Exploring Data Ark Data Commons: A Focus on Accessing MarketScan Dataset

Yiyuan Liu, PhD

April 24, 2024

Icahn School of Medicine at Mount Sinai

Outline

Data Ark Introduction

- Mission
- Available Datasets
- Data Access
- Data Onboarding Procedures and Policy

Accessing MarketScan Dataset

- Data Information
- Data Access Workflow

Introduction to Data Ark

Data Ark Data Commons Increase the Power, Pace and Relevance of Our Science

Challenges

Exhaustive searches for relevant datasets
Repeated downloads of the same files across groups
Difficulty in understanding opaque data structures

How Data Ark Helps

Storage space for frequent-use research datasets
A team managing the resource, simplifying access, training and user support

Data Ark website: https://labs.icahn.mssm.edu/minervalab/resources/data-ark/

Data Ark Offers Mount Sinai Researchers Readily Available Datasets

There are 17 datasets hosted under Data Ark currently

Access within 24 hours after DUA signed

Public Data Sets

- 1,000 Genomes Project
- BLAST
- gnomAD
- eQTLGen
- Genebass
- GTEx
- GWAS Summary Stats
- LDSCORE
- Reference Genome
- The Cancer Genome Atlas (TCGA)
- UKBB-LD

Mount Sinai Generated Data

- CBIPM-BioMe Data (Pending IRB Approval)
- Living Brain Project
- Mount Sinai COVID-19 Biobank
- MSDW COVID-19 EHR Data Set
- MSDW OMOP EHR Data (Pending on approval)
- STOP COVID NYC Cohort

Restricted access

Public Data Sets

- UK Biobank Genotype

User Group-acquired Data Sets

- MarketScan®

How to Rapidly Access Public and Mount Sinai-Generated Datasets Through Data Ark

Visit the dataset webpage on Data Ark
Access instruction is provided in the ‘Access’ section
User completion of dataset-specific **DUA** (data use agreement)
Access is granted within 24 hours

How to Access the Restricted Dataset (UK Biobank Genotype) Through Data Ark

Data Ark hosts a single copy of UK Biobank (UKB) genotype and imputed data.
Since Dec 15, 2023, all UKB phenotype/clinical data has been managed by individual groups and stored at project-specified directories.
To access UKB genotype/imputed data,

- Researchers need to be listed in an approved UKB application by a PI.
- **Independent of Data Ark operations**: the PI of the user group will be responsible for UKB application and adding users to the approved UKB application.
- Data Ark team verifies whether an UKB application had previously been hosted through Data Ark.
  - For new UKB applications, a key file obtained from UKB that provides permissions to genotype and imputed data is required by Data Ark for verification of access.
  - Data Ark holds a record of previously hosted UKB applications.
- Upon successful verification, the user completes the DUA.
- Approval from the PI of the UKB application is required.
- Access is granted upon completion of this workflow.

The instruction for generate a mapping file to link your phenotypic dataset to the genetic dataset is provided by UK Biobank.
UKB officially stated that **generally however the use of shared datasets is both discouraged and will be deprecated as UKB develops its own online access platforms.**

How to Access the Restricted Dataset (MarketScan) Through Data Ark

For MarketScan data

- Providing training certificates
- Sign the DUA
- PI signs the DUA
- License cost applicable after initial 90-day free-of-charge access
- Contact Dr. Parul Agarwal (the Lead PI for the MarketScan user group)

Access

The following documents are required for each member of the research team:

- Mount Sinai HIPAA training certificate. Certificates of completion can be found at Mount Sinai PEAK
- CITI Basic Course and Refresher (if due) training certificates. Information for required certificates is here
- Terms of Use for MarketScan®

To use these data, you must read, agree and sign the Data Use Agreement (you must be logged in with **your Minerva ID and password** through the Mount Sinai campus network or secure remote VPN). If you don't have a Minerva ID, please open a ticket with us on MarketScan data access at hpchelp@hpc.mssm.edu

Data Onboarding Process

User-requested datasets must follow an approval process.

1. Data onboarding requested via the REDCap form (https://redcap.link/data_intake) that asks for the storage needed and prospective users.
2. Data Ark team verifies that the prospective users will use the data set.
3. If the dataset is < 1 TB, the Data Ark team will approve and start the onboarding process (webpage, copying data, verifying consent, build data usage agreement, notify users, etc.)
4. Dataset > 1 TB requires the Advisory Board for approval.

Eligibility for Cost-waived for Data Ark Hosting and Data Retention Policy

The eligibility for cost-waived hosting on Data Ark is based on the number of user groups calibrated to the data size.

Data Size (in Terabytes)	# Of user groups/dataset	Cost waived/year
1 or less	>=2	$100
3	>=3	$300
10	>=10	$1,000 ($500 for 6 months)
20	>=20	$2,000 ($1,000 for 6 months)
30	>=20	$3,000 ($1,500 for 6 months)
100	>=20	$10,000 ($5,000 for 6 months)

Data with annual low usage will be archived and offboarded.

Questions and Support: Contact Data Ark

All service requests must come through the ticket system:

hpchelp@hpc.mssm.edu

MarketScan Data Access

MarketScan Proprietary Data

Longitudinal, retrospective, de-identified medical and prescription drug claims information of millions of Americans across states insured by private health plans, Medicare and Medicaid plans.

For analysis of episodes of illness, health risk factors, treatment patterns, costs, and outcomes.

MarketScan data served by Data Ark is owned by Merative.

By early November each year, Merative, the data owner, settles into an annual use agreement with the school user group to license the MarketScan data use.

MarketScan data hosted by Data Ark spans the years 2013 to 2021.

