A DATA QUALITY ASSESSMENT AND MANAGING TOOL FOR THE OMOP COMMON DATA MODEL

DQUEEN

Department of Biomedical informatics, Ajou University School of Medicine

Junghyun Byun, BE, Dongsu Park, BE ,SongHeui Oh, MS, Rae Woong Park, MD, Ph.D

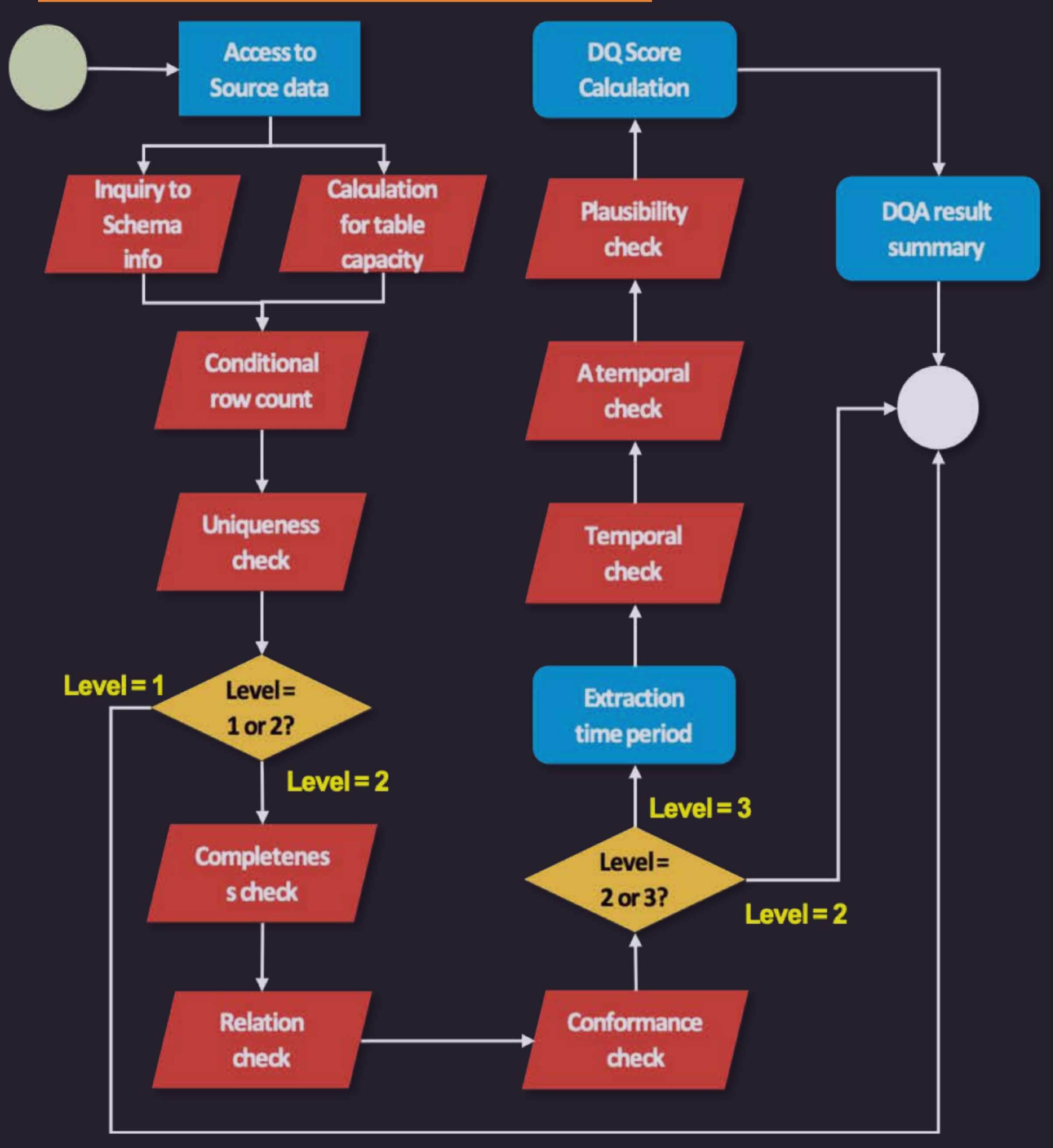
Introduction of DQUEEN

- Each distributed research network (DRN) makes considerable effort to confirm the data quality (DQ) by providing their own DQ tools to ensure that the CDM data is "high-quality" or "ready for research use
  - However, although not significantly different, the existing DQ tools have different terminology, validation scope, and DQ checks and criteria.
  - Furthermore, for most DQ tools, the source data are not regarded well; therefore, they suffer from limitations for confirming the quality of source data.
- The purpose of this study is to confirm **the generalization possibility of DQA** by applying the existing **DQ checks and In-house DQ checks to the source data and OMOP-CDM.**
  - Moreover, we propose **a DQ tool that can perform DQ evaluation between the source data and OMOP-CDM by providing DQA for the source data.**

