CHITA

CONFERENCE ON HEALTH IT AND ANALYTICS

November 3rd – 4th 2017 | Washington, DC

Population Health and Distributed Health Data Networks: Privacy Preserving Menu-Driven Approaches to Querying Electronic Health Data Sources

Jessica M. Malenfant, MPH; Kyle Erickson; Kimberly Barrett, MPH; Adam Paczuski; Zachary Wyner, MPH; Chayim Herzig-Marx, PhD; Jeffrey S. Brown, PhD

Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA

Hosted by

PopMedNet™

DEPARTMENT OF POPULATION MEDICINE
HARVARD MEDICAL SCHOOL
Harvard Pilgrim Health Care Institute

ROBERT H. SMITH SCHOOL OF BUSINESS
CHIDS CENTER FOR HEALTH INFORMATION AND DECISION SYSTEMS

SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MICHIGAN

Outline

- Describe how PopMedNet (PMN) powers distributed health data networks (DDNs)
- Describe PMN software design & features
- Menu-Driven Query (MDQ) tool
  - Problem & use cases
  - Solution & Challenges
- Distributed Regression Analysis
- Current status, opportunities & next steps

Distributed Health Data Networks

- Distributed health data networks are increasingly used to conduct clinical and observational research
- PopMedNet powers efficient, privacy-protecting, public health research and surveillance activities within distributed networks
- PMN is a mature platform that is used by 100s of organizations
- PMN offers a variety of query tools used in several large-scale distributed data networks ,including PCORI's PCORnet and FDA's Sentinel Initiative

PopMedNet (PMN) Platform: Powering Distributed Data & Distributed Analysis

- Mature architecture using an approach shown to be **accepted by health plans, clinical sites and other data holders**
- Data partners **maintain control** over their own data
- **Distribute code** to partners for local execution
- **Sites Provide results**, not data, to the requestor
- Standardize the data using a **common data model**
- All activities **audited** and **secure**
  - Meets the **privacy, proprietary, security, and research integrity** demands of health plans and other data holders institutions' IT departments
- Especially well suited for **multi-site, multi-use networks**
- **Contribute to the Learning Health System** by providing a socio-technical platform to support the people, process, technology contributing to knowledge generation

How it works: A Common Data Model

- Common Data Models (CDM) provide a **mechanism for efficient sharing of health data for secondary uses** – research and public health surveillance
- Agreed upon **structure for capturing data**
- Data owners **map their source data** (e.g. EHR, registry data, administrative claims data) into the CDM format including
  - Table names
  - Variable names
  - Value sets
  - Data formatting specifications
  - Database or data repository implementations
- Typically leverage **health IT standard coding systems and vocabularies**

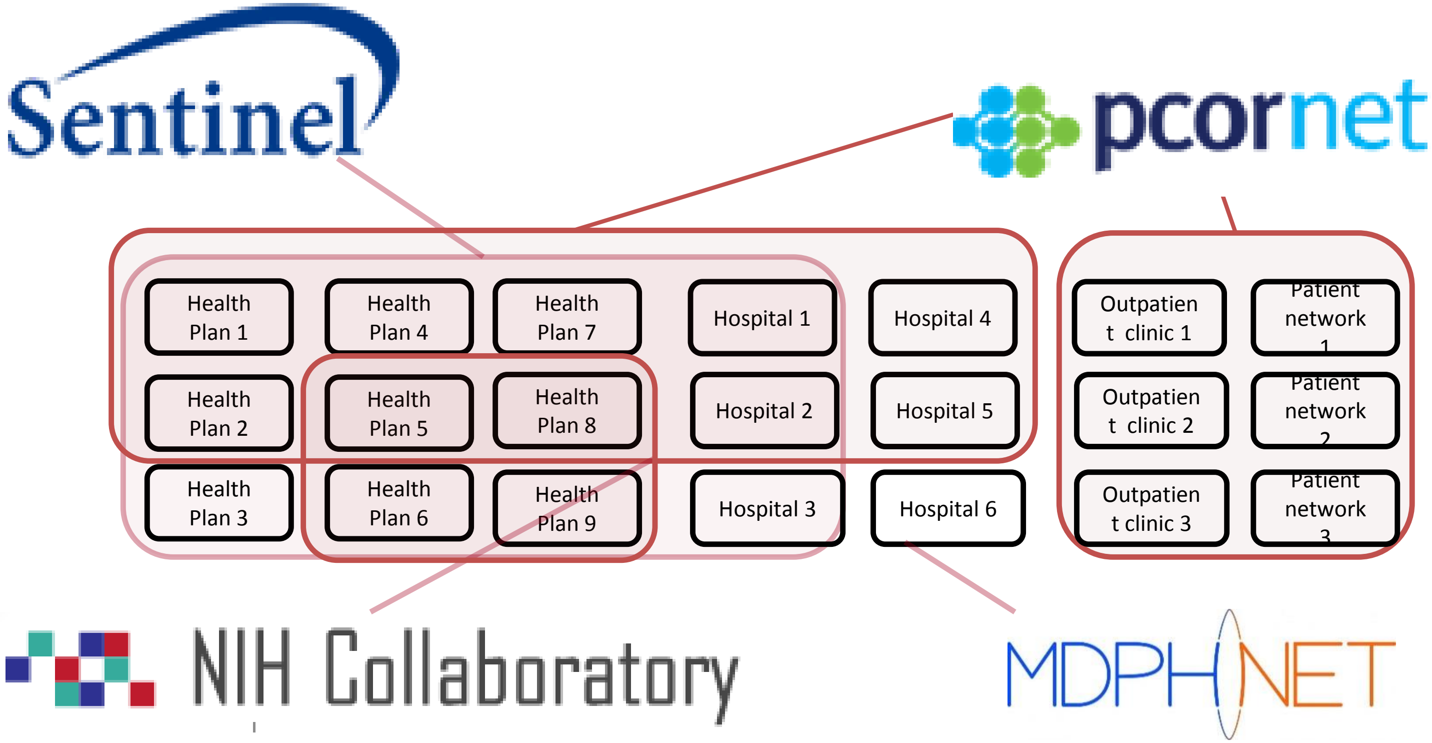
Multiple Networks Sharing PMN Infrastructure

- Each organization can **participate in multiple networks**
- Each network **benefits from architecture and security improvements** while **maintaining their unique governance and policies**
- Networks **share** analytic tools, lessons learned, and system improvements
- Each network **controls its governance** and coordination
- **Funding from each network is leveraged** across initiatives to contribute to the core PMN platform

Key Software and Security Features

- Secure, private multi-center research network
- Open source application
- Data partners maintain control of their data
- Flexible governance, access control, permissions, and auditing
- Mature documentation and set-up procedures
- Scalable: easy to add new data, new partners
- Interoperable with other networks using the same software (PopMedNet)
- FISMA compliant tier III data center
- Annual 3rd-party security audits of software
- Annual FISMA-compliance audits of network operations
- Security regularly tested by partners (software and penetration testing)

