
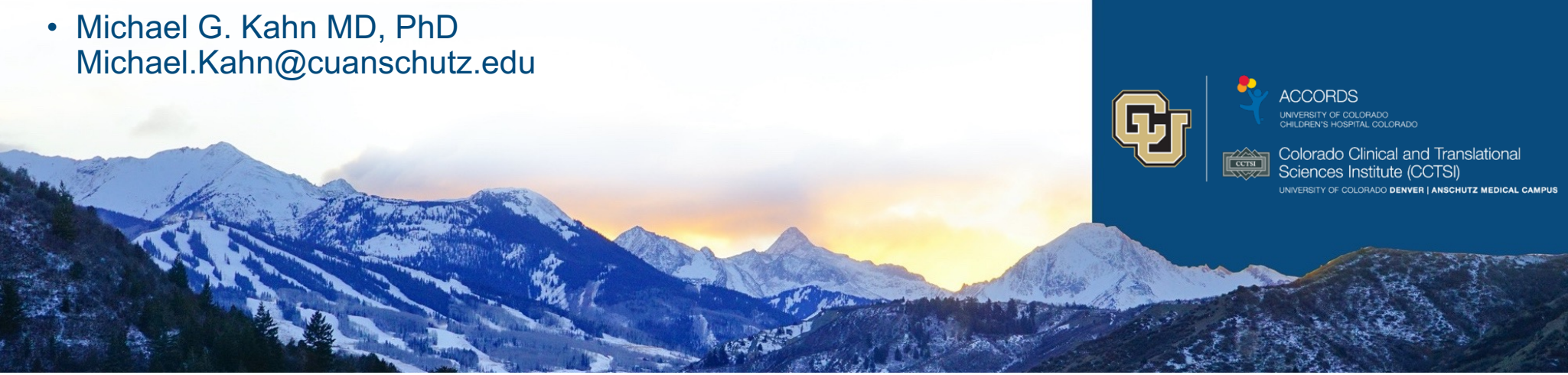




Data Quality Assessment Issues and Methods for Secondary Data Use


- Michael G. Kahn MD, PhD
Michael.Kahn@cuanschutz.edu



COPRH Con
Colorado Pragmatic
Research in Health
Conference



ACCORDS
UNIVERSITY OF COLORADO
CHILDREN'S HOSPITAL COLORADO



Colorado Clinical and Translational
Sciences Institute (CCTSI)
UNIVERSITY OF COLORADO DENVER | ANSCHUTZ MEDICAL CAMPUS

Agenda

- 2:00-2:15 Didactic presentation
- 2:15-2:25 A quick look at the DQD tool & key web sites
- 2:25-2:35 Q&A



ACCORDS



Assessing Data Quality in Secondary Data

Two questions – very different perspectives:

- Q1: Are these data any good?
 - Q2: Can I use these data to answer my question
-
- Q1: **Global data quality** – do I even bother to spend time with these data
 - Q2: **Fitness for use** – do I invest more time drilling into data quality for my study



