A Common Data Model for Europe: Why? Which? How?

Data Quality Checking and Validation of the Sentinel Common Data Model and Tools

European Medicines Agency

December 11, 2017

Jeffrey Brown, PhD

DEPARTMENT OF POPULATION MEDICINE

HARVARD MEDICAL SCHOOL

Harvard Pilgrim Health Care Institute

Data Validation within Research Networks:

From *Ad Hoc* Practice to System Practice

Study-specific versus network data validation approaches

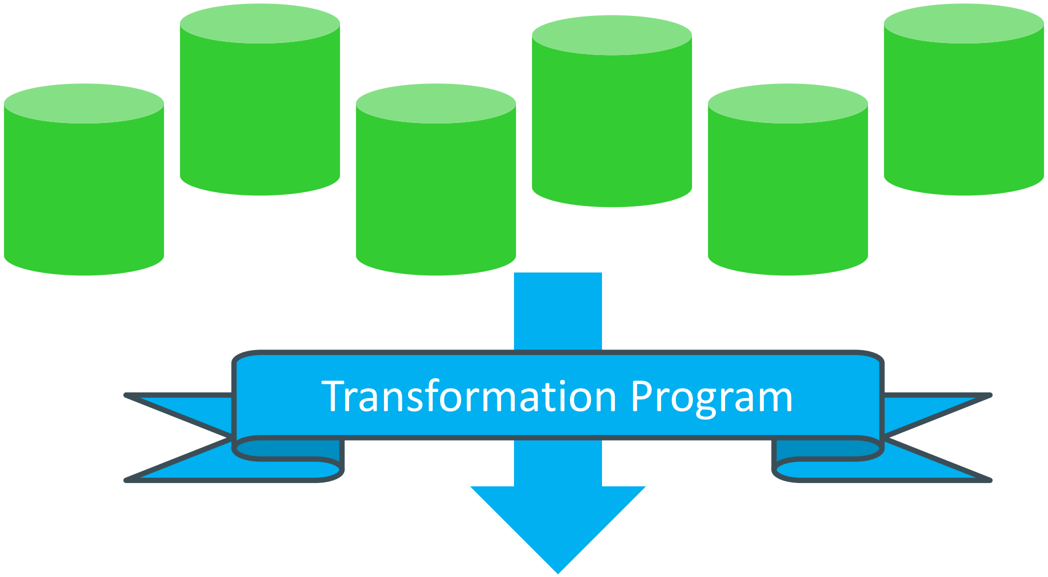
Study	Network
"As needed / as you go"	"Always Ready / Semper Paratus"
Burden on study team	Burden on quality assurance team
Ad hoc	Repeatable, Systematic, Learning
Cost is included in the cost of a study	Cost of 0 studies == cost of 1000+ studies
Variable amount of data cleaning	1400+ checks to pass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

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

Sentinel Data Validation Described

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

Unique Data Partner's Source Database Structure



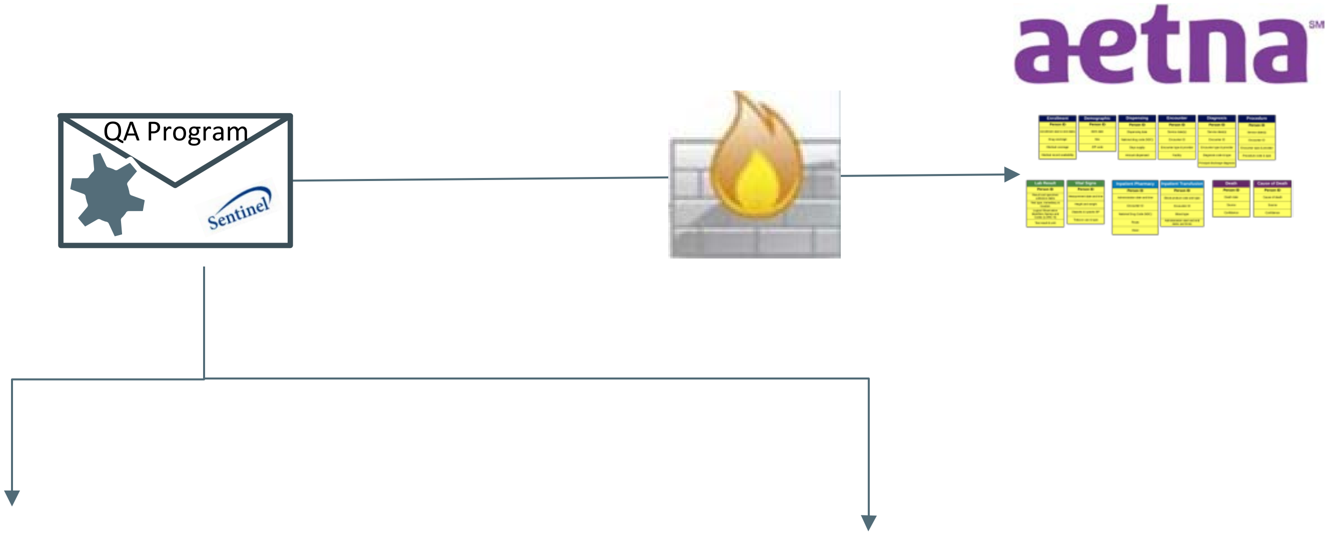
Data Partner's Database Transformed into SCDM Format (DP ETL)

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 (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		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**
Result and specimen collection dates	Measurement 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 (NDC)	Blood type	Confidence	Confidence
Test result & unit	Tobacco use & type	Route	Administration start and end dates and times		
		Dose			

The data validation process

Send a standard QA checking program to check DP’s ETL in waiting



Compliance Checks

Level 1: Completeness, validity, accuracy

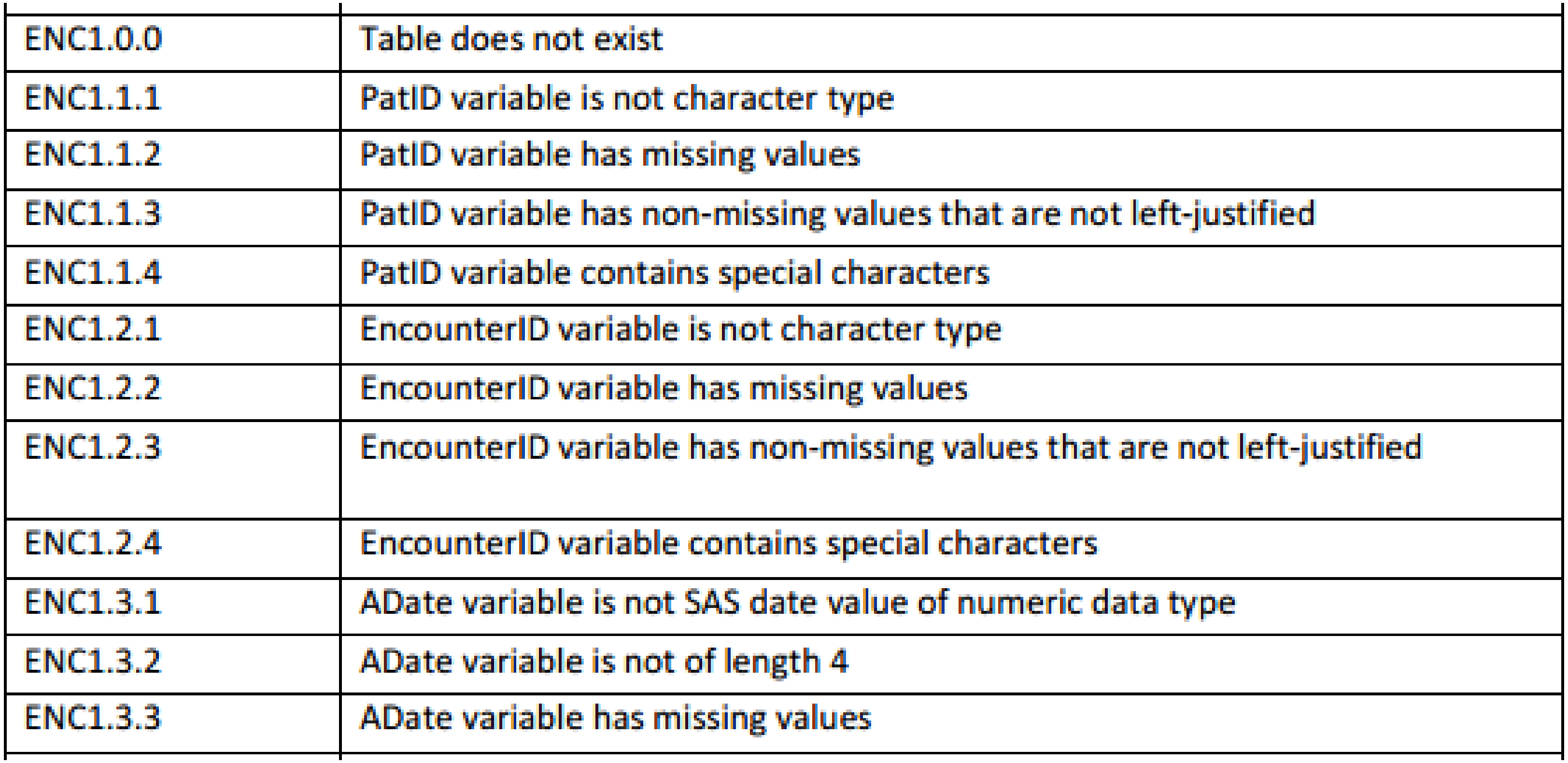
Level 2: Cross-variable and cross-table integrity

