



Methods for Examining Data Quality in Healthcare Integrated Data Repositories

Data Quality Checking and Validation in Distributed Health Data Networks

**Pacific Symposium on Biocomputing
Big Island of Hawaii**

January 4, 2018

Jeffrey Brown, PhD





Conflicts and Disclosures

I have no conflicts of interest related to this presentation.

I am currently funded by FDA, NIH, the Biologics and Biosimilars Collective Intelligence Consortium, Pfizer, PCORI, IBM, and Roche.



Outline

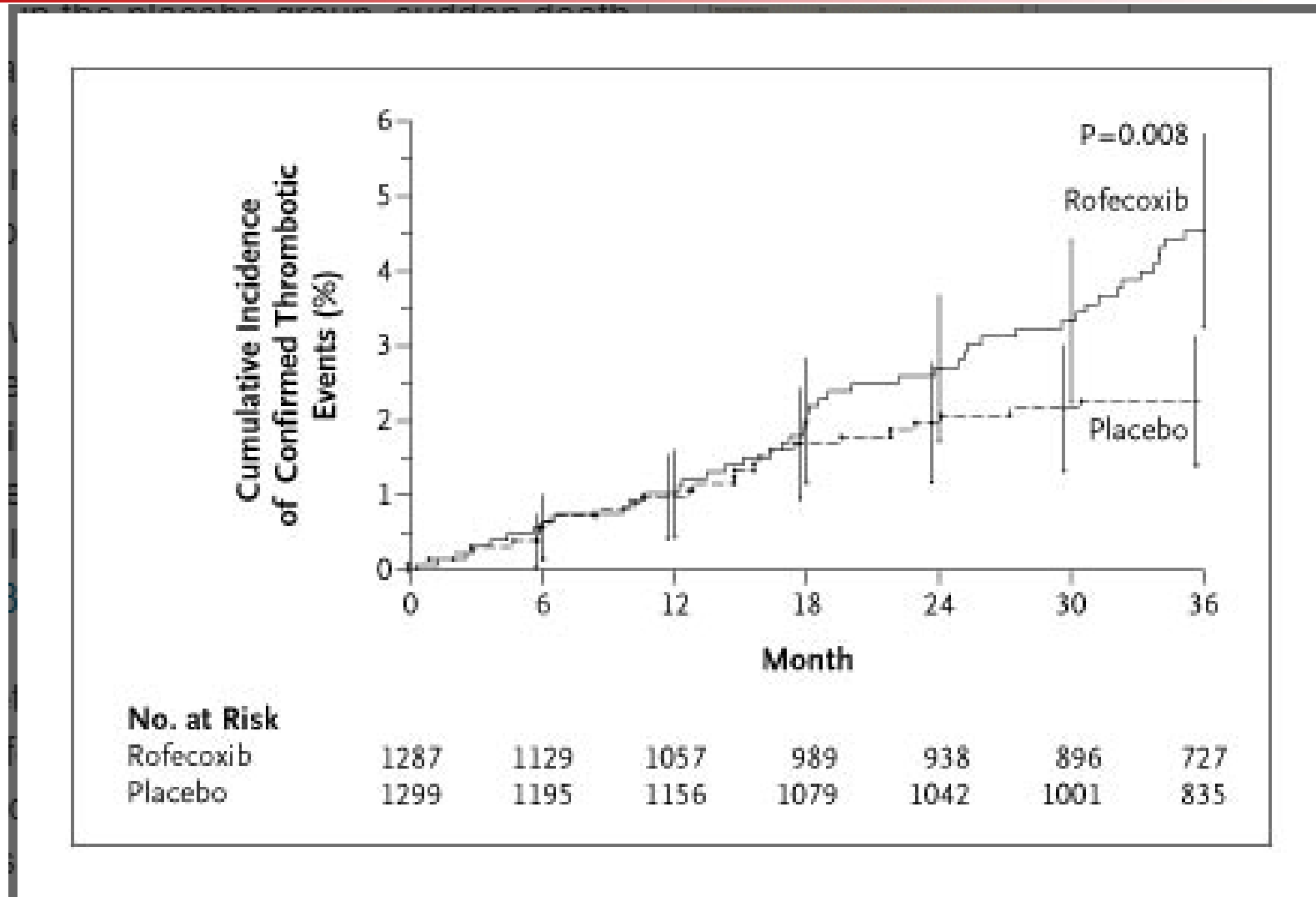
- Need for multisite studies and distributed networks
- FDA Sentinel project as worked example
- Q&A



Summary

- Advanced analytics need stable, well curated, and well characterized data
- Creating and maintaining stable, well curated, and well characterized data is hard and expensive
- Applying work across institutions makes it even harder
- But it can be done





Kaplan–Meier Estimates of the Cumulative Incidence of Confirmed Serious Thrombotic Events.

Bresalier RS, et al. *N Engl J Med.* 2005 Mar 17;352(11):1092-102.



We could have known earlier

PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2007; **16**: 1275–1284

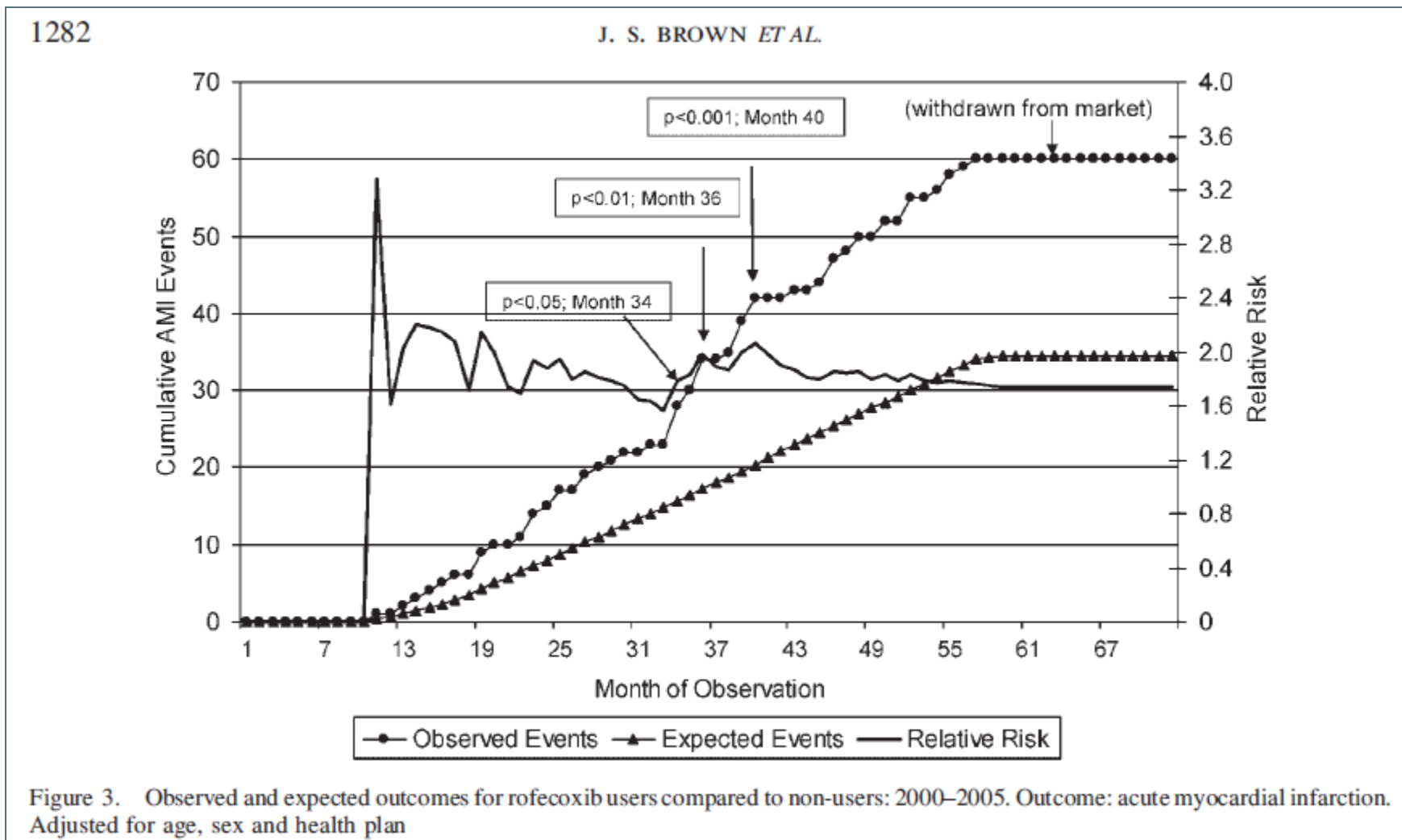
Published online 22 October 2007 in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/pds.1509