ACCORDS



The language of data quality is a mess

Proposed Categories	Johnson 2015 [45] ^a	Zozus 2014 [44] ^b	Liau 2013 [40] ^c	Weiskopf 2013a [42] ^d	Weiskopf 2013b [39] ^e	Kahn 2012 [38] ^f	Nahm 2012 [51] ^g	McGilvray 2008 [56] ^h	Eppler 2006 [57] ⁱ	Wang 1996 [36] ^j
COMPLETENESS										
Density	<ul style="list-style-type: none"> Representation-Correctness Representation-Complete Domain-Complete Relative-Completeness 	<ul style="list-style-type: none"> "Column" Data Value Completeness Information Loss and Degradation 	<ul style="list-style-type: none"> Completeness (Elements of External Consistency and correctness) 	<ul style="list-style-type: none"> Documentation Density 	<ul style="list-style-type: none"> Completeness (Atemporal) 	<ul style="list-style-type: none"> Attribute Domain Constraints 	<ul style="list-style-type: none"> Completeness Attribution 	<ul style="list-style-type: none"> Data Integrity Fundamentals 	<ul style="list-style-type: none"> Traceability 	<ul style="list-style-type: none"> Completeness
FIDELITY										
Metadata	<ul style="list-style-type: none"> Representation-Consistency Domain-Consistency Coding-Consistency Domain-Metadata 	<ul style="list-style-type: none"> Data Element Completeness 	<ul style="list-style-type: none"> Internal Consistency External Consistency 			<ul style="list-style-type: none"> Attribute Domain Constraints Relational Integrity Rules Historical Data Rules State-Dependent Object Rules 	<ul style="list-style-type: none"> Granularity Precision 	<ul style="list-style-type: none"> Data Specifications Data Integrity Fundamentals 		<ul style="list-style-type: none"> Representational Consistency
Measure		<ul style="list-style-type: none"> Consistency 	<ul style="list-style-type: none"> Consistency (Reliability Elements) 		<ul style="list-style-type: none"> Correctness Concordance 	<ul style="list-style-type: none"> Relational Integrity Rules Attribute Dependency Rules 		<ul style="list-style-type: none"> Consistency and Synchronization 		<ul style="list-style-type: none"> Accuracy Representational Consistency
Derivation			<ul style="list-style-type: none"> Correctness (Accuracy Elements) Accuracy 			<ul style="list-style-type: none"> Attribute Dependency Rules 				
Uniqueness		<ul style="list-style-type: none"> Ascertainment Completeness 	<ul style="list-style-type: none"> No Duplication 			<ul style="list-style-type: none"> Relational Integrity Rules 		<ul style="list-style-type: none"> Duplication 		
PLAUSIBILITY										
Measure	<ul style="list-style-type: none"> Representation-Integrity Domain-Consistency Relative-Correctness 	<ul style="list-style-type: none"> Representational Inaccuracy Information Loss and Degradation Consistency 	<ul style="list-style-type: none"> Correctness (Reliability Elements) External Consistency 		<ul style="list-style-type: none"> Correctness Concordance Plausibility 	<ul style="list-style-type: none"> Attribute Domain Constraints Attribute Dependency Rules 	<ul style="list-style-type: none"> Consistency (Internal) 	<ul style="list-style-type: none"> Data Integrity Fundamentals Accuracy 	<ul style="list-style-type: none"> Consistency Correctness Accuracy 	<ul style="list-style-type: none"> Believability Accuracy
Time	<ul style="list-style-type: none"> Domain-Consistency 	<ul style="list-style-type: none"> Consistency 				<ul style="list-style-type: none"> Historical Data Rules Attribute Dependency Rules State-Dependent Object Rules 	<ul style="list-style-type: none"> Accuracy 	<ul style="list-style-type: none"> Data Integrity Fundamentals 		

Informatic Community coalescing around a harmonized set of DQ terms



A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data

Michael G. Kahn, MD, PhD;ⁱ Tiffany J. Callahan, MPH;^j Juliana Barnard, MA;^k Alan E. Bauck;ⁱⁱ Jeff Brown, PhD;ⁱⁱⁱ Bruce N. Davidson, PhD;^{iv} Hossein Estiri, PhD;^v Carsten Goerg, PhD;ⁱ Erin Holve, PhD, MPH, MPP;^{vi} Steven G. Johnson, MS;^{vii} Siaw-Teng Liaw, MBBS, PhD, FRACGP, FACHI;^{viii} Marianne Hamilton-Lopez, PhD, MPA;^{ix} Daniella Meeker, PhD;^x Toan C. Ong, PhD;^{xi} Patrick Ryan, PhD;^{xii} Ning Shang, PhD;^{xiii} Nicole G. Weiskopf, PhD;^{xiv} Chunhua Weng, PhD, FACMI;^{xiii} Meredith N. Zozus, PhD;^{xv} Lisa Schilling, MD^{xi}