Judgment Call Checks

Level 3: Trends: consistency

Level 4: Logical: plausibility, convergence

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

Example: Admission and discharge date

Completeness:

- ADate variable has missing values

Validity:

- ADate variable is not SAS date value of numeric data type
- ADate variable is not of length 4

Accuracy:

- ADate is before DDate (for IP and IS only)
- ADate and DDate variables have values after DP_MinDate

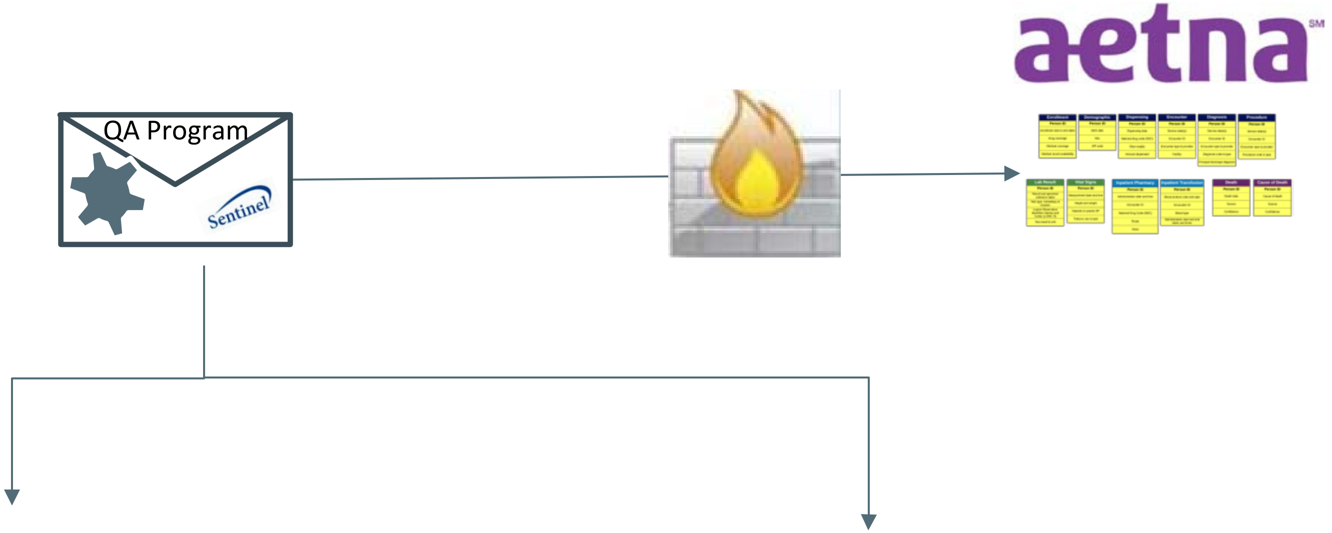
Integrity:

- DDate variable is missing for EncType value "IP"
- DDate variable is populated for EncType values other than "IP" or "IS”

*IP = Inpatient Setting, IS= Institutional Setting like a Skilled Nursing Facility

The data validation process

Send a standard QA checking program to check DP’s ETL in waiting



Compliance Checks

Level 1: Completeness, validity, accuracy

Level 2: Cross-variable and cross-table integrity

Judgment Call Checks

Level 3: Trends: consistency

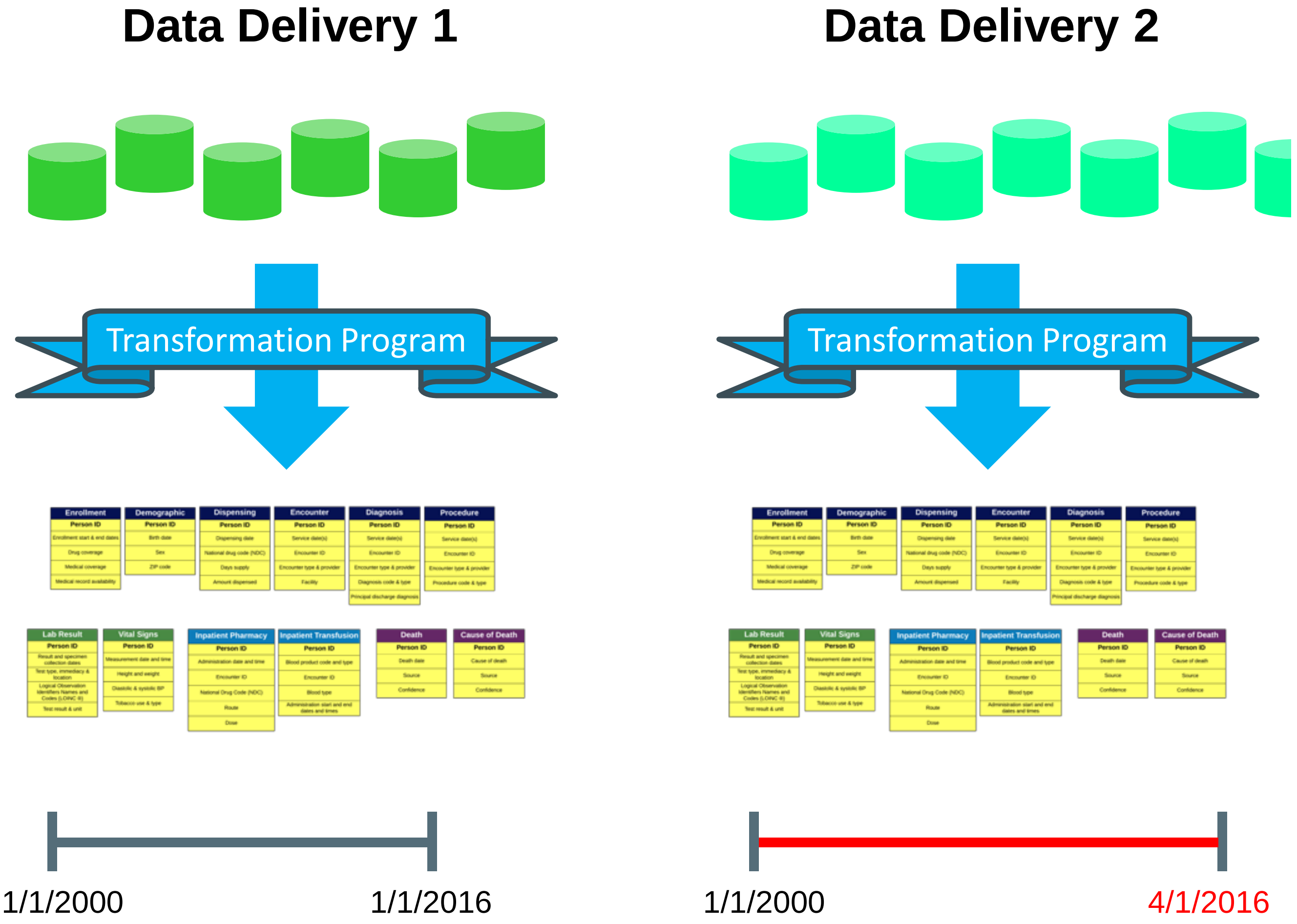
Level 4: Logical: plausibility, convergence

