st diagnostic code of DM	365 days	Age >18 at index date 2nd diagnostic code of DM within 30-365 days of index date
1788567	2	1st diagnostic code of DM	0 days	Age >18 at index date 2nd diagnostic code of DM within 30-365 days of index date
1787425	3	1st diagnostic code of DM	365 days	Age >18 at index date 2nd diagnostic code of DM within 30-365 days of index date At least one immunosuppressive medication prescribed
1788503	4		0 days	Age >18 at index date

		1st diagnostic code of DM		2nd diagnostic code of DM within 30-365 days of index date At least one immunosuppressive medication prescribed
1789031	5	1st diagnostic code of DM	0 days	Age >18 at index date Myositis-specific autoantibody test ordered
1789032	6	1st diagnostic code of DM	0 days	Age >18 at index date Myositis-specific autoantibody test positive result
1788875	7	1st diagnostic code of DM	0 days	Age >18 at index date 2nd diagnostic code of DM within 30-365 days of index date At least one immunosuppressive medication prescribed Diagnosed by rheumatology, dermatology, or neurology
1789289	8	1st diagnostic code of DM	365 days	Age >18 at index date At least one immunosuppressive medication prescribed

5.6 Analysis: Characterizing cohorts

All analyses will be performed using code developed for HADES (Health Analytics Data-to-Evidence Suite), formerly known as the OHDSI Methods library. A diagnostic package, built off the OHDSI Cohort Diagnostics (https://ohdsi.github.io/CohortDiagnostics/) library, is included in the base package as a preliminary step to assess the fitness of use of phenotypes on your database. If a database passes cohort diagnostics, the full study package will be executed. Baseline covariates will be extracted using an optimized SQL extraction script based on principles of the FeatureExtraction package (http://ohdsi.github.io/FeatureExtraction/) to quantify Demographics (Gender, Prior Observation Time, Age Group), Condition Group Eras and Drug Group Eras (at and within 30 days after index date, at index date, within 30 days before index date, and within 365 days before index date). Additional cohort-specific covariates will be constructed using OMOP standard vocabulary concepts.

Data quality metrics generated by the HADES DbDiagnostics (https://ohdsi.github.io/DbDiagnostics/) package will be included to evaluate and characterize potential data quality failures that could impact deployment of the phenotypes in future network studies.

5.7 Logistics of Executing a Federated Analysis

Sites will run the study analysis package locally on their data coded according to OMOP-CDM. Only aggregate results will be shared with the study coordinator. Result files will be automatically staged into a ZIP file that can be transmitted using the OhdsiSharing R Library (http://ohdsi.github.io/OhdsiSharing/) or through a site's preferred SFTP client using a site-specific key provisioned by the OHDSI Study Coordinator. Other methods of securely transferring files may be utilized to meet operational requirements on a case-by-case basis as determined by the study coordinator. Local data stewards are encouraged to review study parameters to ensure minCellCount function follows local governance. At a minimum, it is encouraged to keep this value to >5 to avoid any potential issues with re-identification of patients.

6. Strengths and Limitations

6.1 Strengths

Leveraging the OHDSI network/the OMOP CDM offers a promising solution for studying rare diseases like dermatomyositis (DM) by utilizing real-world data from diverse sources. This approach can improve the generalizability and reproducibility of research findings, ultimately contributing to a better understanding of DM and the development of more effective treatment strategies.

6.2 Limitations

While leveraging multiple databases in network studies is appealing from a sample size perspective, concerns remain regarding the accuracy of these phenotypes. For example, in EHR OMOP CDMs, the index date may not reflect the actual DM onset date, as the interval from DM diagnosis to referral to a tertiary referral center can sometimes take months to years. Further, additional limitations of this study stem from suboptimal ETL practices from source data to the OMOP CDM. This study will highlight the heterogeneity of the ETL process for uncommon concepts (such as dermatomyositis-specific autoantibody measurement), which may inform future research agendas.

7. Protection of Human Subjects

The study uses only de-identified data. Confidentiality of patient records will be maintained at all times. Data custodians will remain in full control of executing the analysis and packaging results. There will be no transmission of patient-level data at any time during these analyses. Only aggregate statistics will be captured. Study packages will contain minimum cell count parameters to obscure any cells which fall below allowable reportable limits. All study reports will contain aggregate data only and will not identify individual patients or physicians.

8. Plans for Disseminating and Communicating Study Results

All results will be posted on a freely available and accessible website such as the OHDSI website (evidence.ohdsi.org) after completion of the study. Results are aimed for publication in a clinically focused peered reviewed scientific journal to inform future network studies.