ORIGINAL REPORT

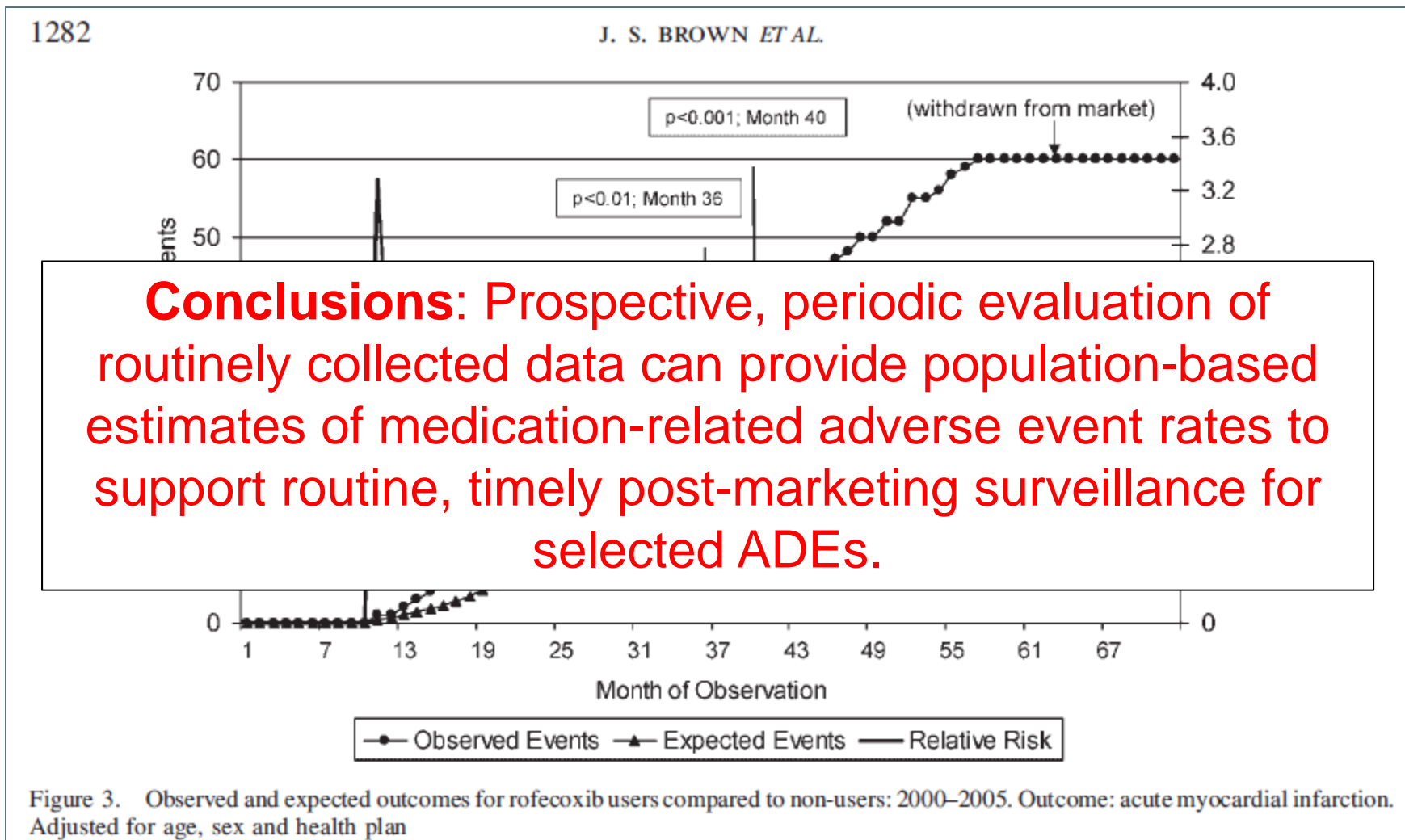
Early detection of adverse drug events within population-based health networks: application of sequential testing methods^{†,‡}

Jeffrey S. Brown PhD^{1,2*}, Martin Kulldorff PhD¹, K. Arnold Chan MD, MPH, ScD^{3,4}, Robert L. Davis MD, MPH⁵, David Graham MD⁶, Parker T. Pettus MS^{1,2}, Susan E. Andrade ScD^{2,7}, Marsha A. Raebel PharmD^{2,8}, Lisa Herrinton PhD^{2,9}, Douglas Roblin PhD^{2,10}, Denise Boudreau PhD^{2,11}, David Smith PhD^{2,12}, Jerry H. Gurwitz MD^{2,7}, Margaret J. Gunter PhD^{2,13} and Richard Platt MD, MSc^{1,2}

We could have known earlier



We could have known earlier



FDA Sentinel System: Background

- **2007: FDA Amendments Act**
 - A mandate to create an active surveillance system
 - Access data from **25 million** individuals by July 2010
 - Access data from **100 million** individuals by July 2012

- 2008: FDA launched the Sentinel Initiative
- 2009: Mini-Sentinel funded under Sentinel Initiative
- 2010: PRISM incorporated into Mini-Sentinel
- 2014: Funding awarded for Sentinel System
- *Operates under FDA's public health authority*




Sentinel is a National Medical Product Monitoring System

LEARN MORE



ABOUT

- Background
- Coordinating Center
- Privacy and Security
- The Sentinel System Story
- Reagan-Udall Foundation and IMEDS



MEDICAL PRODUCT ASSESSMENTS

- Active Risk Identification and Analysis System
- Ongoing ARIA Assessments
- Assessments of Drugs
- Assessments of Vaccines, Blood, & Biologics
- FDA-Catalyst



DATA & SURVEILLANCE TOOLS



COMMUNICATIONS

Latest Postings

SPOTLIGHT

- CDER Conversation: The FDA's Sentinel Initiative
Mon, 11/27/2017

PUBLICATIONS AND PRESENTATIONS

- Development of Metrics to Assess Appropriate Prescribing of Opioids in the Mini-Sentinel Distributed Database (MSDD)
Mon, 11/20/2017
- Prospective Postmarketing Surveillance of Acute Myocardial Infarction in New Users of Saxagliptin: A Population-Based Study
Fri, 11/10/2017
- Safety Assessment of Niacin in the U.S. Food and

<https://www.sentinelinitiative.org/>

Jeff_Brown@harvardpilgrim.org



Sentinel as a distributed data network

- Rare exposures
- Rare outcomes
- Sample size (speed)
- Sub-group analyses
- Analytic flexibility



What is a distributed data network?

Coordinating Center



These institutions have no interest in sharing data with each other



What is a distributed data network?

Coordinating Center

Send question to the data



These institutions have no interest in sharing data with each other

What is a distributed data network?

Coordinating Center

Return answers (not data)



These institutions have no interest in sharing data with each other

Characteristics of distributed networks

- Data sits behind data partner's firewall
- Data remains under local control
- Only minimally necessary info is shared in a given analysis
- Preserve patient privacy & institutional proprietary interests
- Enables rapid creation of multiple networks that leverage the architecture
- Avoids complex contracting and institutional agreements

Data networks have different goals

- Exchange of patient data for patient care at the point of care
- Public health surveillance
- Research
- Clinical trial planning and enrollment

Keep in mind: These goals have different data quality requirements



Data networks introduce complexity

- Data access approach
- Data standardization
- Data quality
- Query standardization
- Governance and policy
- Privacy and security
- Trust

FDA Sentinel's charge

Assess the use, safety, and effectiveness of regulated medical products by using electronic healthcare data plus other resources

Create data, informatics, and methodologic capabilities to support these activities

Quickly!



Sentinel partner organizations

DEPARTMENT OF POPULATION MEDICINE

Lead – HPHC Institute



HARVARD MEDICAL SCHOOL



Harvard Pilgrim Health Care Institute

Data and scientific partners



Hospital Corporation of America™



Scientific partners



America's Health Insurance Plans



Institute for Health

Sentinel distributed database*

- Populations with well-defined person-time for which most medically-attended events are known
- 425 million person-years of observation time
- 43 million people currently accruing new data
- 5.9 billion pharmacy dispensings
- 7.2 billion unique medical encounters
- 42 million people with at least one laboratory test result

<https://www.sentinelinitiative.org/sentinel/snapshot-database-statistics>

* As of January 2017



Sentinel Common Data

Medical Encounters					
Enrollment	Demographic	Dispensing	Encounter	Diagnosis	Procedure
Person ID	Person ID	Person ID	Person ID	Person ID	Person ID
Enrollment start & end dates	Birth date	Dispensing date	Service date(s)	Service date(s)	Service date(s)
Drug coverage	Sex	National drug code (NDC)	Encounter ID	Encounter ID	Encounter ID
Medical coverage	ZIP code	Days supply	Encounter type & provider	Encounter type & provider	Encounter type & provider
Medical record availability	Etc.	Amount dispensed	Facility	Diagnosis code & type	Procedure code & type
			Etc.	Principal discharge diagnosis	Etc.

Clinical		Registry			Inpatient	
Lab Result	Vital Signs	Death	Cause of Death	State Vaccine	Inpatient Pharmacy	Inpatient Transfusion
Person ID	Person ID	Person ID	Person ID	Person ID	Person ID	Person ID
Result and specimen collection dates	Measurement date and time	Death date	Cause of death	Vaccination date	Administration date and time	Administration start and end date and time
Test type, immediacy & location	Height and weight	Source	Source	Admission Type	Encounter ID	Encounter ID
Logical Observation Identifiers Names and Codes (LOINC ®)	Diastolic & systolic BP	Confidence	Confidence	Vaccine code & type	National Drug Code (NDC)	Transfusion administration ID
Test result & unit	Tobacco use & type	Etc.	Etc.	Provider	Route	Transfusion product code
Etc.	Etc.			Etc.	Dose	Blood Type
					Etc.	Etc.



Data Validation within Research Networks:

From *Ad Hoc* Practice to System Practice

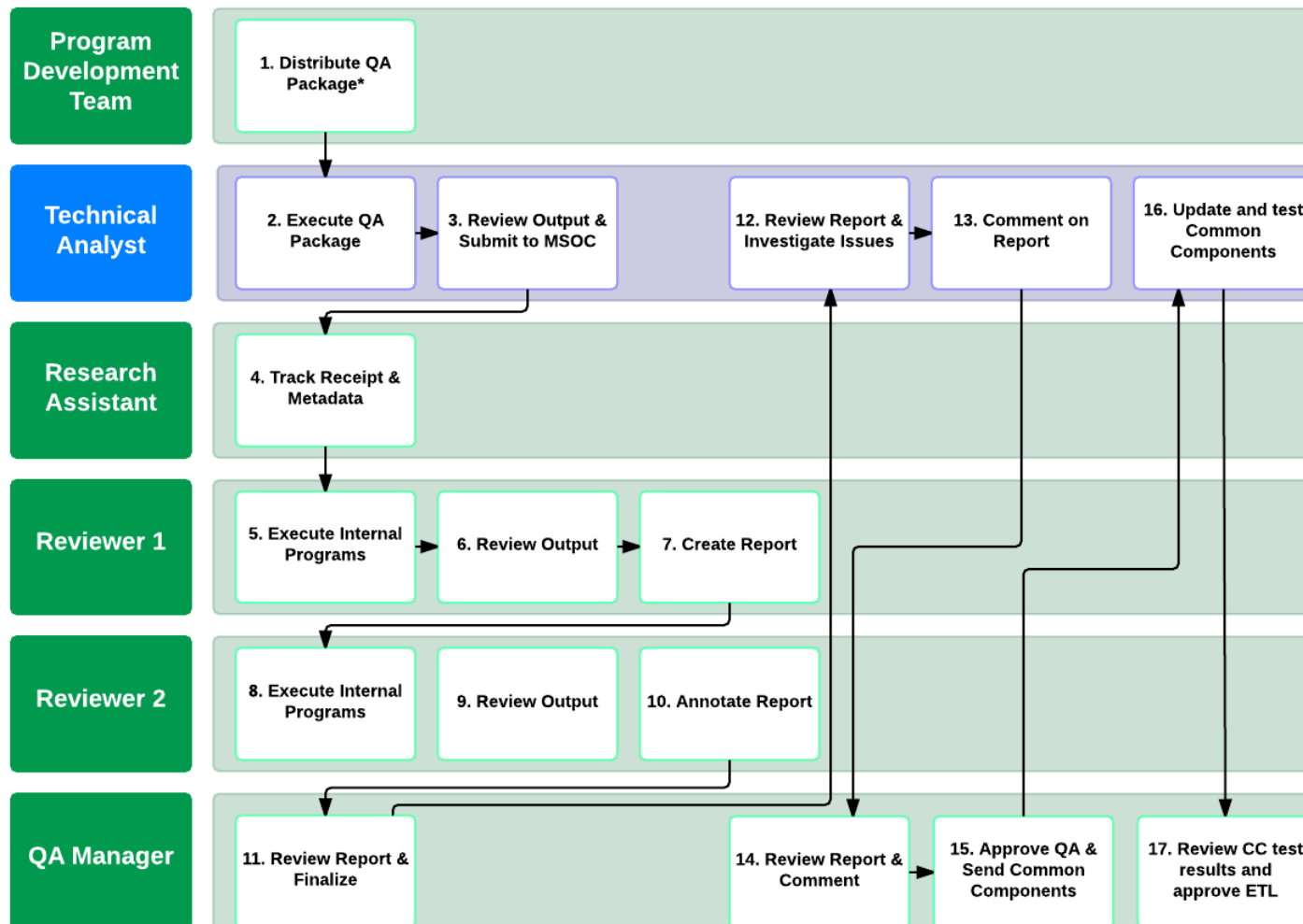
Study-specific versus network data validation approaches

Study	Network
“As needed / as you go”	“Always Ready / Semper Paratus”
Burden on study team	Burden on quality assurance team
<i>Ad hoc</i>	Repeatable, Systematic, Learning
Cost is included in the cost of a study	Cost of 0 studies = cost of 1000+ studies
Variable amount of data cleaning	1400+ checks to pass a site’s QA

Sentinel quality assurance avoids the costs and delays of having individual projects devote significant resources to data investigation and cleaning



Data quality assurance process



*Program Development Team Follows MS SAS Program Development SOP to Create QA Package



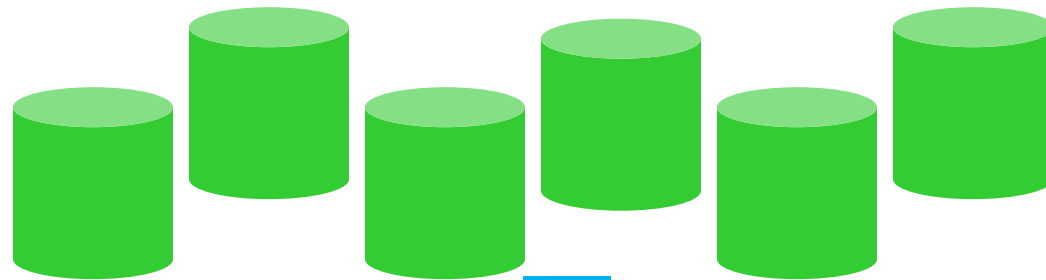
Data Partner



MSOC

Every Data Partner transforms their data into the Sentinel Common Data Model

Unique Data Partner's Source Database Structure



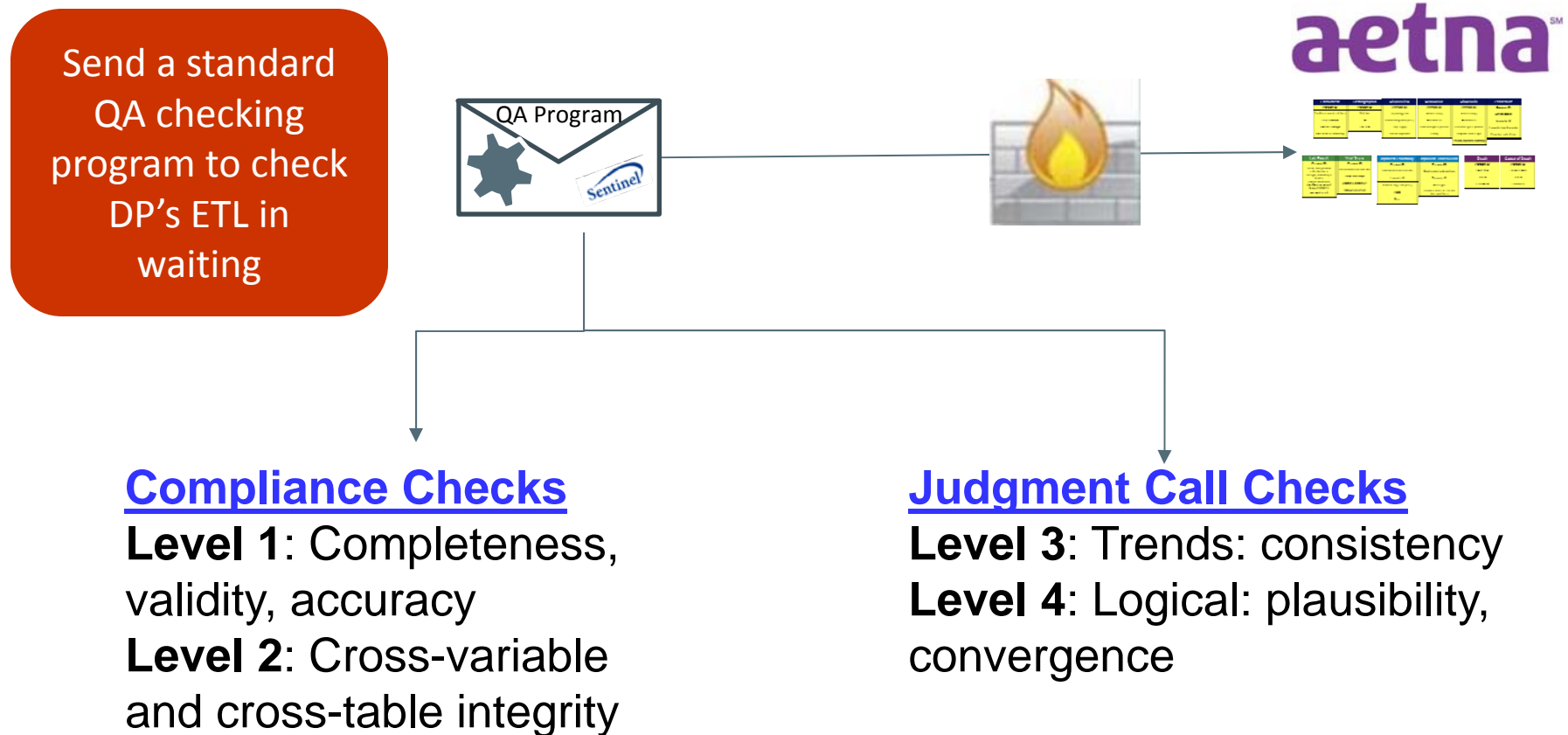
Data Partner's Database Transformed into SCDM Format (DP ETL)

Enrollment	Demographic	Dispensing	Encounter	Diagnosis	Procedure
Person ID	Person ID	Person ID	Person ID	Person ID	Person ID
Enrollment start & end dates	Birth date	Dispensing date	Service date(s)	Service date(s)	Service date(s)
Drug coverage	Sex	National drug code (NUC)	Encounter ID	Encounter ID	Encounter ID
Medical coverage	ZIP code	Days supply	Encounter type & provider	Encounter type & provider	Encounter type & provider
Medical record availability		Amount dispensed	Facility	Diagnosis code & type	Procedure code & type
				Principal discharge diagnosis	

Lab Result	Vital Signs	Inpatient Pharmacy	Inpatient Transfusion	Death	Cause of Death
Person ID	Person ID	Person ID	Person ID	Person ID	Person ID
Results and specimen collection dates	Measurements date and time	Administration date and time	Blood product code and type	Death date	Cause of death
Test type, immediacy & location	Height and weight	Encounter ID	Encounter ID	Source	Source
Logical Observation Identifiers Names and Codes (LOINC®)	Diastolic & systolic BP	National Drug Code (NUC)	Blood type	Confidence	Confidence
Test result & unit	Tobacco use & type	Route	Administration start and end dates and times		
		Dose			



The data validation process



What do the checks look like

ENC1.0.0	Table does not exist
ENC1.1.1	PatID variable is not character type
ENC1.1.2	PatID variable has missing values
ENC1.1.3	PatID variable has non-missing values that are not left-justified
ENC1.1.4	PatID variable contains special characters
ENC1.2.1	EncounterID variable is not character type
ENC1.2.2	EncounterID variable has missing values
ENC1.2.3	EncounterID variable has non-missing values that are not left-justified
ENC1.2.4	EncounterID variable contains special characters
ENC1.3.1	ADate variable is not SAS date value of numeric data type
ENC1.3.2	ADate variable is not of length 4
ENC1.3.3	ADate variable has missing values