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

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

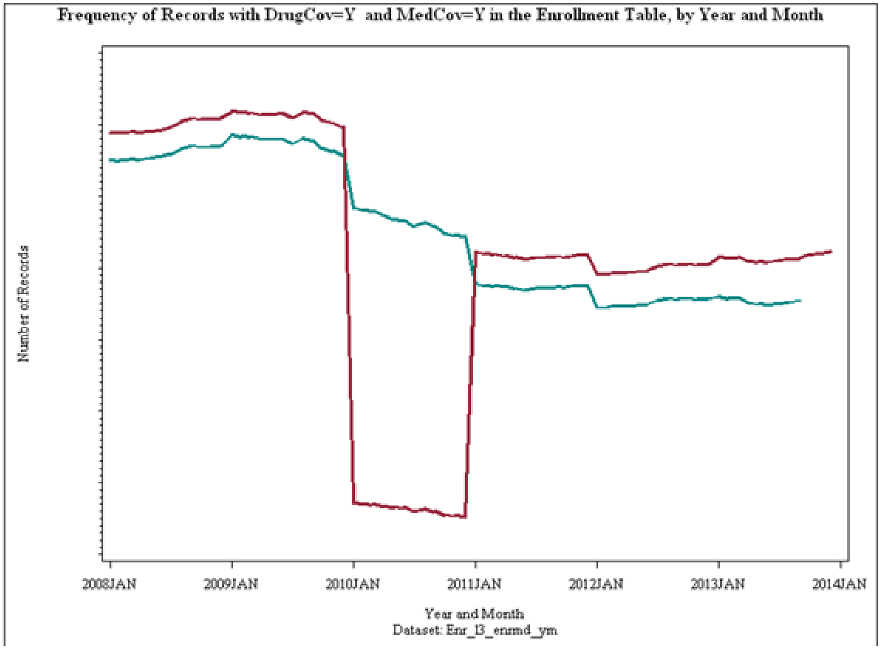
Example: Admission and discharge date

Check distributions and patterns for significant changes

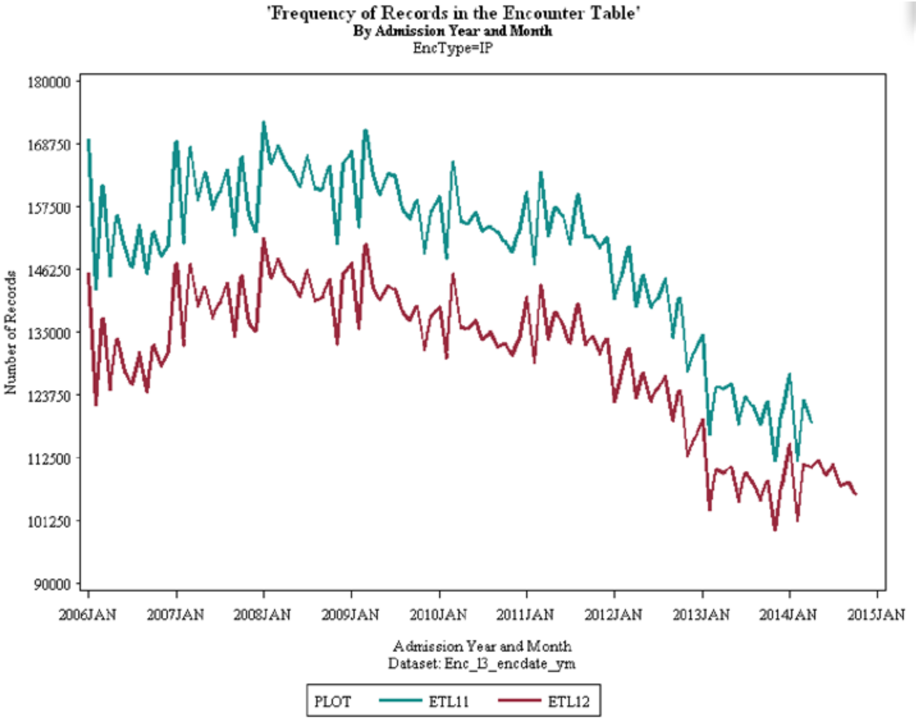
- Problem with distribution of ADate (e.g., records per year) within the ETL
- Problem with distribution of ADate (e.g., records per year-month) within the ETL
- Problem with distribution of ADate across ETLs
- Significant change in records per ADate (year) across ETLs
- Significant change in records 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

Example: Consistency Checks

- Is source of inconsistency clear error or Data Partner changes / improvements?



Incorrect Data Load

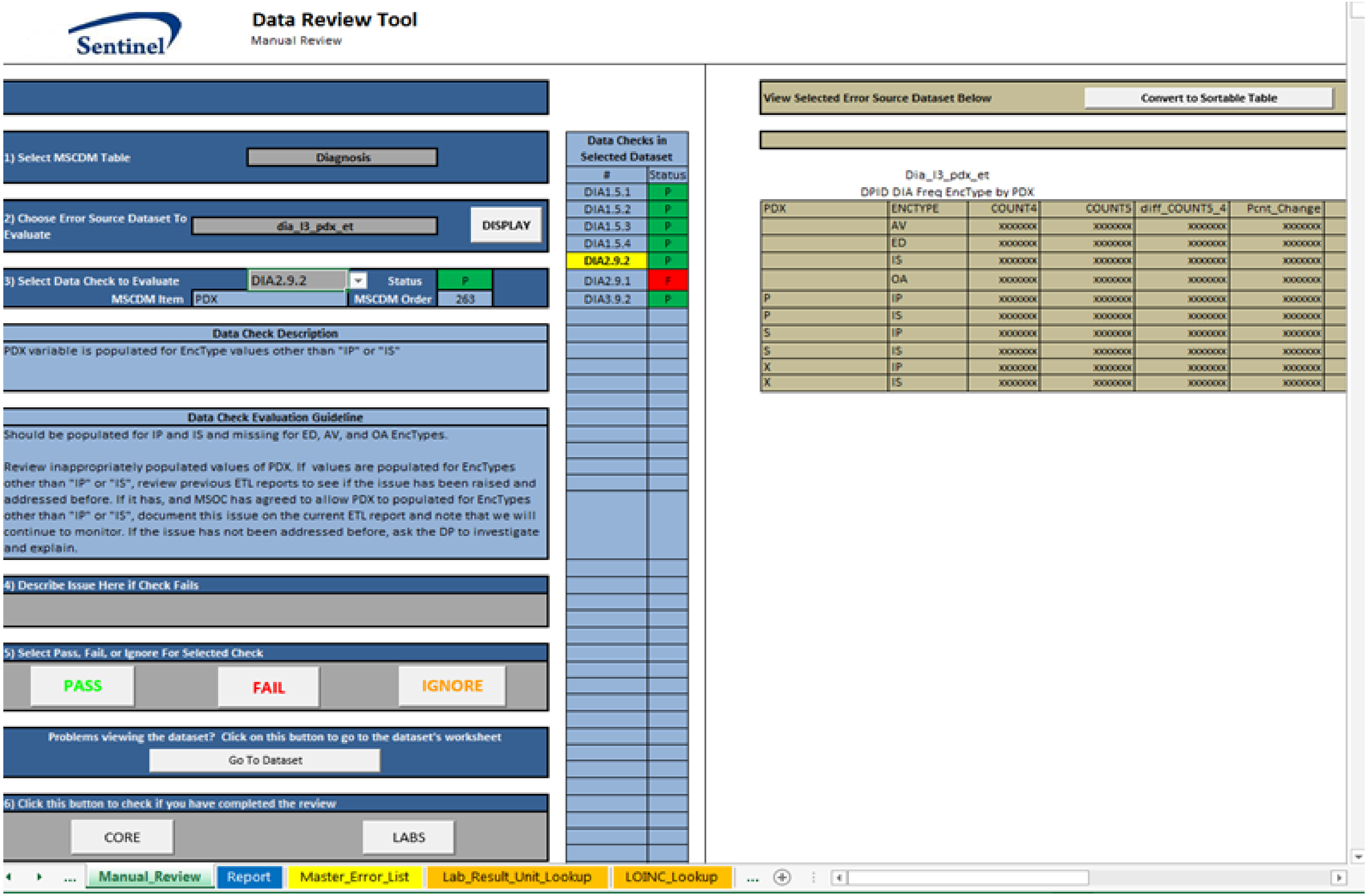


Reclassification of Encounter Type

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

Data Review Tool: Review and documentation of issues



Data Review Tool

Manual Review

1) Select MSCDM Table: Diagnosis

2) Choose Error Source Dataset To Evaluate: dia_l3_pdx_et

3) Select Data Check to Evaluate: DIA2.9.2 — Status: P

MSCDM Item: PDX — MSCDM Order: 263

Data Check Description

PDX variable is populated for EncType values other than "IP" or "IS"

Data Check Evaluation Guideline

Should be populated for IP and IS and missing for ED, AV, and OA EncTypes.

Review inappropriately populated values of PDX. If values are populated for EncTypes other than "IP" or "IS", review previous ETL reports to see if the issue has been raised and addressed before. If it has, and MSOC has agreed to allow PDX to populated for EncTypes other than "IP" or "IS", document this issue on the current ETL report and note that we will continue to monitor. If the issue has not been addressed before, ask the DP to investigate and explain.