Data Elements in MarketScan

MarketScan data contains information for millions of enrollees annually on

Dataset	Key Data
Enrollment	• Demographics (age, gender, geographic region) • Plan type (HMO, PPO, etc.) • Enrollment start and end dates
Medical Claims	• Inpatient admissions • Outpatient visits • Diagnoses (ICD codes) • Procedures (CPT/HCPCS codes) • Dates and place of service • Payments and charges
Outpatient Prescription Drug Claims	• NDC drug codes • Dispense date • Days supply • Payments and charges
Provider	• Type of provider (physician, facility, etc.) • Specialty for professionals • Geographic region

Documentation and Training Materials Available for MarketScan Data

Documentation on MarketScan data and access is available on the Data Ark webpage **MarketScan Data** and **MarketScan User Handbook**.
Additional meta-data including data dictionary provided by Merative, the data owner, is available and **restricted to existing MarketScan users** through a direcotry **meta-data on One Drive**.
Slide deck and video recording for the MarketScan data training provided by Merative on January 31, 2024 is available and also **restricted to existing MarketScan users** through a separate directory **training materials on One Drive**.

Data Ark and MarketScan User Group Leader Jointly Ensure Our MarketScan Services Meet Users' Needs

What Data Ark does

Access permission granting according to the agreement with the user group,
Technical support,

a. SAS license renewal;
b. Computational resource and infrastructure for users' MarketScan data analyses.

What the user group does

License cost and terms negotiation with the data provider and among member groups;
Setting out Data Use Agreement;
User access authorization;
Limited capacity for consultation on data analysis and SAS programming;
Leader Dr. Parul Agarwal (effective Nov 2023- Nov 2024).

Users' responsibilities include learning SAS programming and understanding MarketScan data elements.

License Cost Imposed by Data Owner for the MarketScan Data

For internally funded studies, users' access to the MarketScan data incurs **no cost for the first 90 days from the date of access granted**.
Thereafter, there is an associated cost, determined by the MarketScan user group.
For **externally funded studies**, the additional fee to use the school-acquired data sets in non-commercial projects will be **$30,000 per study**.
The additional fee to use the data in a **commercially funded study** will be **$60,000 per study**.
License cost-related inquiries shall be directed to the MarketScan user group leader, **Dr. Parul Agarwal (effective Nov 2023- Nov 2024).**

MarketScan Data Access via Minerva HPC and Windows-based Server

MarketScan data is now accessible through Linux-based Minerva HPC (high-performance computing) and the legacy Windows-based server.

Advantages provided by Minerva HPC:

- Guaranteed computational resources (CPU capacity, memory and disk storage for user processed data) dedicated to each user job;
- Concurrent sessions from multiple users without compromising on the computational performance;
- Potentially, computational resources expandable based on users' needs;
- User data import from or export to project directories, user account-specific directories.

Data Ark/HPC team provides SAS program IDE (Integrated Development Environment) on **OnDemand of Minerva HPC** , and will announce additional programming (for example, R) environment deployed on **OnDemand of Minerva HPC**.

Windows-based server is intended to be retired by Nov 2024.

Notification scheduled for dissemination today to all MarketScan users regarding the rollout of MarketScan data computing through **OnDemand of Minerva HPC**.

All MarketScan users are required to assign the revised DUA effective May 1, 2024.

MarketScan Data Access Prerequisite

1. **Obtain an active account of the school (Icahn School of Medicine at Mount Sinai)**
    a. Register an Icahn School of Medicine account through **Sailpoint**, if not already.
2. **School VPN tunnel setup when remotely and VIP two-factor authentication**
    a. Setup of VPN tunnel and VIP token is through **ASCIT**, if not already.

Access to the following webpages requires campus network or school/hospital VPN tunnel enabled when remote.

1. **Minerva account granted upon request and account activation**
    a. **Request a Minerva account**, if not already.
    b. Annual NIH and HIPAA forms and OnDemand profile completion through **forms** (authentication using either school or hospital ID and associated password **without** the VIP token).
2. **Login to Minerva HPC once in order for the home directory for a new user to be created**
    a. Documentation on **logging in to Minerva** (authentication using only school ID and associated password **with** the VIP token).

MarketScan Data Access

5. **Access MarketScan data**
    a. Complete MarketScan DUA (data use agreement) through the **Data Ark DUA website** (authentication using only school ID and associated password **without** the VIP token)**.**
    b. Submit training certificates required in the DUA to **hpchelp@hpc.mssm.edu**.
    c. PI completes a separate MarketScan DUA requested by the Data Ark/HPC team.

Please proceed to the next step only after receiving confirmation through the HPC ticketing system if you are a new user or access Minerva for the first time.

5. **MarketScan data and SAS IDE is accessible via**
    a. The legacy Windows-based server through **Storefront** (local installation of Citrix Workspace app required) scheduled for retirement before Nov 2024(or maybe sooner).
    b. **OnDemand of Minerva HPC**.

Demo: Access Market Data via OnDemand (ondemand.hpc.mssm.edu)

Contact for Help

Computational infrastructure-related queries

- HPC ticket system **hpchelp@hpc.mssm.edu**.

Limited consultation on data analysis and SAS programming

- MarketScan user group leader
  - Dr. Parul Agarwal
  - Associate Professor at Population Health Science and Policy
  - **parul.agarwal@mountsinai.org**.

Please Acknowledge CTSA in Your Publications

Please acknowledge the support from Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai by including the following acknowledgement in a publication of any material, whether copyrighted or not, based on or developed with Minerva HPC resources:

"This work was supported in part through the computational resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences."


DATA ARK


Thank You!