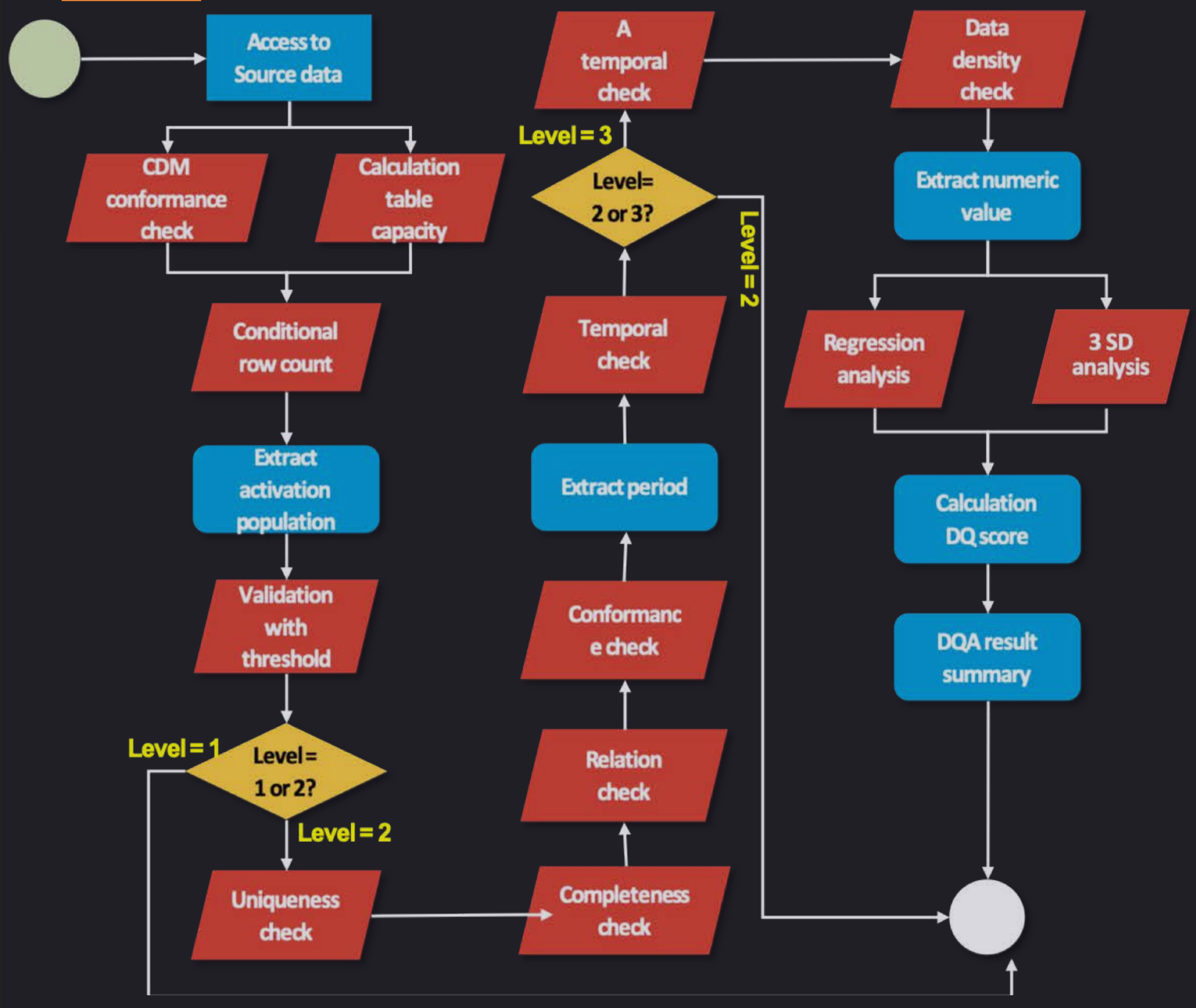
Data Quality Assessment process of DQUEEN

- The system process is designed from the low to high complexity DQ concept, which identifies data errors.
- This tool provides three levels of DQ , and the user can perform a desired level of DQA.