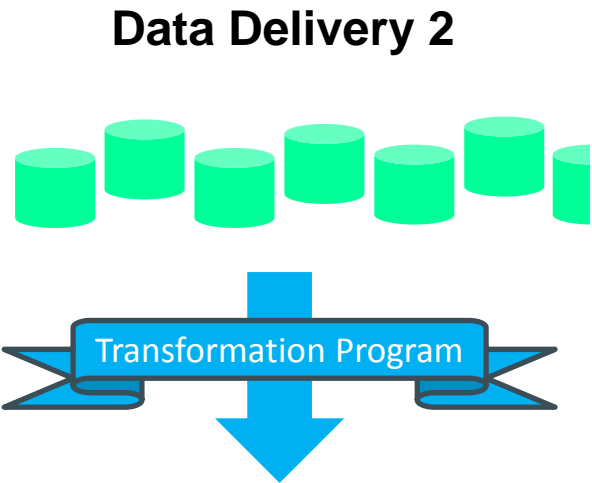
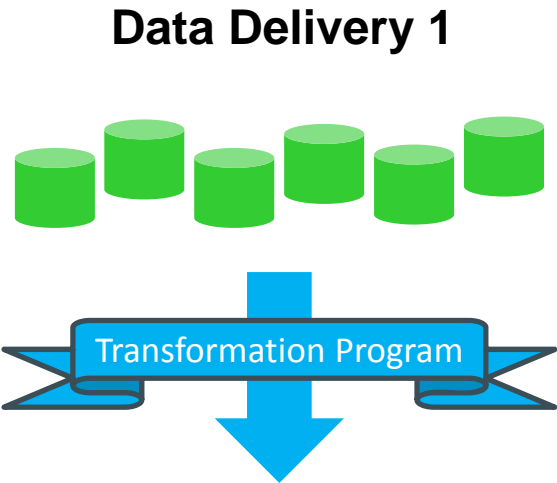
Standardized check codes

Check code: Table, Level, Variable Number, and Check Number

Check code “DEM1.3.2” denotes the second level 1 check performed on the variable SEX in the Demographic table

Recall: We have a dynamic database – new refreshes overwrite old data

Unique Data Partner Source Database Structure



Data Partner's Database Transformed into SCDM Format

Enrollment	Demographic	Dispositivity	Encounter	Diagnosis	Procedure
PERSON ID Conditions seen & met dates Drug message TREATMENT Current enrolment availability	PERSON ID DOB date Age JAF CODE Current enrolment availability	PERSON ID Shipping date HOSPITAL STAY CODE Phys visits Current enrolment availability	PERSON ID Encounter date ENCOUNTER TYPE Primary Current enrolment availability	PERSON ID Encounter date ICD9 CODE Diagnosis code & type ICD9CM CODE & TYPE Current enrolment availability	PERSON ID ENCOUNTER DATE ICD9 CODE Procedure code & type ICD9CM CODE & TYPE Current enrolment availability

Lab Result	Vital Signs	Immunization/Pharmacy	Regional Transportation	Death	Cause of Death
PERSON ID Result and specimen collection date Test type, laboratory & location LABORATORY REPORTING STATUS DATE LABORATORY ID Current enrolment availability	PERSON ID Components date and time Height and weight LABORATORY ID LABORATORY REPORTING STATUS DATE LABORATORY ID Current enrolment availability	PERSON ID Administration date and time LABORATORY ID National Drug Code (NDC) Drug Current enrolment availability	PERSON ID Mailed product code and type LABORATORY ID Mailed type HOSPITALIZATION DATE AND TIME Current enrolment availability	PERSON ID LABORATORY ID Cause Classification Current enrolment availability	PERSON ID LABORATORY ID Cause Classification Current enrolment availability

Enrollment	Demographic	Dispositivity	Encounter	Diagnosis	Procedure
PERSON ID Conditions seen & met dates Drug message TREATMENT Current enrolment availability	PERSON ID DOB date Age JAF CODE Current enrolment availability	PERSON ID Shipping date HOSPITAL STAY CODE Phys visits Current enrolment availability	PERSON ID Encounter date ENCOUNTER TYPE Primary Current enrolment availability	PERSON ID Encounter date ICD9 CODE Diagnosis code & type ICD9CM CODE & TYPE Current enrolment availability	PERSON ID ENCOUNTER DATE ICD9 CODE Procedure code & type ICD9CM CODE & TYPE Current enrolment availability

Lab Result	Vital Signs	Immunization/Pharmacy	Regional Transportation	Death	Cause of Death
PERSON ID Result and specimen collection date Test type, laboratory & location LABORATORY REPORTING STATUS DATE LABORATORY ID Current enrolment availability	PERSON ID Components date and time Height and weight LABORATORY ID LABORATORY REPORTING STATUS DATE LABORATORY ID Current enrolment availability	PERSON ID Administration date and time LABORATORY ID National Drug Code (NDC) Drug Current enrolment availability	PERSON ID Mailed product code and type LABORATORY ID Mailed type HOSPITALIZATION DATE AND TIME Current enrolment availability	PERSON ID LABORATORY ID Cause Classification Current enrolment availability	PERSON ID LABORATORY ID Cause Classification Current enrolment availability

Timeframe of Data Available in Database





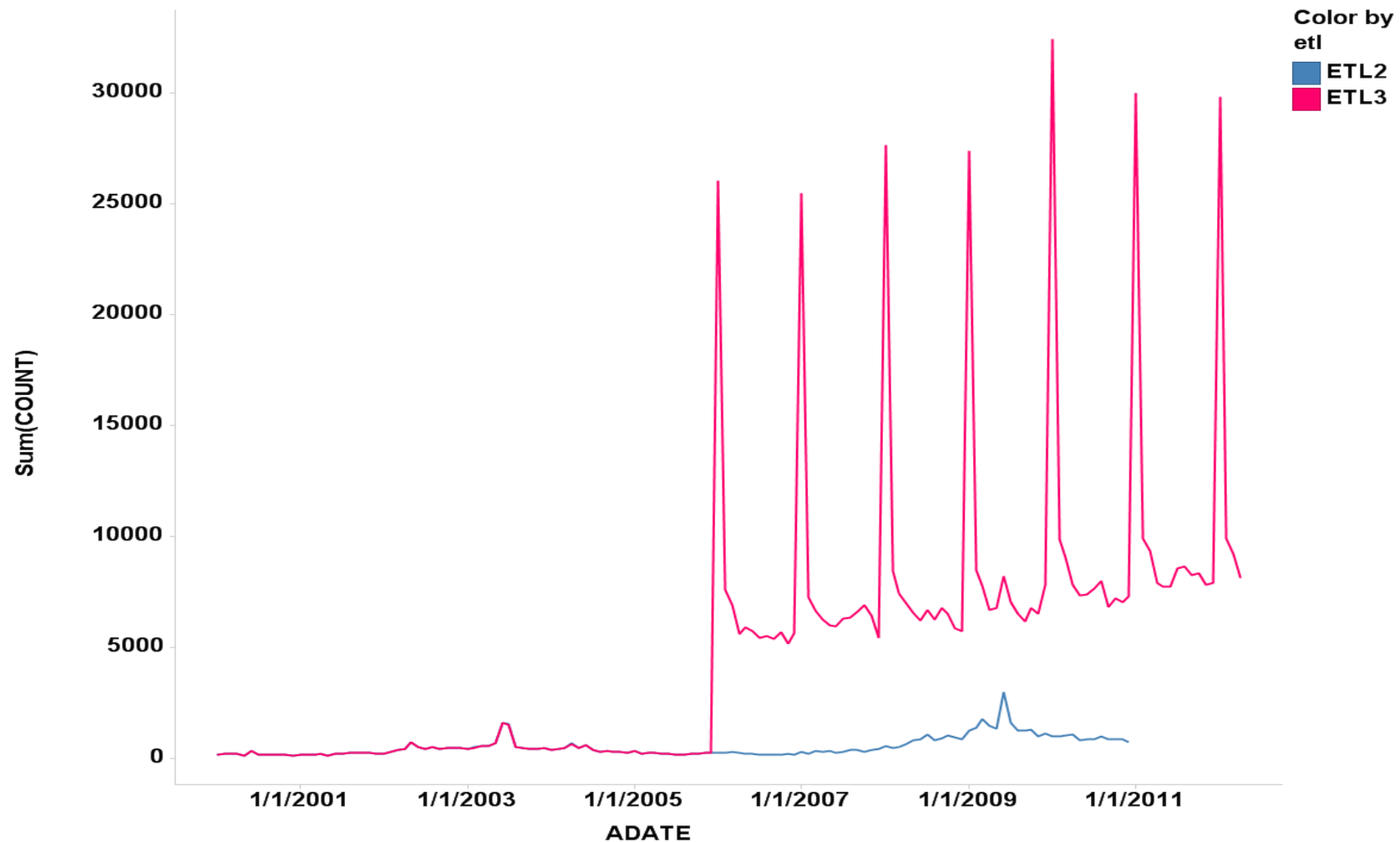
Why check after every refresh?

- Analytic tools depend on data model compliance
- Underlying data sources are dynamic
- Identify changes in trends, others issues or difference across sites
- Ongoing studies expect consistency in data refreshes

Communicate data validity findings with stakeholders



Visits per month, 2 refreshes



Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care* 2013 Aug;51(8 Suppl 3):S22-9.



Why check after every refresh?