Multiple Networks Sharing PMN Infrastructure



Sentinel

pcornet

Health Plan 1

Health Plan 4

Health Plan 7

Hospital 1

Hospital 4

Outpatient clinic 1

Patient network 1

Health Plan 2

Health Plan 5

Health Plan 8

Hospital 2

Hospital 5

Outpatient clinic 2

Patient network 2

Health Plan 3

Health Plan 6

Health Plan 9

Hospital 3

Hospital 6

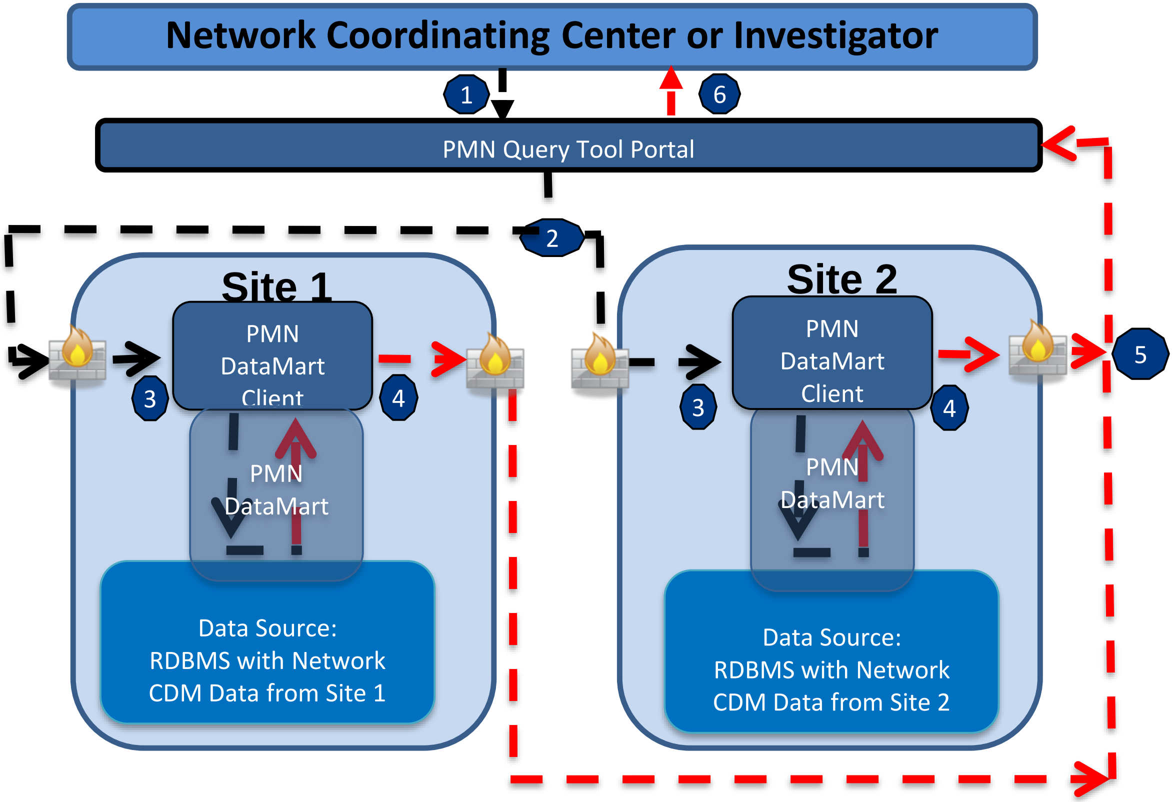
Outpatient clinic 3

Patient network 3

NIH Collaboratory

MDPHNET

PMN Request Cycle: Menu Driven Query



1. Investigator creates and submits query to selected sites

2. Individual sites retrieve query

3. Sites review and run query directly against the CDM via the PMN DataMart Client

4. Sites review results

5. Individual site returns results via secure network

6. Requestor views results in PMN Portal

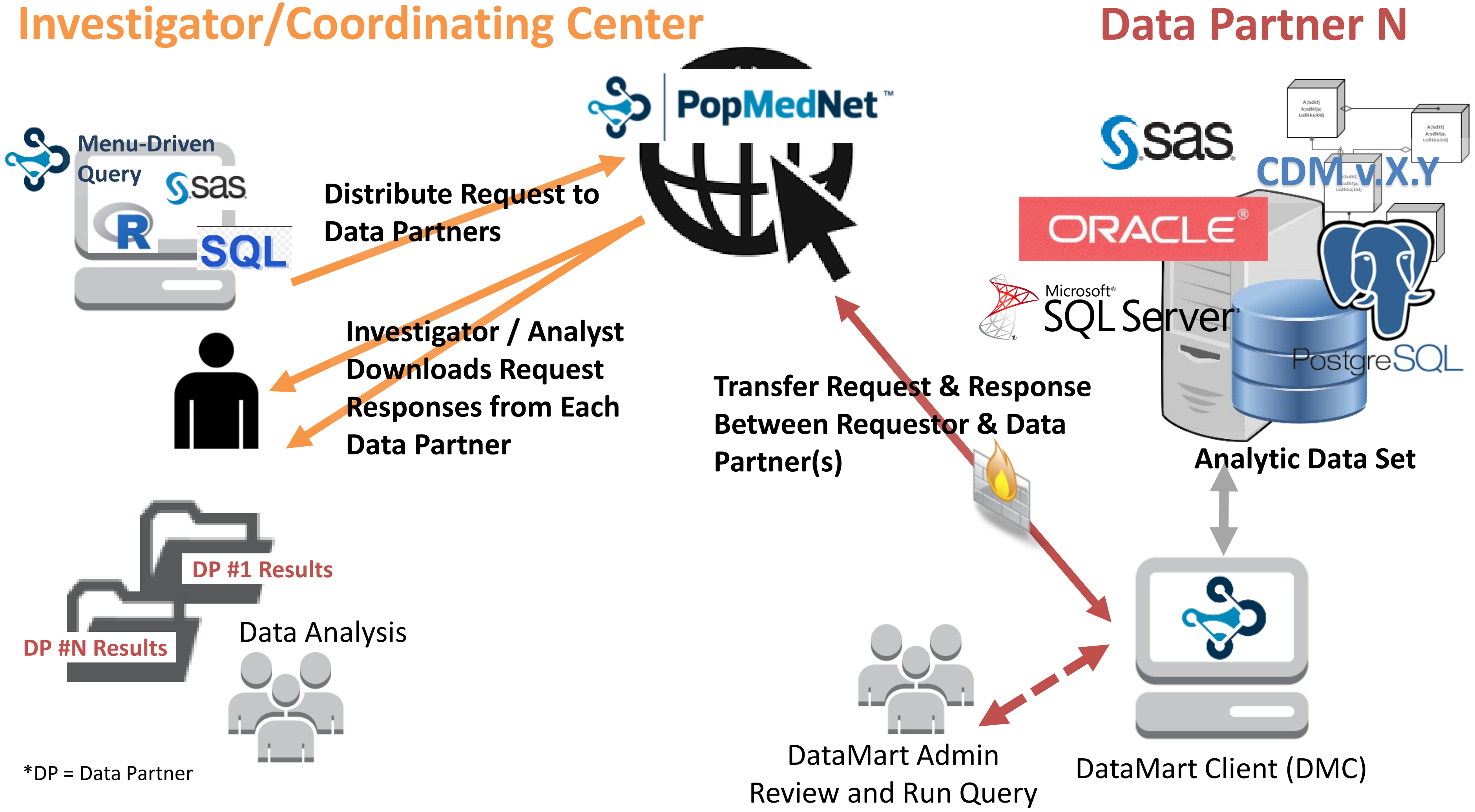
--Users have options to receive notifications throughout request cycle; various automation and approval workflows available

PopMedNet™

Problems Identified with the Initial MDQ Tool

- Legacy Query Composer: Developed for limited use resulting in scalability issues
    - Each query tool was hardcoded for use against a single CDM and RDBMS
        - The MDPHnet network's data model and PostgreSQL
        - FDA's Sentinel System Summary Table data model and MS Access database
    - All changes required manual and redundant hard-coding
    - Queryable terms could not be shared across networks (e.g. if 2 networks wanted to query race data, each query tool needed to be developed separately, even if the field names and value sets were the same)
    - Changes required the sites to download a new version of the PMN DataMart Client software in order to respond to a query

PMN Request Cycle: Menu Driven Query



Investigator/Coordinating Center

Data Partner N

Distribute Request to Data Partners

Investigator / Analyst Downloads Request Responses from Each Data Partner

Transfer Request & Response Between Requestor & Data Partner(s)

Analytic Data Set

Data Analysis

DataMart Admin Review and Run Query

DataMart Client (DMC)

*DP = Data Partner

Challenges to Distributed Querying

- Heterogeneity of technical environments (e.g. Windows, Linux/Unix)
- Source data systems and refresh cycles populating the CDMs vary
- Database management system (i.e. RDBMS) flavors and versions that store the CDM data vary across sites
- Data holders have local IT policies and procedures for how and where data are stored and accessed

(of course these are just a select list of challenges that need to be considered)

Objective

Demonstrate a new architecture and framework for an extensible point-and-click query interface in PopMedNet (PMN). These tools:

- Address challenges in platform and software heterogeneity in PCORnet, the largest PMN network
- Are modularized and can successfully target multiple data models and various technical ecosystems
- Utilize widely adopted standard data exchange formats e.g. JSON, LINQ, Microsoft Entity Framework, and SQL
- Produce consistent and valid results
- Provide a simple query tool interface and workflow
- Consider workflows for full request lifecycle including integration points with external systems

Tools Developed

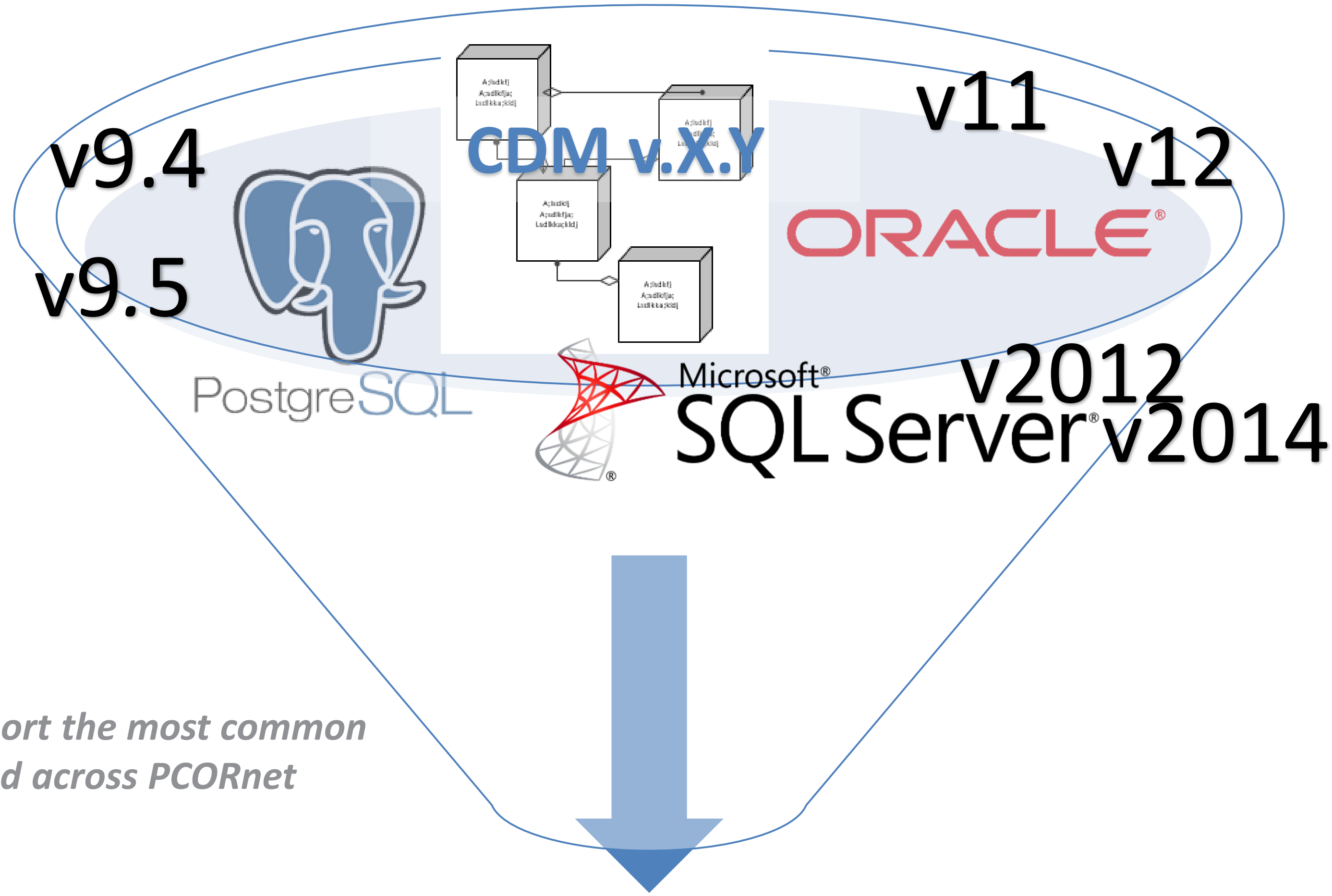
Menu-Driven Queries (MDQs):