Meta(source data)



CDM



What is metadata?

- Metadata is a structure of elements for each CDM table attribute of source data required for CDM conversion.
  - This structure is not used CDM properties, such as the CDM concept id so It can be assessed the source data

<Source data>

<Meta data>

<Meta_person>

-> person_id, birthday, Sex, location info ..

<Meta_visit>

-> person_id, visit_type, visit date, discharge date ..

<Meta_condition>

-> person_id, diagnosis date, diagnosis code ..

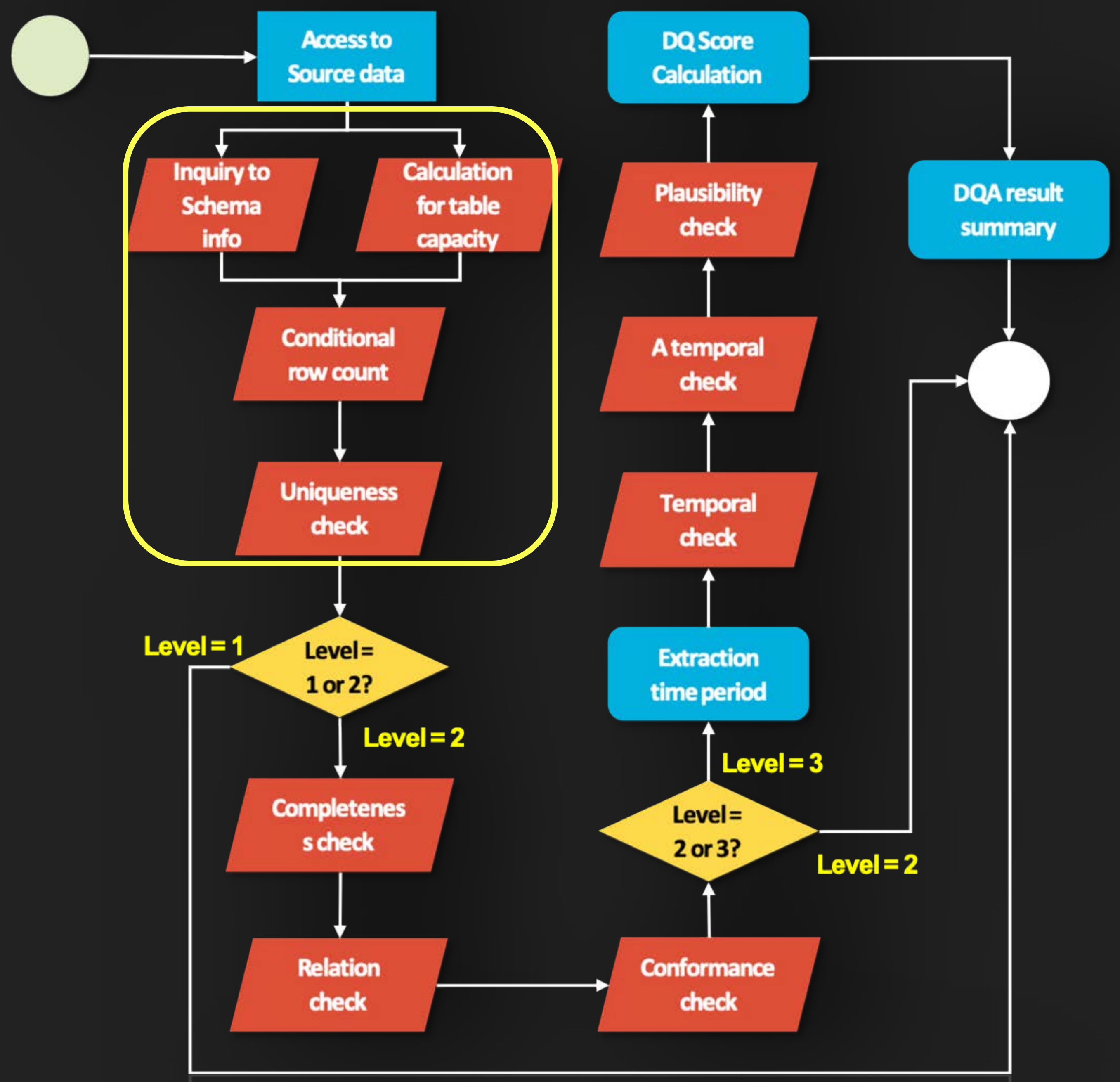
<Meta_procedure>

-> person_id, procedure date, procedure code ..

