

RE06: Leveraging Big Data in Real-World Evidence Generation:

Data Engineering Challenges and Possibilities

Samiul Haque / Samiul.Haque@sas.com

Robert Collins / Robert.Collins@sas.com

SAS Institute, Cary, NC, USA



Real-World Evidence

Clinical evidence regarding usage and benefits or risks of a medical product derived from RWD

Real-World Data

Data relating to patient health status/delivery of care routinely collected

- Claims and billing activities
- Electronic health records
- Product and disease registries
- Patient-generated data Including in-home
- Other sources, incl. mobile devices

Drug Development Path

Where RWD Can Be Applied

Market opportunity
Unmet need

External control arms
New indications

Molecule
Discovery

Preclinical

RCT
(I – III)

Approval

Market

Trial feasibility
Trial planning
Site & patient selection

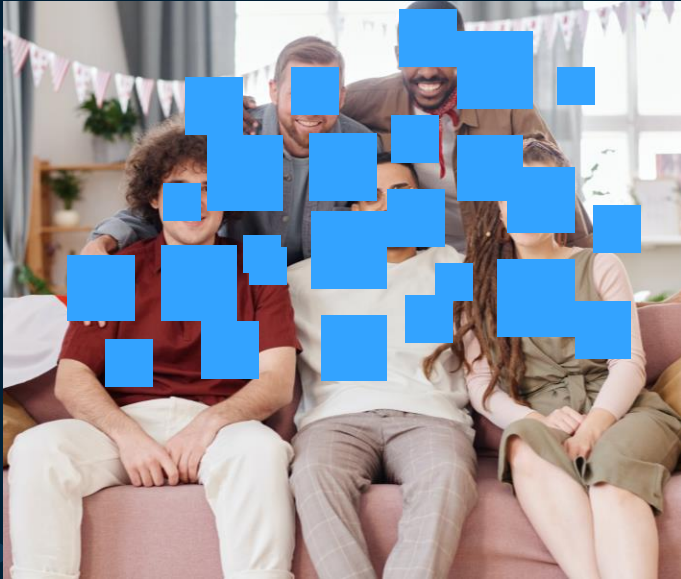
Marketing
Safety/PV
Formulary access
Subpopulations
Value

RCT vs. RWD

- Narrow population
 - Safety and efficacy
- All patients' experience
 - All patient care activities
- *Data sizes – millions of patients; years of data*
- *Data quality - no “gold standard” or CRF to resolve issues*
- *No enrollment date – index events*
- *No randomization – propensity score matching, etc.*
- *No defined follow-up intervals/activities – derive outcomes*

*Instead of complete information for hundreds of patients,
you have incomplete information for millions of patients*

We Can Infer from Incomplete Data If..



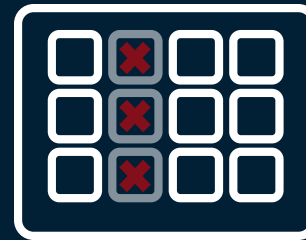
Photograph: Cedric Fauntleroy

- Mine data for millions of patients over time
- Mining through big data requires special considerations
- Requires close collaboration between data engineer, data scientist, clinical programmers and clinicians

Technical Approaches

Only Keep Data Required for Analyses

- Keep minimal columns and rows
 - Reduces I/O, memory use, computation time, etc.
 - Review data documentation
 - Include clinicians in discussions
 - Restrict by date or other factors



KEEP / DROP
SELECT

Technical Approaches

Considerations With VARCHAR

- Often DBMS ignores maximum length of VARCHAR columns
- Careful consideration is required pulling in those data into SAS environment (SAS trusts metadata)
- Don't panic if you see 32,767 length CHAR variables
 - `default_varchar_size`
 - Cast length using explicit passthrough

Technical Approaches

Exploratory Data Analysis

- Examine values, ranges and missingness of key variables
 - PROC SQL vs. PROC FREQ
 - Representative sample vs. full data

```
select distinct num_rows,  
               count(*) as freq  
  from (select distinct patient_id,  
                    count(*) as num_rows  
        from join_result)  
 group by num_rows  
 order by num_rows;
```

Technical Considerations

Missing or Erroneous Values

There will be bad values, regardless of source

How to address?