ACCORDS



<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5051581/pdf/egems1244.pdf>

Three DQ Dimensions that build on each other

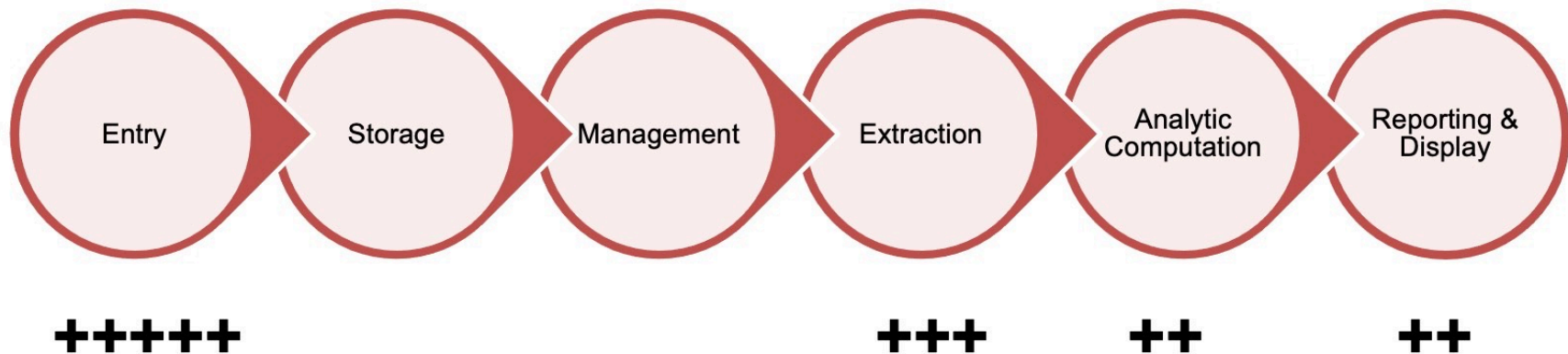
- **Completeness: Are data values present?**
 - Values are there or not: Missingness, Density, Domains
 - Does not evaluate if values make sense.
 - No need to proceed if answer is “No”
- **Fidelity: Are the data dependable?**
 - Do values align together as expected? Temporal trends/discontinuities, interdependencies
 - Does not evaluate if the values are believable
 - No need to proceed if answer is “No”
- **Plausibility: Are the data believable?**
 - Does not require existence of an absolute universal, never-changing truth: Congruence with expectations
 - **Avoided the use of accuracy/precision:** These terms have explicit meanings with psychometrics.
 - What is “believable” may be context-dependent