4) Describe Issue Here if Check Fails

5) Select Pass, Fail, or Ignore For Selected Check

Problems viewing the dataset? Click on this button to go to the dataset's worksheet

6) Click this button to check if you have completed the review

Data Checks in Selected Dataset

#	Status
DIA1.5.1	P
DIA1.5.2	P
DIA1.5.3	P
DIA1.5.4	P
DIA2.9.2	P
DIA2.9.1	F
DIA3.9.2	P

View Selected Error Source Dataset Below

Dia_l3_pdx_et

DPID DIA Freq EncType by PDX

PDX	ENCTYPE	COUNT4	COUNT5	diff_COUNT5_4	Pcnt_Change
	AV	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx
	ED	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx
	IS	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx
	OA	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx
P	IP	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx
P	IS	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx
S	IP	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx
S	IS	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx
X	IP	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx
X	IS	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx

Data Validity and Quality Assurance Require Knowledge Management

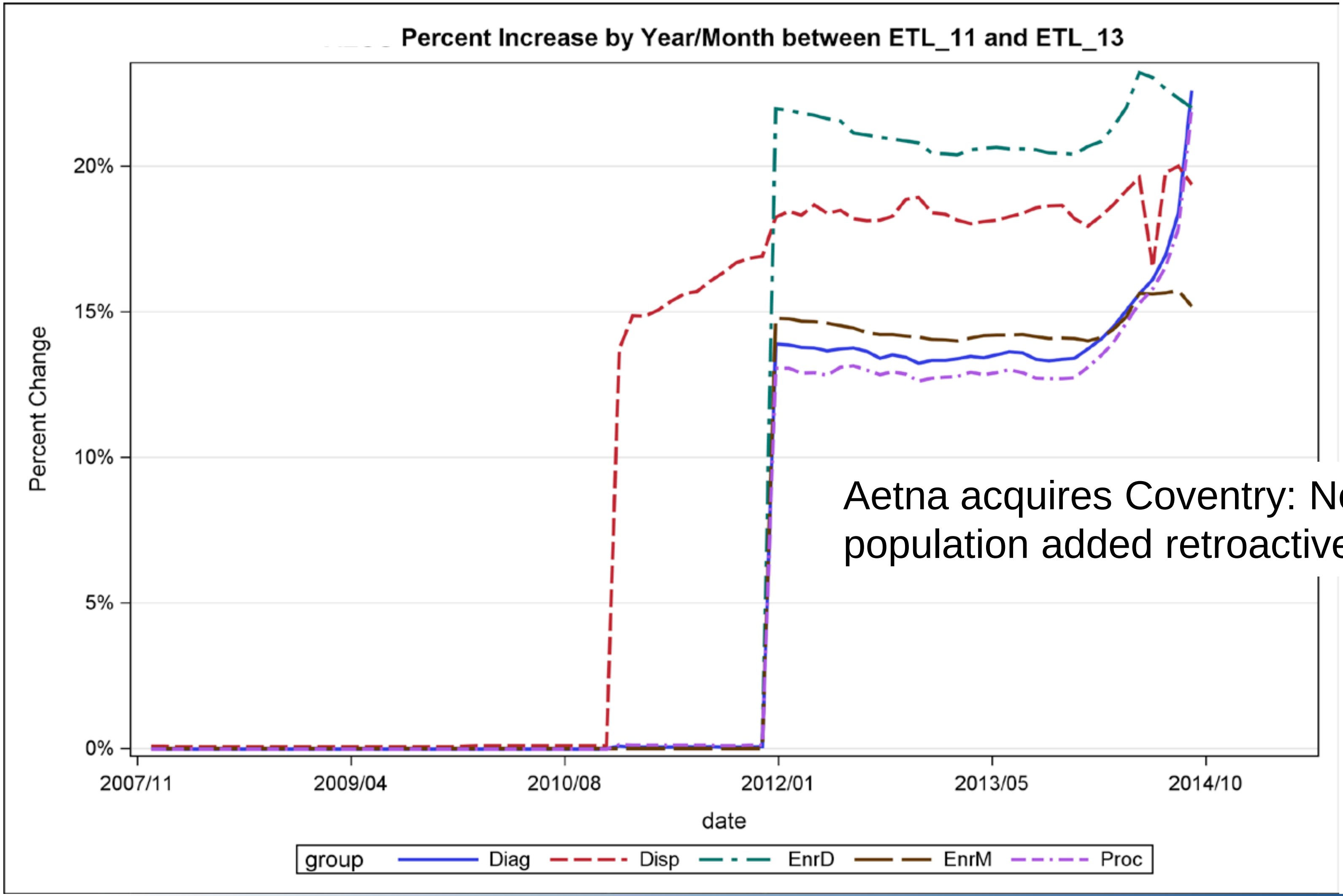
Knowledge management: Documenting and communicating changes

- Searchable internal wiki documents all data issues
- Every issue is logged and resolution documented
- QA team has regular interaction with programming and query fulfillment teams to communicate issues
- Coordination across activities is critical
  - Analytic tool development team that builds new tools
  - Software development team that maintains and enhances core software tools
  - Ongoing analyses, especially sequential studies
  - Planned projects

Other data validation activities

- Use of data validation query results to answer questions about the data
  - Investigate the uptake of new ICD-10-CM codes
  - Use of codes across the network
  - Utilization trends and missingness
  - Questions about demographics by site
  - Data availability at previous time points
- Data validation team included in data interpretation, as needed

Example: Review identifies an anomoly



Aetna acquires Coventry: New population added retroactively.

Responses to data validation findings

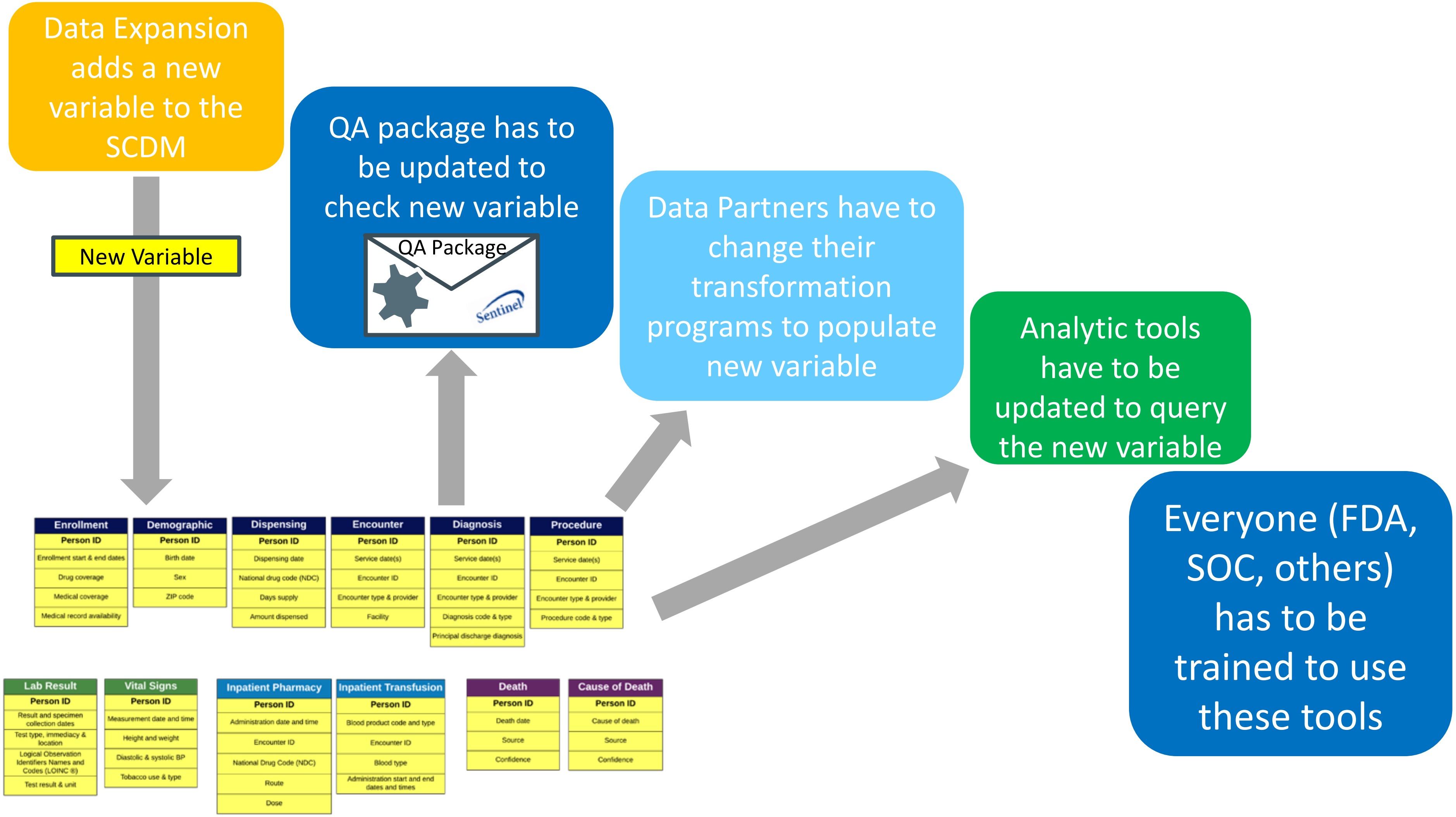
- Sequential study: Use the “partial lock” mode so new users appearing in prior periods are ignored.
- Use a prior extract to avoid issue of “new old data”
- Develop sensitivity analyses to ensure enhanced refreshes are not introducing error

When are updates expected? Are the data reasonably complete?