- PMN interface supports querying terms and stratifications (e.g. Race field) to be easily re-purposed for use against multiple data models and in multiple networks
- Investigators can compose a simple or complex MDQ that includes logical operators: “OR”; “AND”; “AND NOT” to define a cohort of interest via a user interface
- Include software-enabled governance to determine what users can query
- Support electronic workflows and embedded analytics
- Include data model adapters that make the MDQs Common Data Model (CDM) aware
- Modular design for sharing queryable terms regardless of data source

Test Case Inserter (TCI):

- Generates databases according to CDM specifications
- Custom program that enables users to easily insert synthetic data into a relational database management system (RDBMS) without requiring the user to have SQL programming skills
- Supports MDQ validation and MDQ prototypes for targeting new data sources

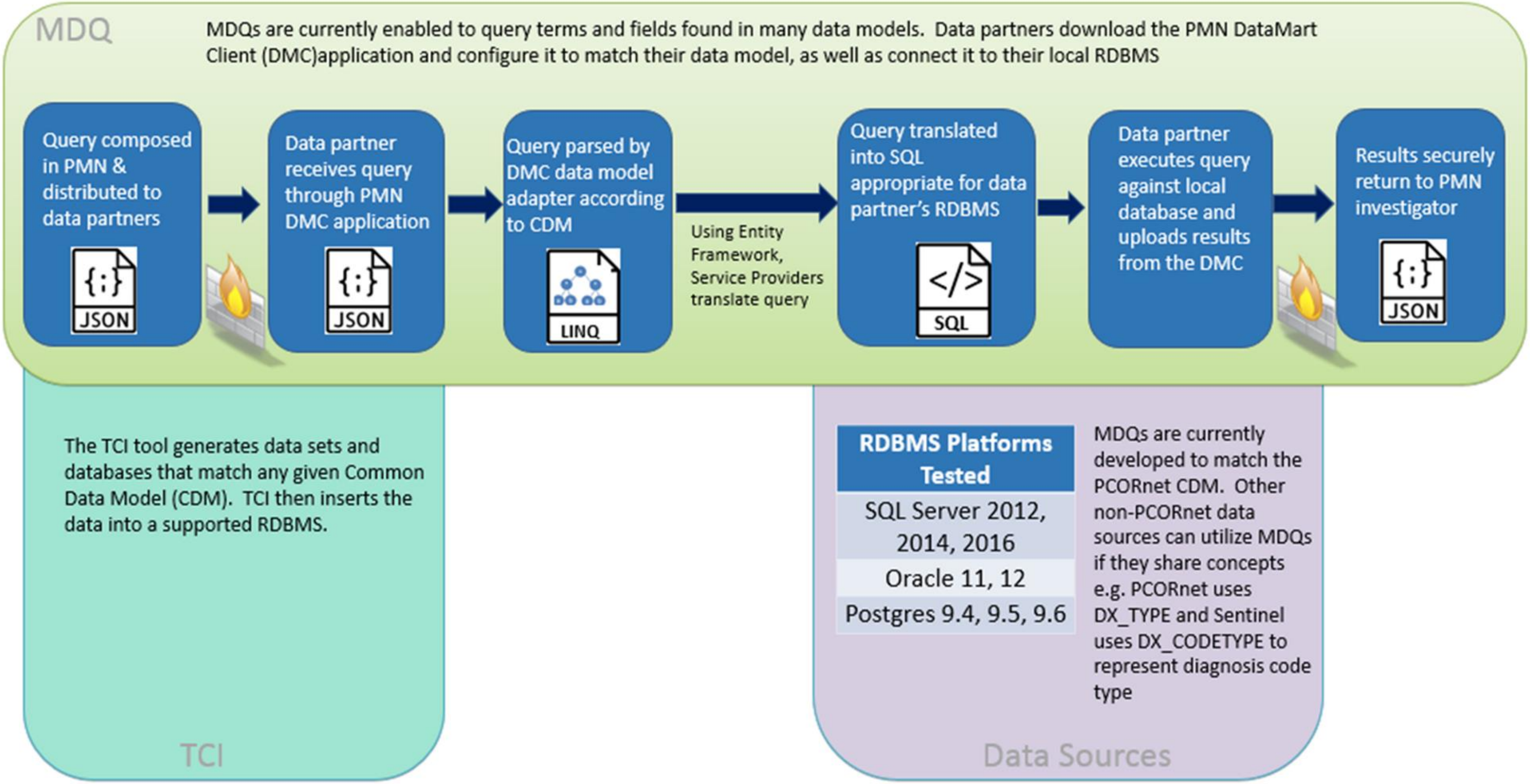
One Size fits Most* MDQ Tool



Single MDQ Tool

Committed to support the most common RDMBS versions used across PCORnet

Validating MDQs



MDQ

MDQs are currently enabled to query terms and fields found in many data models. Data partners download the PMN DataMart Client (DMC)application and configure it to match their data model, as well as connect it to their local RDBMS

Query composed in PMN & distributed to data partners

Data partner receives query through PMN DMC application

Query parsed by DMC data model adapter according to CDM

Using Entity Framework, Service Providers translate query

Query translated into SQL appropriate for data partner's RDBMS

Data partner executes query against local database and uploads results from the DMC

Results securely return to PMN investigator

The TCI tool generates data sets and databases that match any given Common Data Model (CDM). TCI then inserts the data into a supported RDBMS.

TCI

RDBMS Platforms Tested
SQL Server 2012, 2014, 2016
Oracle 11, 12
Postgres 9.4, 9.5, 9.6

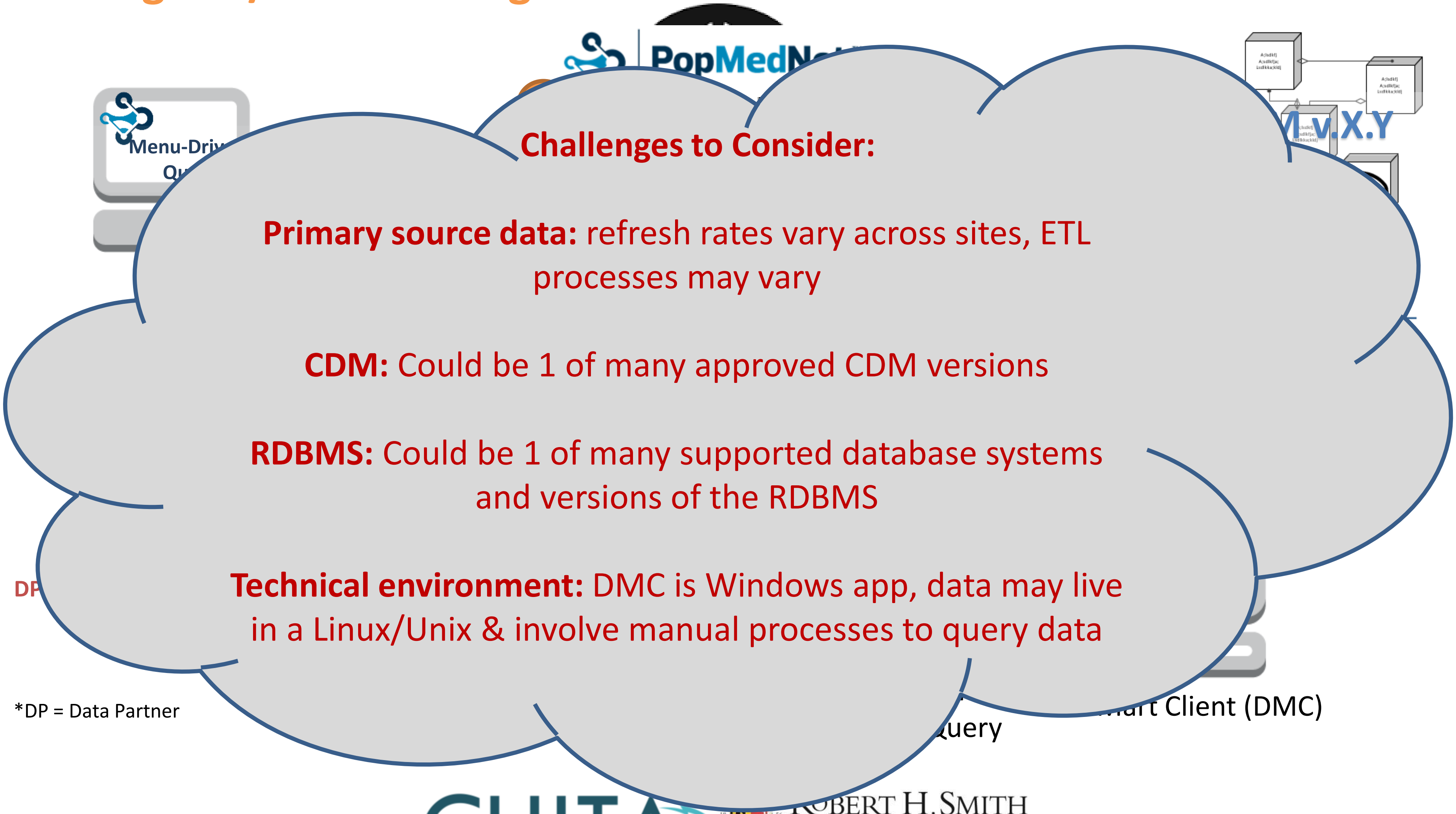
MDQs are currently developed to match the PCORnet CDM. Other non-PCORnet data sources can utilize MDQs if they share concepts e.g. PCORnet uses DX_TYPE and Sentinel uses DX_CODETYPE to represent diagnosis code type