⋮ ⋮

Data Quality Assessment process (meta part)-1

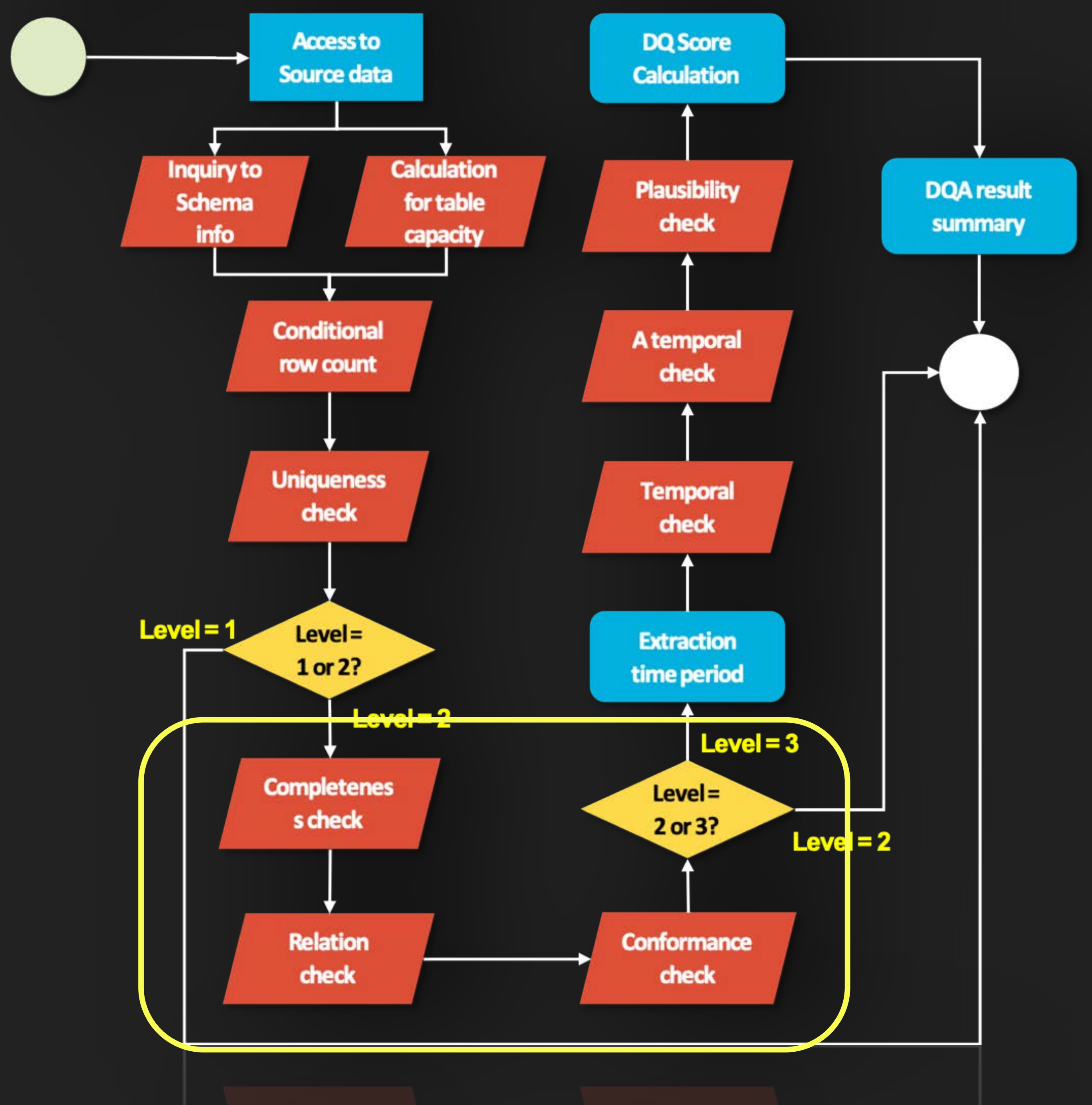
Meta module (Source data)