- Networks have to manage and coordinate data updates
- A must for all sequential analysis
- A must for time-sensitive queries

Cascade Effects of Data Expansion

Adding a variable to the data model



Data Expansion adds a new variable to the SCDM

New Variable

QA package has to be updated to check new variable

Data Partners have to change their transformation programs to populate new variable

Analytic tools have to be updated to query the new variable

Everyone (FDA, SOC, others) has to be trained to use these tools

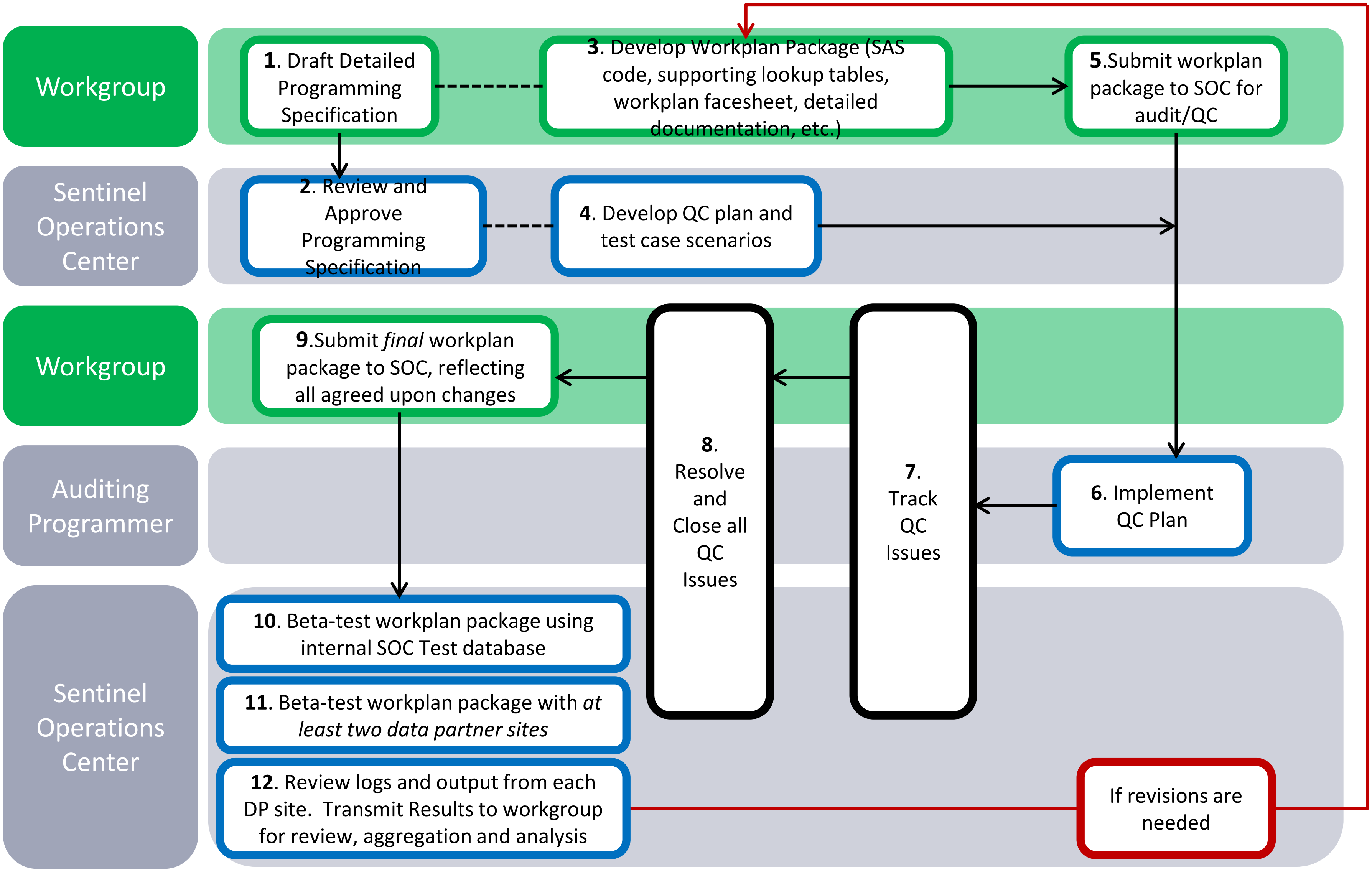
Enrollment	Demographic	Dispensing	Encounter	Diagnosis	Procedure
Person ID	**Person ID**	**Person ID**	**Person ID**	**Person ID**	**Person ID**
Enrollment start & end dates	Birth date	Dispensing date	Service date(s)	Service date(s)	Service date(s)
Drug coverage	Sex	National drug code (NDC)	Encounter ID	Encounter ID	Encounter ID
Medical coverage	ZIP code	Days supply	Encounter type & provider	Encounter type & provider	Encounter type & provider
Medical record availability		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**
Result and specimen collection dates	Measurement 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 (NDC)	Blood type	Confidence	Confidence
Test result & unit	Tobacco use & type	Route	Administration start and end dates and times		
		Dose			

Data Validity in Analytics

Validate the tools before use
Validate the data (again) at each use

Programming SOP for tool development



Workgroup

1. Draft Detailed Programming Specification

3. Develop Workplan Package (SAS code, supporting lookup tables, workplan facesheet, detailed documentation, etc.)

5.Submit workplan package to SOC for audit/QC

Sentinel Operations Center

2. Review and Approve Programming Specification

4. Develop QC plan and test case scenarios

Workgroup

9.Submit *final* workplan package to SOC, reflecting all agreed upon changes

8. Resolve and Close all QC Issues

7. Track QC Issues

Auditing Programmer

6. Implement QC Plan

Sentinel Operations Center

10. Beta-test workplan package using internal SOC Test database

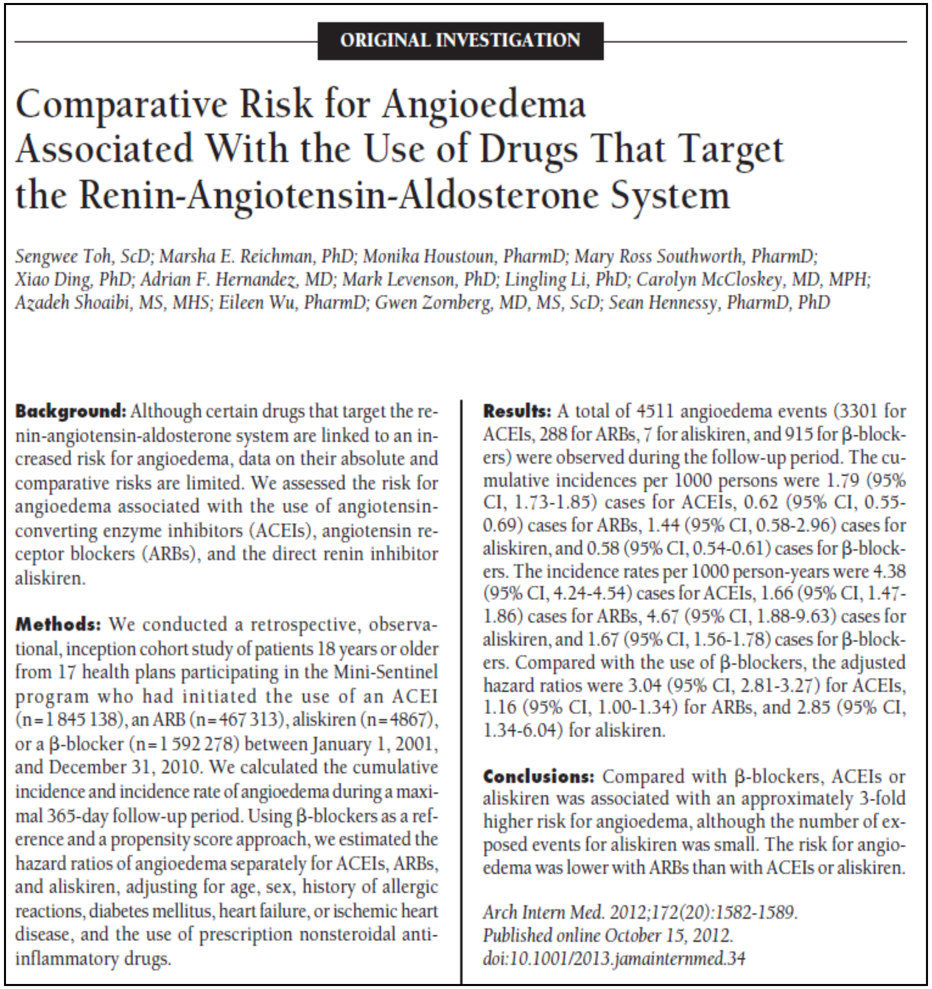
11. Beta-test workplan package with *at least two data partner sites*

12. Review logs and output from each DP site. Transmit Results to workgroup for review, aggregation and analysis

If revisions are needed

Validity of the re-usable tools

- Protocol-based analysis from Toh *et al*
- ACEIs vs β-blockers: **Adjusted hazard ratio: 3.0** (95% CI: 2.8-3.3)



ORIGINAL INVESTIGATION

Comparative Risk for Angioedema Associated With the Use of Drugs That Target the Renin-Angiotensin-Aldosterone System

Sengwee Toh, ScD; Marsha E. Reichman, PhD; Monika Houstoun, PharmD; Mary Ross Southworth, PharmD; Xiao Ding, PhD; Adrian F. Hernandez, MD; Mark Levenson, PhD; Lingling Li, PhD; Carolyn McCloskey, MD, MPH; Azadeh Shoaibi, MS, MHS; Eileen Wu, PharmD; Gwen Zornberg, MD, MS, ScD; Sean Hennessy, PharmD, PhD

Background: Although certain drugs that target the renin-angiotensin-aldosterone system are linked to an increased risk for angioedema, data on their absolute and comparative risks are limited. We assessed the risk for angioedema associated with the use of angiotensin-converting enzyme inhibitors (ACEIs), angiotensin receptor blockers (ARBs), and the direct renin inhibitor aliskiren.