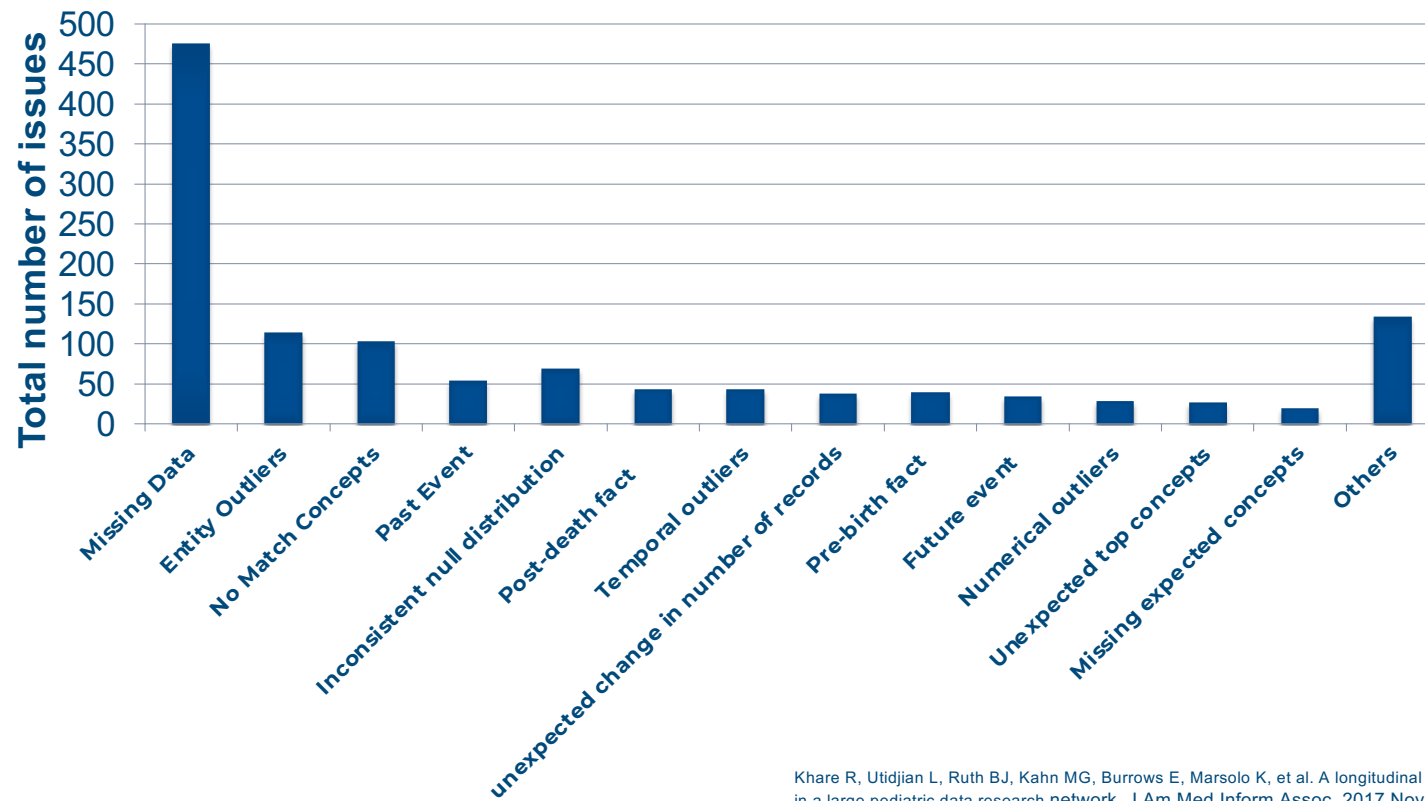
ACCORDS



Sources of data quality issues in EHR data lifecycle



Missing data always dominates in observational data



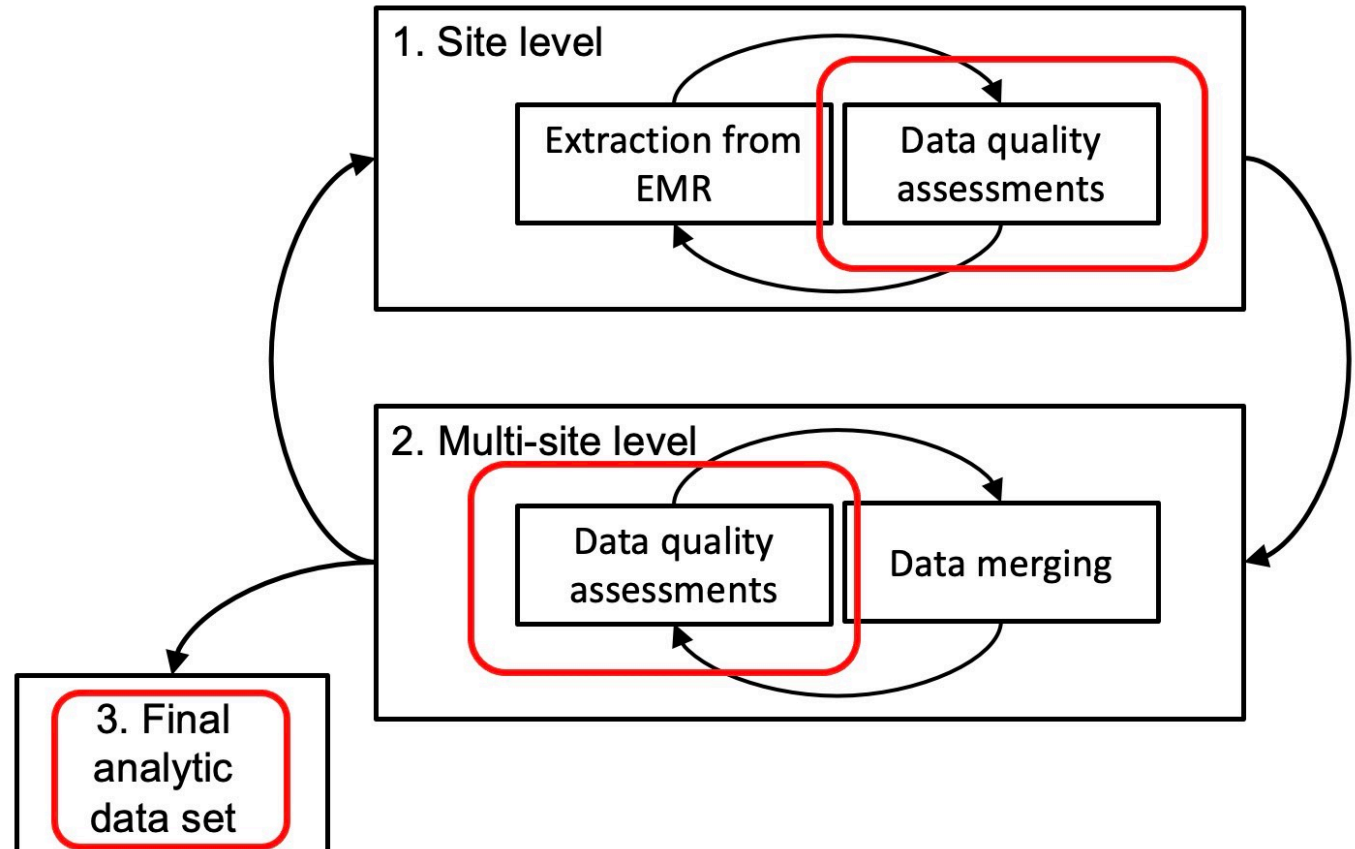
Khare R, Utidjian L, Ruth BJ, Kahn MG, Burrows E, Marsolo K, et al. A longitudinal analysis of data quality in a large pediatric data research network. *J Am Med Inform Assoc.* 2017 Nov 1;24(6):1072–9.



