Overall Architecture and Distributed Analysis Tools

Daniella Meeker, PhD
Department of Preventive Medicine
University of Southern California

Michael E. Matheny, MD, MS, MPH
Tennessee Valley Healthcare System VA
Departments of Biomedical Informatics
Vanderbilt University
The overall goal of the development of the analytic tools and user interfaces in pScanner is to:

› Conduct analyses in a distributed national network that achieves a user experience comparable to execution on a single local node

› Provide menu and graphical user interface to
  » Manage study meta-data and provenance
  » user role drive permissions at each node
  » Study design
  » Analytic cohort development
  » Support common comparative effectiveness statistical methods
  » Provide result tables and graphs typically useful for reports and manuscripts
PCORNet Coverage

With coverage in every state — PCORnet represents thousands of conditions

This map depicts the number of PCORI-funded Patient-Powered or Clinical Data Research Networks that have coverage in each state.
Requirements

- Privacy Preserving – No patient row level data sharing
- Service Oriented Architecture – allows for tools and modules to be stored in a different location from the end user
- Extensible – established guidance and interface to allow anyone to contribute methods / user interfaces within the overall framework
- Open Source – promotes transparency, independent validation of functionality, and larger user community of use
- Leveraging large initiatives / established user community – PopMedNet data mart client (version proximity to Mini-Sentinel)
pSCANNER Architecture

- GitHub
- R Studio
- Shiny
- pSCANNER portal
- pSCANNER aggregator
- PopMedNet API
- PopMedNet Data Mart Client
- pSCANNER General to Site Specific Translator
- pSCANNER DAN (Site Analytic Coordination)
- MSSQL 2016
- SITE NODE(S)

Supported by the Patient-Centered Outcomes Research Institute (PCORI) Contract CDRN-1306-04819
pSCANNER Process Model Separation

Platform Independent Specification in pSCANNER web portal
- HQMF (Data Processing Query)
- QRDA (Result Data Set Specification)
- PPML (Analysis Query Specification)

Platform Specific Implementation in pSCANNER nodes
- OMOP
  - CDAMDB
  - PCORNet
- CSV
  - HCL
  - swETL
- OCTANS-GLR
  - R
  - Madlib
  - SAS

Result Display in pSCANNER web portal
- OCTANS-SCANNER V1
- HTML/JS Result Display Library

Supported by the Patient-Centered Outcomes Research Institute (PCORI) Contract CDRN-1306-04819
<table>
<thead>
<tr>
<th>Process we need to represent</th>
<th>Standard</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data processing rules</td>
<td>HQMF</td>
<td>CMS, ONC, HL7 endorsed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Part of EHR certification process</td>
</tr>
<tr>
<td>Cohort definition rules</td>
<td>HQMF</td>
<td>100’s of established data sets</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000s of cohort criteria</td>
</tr>
<tr>
<td>Data set description</td>
<td>QRDA</td>
<td>QRDA – Quality Measurement EHR</td>
</tr>
<tr>
<td></td>
<td>PMML</td>
<td>Certification &amp; CMS</td>
</tr>
<tr>
<td>Data Analysis Methods</td>
<td>PMML</td>
<td>UCSD Data Mining Group</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extensible to support model specifications</td>
</tr>
<tr>
<td>Data Analysis Results</td>
<td>PMML</td>
<td>Developed to represent results</td>
</tr>
<tr>
<td>Process Workflow</td>
<td>BPML?</td>
<td>Ad-Hoc at present, in development</td>
</tr>
</tbody>
</table>
Architecture Drill-Down

- PopMedNet – pScanner API
  - Communication Layer Between:
    - PopMedNet Client
    - pScanner Portal

- PopMedNet DataMart Adapters
  - Site-Specific Analytic Cluster (DAN/OCEANS)

- Contract with Lincoln Peak to extend existing API / DataMart adapters for use with pScanner
PopMedNet → DAN

- **DataMart Client**
  - Requests Analysis from Portal
  - Sends Standardized Analytic Request to DAN via DataAdapter
  - On completion returns result to Portal

- **DataAdapter**
  - Pulls standardized requests from PopMedNet
  - Calls DAN (Site specific analytics coordination) with request to queue for processing
Three-tier Distributed Network

- Broker accepts External Requests (e.g., DataMartClient)
- Cluster accepts Broker Requests based on
  - Analysis Type
  - Analysis Engine
  - Availability
- Processing Node accepts Cluster Requests

Design

- Simple Deployment as resource need increases
- Fault-Tolerance – A node fails, and the analysis continues
- High-availability – Distribution of load via routing algorithm
- Analysis Engine Agnostic – Workflow supports integration into widely used engines (e.g., R, Spark, SAS, OCEANS, etc.)
- Asynchronous – Queues requests for processing as resources become available
DAN – Components

- **Subscription-based network**
  - Child Node plugs into network
  - Informs Parent of capabilities
- **Use Case**
  - Request from PMN passes from DataAdapter (.NET) to DAN (Java) Webservice
  - Datasource defined in PMML retrieved by DAN
  - PMML/Raw request and data passed into Request Node
  - Request Node determines Cluster Node destination
  - Cluster Node determines available Processing Node destination
  - Processing Node performs analysis
  - Results are posted child-to-parent
  - Each stage of processing (e.g., Request, Response, Status etc.) logged in Node Storage for fault tolerance.

---

**PopMedNet**

- DataAdapter (Proxy to DAN)

---

**Distributed Analysis Network (DAN) - Example in R**

- Request Node
- R Cluster Node
- Node Storage
- JRI
- R Srvr
- Processing Nodes

---

Supported by the Patient-Centered Outcomes Research Institute (PCORI) Contract CDRN-1306-04819
Local Analytic Engines

- R/RServe (active)
- Revolution Analytics (in exploration)
- Potential Analytic Engines
  - Oracle In-Database R
  - Microsoft SQL 2016 In-Database R
  - Apache Spark (future)
  - SAS / SAS Grid (future)
Development Team

- Bill Clarke (Lincoln Peak)
- Claudio Farcas (UCSD)
- Josh Gieringer (VA/VU)
- Michael Matheny (VA/VU)
- Daniella Meeker (USC)
- Laura Perlman (USC)
- Dax Westerman (VA/VU)
- Bruce Swan (Lincoln Peak)