Data Sources

PMN Request Cycle: Menu Driven Query

Investigator/Coordinating Center

Data Partner N



Challenges to Consider:

Primary source data: refresh rates vary across sites, ETL processes may vary

CDM: Could be 1 of many approved CDM versions

RDBMS: Could be 1 of many supported database systems and versions of the RDBMS

Technical environment: DMC is Windows app, data may live in a Linux/Unix & involve manual processes to query data

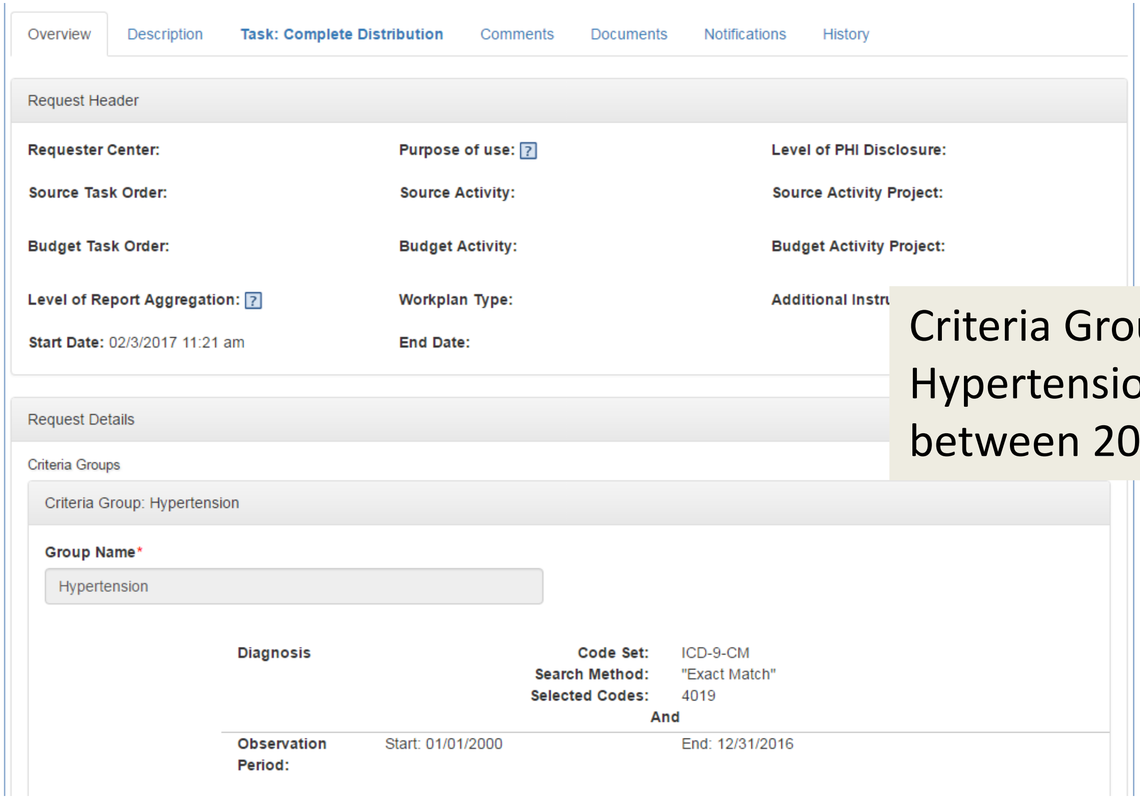
*DP = Data Partner

Use Case 1: Investigator Composes the MDQ Query: Why don’t all people with high blood cholesterol and blood pressure get heart disease?

Use MDQ to find patients of interest

Terms are added to the PMN MDQ interface according to the data model. Terms can be re-purposed for other data models.

Note that these example queries are based on the PCORnet Common Data Model



Request Header

Requester Center:	**Purpose of use:** [?]	**Level of PHI Disclosure:**
Source Task Order:	**Source Activity:**	**Source Activity Project:**
Budget Task Order:	**Budget Activity:**	**Budget Activity Project:**
Level of Report Aggregation: [?]	**Workplan Type:**	**Additional Instru**
Start Date: 02/3/2017 11:21 am	**End Date:**	

Request Details

Criteria Groups

Criteria Group: Hypertension

Group Name*

Hypertension

Diagnosis

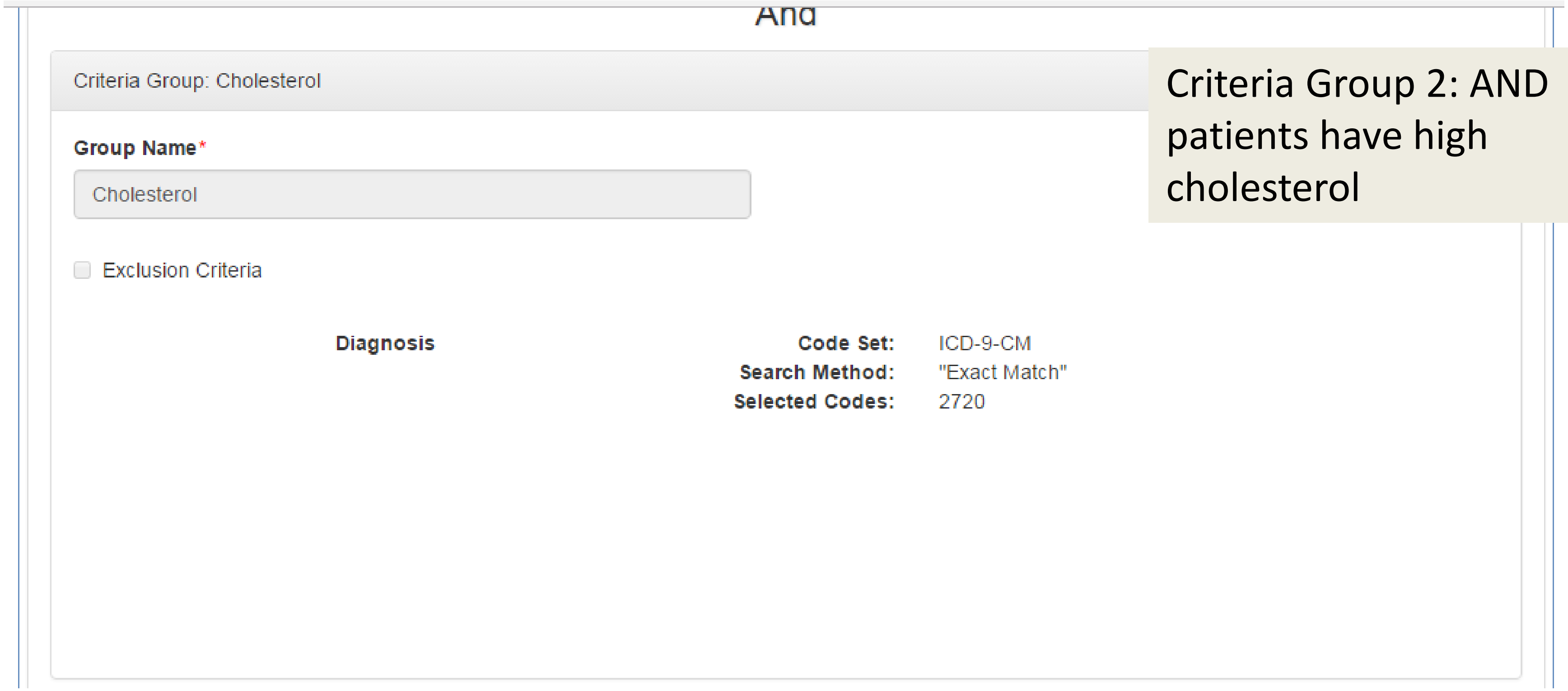
Code Set: ICD-9-CM
Search Method: "Exact Match"
Selected Codes: 4019

And

Observation Period: Start: 01/01/2000 End: 12/31/2016

Criteria Group 1: Hypertension with visits between 2000-2016

Use Case 1: Investigator Composes the MDQ Query: Why don't all people with high blood cholesterol and blood pressure get heart disease?