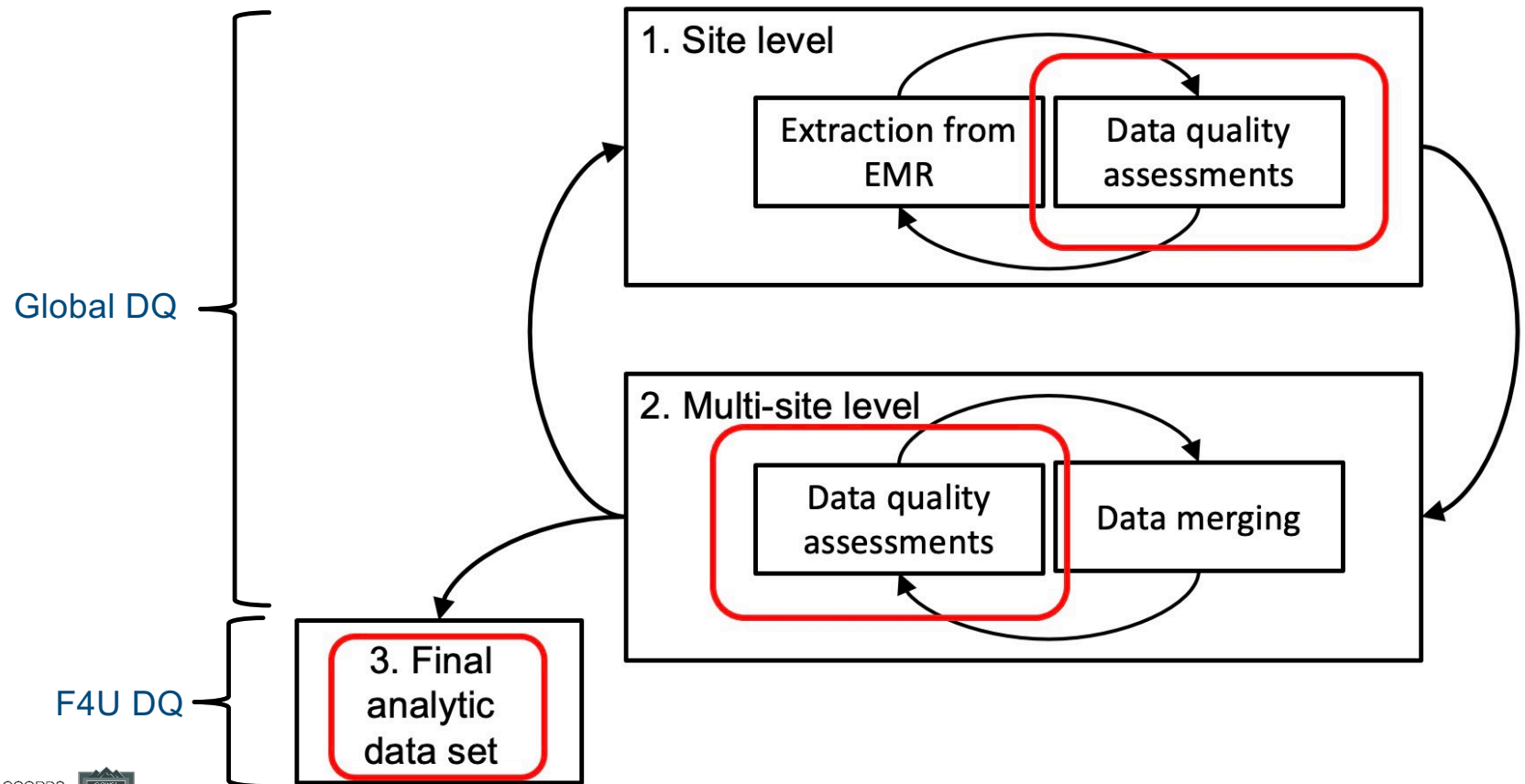
ACCORDS



When should data quality be assessed in multi-site pragmatic trials?



When should data quality be assessed in multi-site pragmatic trials?



What do global data quality reports look like?

PCORNet Data Characterization: <https://github.com/PCORnet-DRN-OC/PCORnet-Data-Curation>

IV. Data Curation Query Output Tables

For table shells of each dataset, please refer to the Technical Specifications available on the Data Curation page on [iMeet](#).