- Analytic tools depend on data model compliance

Your really cool analytics won't work within your site, and especially across sites, unless the data are stable and well curated

refreshes

Communicate data validity findings with stakeholders

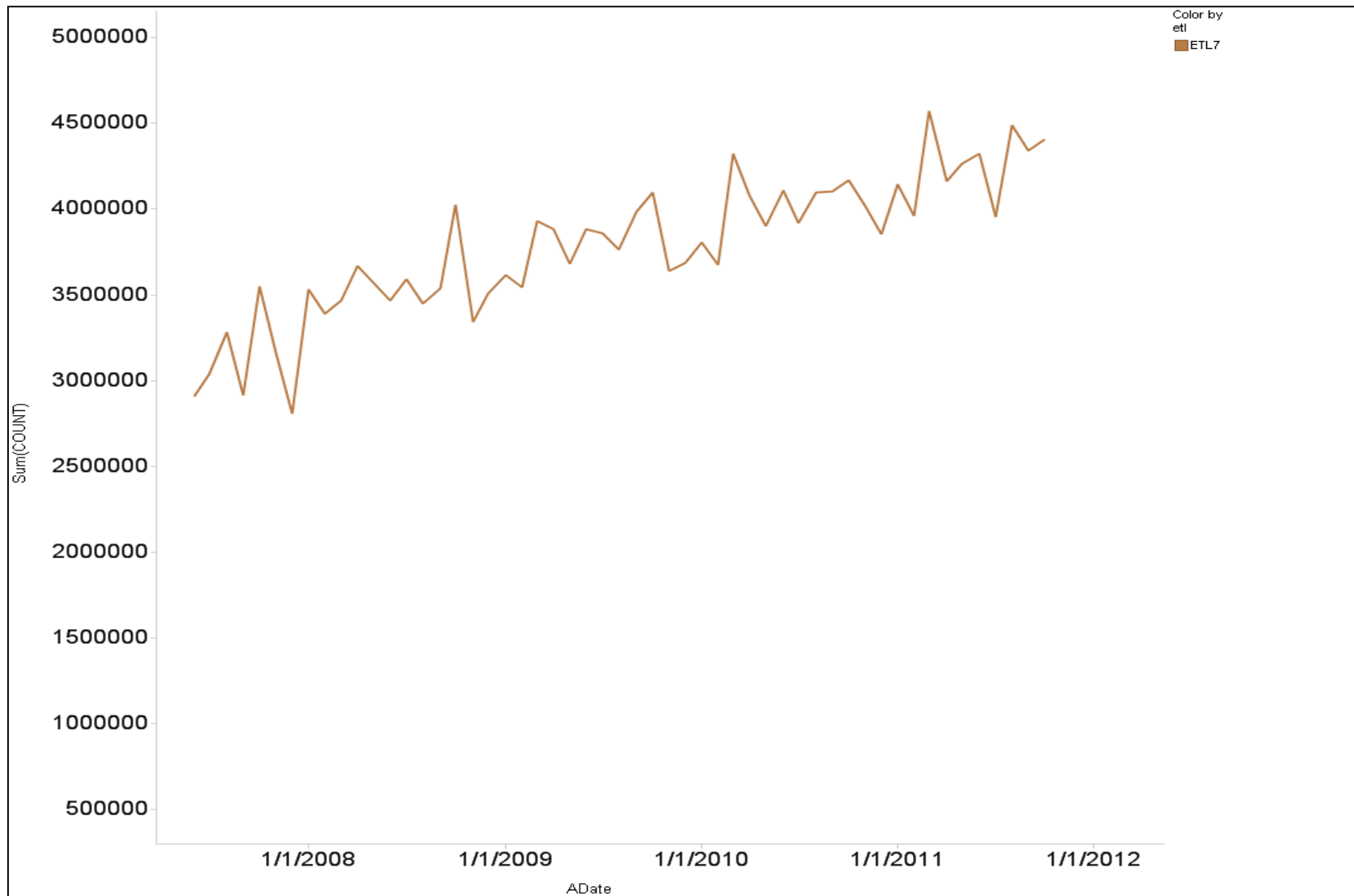
Admission and discharge date

Check distributions and patterns for significant changes

- Problem with distribution of ADate (e.g., records per year) within the ETL
- Problem with distribution of ADate (e.g., records per year-month) within the ETL
- Problem with distribution of ADate across ETLs
- Significant change in records per ADate (year) across ETLs
- Significant change in records per ADate (year-month) across ETLs
- Problem with distribution of DDate variable by encounter type per year-month
- Problem with distribution of length of stay (DDate-ADate + 1) by encounter type per year