And

Criteria Group: Cholesterol

Group Name*

Cholesterol

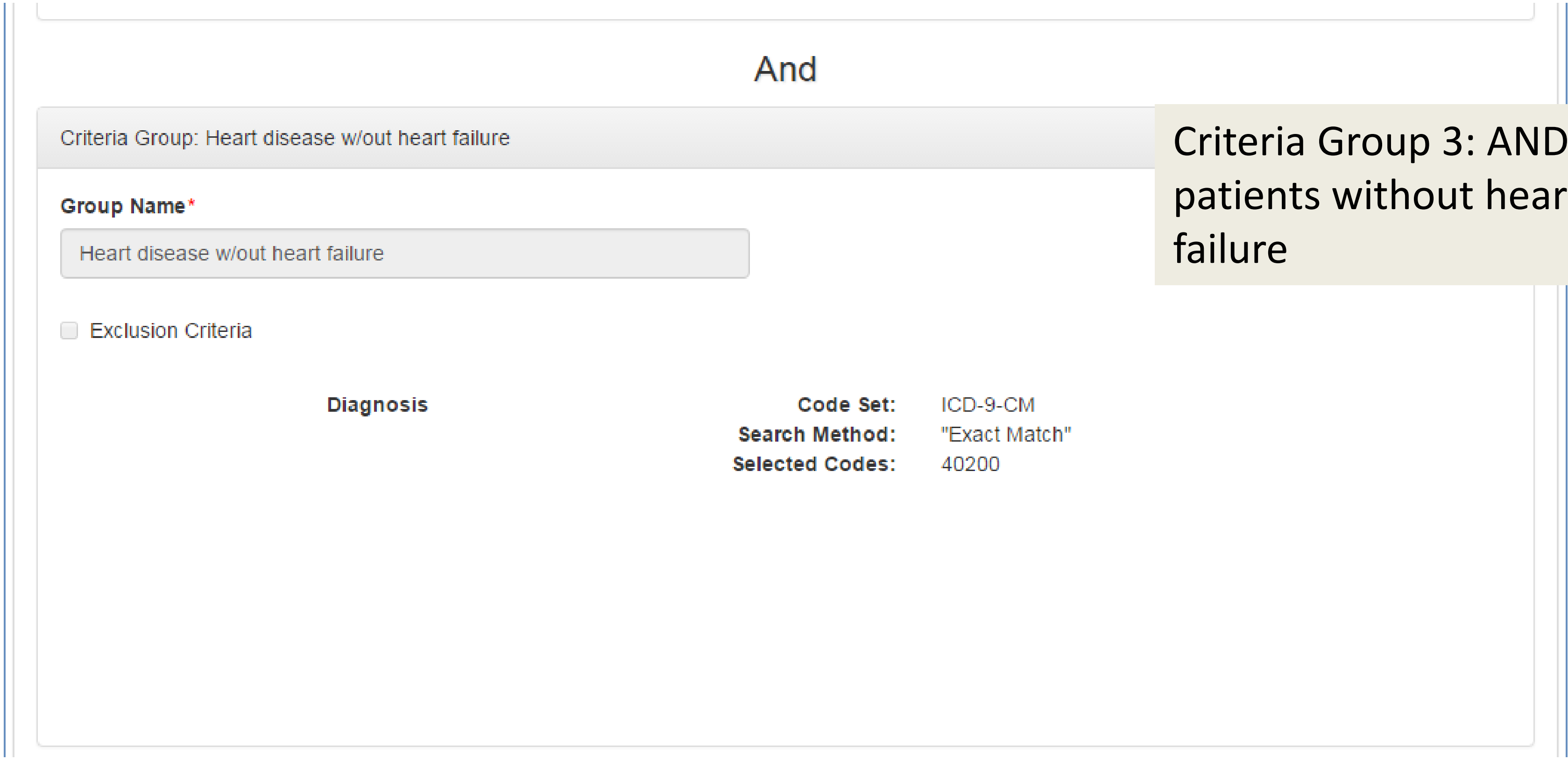
□ Exclusion Criteria

Diagnosis

Code Set:	ICD-9-CM
Search Method:	"Exact Match"
Selected Codes:	2720

Criteria Group 2: AND patients have high cholesterol

Use Case 1: Investigator Composes the MDQ Query: Why don't all people with high blood cholesterol and blood pressure get heart disease?



And

Criteria Group: Heart disease w/out heart failure

Group Name*

Heart disease w/out heart failure

□ Exclusion Criteria

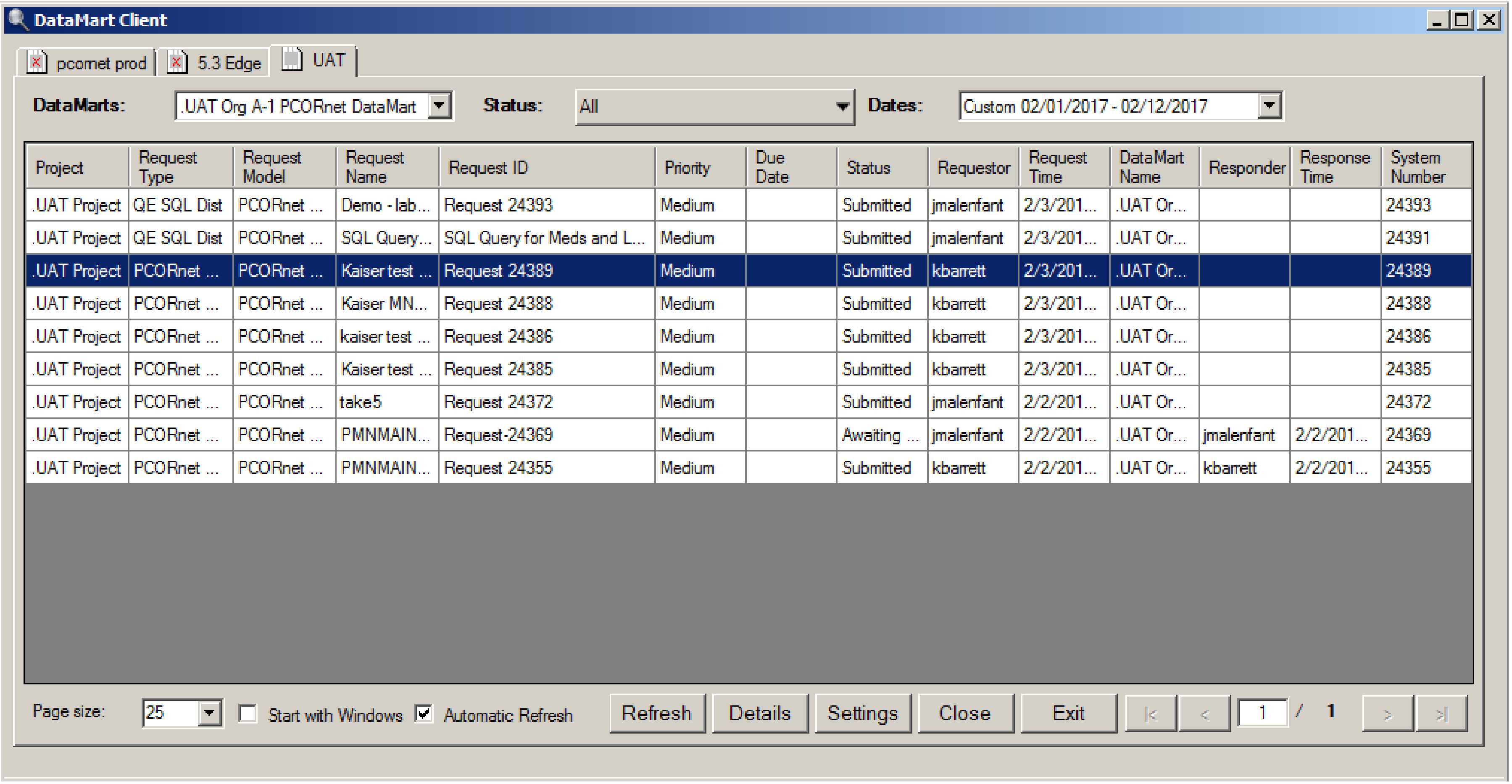
Diagnosis

Code Set: ICD-9-CM
Search Method: "Exact Match"
Selected Codes: 40200

Criteria Group 3: AND patients without heart failure

DataMart Administrator Receives the Query

DataMart Administrator Inbox – locally installed app at each site



DataMarts: .UAT Org A-1 PCORnet DataMart

Status: All

Dates: Custom 02/01/2017 - 02/12/2017

Project	Request Type	Request Model	Request Name	Request ID	Priority	Due Date	Status	Requestor	Request Time	DataMart Name	Responder	Response Time	System Number
.UAT Project	QE SQL Dist	PCORnet ...	Demo - lab...	Request 24393	Medium		Submitted	jmalenfant	2/3/201...	.UAT Or...			24393
.UAT Project	QE SQL Dist	PCORnet ...	SQL Query...	SQL Query for Meds and L...	Medium		Submitted	jmalenfant	2/3/201...	.UAT Or...			24391
.UAT Project	PCORnet ...	PCORnet ...	Kaiser test ...	Request 24389	Medium		Submitted	kbarrett	2/3/201...	.UAT Or...			24389
.UAT Project	PCORnet ...	PCORnet ...	Kaiser MN...	Request 24388	Medium		Submitted	kbarrett	2/3/201...	.UAT Or...			24388
.UAT Project	PCORnet ...	PCORnet ...	kaiser test ...	Request 24386	Medium		Submitted	kbarrett	2/3/201...	.UAT Or...			24386
.UAT Project	PCORnet ...	PCORnet ...	Kaiser test ...	Request 24385	Medium		Submitted	kbarrett	2/3/201...	.UAT Or...			24385
.UAT Project	PCORnet ...	PCORnet ...	take5	Request 24372	Medium		Submitted	jmalenfant	2/2/201...	.UAT Or...			24372
.UAT Project	PCORnet ...	PCORnet ...	PMNMAIN...	Request-24369	Medium		Awaiting ...	jmalenfant	2/2/201...	.UAT Or...	jmalenfant	2/2/201...	24369
.UAT Project	PCORnet ...	PCORnet ...	PMNMAIN...	Request 24355	Medium		Submitted	kbarrett	2/2/201...	.UAT Or...	kbarrett	2/2/201...	24355