ID	PCORnet Table(s)	Output table	Output table description
1	CONDITION	cond_l3_condition	CONDITION frequency
2	CONDITION	cond_l3_n	Counts PATID, ENCOUNTERID, and CONDITIONID
3	CONDITION	cond_l3_rdate_y	REPORT_DATE year frequency
4	CONDITION	cond_l3_rdate_ym	REPORT_DATE year month frequency
5	CONDITION	cond_l3_source	CONDITION_SOURCE frequency
6	CONDITION	cond_l3_status	CONDITION_STATUS frequency
7	CONDITION	cond_l3_type	CONDITION_TYPE frequency
8	DEATH	death_l3_date_y	DEATH_DATE year frequency
9	DEATH	death_l3_date_ym	DEATH_DATE year month frequency
10	DEATH	death_l3_impute	DEATH_DATE_IMPUTE frequency
11	DEATH	death_l3_match	DEATH_MATCH_CONFIDENCE frequency
12	DEATH	death_l3_n	Counts non-missing, distinct, and missing PATID and DEATHID
13	DEATH	death_l3_source	DEATH_SOURCE frequency
14	DEATH	death_l3_source_ym	DEATH_SOURCE and DEATH_DATE year month crosstab
15	DEATH_CAUSE	deathc_l3_code	DEATH_CAUSE_CODE frequency
16	DEATH_CAUSE	deathc_l3_conf	DEATH_CAUSE_CONFIDENCE frequency
17	DEATH_CAUSE	deathc_l3_n	Counts PATID, DEATH_CAUSE, and DEATHCID
18	DEATH_CAUSE	deathc_l3_source	DEATH_CAUSE_SOURCE frequency
19	DEATH_CAUSE	deathc_l3_type	DEATH_CAUSE_TYPE frequency
			Descriptive statistics for age. Age is calculated as current age or age at

267	LAB_HISTORY	labhist_l3_sexdist	SEX frequency
268	LAB_HISTORY	labhist_l3_racedist	RACE frequency
269	LAB_HISTORY	labhist_l3_min_wks	Descriptive statistics for AGE_MIN_WKS
270	LAB_HISTORY	labhist_l3_max_wks	Descriptive statistics for AGE_MAX_WKS

PCORnet Data Curation Work Plan v6.01

Page 21

ID	PCORnet Table(s)	Output table	Output table description
271	LAB_HISTORY	labhist_l3_unit	RESULT_UNIT frequency
272	LAB_HISTORY	labhist_l3_low	NORM_MODIFIER_LOW frequency
273	LAB_HISTORY	labhist_l3_high	NORM_MODIFIER_HIGH frequency
274	LAB_HISTORY	labhist_l3_pdstart_y	PERIOD_START year frequency
275	LAB_HISTORY	labhist_l3_pdend_y	PERIOD_END year frequency
276	LAB_HISTORY	labhist_l3_rlow_dist	Descriptive statistics for NORM_RANGE_LOW
277	LAB_HISTORY	labhist_l3_rhigh_dist	Descriptive statistics for NORM_RANGE_HIGH

University of Kansas Data Characterization Summary Report

Full report available @ <https://drive.google.com/drive/folders/11iJIG0AoS1KwgOjJM5RxPzqVaunZ6Kt9>

Table IA. Demographic Summary

This table contains general descriptive information about the patients in the DEMOGRAPHIC table. These patients may or may not be represented in other CDM tables.

	N	%	Source table
Patients	2,463,453		DEM_L3_N
Age			DEM_L3_AGEYRSDIST1
Mean	61		
Median	59		
Age group			DEM_L3_AGEYRSDIST2
0-4	22,435	0.9	
5-14	81,297	3.3	
15-21	98,019	4.0	
22-64	1,244,430	50.5	
65+	1,002,613	40.7	
Missing	14,659	0.6	
Hispanic			DEM_L3_HISPDIST
N (No)	858,331	34.8	
Y (Yes)	81,766	3.3	
Missing or Refused	1,523,356	61.8	
Sex			DEM_L3_SEXDIST
F (Female)	1,356,283	55.1	
M (Male)	1,091,721	44.3	
Missing or Ambiguous	15,449	0.6	
Race			DEM_L3_RACEDIST
White	725,689	29.5	
Non-White	141,070	5.7	
Missing or Refused	1,596,694	64.8	



University of Kansas Data Characterization Summary Report

Table IB. Potential Pools of Patients

This table illustrates the number of patients meeting different inclusion criteria and supports Data Check 3.04 (less than 50% of patients with encounters have DIAGNOSIS records) and Data Check 3.05 (less than 50% of patients with encounters have PROCEDURES records). Data check exceptions are highlighted in red and must be corrected.