- Omit value / set to missing
- Impute value
- Omit observation
- Omit patient

Decisions should be:

- Made as team
- Documented
- Be reflected appropriately in outputs

Technical Approaches

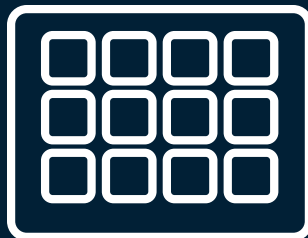
Perform Efficient Sampling

- Patient identifiers are typically strings of numeric characters and are in every file/table
- Use last one or two characters of patient identifier to sample
 - One character = 10% sample; two characters = 1%
 - Rightmost characters can usually be assumed to be random
- Guaranteed reproducibility of sample; easy to make different samples
- No joins or merges required to find data across tables
- Use samples to test code and perform exploratory work

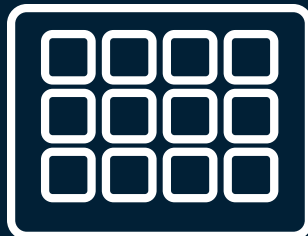
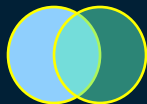
Technical Approaches

Market Definition Tables to Subset Data

Market	NDC_Code
Diabetes	00024592605
Diabetes	50904750000
Diabetes	73070010011
Diabetes	



Market	ICD_10
Diabetes	E08
Diabetes	E09
Diabetes	E10
Diabetes	



AND

Study Start Date $\leq$ Service Date $\leq$ Study End Date

Technical Approaches

Leverage SQL Pass-Through

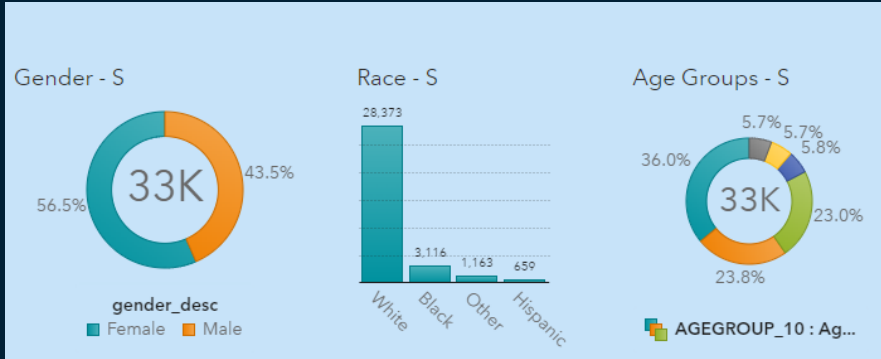
```
Connect to oracle as myconn (user=smith password=secret  
Path='myoracleserver');
```

```
Select *  
    from connection to myconn  
        (select patient_id, drug_id, servc_date from  
t_rx where servc_date < to_date('31-Dec-2022', 'DD-MON-YY') );
```