Page size: 25

☐ Start with Windows ☑ Automatic Refresh

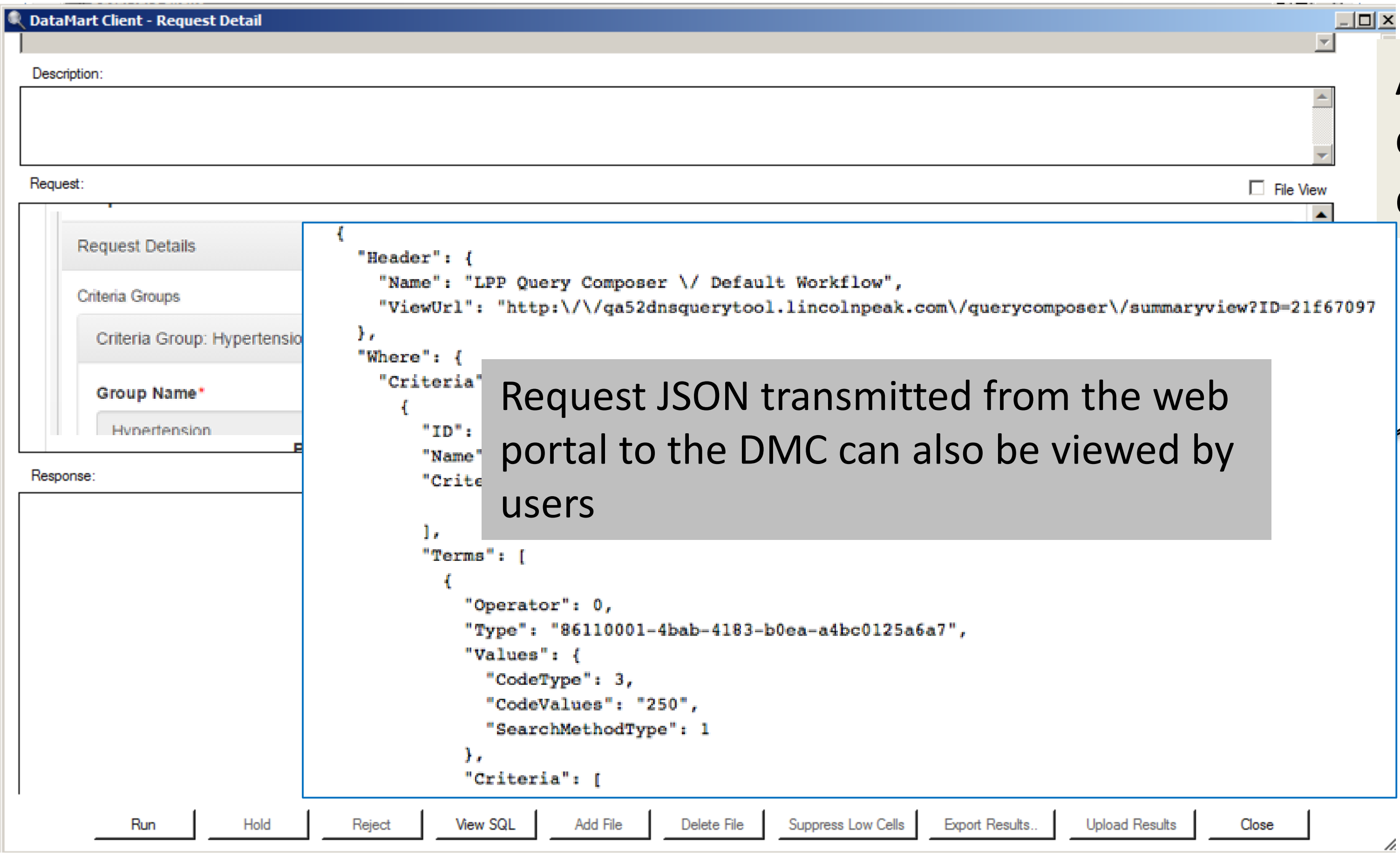
1 / 1

CHITA CONFERENCE ON HEALTH IT AND ANALYTICS

ROBERT H. SMITH SCHOOL OF BUSINESS

CHIDS CENTER FOR HEALTH INFORMATION AND DECISION SYSTEMS

DataMart Administrator Reviews Query Details



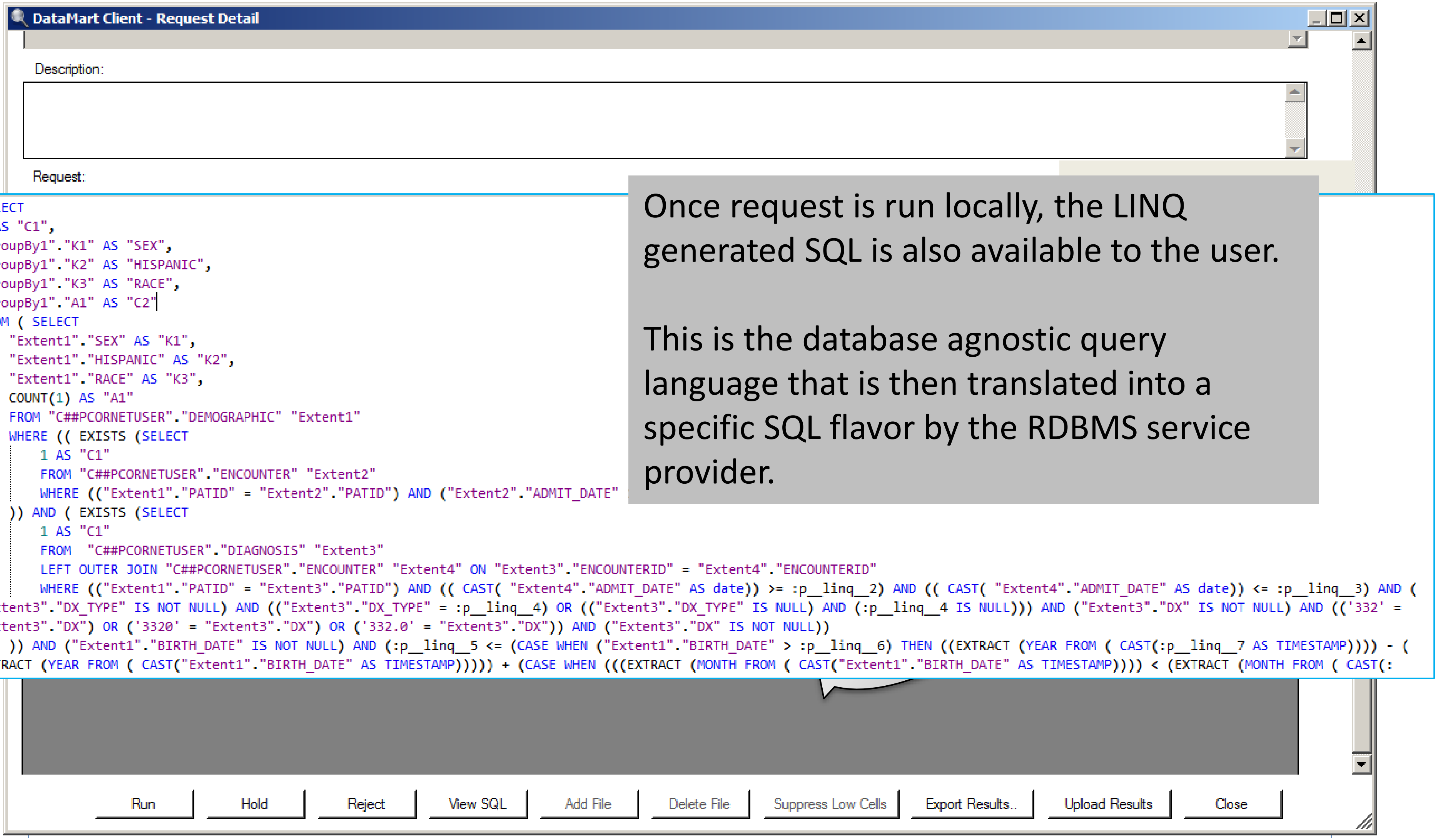
```
{
  "Header": {
    "Name": "LPP Query Composer \/ Default Workflow",
    "ViewUrl": "http:\/\/qa52dnsquerytool.lincolnpeak.com\/querycomposer\/summaryview?ID=21f67097
  },
  "Where": {
    "Criteria"
      {
        "ID":
        "Name"
        "Crite

        ],
        "Terms": [
          {
            "Operator": 0,
            "Type": "86110001-4bab-4183-b0ea-a4bc0125a6a7",
            "Values": {
              "CodeType": 3,
              "CodeValues": "250",
              "SearchMethodType": 1
            },
            "Criteria": [
```

Request JSON transmitted from the web portal to the DMC can also be viewed by users