Metric	Metric Description	Result	Source table
Potential pool of patients for observational studies	Number of unique patients with at least 1 ED, EI, IP, OS, or AV encounter within the past 5 years	598,018	ENC_L3_DASH2
Potential pool of patients for trials	Number of unique patients with at least 1 ED, EI, IP, OS, or AV encounter within the past 1 year	264,305	ENC_L3_DASH2
Potential pool of patients for studies requiring data on diagnoses, vital measures and (a) medications or (b) medications and lab results	Number of unique patients with at least 1 encounter and DIAGNOSIS and VITAL records within the past 5 years	621,539	XTBL_L3_DASH1
	Number of unique patients with at least 1 encounter and DIAGNOSIS, VITAL, and PRESCRIBING or DISPENSING records within the past 5 years	587,749	XTBL_L3_DASH2
	Number of unique patients with at least 1 encounter and DIAGNOSIS, VITAL, PRESCRIBING or DISPENSING, and LAB_RESULT_CM records within the past 5 years	383,993	XTBL_L3_DASH3
Patients with diagnosis data	Percentage of patients with encounters who have at least 1 diagnosis	80%	ENC_L3_N; DIA_L3_N
Patients with procedure data	Percentage of patients with encounters who have at least 1 procedure	75%	ENC_L3_N; PRO_L3_N



University of Kansas Data Characterization Summary Report

Table IG. Lab Results For Selected Lab Tests

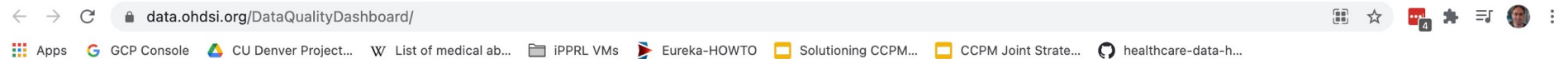
This table illustrates the number of records and number of unique patients for 30 high volume data curation lab groups, and the percentage of patients in the ENCOUNTER table who have these results. Although there is not a required relationship between the ENCOUNTER and LAB_RESULT_CM tables, patients with encounters are the most relevant denominator for this table. Version 3.2 of the data curation lab groups includes 490 concepts of interest to the Collaborative Research Groups (CRGs). Groups were constructed based on the LOINC attributes of COMPONENT, SYSTEM, and, if necessary, TIME, METHOD and CLASS. More information about the data curation lab groups is available on the Data Curation home page (<https://pconet.imeetcentral.com/p/aQAAAAACjjsH>).

DC_LAB_GROUP	Records	Percentage of records in the LAB_RESULT_CM table with a LAB_LOINC code	Patients	Percentage of patients in the ENCOUNTER table	Source tables
ALBUMIN B/S/P	2,974,079	1.8	355,113	32.3	LAB_L3_DCGROUP;ENC_L3_N
ALP TOTAL	2,972,884	1.8	354,578	32.3	LAB_L3_DCGROUP;ENC_L3_N
ALT	3,038,888	1.8	356,612	32.5	LAB_L3_DCGROUP;ENC_L3_N
AST	3,002,984	1.8	356,333	32.4	LAB_L3_DCGROUP;ENC_L3_N
BASOPHILS ABSOLUTE	2,509,728	1.5	315,946	28.8	LAB_L3_DCGROUP;ENC_L3_N
BILIRUBIN TOTAL B/S/P	2,982,962	1.8	359,325	32.7	LAB_L3_DCGROUP;ENC_L3_N
BUN	4,356,222	2.6	392,459	35.7	LAB_L3_DCGROUP;ENC_L3_N
CALCIUM B/S/P	4,323,557	2.6	391,683	35.7	LAB_L3_DCGROUP;ENC_L3_N
CHLORIDE B/S/P	4,322,916	2.6	391,527	35.6	LAB_L3_DCGROUP;ENC_L3_N
CHOLESTEROL-LDL ABSOLUTE	580,032	0.3	176,985	16.1	LAB_L3_DCGROUP;ENC_L3_N
CREATININE B/S/P	4,502,392	2.7	399,604	36.4	LAB_L3_DCGROUP;ENC_L3_N
EGFR	13,986,405	8.3	383,685	34.9	LAB_L3_DCGROUP;ENC_L3_N
GLUCOSE B/S/P	4,632,580	2.7	204,454	18.6	LAB_L3_DCGROUP;ENC_L3_N
HEMATOCRIT	4,342,796	2.6	407,737	37.1	LAB_L3_DCGROUP;ENC_L3_N
HEMOGLOBIN A1C	306,928	0.2	119,198	10.9	LAB_L3_DCGROUP;ENC_L3_N
HEMOGLOBIN B/S/P	4,363,101	2.6	407,938	37.1	LAB_L3_DCGROUP;ENC_L3_N
INR	969,729	0.6	140,027	12.7	LAB_L3_DCGROUP;ENC_L3_N
LYMPHOCYTES ABSOLUTE	2,557,796	1.5	321,987	29.3	LAB_L3_DCGROUP;ENC_L3_N
MCH	4,148,044	2.5	401,165	36.5	LAB_L3_DCGROUP;ENC_L3_N
MCHC	4,147,449	2.5	401,132	36.5	LAB_L3_DCGROUP;ENC_L3_N
MCV	4,150,217	2.5	401,436	36.6	LAB_L3_DCGROUP;ENC_L3_N
MONOCYTES ABSOLUTE	2,512,194	1.5	316,026	28.8	LAB_L3_DCGROUP;ENC_L3_N
NEUTROPHILS ABSOLUTE	2,558,425	1.5	321,948	29.3	LAB_L3_DCGROUP;ENC_L3_N