Methods: We conducted a retrospective, observational, inception cohort study of patients 18 years or older from 17 health plans participating in the Mini-Sentinel program who had initiated the use of an ACEI (n=1 845 138), an ARB (n=467 313), aliskiren (n=4867), or a β-blocker (n=1 592 278) between January 1, 2001, and December 31, 2010. We calculated the cumulative incidence and incidence rate of angioedema during a maximal 365-day follow-up period. Using β-blockers as a reference and a propensity score approach, we estimated the hazard ratios of angioedema separately for ACEIs, ARBs, and aliskiren, adjusting for age, sex, history of allergic reactions, diabetes mellitus, heart failure, or ischemic heart disease, and the use of prescription nonsteroidal anti-inflammatory drugs.

Results: A total of 4511 angioedema events (3301 for ACEIs, 288 for ARBs, 7 for aliskiren, and 915 for β-blockers) were observed during the follow-up period. The cumulative incidences per 1000 persons were 1.79 (95% CI, 1.73-1.85) cases for ACEIs, 0.62 (95% CI, 0.55-0.69) cases for ARBs, 1.44 (95% CI, 0.58-2.96) cases for aliskiren, and 0.58 (95% CI, 0.54-0.61) cases for β-blockers. The incidence rates per 1000 person-years were 4.38 (95% CI, 4.24-4.54) cases for ACEIs, 1.66 (95% CI, 1.47-1.86) cases for ARBs, 4.67 (95% CI, 1.88-9.63) cases for aliskiren, and 1.67 (95% CI, 1.56-1.78) cases for β-blockers. Compared with the use of β-blockers, the adjusted hazard ratios were 3.04 (95% CI, 2.81-3.27) for ACEIs, 1.16 (95% CI, 1.00-1.34) for ARBs, and 2.85 (95% CI, 1.34-6.04) for aliskiren.

Conclusions: Compared with β-blockers, ACEIs or aliskiren was associated with an approximately 3-fold higher risk for angioedema, although the number of exposed events for aliskiren was small. The risk for angioedema was lower with ARBs than with ACEIs or aliskiren.

*Arch Intern Med. 2012;172(20):1582-1589.
Published online October 15, 2012.
doi:10.1001/2013.jamainternmed.34*

Results

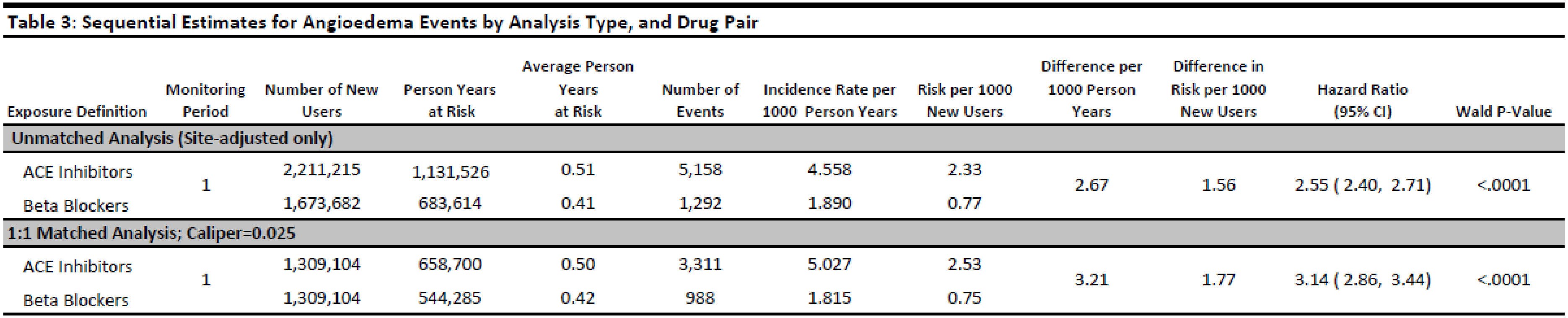


Table 3: Sequential Estimates for Angioedema Events by Analysis Type, and Drug Pair

Exposure Definition	Monitoring Period	Number of New Users	Person Years at Risk	Average Person Years at Risk	Number of Events	Incidence Rate per 1000 Person Years	Risk per 1000 New Users	Difference per 1000 Person Years	Difference in Risk per 1000 New Users	Hazard Ratio (95% CI)	Wald P-Value
Unmatched Analysis (Site-adjusted only)											
ACE Inhibitors	1	2,211,215	1,131,526	0.51	5,158	4.558	2.33	2.67	1.56	2.55 (2.40, 2.71)	<.0001
Beta Blockers		1,673,682	683,614	0.41	1,292	1.890	0.77				
1:1 Matched Analysis; Caliper=0.025											
ACE Inhibitors	1	1,309,104	658,700	0.50	3,311	5.027	2.53	3.21	1.77	3.14 (2.86, 3.44)	<.0001
Beta Blockers		1,309,104	544,285	0.42	988	1.815	0.75				

From protocol-based analysis with ad hoc program

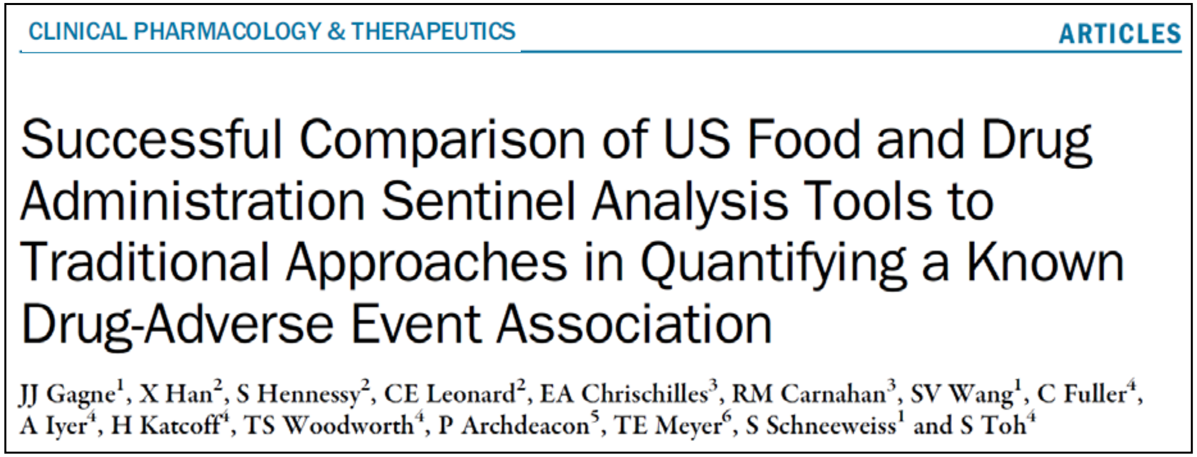
- **HR: 3.0** (95% CI: 2.8, 3.3)

From PS-matched analysis with re-usable analytic tools

- **HR: 3.1** (95% CI: 2.9, 3.4)

https://www.sentinelinitiative.org/sites/default/files/Drugs/Assessments/Mini-Sentinel_Methods_Known-Positives-ACEI-Angioedema.pdf