- DQA Level 1
  - Inquiry to Schema information (E.g. Table list, Table rows, Data type, etc.)
  - Calculation of data table volume size
  - Conditional row count (E.g. Duplicate row check, Missing data, Special char)
  - Data uniqueness check

Data Quality Assessment process (meta part)-2

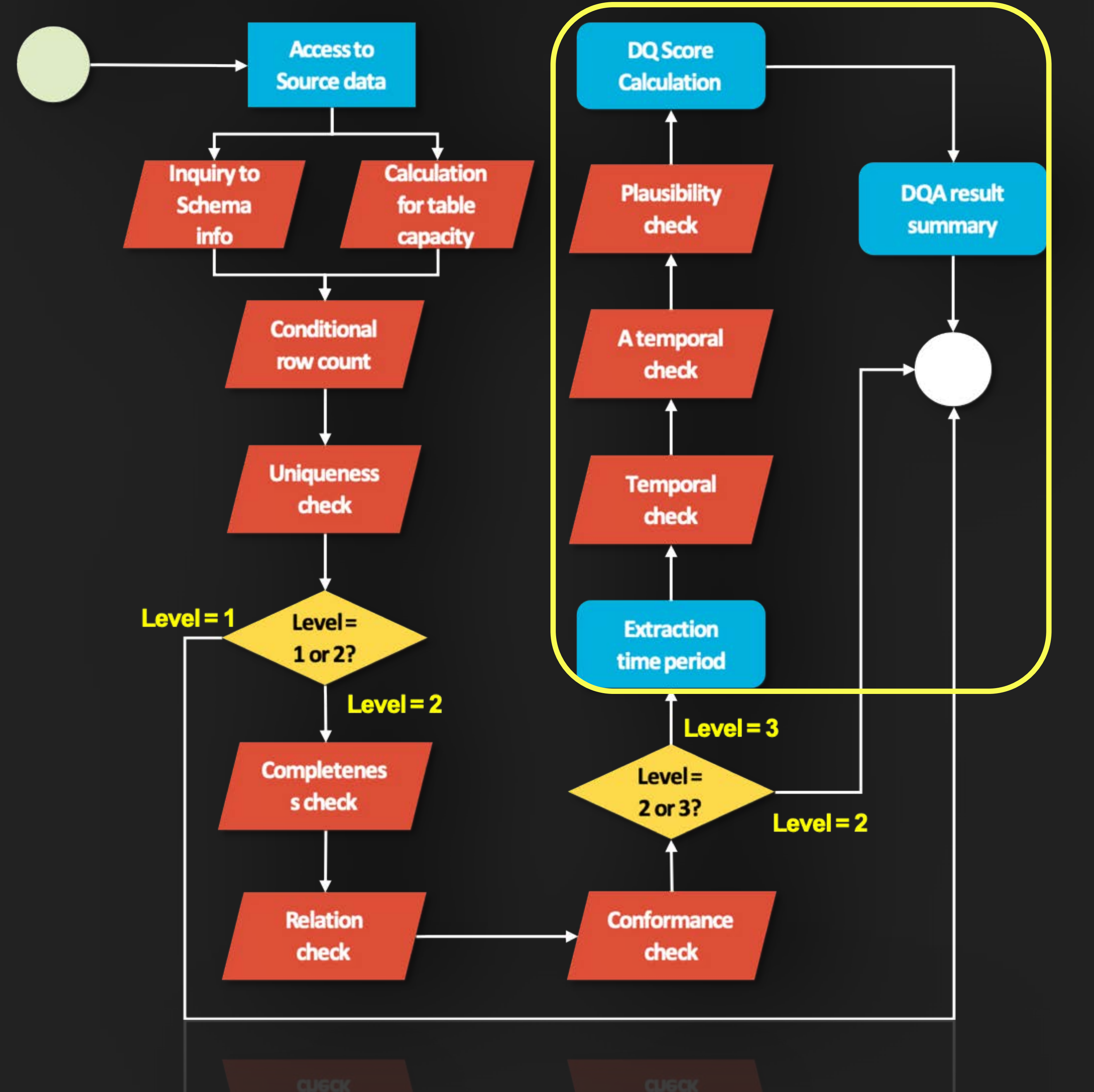
Meta module (Source data)



- DQA Level 2
  - Data completeness check (E.g. Missing data, measurement result value, etc.)
  - Data relation check (E.g. Primary key, Foreign key, etc.)
  - Data conformance check (E.g. Valid with Data type and column type, etc.)