OHDSI Data Quality Dashboard

Online demo @ <https://data.ohdsi.org/DataQualityDashboard/>



SYNTHEA SYNTHETIC HEALTH DATABASE

OVERVIEW

METADATA

RESULTS

ABOUT

DATA QUALITY ASSESSMENT

SYNTHEA SYNTHETIC HEALTH DATABASE

Results generated at 2019-08-22 14:15:06 in 29 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	159	21	180	88%	283	0	283	100%	442	21	463	95%
Conformance	637	34	671	95%	104	0	104	100%	741	34	775	96%
Completeness	369	17	386	96%	5	10	15	33%	374	27	401	93%
Total	1165	72	1237	94%	392	10	402	98%	1557	82	1639	95%



OHDSI Data Quality Dashboard

Online demo @ <https://data.ohdsi.org/DataQualityDashboard/>

RESULTS

SYNTHEA SYNTHETIC HEALTH DATABASE

Results generated at 2019-08-22 14:15:06 in 29 mins

Show entries Column visibility CSV Search:

STATUS	CONTEXT	CATEGORY	SUBCATEGORY	LEVEL	DESCRIPTION	% RECORDS	
FAIL		Conformance		FIELD			
+	FAIL	Verification	Conformance	Relational	FIELD	The number and percent of records that have a value in the person_id field in the OBSERVATION_PERIOD table that does not exist in the PERSON table. (Threshold=0%).	49.58%
+	FAIL	Verification	Conformance	Relational	FIELD	The number and percent of records that have a value in the ethnicity_concept_id field in the PERSON table that does not exist in the CONCEPT table. (Threshold=0%).	16.15%
+	FAIL	Verification	Conformance	Relational	FIELD	The number and percent of records that have a duplicate value in the measurement_id field of the MEASUREMENT. (Threshold=0%).	9.91%
+	FAIL	Verification	Conformance	Relational	FIELD	The number and percent of records that have a duplicate value in the observation_id field of the OBSERVATION. (Threshold=0%).	0.81%
+	FAIL	Verification	Conformance	Value	FIELD	A yes or no value indicating if the visit_occurrence_id in the DRUG_EXPOSURE is the expected data type based on the specification. (Threshold=0%).	0.77%

Showing 26 to 30 of 34 entries (filtered from 1,639 total entries) Previous 1 2 3 4 5 6 7 Next



More Resources

Everything here on COPRHCON 2021 site

More stuff available @

<https://drive.google.com/drive/folders/11iJIG0AoS1KwgOjJM5RxPzqVaunZ6Kt9>

- Everything from COPRHCON 2021
 - Handout
 - My Slides
 - Example DQ report output
- Slides from Ajit Londhe on OHDSI Data Quality Dashboard (four videos about OHDSI DQ @ <https://www.ohdsi.org/2019-tutorial-data-quality>)
- Slides from Keith Marsolo on PCORNet Data Characterization (video @ <https://www.youtube.com/watch?v=4c90VWmXYQc>)
- List of some publications focused on health care data quality
- An old lecture with some fun examples of real-world data quality findings



ACCORDS