Tool validation studies

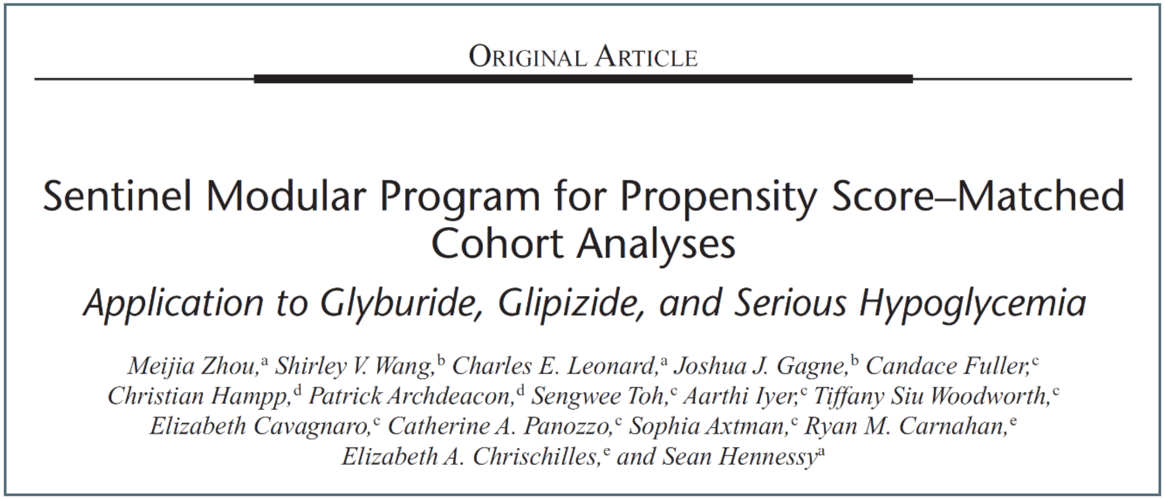


CLINICAL PHARMACOLOGY & THERAPEUTICS

ARTICLES

Successful Comparison of US Food and Drug Administration Sentinel Analysis Tools to Traditional Approaches in Quantifying a Known Drug-Adverse Event Association

JJ Gagne[1], X Han[2], S Hennessy[2], CE Leonard[2], EA Chrischilles[3], RM Carnahan[3], SV Wang[1], C Fuller[4], A Iyer[4], H Katcoff[4], TS Woodworth[4], P Archdeacon[5], TE Meyer[6], S Schneeweiss[1] and S Toh[4]



ORIGINAL ARTICLE

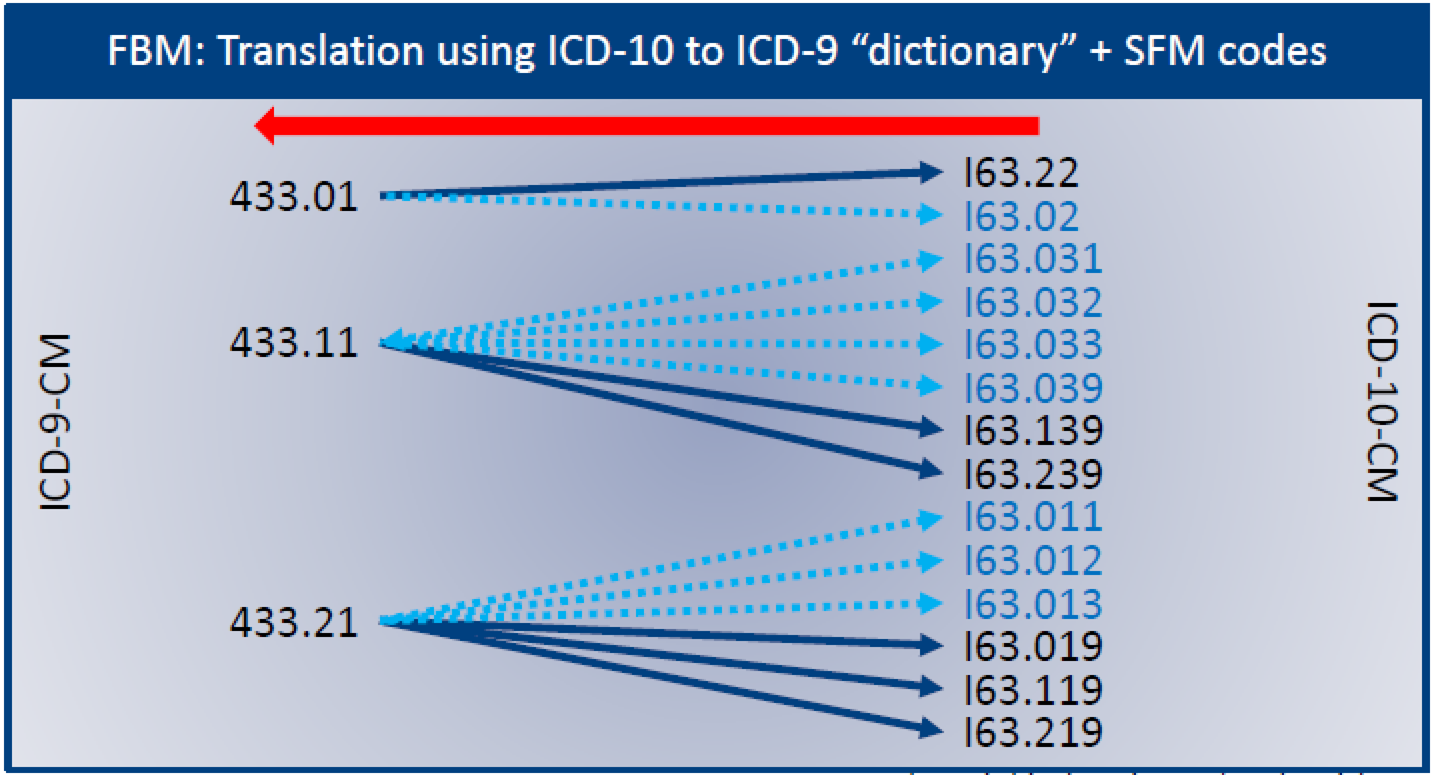
Sentinel Modular Program for Propensity Score–Matched Cohort Analyses

Application to Glyburide, Glipizide, and Serious Hypoglycemia

Meijia Zhou,[a] Shirley V. Wang,[b] Charles E. Leonard,[a] Joshua J. Gagne,[b] Candace Fuller,[c] Christian Hampp,[d] Patrick Archdeacon,[d] Sengwee Toh,[c] Aarthi Iyer,[c] Tiffany Siu Woodworth,[c] Elizabeth Cavagnaro,[c] Catherine A. Panozzo,[c] Sophia Axtman,[c] Ryan M. Carnahan,[e] Elizabeth A. Chrischilles,[e] and Sean Hennessy[a]

Validation in analytics: ICD-9 to ICD-10 transition

- Ischemic Stroke algorithm* *(total 91 codes)*:
  - Utilize both forward mapping and backward mapping files