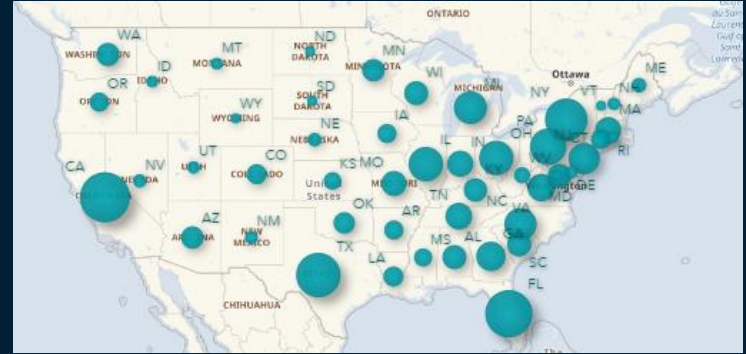
```
Disconnect from myconn;  
quit;
```

Technical Approaches

Interactive Visualization of Data – Speed to Insight



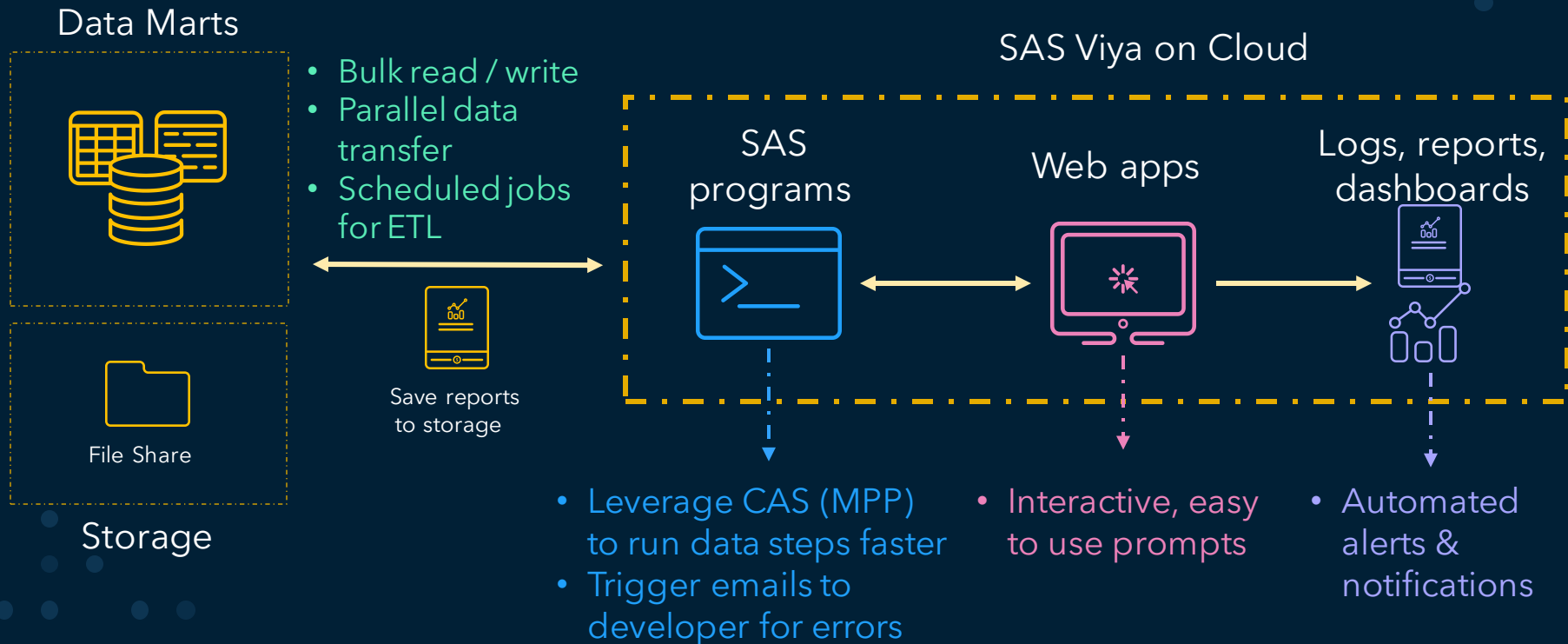
Cohort Demography



Geographic Distribution

Analytic Use Cases

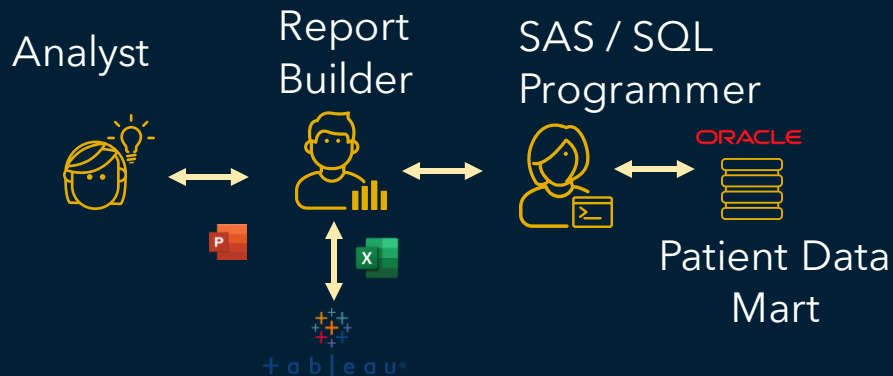
Adverse Event Signal Detection



Analytic Use Cases

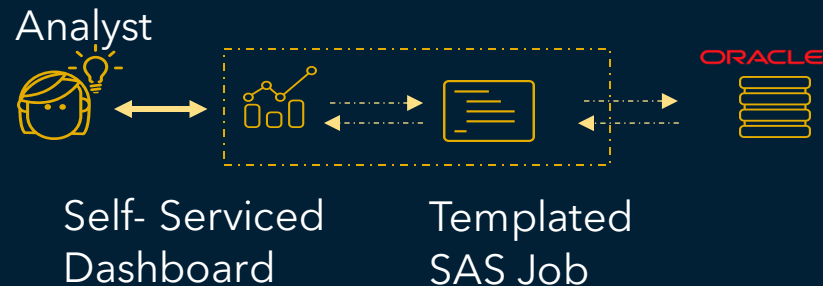
Self-Service Cohort Analysis

Before



- Multiple platforms
- Programmers spending time for repetitive work
- Longer turnaround time
- Multiple copies of data

Now



- One platform with cohort building functionality
- Programmers can focus on innovation
- Faster turnaround time
- Fewer tools

Graphical Cohort Query Tool

The screenshot displays the SAS Health Cohort Builder interface. The top blue header bar contains the text "SAS® Health Cohort Builder" and a search icon. On the left, a sidebar lists various data categories: Demographics, Diagnoses (selected), CCS Codes, ICD-10CM Diagnosis Code, ICD-9CM Diagnosis Codes, MEDDRA Codes, SNOMED CT Codes, Diagnosis Related Groups, EPIC, Micromedex, Payments, Prescriptions, Procedures, and Revenue Codes. The main workspace is titled "Expressions > Diabetes Diagnosis" and includes a "Filter" input field. Below the filter, a tree view shows the "Diabetes Diagnosis" expression with a green checkmark. Under this expression, the "Codes" section is expanded, showing a table of ICD-10CM codes. The table has columns for Code, Description, Type, and Vocabulary. The "Occurrences" section is also expanded, showing a "Number of occurrences" set to 1 with a dropdown menu. A "Specify code location" dropdown is set to "Primary". A "Close" button is visible in the top right corner of the workspace.

Expressions > Diabetes Diagnosis

Filter

- Demographics
- Diagnoses
 - CCS Codes
 - ICD-10CM Diagnosis Code
 - ICD-9CM Diagnosis Codes
 - MEDDRA Codes
 - SNOMED CT Codes
- Diagnosis Related Groups
- EPIC
- Micromedex
- Payments
- Prescriptions
- Procedures
- Revenue Codes

Diabetes Diagnosis

Codes

Label: Diabetes Diagnosis (based on "Diagnoses")

Code	Description	Type	Vocabulary
E08	Diabetes mellitus d...	DX	ICD10CM
E09	Drug or chemical in...	DX	ICD10CM
E10	Type 1 diabetes me...	DX	ICD10CM
E13	Other specified dia...	DX	ICD10CM

Specify code location: Primary

Occurrences ⓘ

Number of occurrences: >= 1

Use earliest date identified

Close

Analytic Use Cases

Trial Feasibility



- Explore actual populations and their behavior/outcomes
- Test different hypotheses rapidly
- Apply predictive analytics to shorten the time from hypothesis generation to optimized trial design
- Use of RWD to optimize eligibility criteria and recruitment will become the “new normal” across the clinical trials enterprise