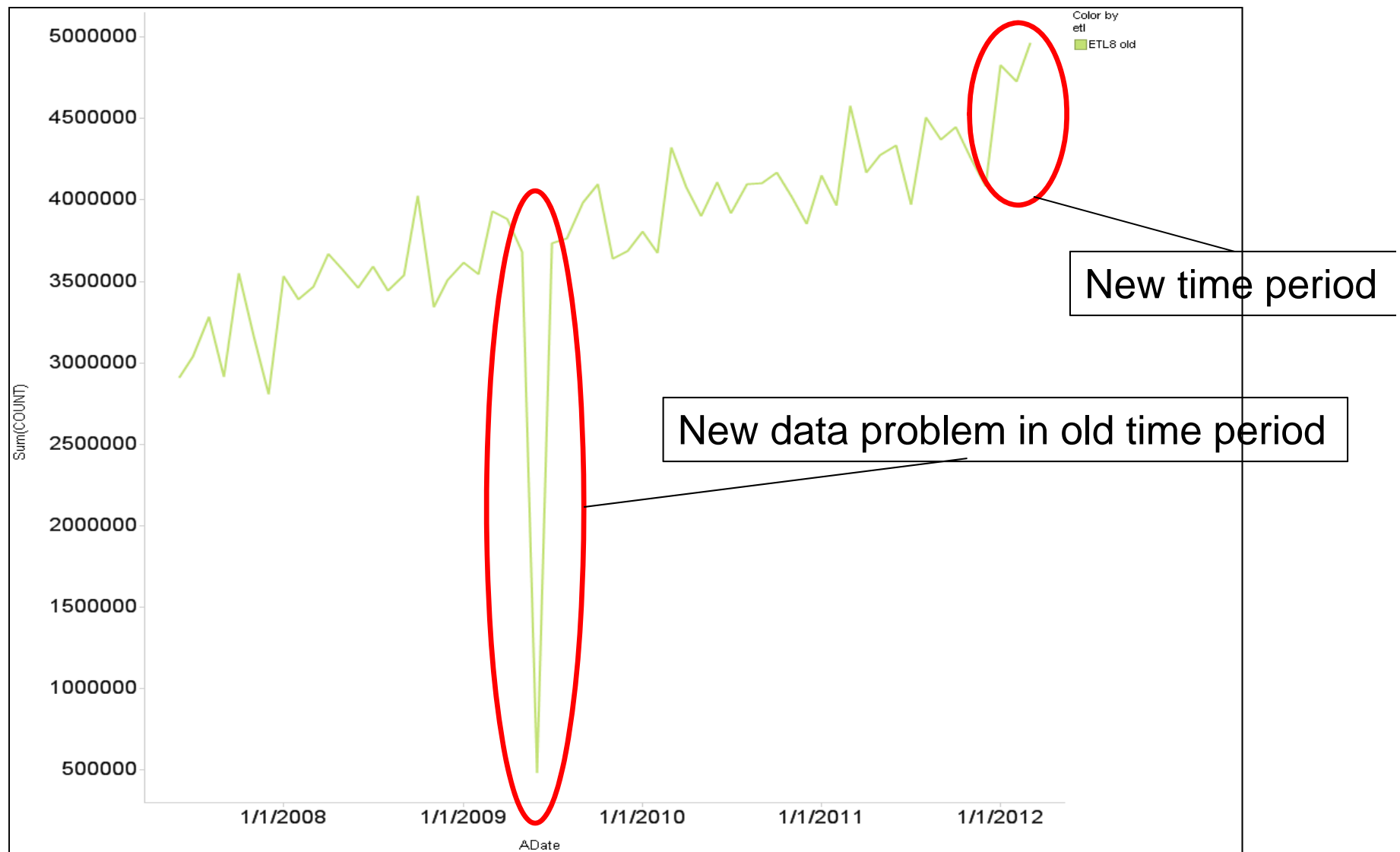


Data visualization: After 7th refresh, partner A



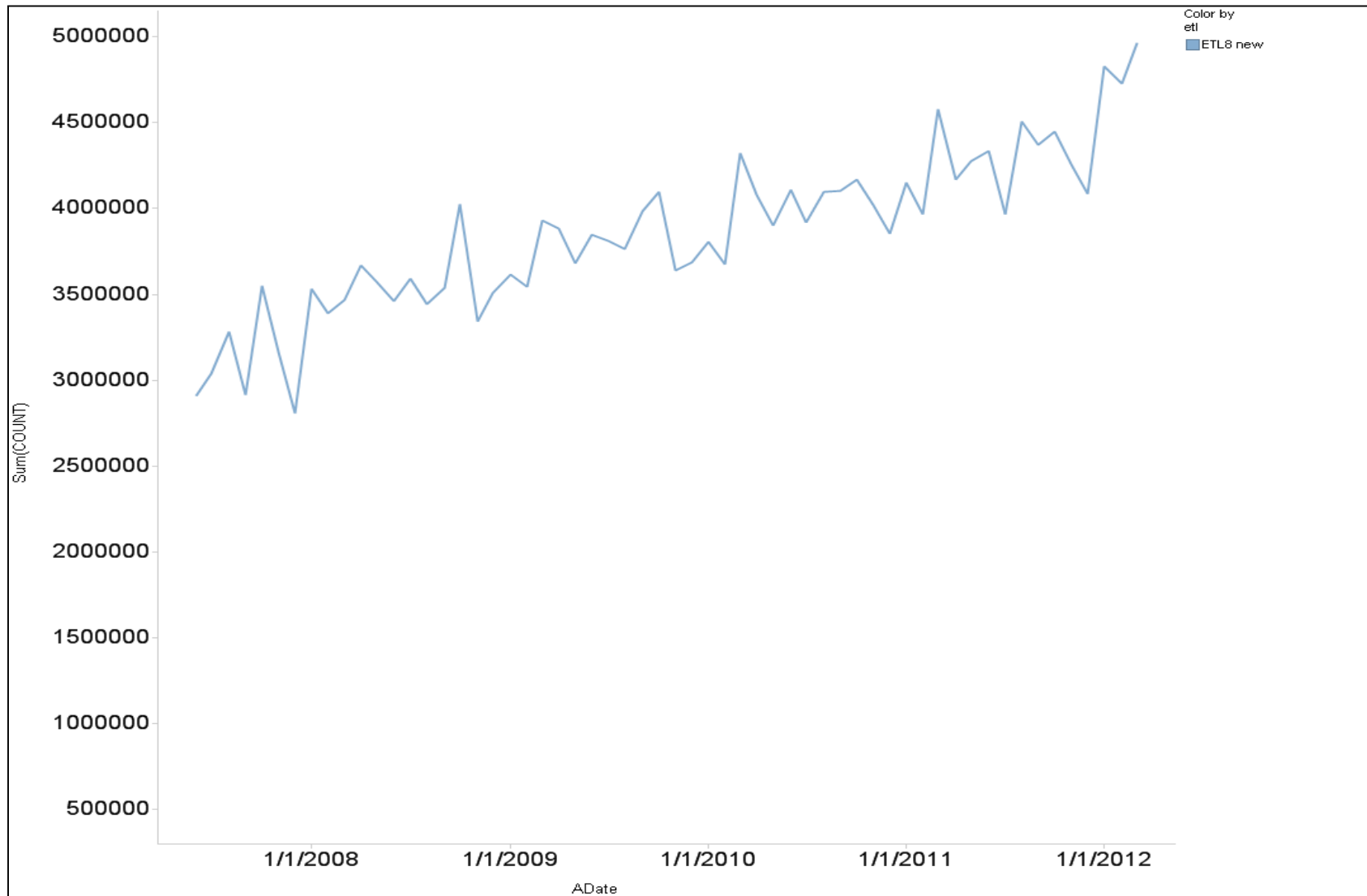


Data visualization: After 8th refresh, partner A



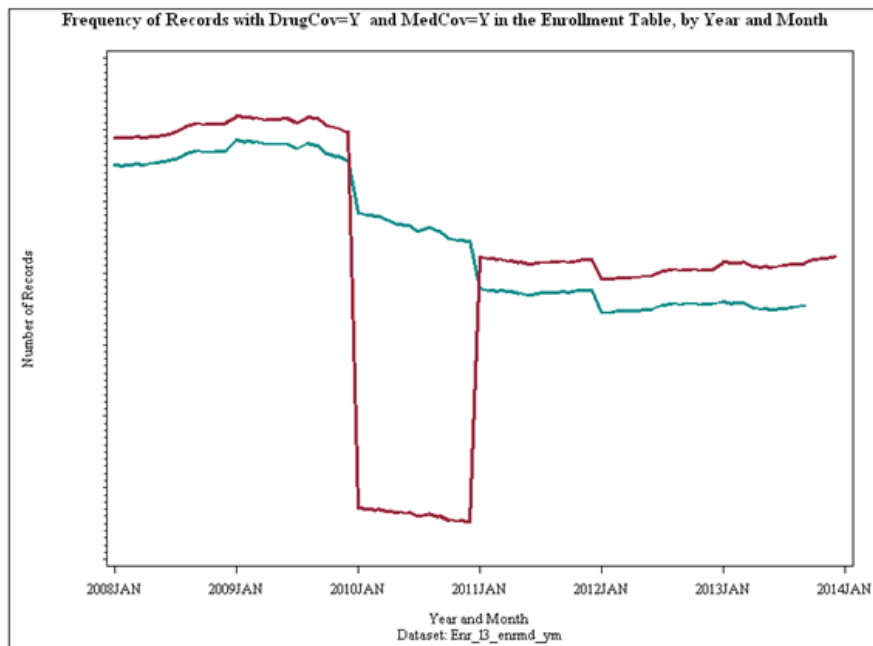


Data visualization: After 8th refresh, fixed

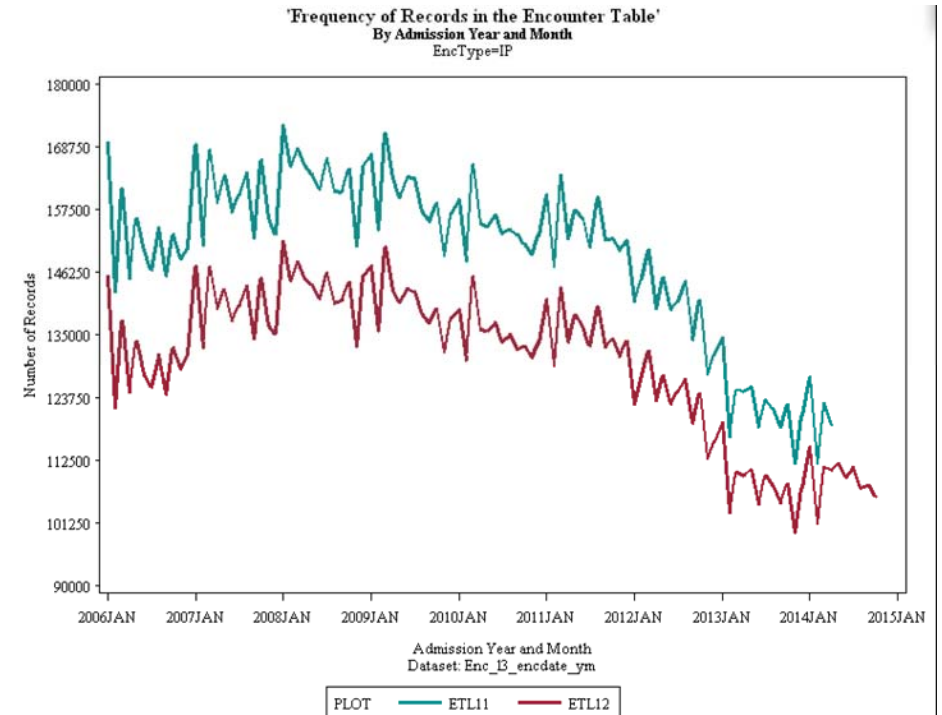




Consistency checks



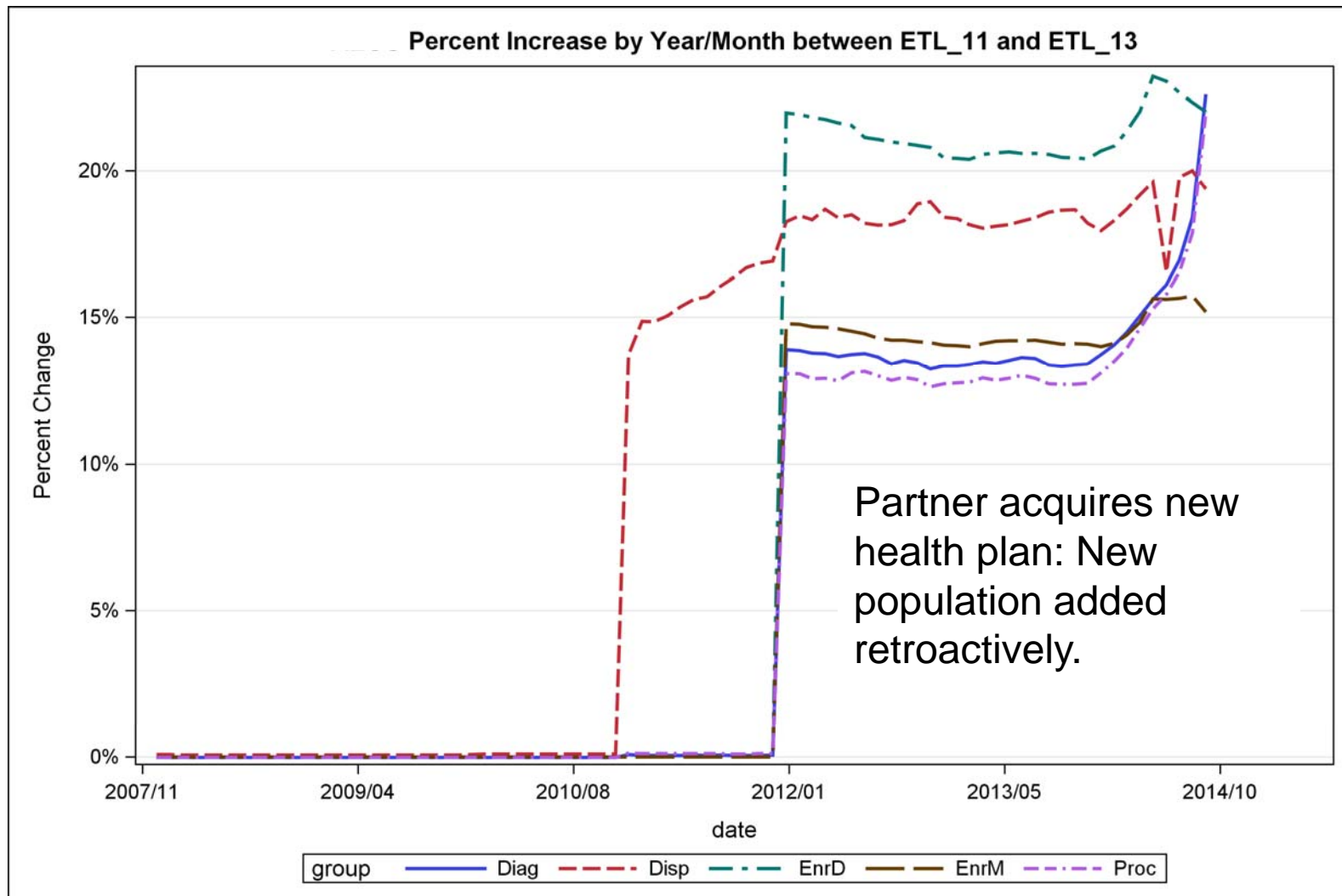
Incorrect Data Load



Reclassification of Encounter Type



Review identifies an anomaly





Platelet count units of measure across Sentinel

Platelet count original result units[‡]

Blank	FL	TH/UL	X10(3)
%	K/CMM	THOU/CMM	1000/UL
/100 W	k/cmm	thou/cmm	X10(3)/MCL
/CMM	K/CU MM	thou/mm ³	X10(3)/UL
CMM	K/CUMM	THOU/UL	X10(6)/MCL
10 3 L	K/MCL	THOUS/CU.MM	X10*9/L
10X3UL	K/mcL	THOUS/MCL	X10E3/UL
10^3/UL	K/UL	THOU/mcL	X1000
10*3/uL	k/uL	THOUS/UL	X10X3
10?3/uL	KU/L	Thou/uL	X10^3/UL
10E3/uL	K/MM3	THOUSA	x10
10e3/uL	K/mm3	THOUSAND	X10?3/ul
10e9/L	LB	THOUSAND/UL	X10E3/UL
E9/L	PLATELET CO	U	X10E3
BIL/L	T/CMM	X 10-3/UL	K/A?L
bil/L	TH/MM3	X 10(3)/UL	K/B5L
CU MM	th/mm ³	X10 3	

Raebel MA, Haynes K, Woodworth TS, Saylor G, Cavagnaro E, Coughlin KO, Curtis LH, Weiner MG, Archdeacon P, and Brown JS. Electronic Clinical Laboratory Test Results Data Tables: Lessons from Mini-Sentinel. *Pharmacoepidemiol Drug Saf.* 2014 Feb;23(6):609-18.