Administrator can review query input

DataMart Administrator Executes the Query and Reviews Results

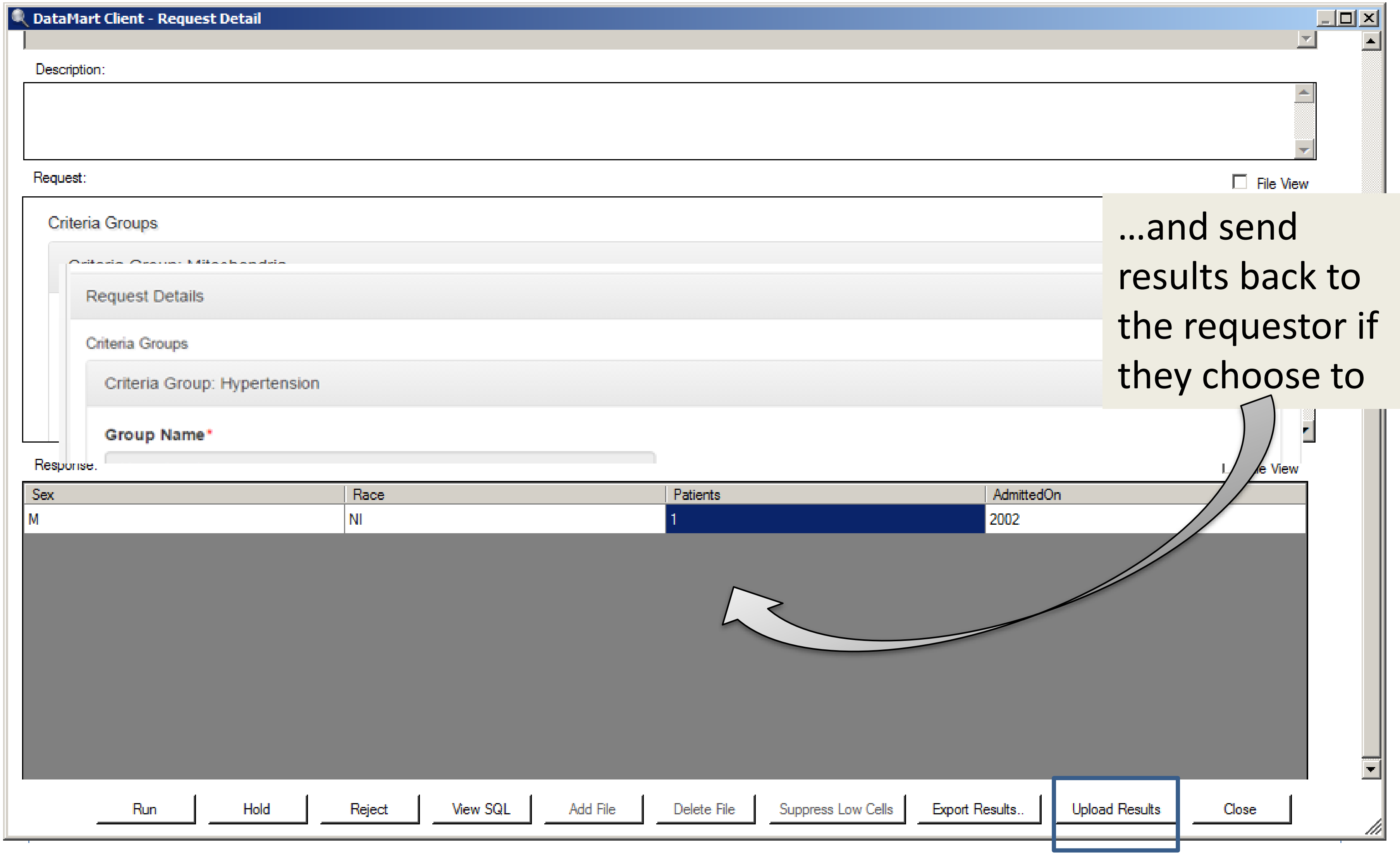


```
SELECT
1 AS "C1",
"GroupBy1"."K1" AS "SEX",
"GroupBy1"."K2" AS "HISPANIC",
"GroupBy1"."K3" AS "RACE",
"GroupBy1"."A1" AS "C2"
FROM ( SELECT
    "Extent1"."SEX" AS "K1",
    "Extent1"."HISPANIC" AS "K2",
    "Extent1"."RACE" AS "K3",
    COUNT(1) AS "A1"
    FROM "C##PCORNETUSER"."DEMOGRAPHIC" "Extent1"
    WHERE (( EXISTS (SELECT
        1 AS "C1"
        FROM "C##PCORNETUSER"."ENCOUNTER" "Extent2"
        WHERE (("Extent1"."PATID" = "Extent2"."PATID") AND ("Extent2"."ADMIT_DATE"
    )) AND ( EXISTS (SELECT
        1 AS "C1"
        FROM  "C##PCORNETUSER"."DIAGNOSIS" "Extent3"
        LEFT OUTER JOIN "C##PCORNETUSER"."ENCOUNTER" "Extent4" ON "Extent3"."ENCOUNTERID" = "Extent4"."ENCOUNTERID"
        WHERE (("Extent1"."PATID" = "Extent3"."PATID") AND (( CAST( "Extent4"."ADMIT_DATE" AS date)) >= :p__linq__2) AND (( CAST( "Extent4"."ADMIT_DATE" AS date)) <= :p__linq__3) AND (
"Extent3"."DX_TYPE" IS NOT NULL) AND (("Extent3"."DX_TYPE" = :p__linq__4) OR (("Extent3"."DX_TYPE" IS NULL) AND (:p__linq__4 IS NULL))) AND ("Extent3"."DX" IS NOT NULL) AND (('332' =
"Extent3"."DX") OR ('3320' = "Extent3"."DX") OR ('332.0' = "Extent3"."DX")) AND ("Extent3"."DX" IS NOT NULL))
    )) AND ("Extent1"."BIRTH_DATE" IS NOT NULL) AND (:p__linq__5 <= (CASE WHEN ("Extent1"."BIRTH_DATE" > :p__linq__6) THEN ((EXTRACT (YEAR FROM ( CAST(:p__linq__7 AS TIMESTAMP)))) - (
EXTRACT (YEAR FROM ( CAST("Extent1"."BIRTH_DATE" AS TIMESTAMP)))) + (CASE WHEN (((EXTRACT (MONTH FROM ( CAST("Extent1"."BIRTH_DATE" AS TIMESTAMP)))) < (EXTRACT (MONTH FROM ( CAST(:
```

Once request is run locally, the LINQ generated SQL is also available to the user.

This is the database agnostic query language that is then translated into a specific SQL flavor by the RDBMS service provider.

DataMart Administrator Uploads Results



DataMart Client - Request Detail

Description:

Request:

□ File View

Criteria Groups

Criteria Group: Mitochondria

Request Details

Criteria Groups

Criteria Group: Hypertension

Group Name*

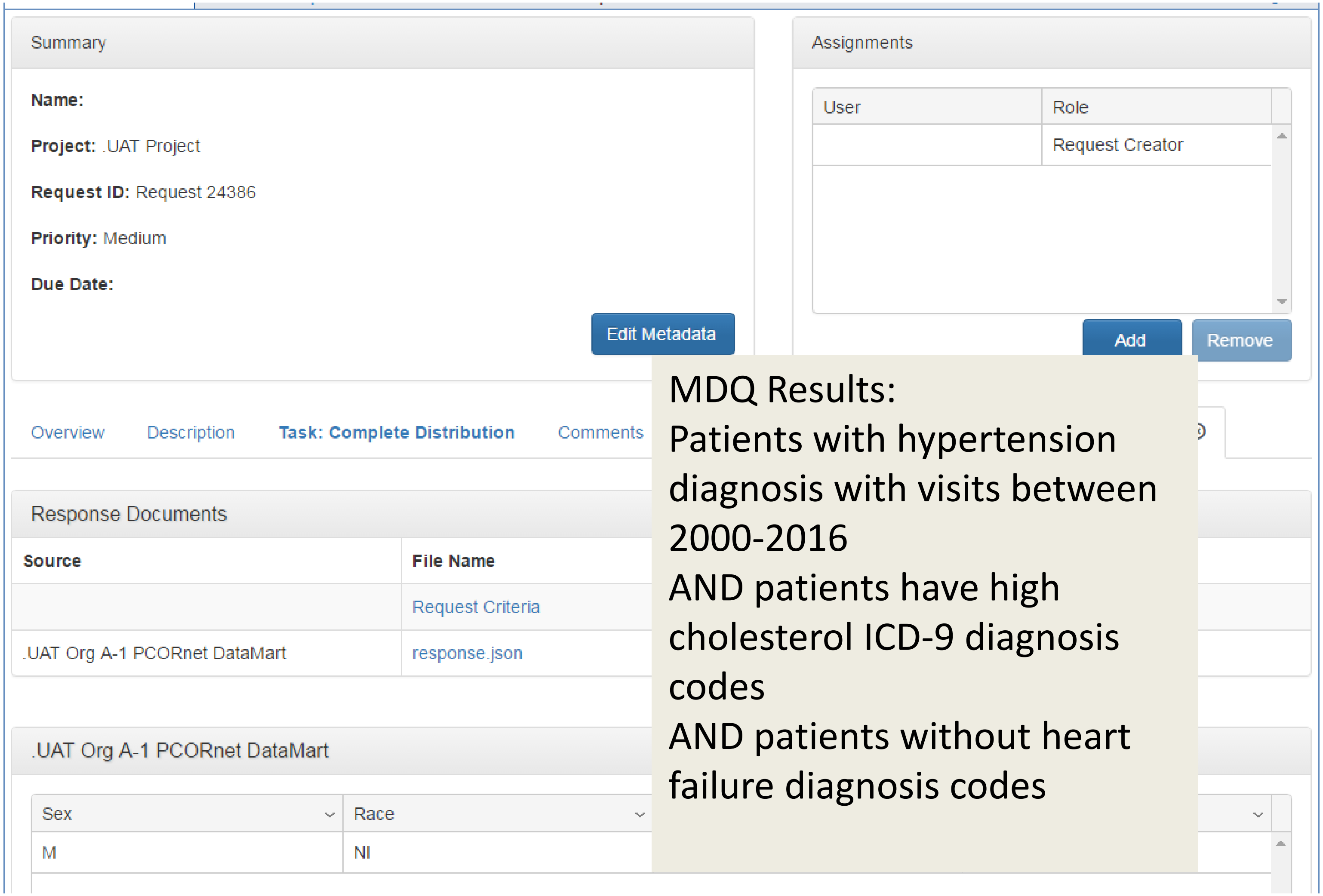
Response:

File View

Sex	Race	Patients	AdmittedOn
M	NI	1	2002

…and send results back to the requestor if they choose to

Investigator Reviews Site-Specific Results on Web Portal



Summary

Name:

Project: .UAT Project

Request ID: Request 24386

Priority: Medium

Due Date:

Assignments

User	Role
	Request Creator

Response Documents

Source	File Name
	Request Criteria
.UAT Org A-1 PCORnet DataMart	response.json

.UAT Org A-1 PCORnet DataMart

Sex	Race
M	NI

MDQ Results:

Patients with hypertension diagnosis with visits between 2000-2016

AND patients have high cholesterol ICD-9 diagnosis codes

AND patients without heart failure diagnosis codes

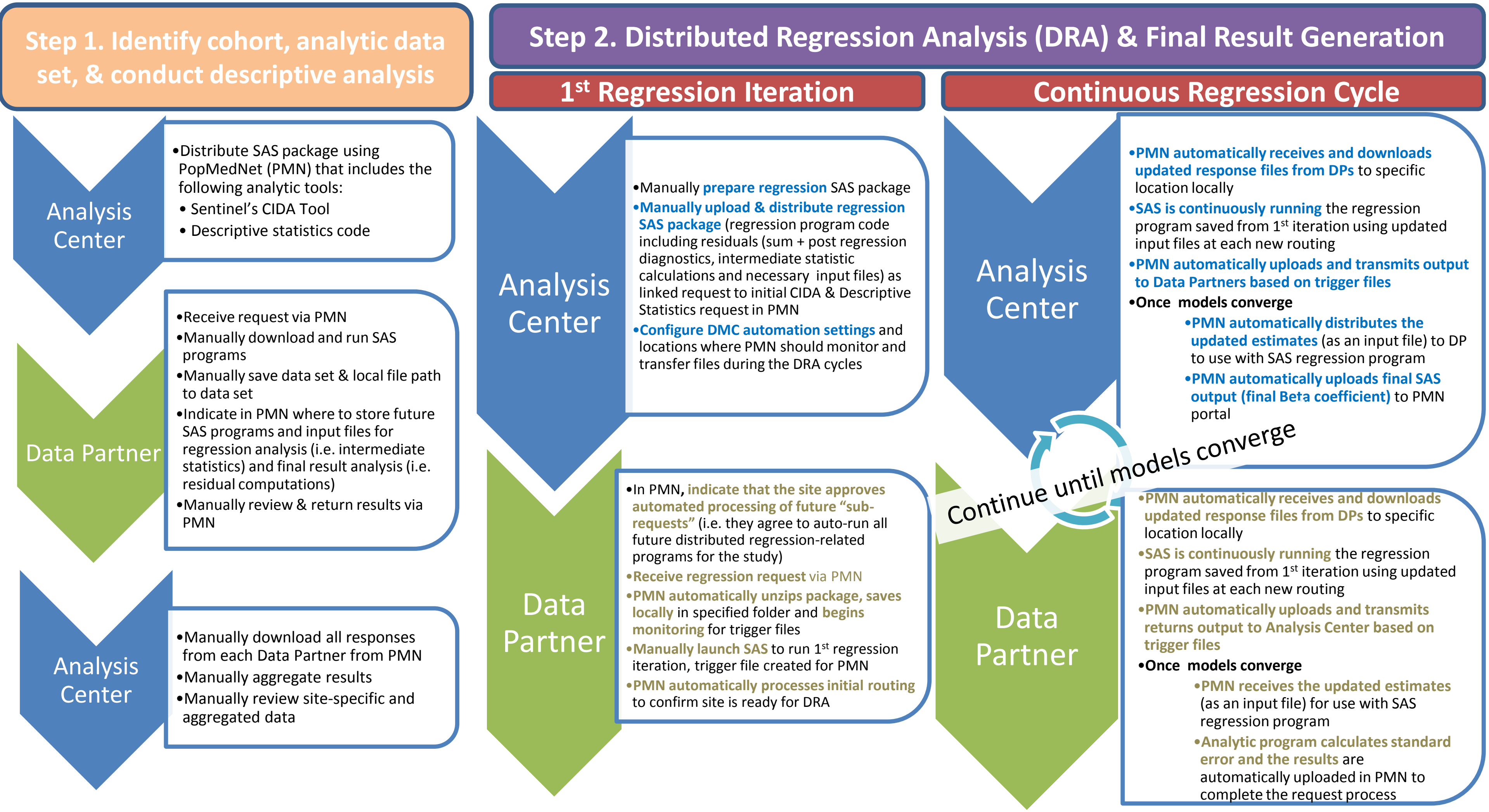
Current Status

- **Multiple terms have been added to the MDQ tool** for several fields including Race, Sex, Observation Period, Diagnosis and Procedure Codes, Height, Weight, Age, etc., more planned
- The **PCORnet data adapter has been updated** to process queries with the new terms and stratification options
- Testing with the TCI tool has verified that **ad hoc data models that share PCORnet CDM fields can use the MDQ out-of-the box, continue to explore**

Current Status

- Enhancing automation functionality, including expanding distributed regression analysis functionality
- Ability to **expose the actual SQL to a user prior to running a query** is under investigation. The request JSON and the LINQ code are currently available to end users but require manual steps to piece the query languages together, for example:

Distributed Regression Analysis



Step 1. Identify cohort, analytic data set, & conduct descriptive analysis

Analysis Center

- Distribute SAS package using PopMedNet (PMN) that includes the following analytic tools:
  - Sentinel's CIDA Tool
  - Descriptive statistics code

Data Partner

- Receive request via PMN
- Manually download and run SAS programs
- Manually save data set & local file path to data set
- Indicate in PMN where to store future SAS programs and input files for regression analysis (i.e. intermediate statistics) and final result analysis (i.e. residual computations)
- Manually review & return results via PMN

Analysis Center

- Manually download all responses from each Data Partner from PMN
- Manually aggregate results
- Manually review site-specific and aggregated data

Step 2. Distributed Regression Analysis (DRA) & Final Result Generation

1st Regression Iteration

Analysis Center

- Manually **prepare regression** SAS package
- **Manually upload & distribute regression SAS package** (regression program code including residuals (sum + post regression diagnostics, intermediate statistic calculations and necessary input files) as linked request to initial CIDA & Descriptive Statistics request in PMN
- **Configure DMC automation settings** and locations where PMN should monitor and transfer files during the DRA cycles

Data Partner

- In PMN, **indicate that the site approves automated processing of future "sub-requests"** (i.e. they agree to auto-run all future distributed regression-related programs for the study)
- **Receive regression request** via PMN
- **PMN automatically unzips package, saves locally** in specified folder and **begins monitoring** for trigger files
- **Manually launch SAS** to run 1st regression iteration, trigger file created for PMN
- **PMN automatically processes initial routing** to confirm site is ready for DRA

Continuous Regression Cycle

Analysis Center

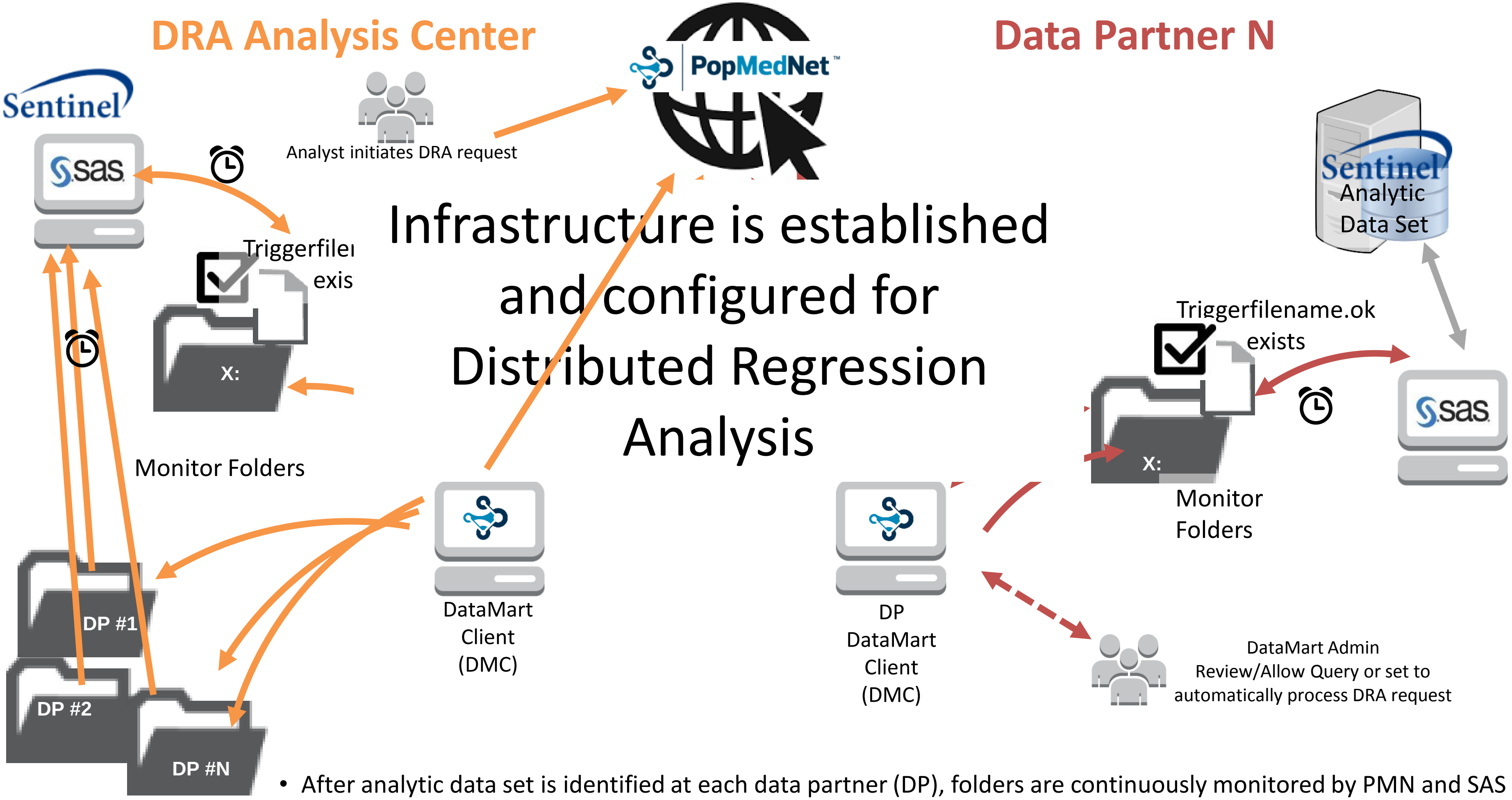
- **PMN automatically receives and downloads updated response files from DPs** to specific location locally
- **SAS is continuously running** the regression program saved from 1st iteration using updated input files at each new routing
- **PMN automatically uploads and transmits output to Data Partners based on trigger files**
- **Once models converge**
  - **PMN automatically distributes the updated estimates** (as an input file) to DP to use with SAS regression program
  - **PMN automatically uploads final SAS output (final Beta coefficient)** to PMN portal

Continue until models converge

Data Partner

- **PMN automatically receives and downloads updated response files from DPs** to specific location locally
- **SAS is continuously running** the regression program saved from 1st iteration using updated input files at each new routing
- **PMN automatically uploads and transmits returns output to Analysis Center based on trigger files**
- **Once models converge**
  - **PMN receives the updated estimates** (as an input file) for use with SAS regression program
  - **Analytic program calculates standard error and the results** are automatically uploaded in PMN to complete the request process

Distributed Regression Analysis



Infrastructure is established and configured for Distributed Regression Analysis

- After analytic data set is identified at each data partner (DP), folders are continuously monitored by PMN and SAS
- SAS deposits Output Files in predetermined directory that are picked up by DMC based on file manifest rules
- Trigger files determine if process continues or stops

✉ Jessica_malenfant@harvardpilgrim.org

in linkedin.com/in/jessicamalenfant

www. popmednet.org
populationmedicine.org