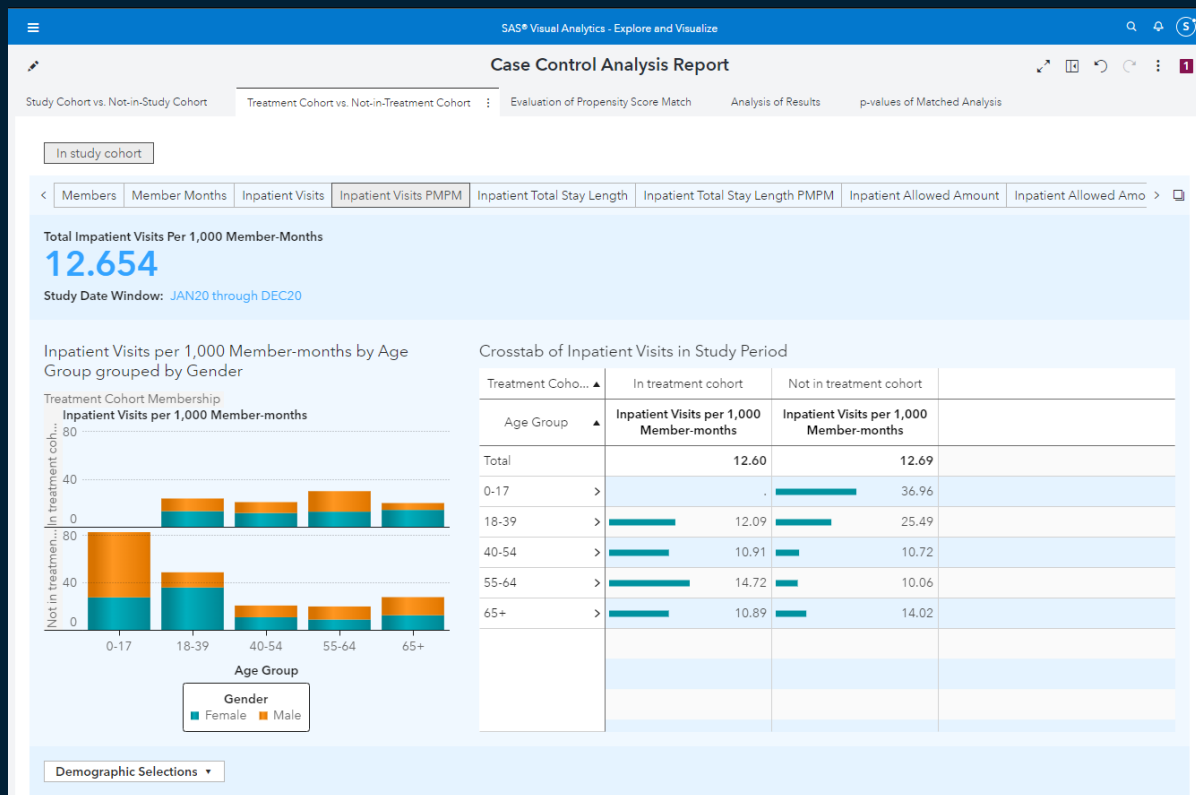
Analytic Use Cases

Synthetic or External Control Arms



Analytic Use Cases

Population Health / Disease Management



Analytic Use Cases

Other Examples

- Safety / pharmacovigilance
- Other patient outcomes
 - QoL, economic
 - Comparative effectiveness
 - Identify early / targeted interventions
 - Persistence / compliance / switching
- Expanded or new indications / drug repurposing
- Market performance
 - Establish value / gain formulary access
 - Accessibility / supply chain

A series of horizontal bars of varying lengths and colors (teal, blue, green) on the left side of the slide.

Samiul.Haque@sas.com
Robert.Collins@sas.com

sas.com