Observed result units for HbA1c across Sentinel

*Glycosylated hemoglobin (HbA1c) original result units**

%	%T.HGB	% TL HGB	% HGB
HEMOGLOBIN	%T.Hgb	% OF TOTAL	PERCENT
U	%T.Hgb	% of Hgb	Percent
%HB	% NGSP	% of total	HbA1c%
% OF T	%NGSP	%THb	%HbA1c
%A1C	% TOTAL HGB	%NGSP	% A1C
MG/DL	G/DL	mmol/mol [†]	Blank
% A1C	% A1c	%Hb	g/dL
NULL	%THb		

Raebel MA, Haynes K, Woodworth TS, Saylor G, Cavagnaro E, Coughlin KO, Curtis LH, Weiner MG, Archdeacon P, and Brown JS. Electronic Clinical Laboratory Test Results Data Tables: Lessons from Mini-Sentinel. *Pharmacoepidemiol Drug Saf.* 2014 Feb;23(6):609-18.



NEGATIVE	.	NEGATTVE
POSITIVE	820	NEGATVIE
UNDETERMINED	840	NEGAVTIV
BORDERLINE	1615	NEGITIVE
BORDERLI	ABNORMAL	NEGTIME
252.3	BOARDERL	NETGATIV
278	BODERLIN	NORM
28	CANCELLE	NORMAL
3178.2	DUPLICAT	POA
5 Int	EQIVOCAL	POPSITIV
DETECTED	EQUIVOCA	POSIIIV
INDETERM	NE-CHECK	POSITIFV
N	NEAGTIVE	POSITTVE
NOT DETE	NEG (-)	POSITIVE
Neg	NEGA	POSOTIVE
Negative	NEGA T I	POSTIVE
Negatvie	NEGA TIV	PSOITIVE
P	NEGAT IV	REPEAT
Positive	NEGATAIV	STAT
SPRCS	NEGATIAV	URINE
TNP	NEGATIBE	
N	NEGATIE	
Neg	NEGATRIV	
Negative		

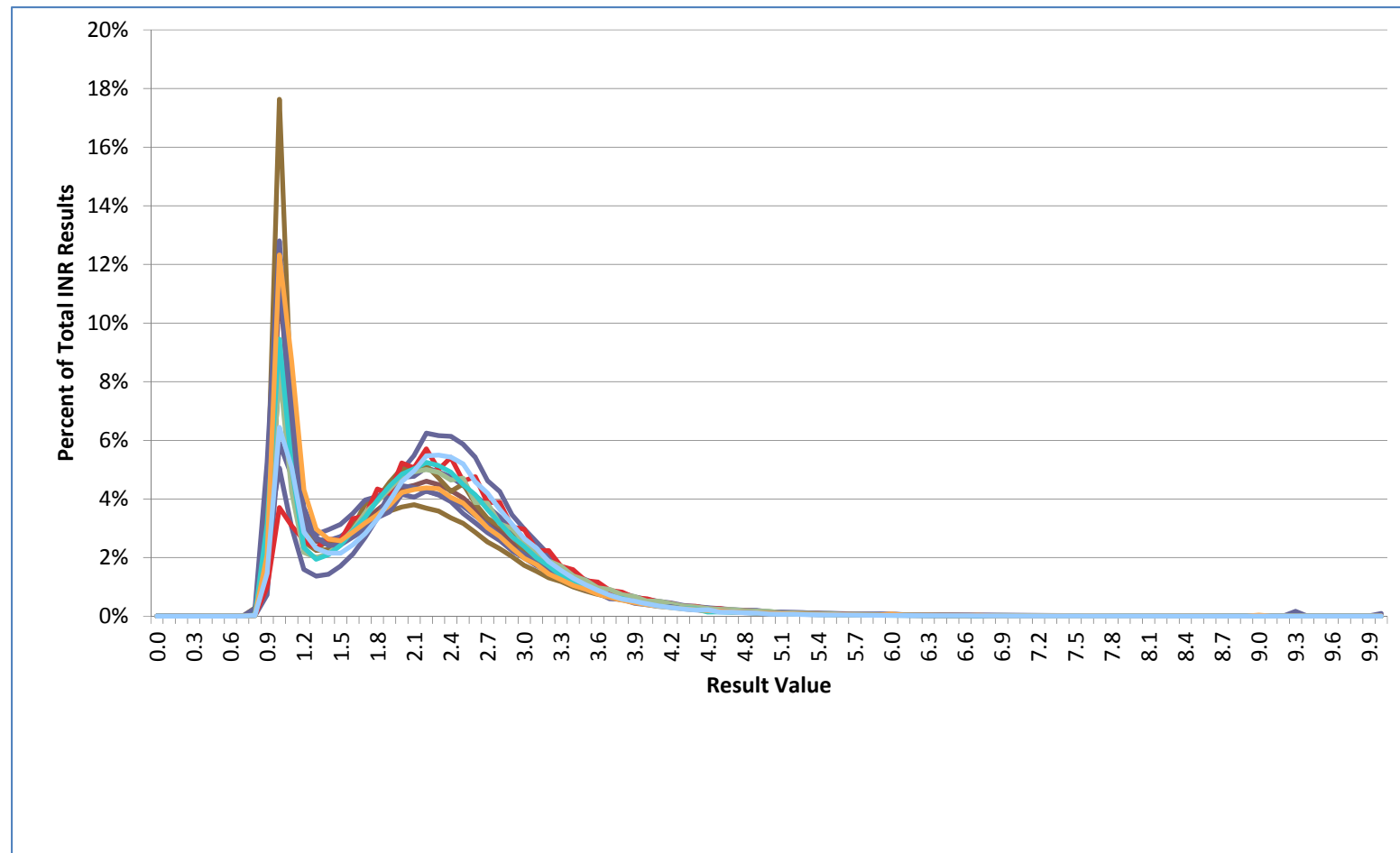
Examples of variations in qualitative pregnancy result units in source data across Sentinel

(I removed some rows...)



Standardizing clinical lab data

Percent of INR Results by Data Partner



Data validation statistics

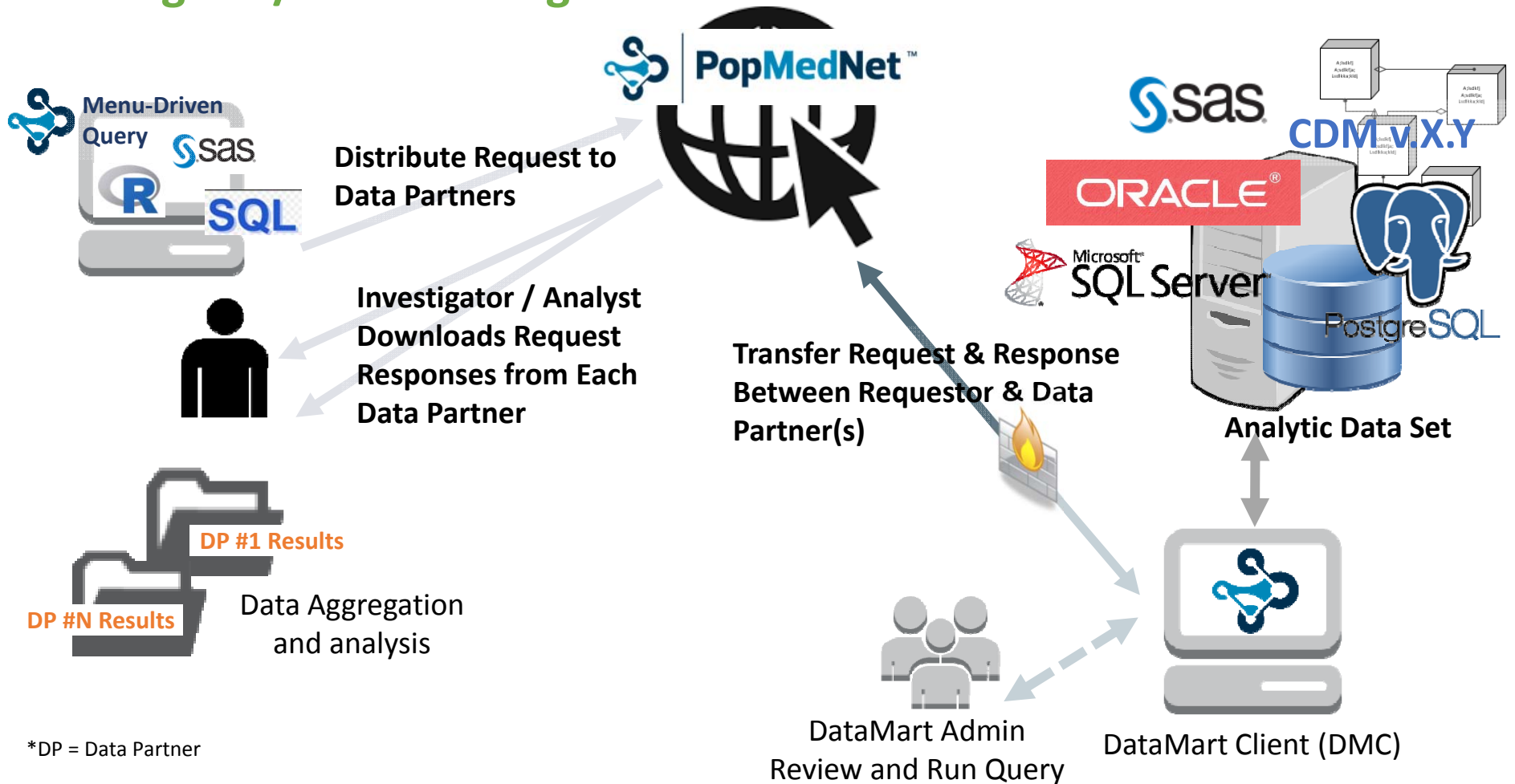
- Annually, the data quality assurance (QA) team reviews for over 50 data deliveries across the network
- Since 1/1/2016, a site has had to re-run the QA package in 16 instances to fix an issue
- In recent data deliveries from the 5 largest sites, 25 checks were reported in QA that required follow-up from the DP
 - 22 of the 25 were Level 3 checks