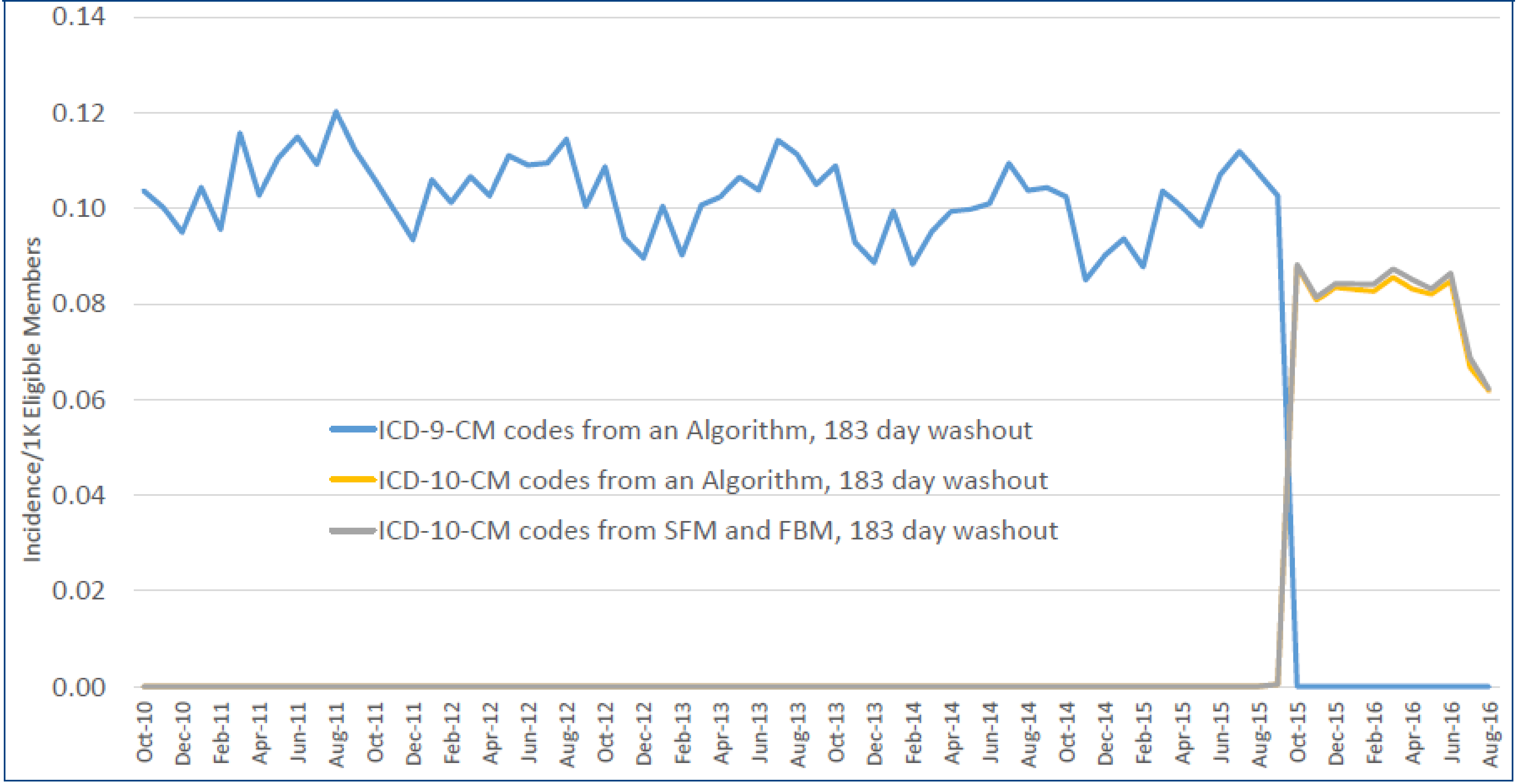


*partial ischemic stroke algorithm

Source: Woodworth, et al. ICPE, 2017

Angioedema: trend analysis

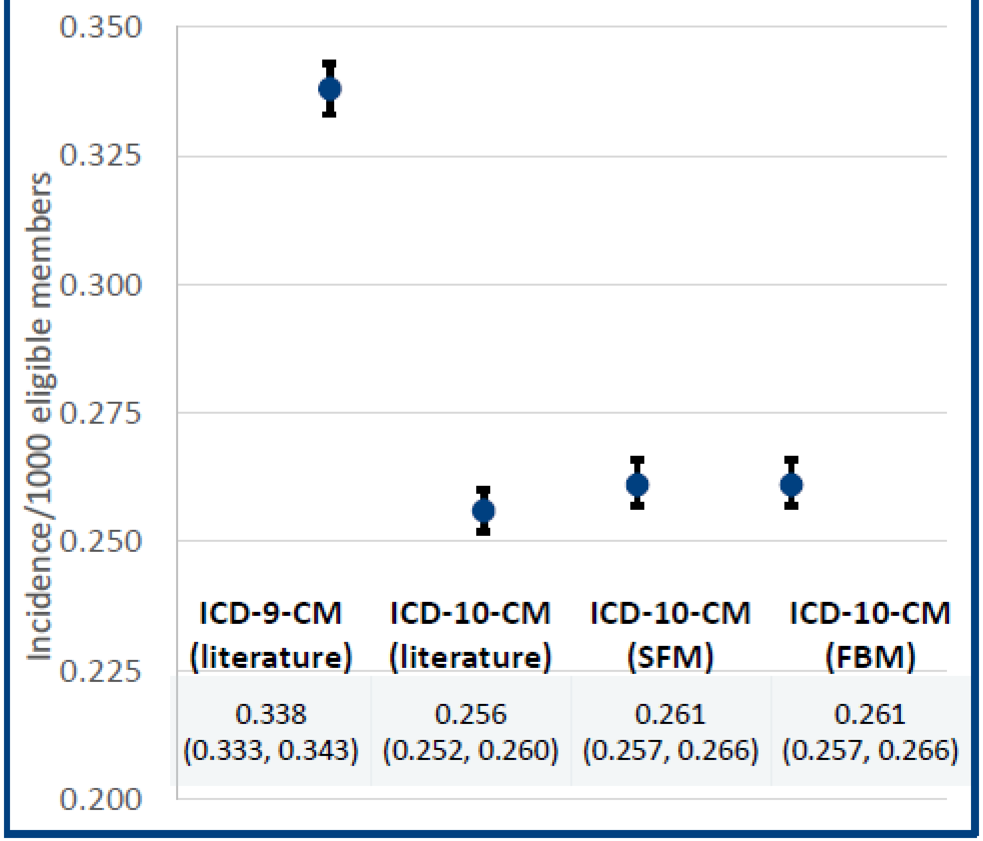
Incidence per 1,000 Eligible Members of Angioedema between October 2010- August 2016, by Outcome Definition



Source: Woodworth, et al. ICPE, 2017

Coding era analysis example (Angioedema)

Incidence of various angioedema definitions per 1000 eligible members using a *90 day washout*, Jan-Mar 2015 vs. Jan-Mar 2016



ICD-9-CM (literature)	ICD-10-CM (literature)	ICD-10-CM (SFM)	ICD-10-CM (FBM)
0.338 (0.333, 0.343)	0.256 (0.252, 0.260)	0.261 (0.257, 0.266)	0.261 (0.257, 0.266)

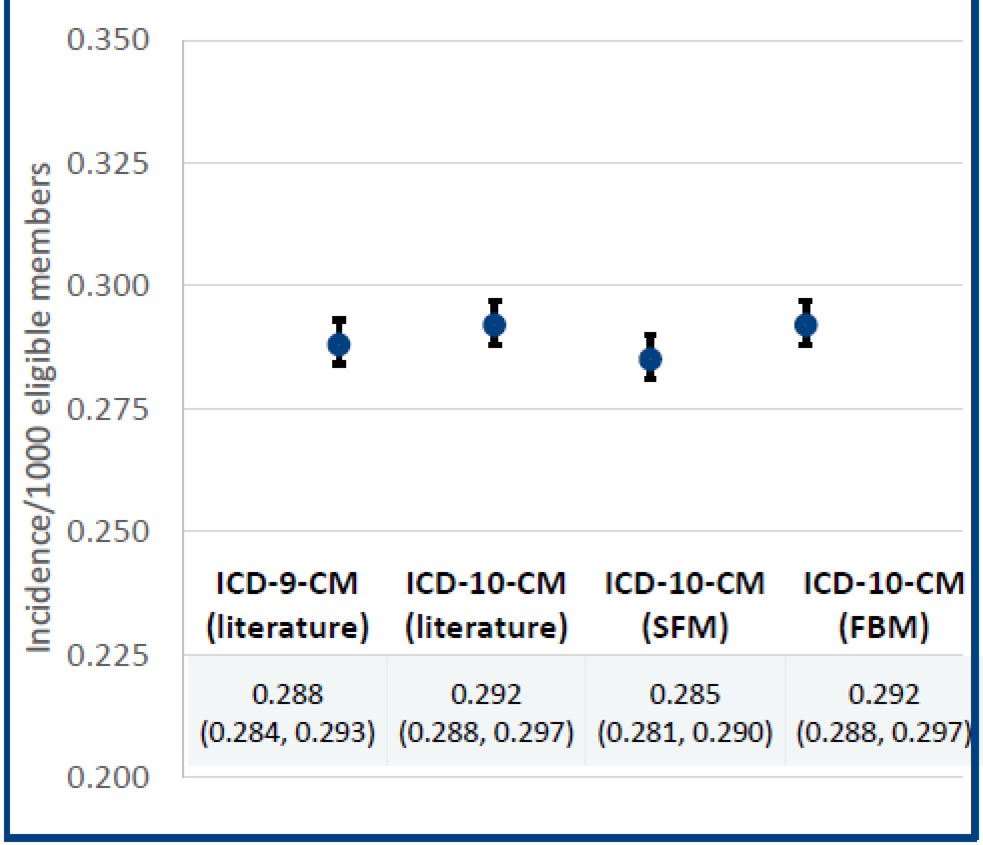
Definition	ICD-9-CM Code Count	ICD-10-CM Code Count
Algorithm*	1	1
Washout	1	3
SFM	1	1
FBM	1	1

- One ICD-9-CM code (**19K events**)
  - *995.1: Angioneurotic edema not elsewhere classified*
- Three ICD-10-CM codes (**15K events**)
  - *T78.3XXA: Angioneurotic edema, initial encounter*
  - *T78.3XXD: Angioneurotic edema, subsequent encounter*
  - *T78.3XXS: Angioneurotic edema, sequela*

**Toh S et al, Johnsen SP et al, Gupta R et al*

Coding era analysis example (Acute MI)

Incidence of various AMI definitions per 1000 eligible members using a *90 day washout*, Jan-Mar 2015 vs. Jan-Mar 2016



ICD-9-CM (literature)	ICD-10-CM (literature)	ICD-10-CM (SFM)	ICD-10-CM (FBM)
0.288 (0.284, 0.293)	0.292 (0.288, 0.297)	0.285 (0.281, 0.290)	0.292 (0.288, 0.297)

Definition	ICD-9-CM Code Count	ICD-10-CM Code Count
Algorithm*	20	12
SFM	20	6
FBM	20	14

- Each algorithm identified ~16.5K events
- ICD-10-CM codes identified by the three approaches all included the most frequently used codes

*Cutrona et al 2013

Source: Woodworth, et al. ICPE, 2017

Thank You