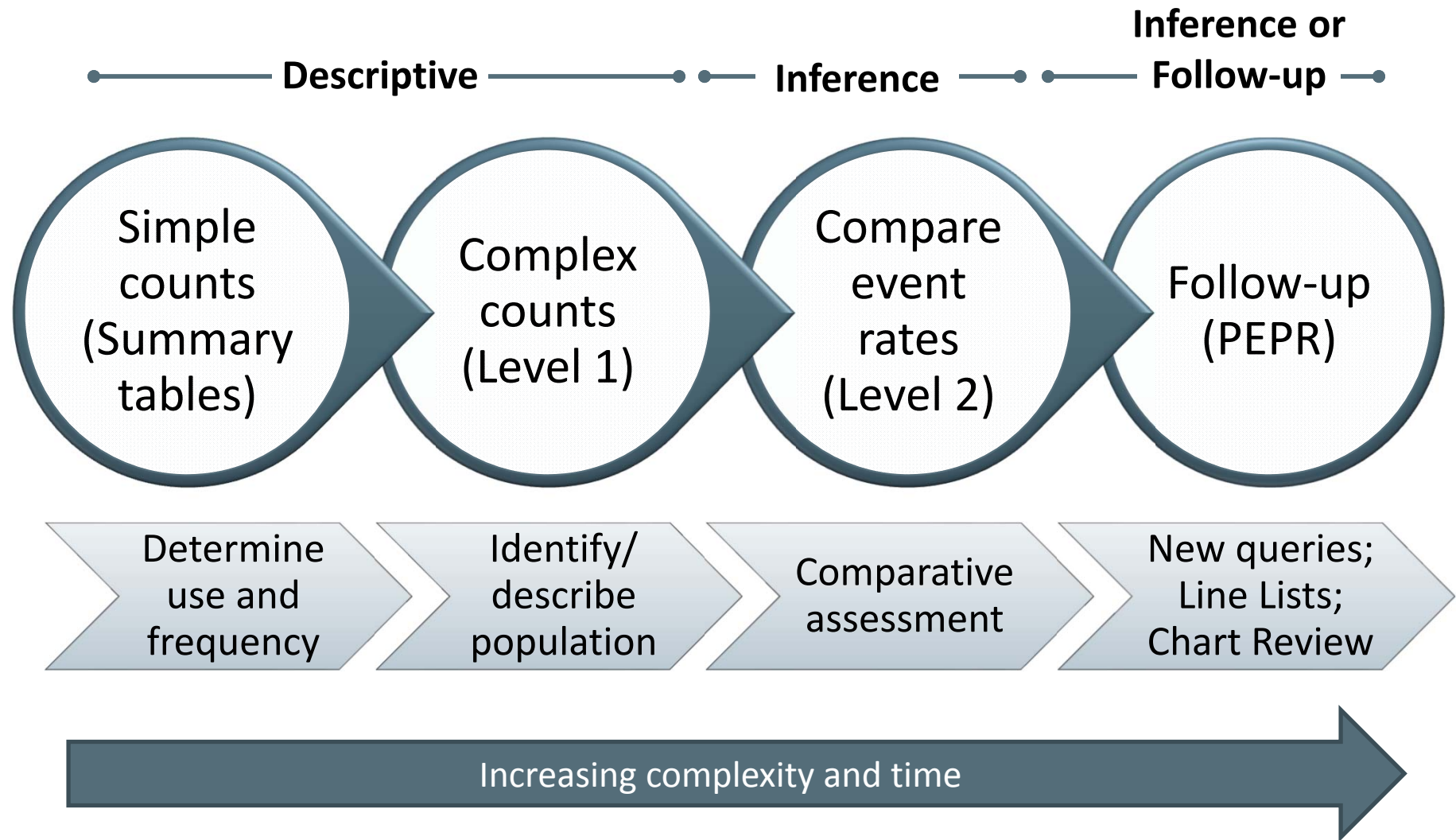
Distributed Querying Framework

Distributed querying

Investigator/Coordinating Center



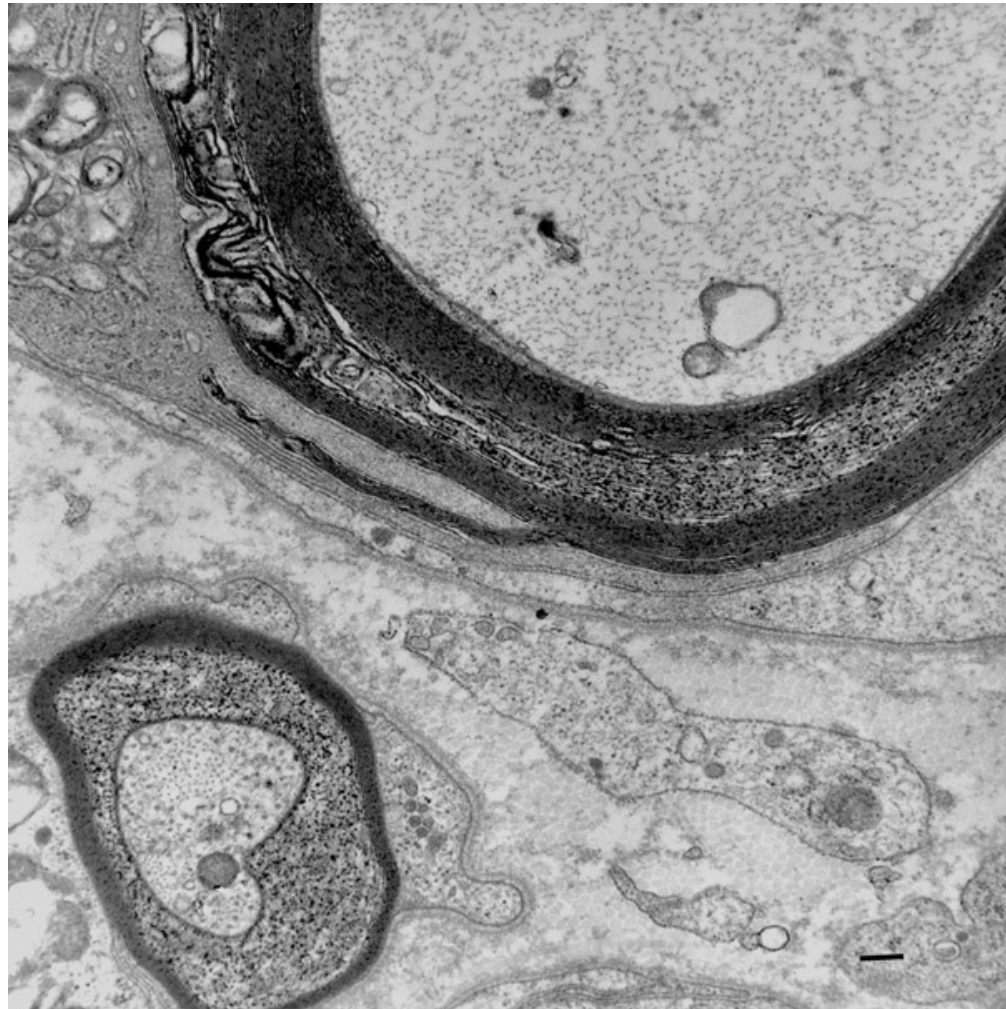
Rapid analysis querying sequence





**Every query includes detailed data
quality assurance steps and output**

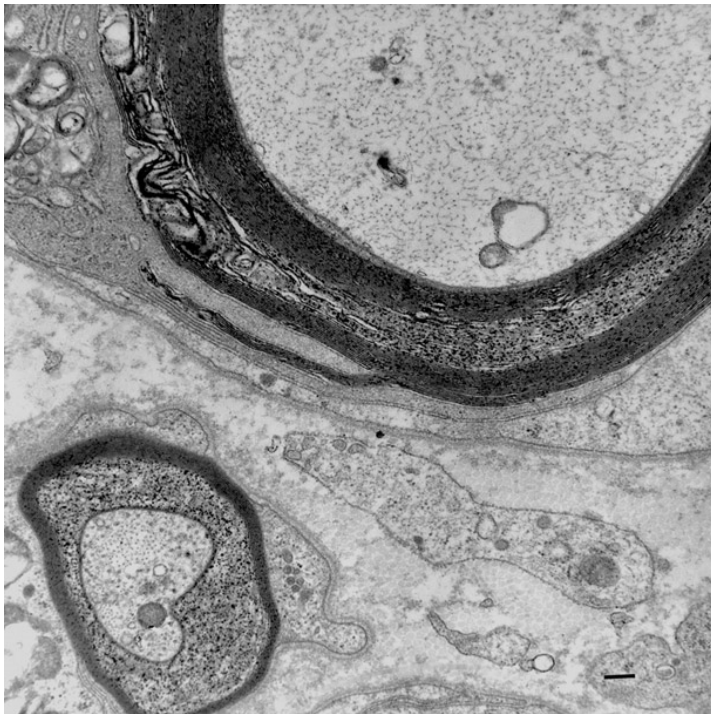
Guillain Barre Syndrome in Pregnancy?



<http://www.unilim.fr/neurolim/Images/GuillainBarre01.jpg>

There's more than one kind of GBS

Guillain Barre – **GBS**



Group B Streptococcus – **GBS**



<http://www.unilim.fr/neurolim/Images/GuillainBarre01.jpg>

<http://www.syracusemedicalmalpracticelawyerblog.com/2011/03/new-york-group-b-strep-infecti.html>



doveryai, no proveryai

(Trust, but verify)

Use the data, but be humble

The right data

The right study design

The right method

The right implementation

And always include sensitivity analysis



Thank You