Data Quality Assessment process (meta part)-3

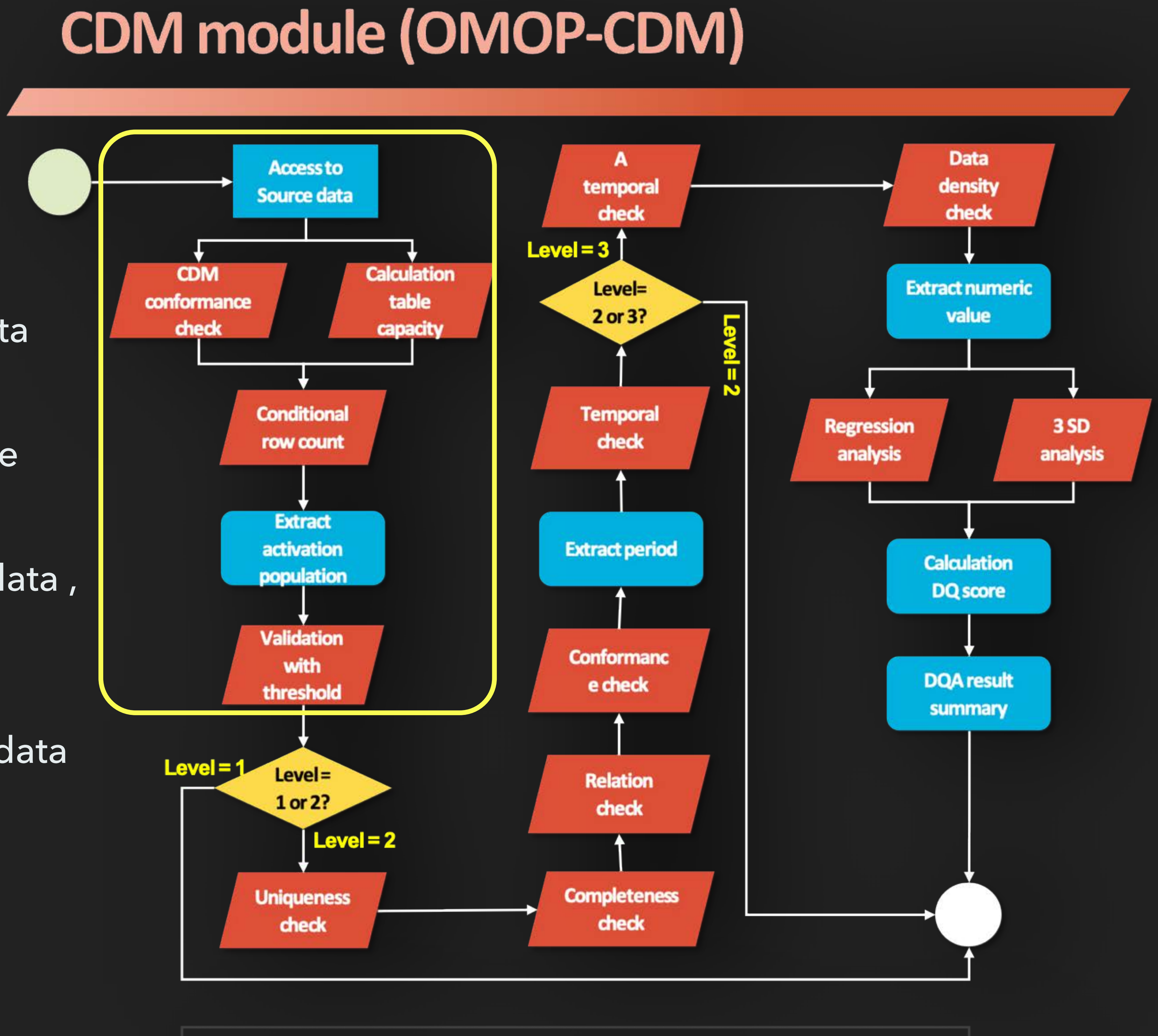
Meta module (Source data)



- DQA Level 3
  - Temporal data error check (E.g. Other suppliers cannot prescribe any order, etc.)
  - Atemporal data error check (E.g. Start date > end date, death date < birth date, etc.)
  - Data quality score calculation with DQA result

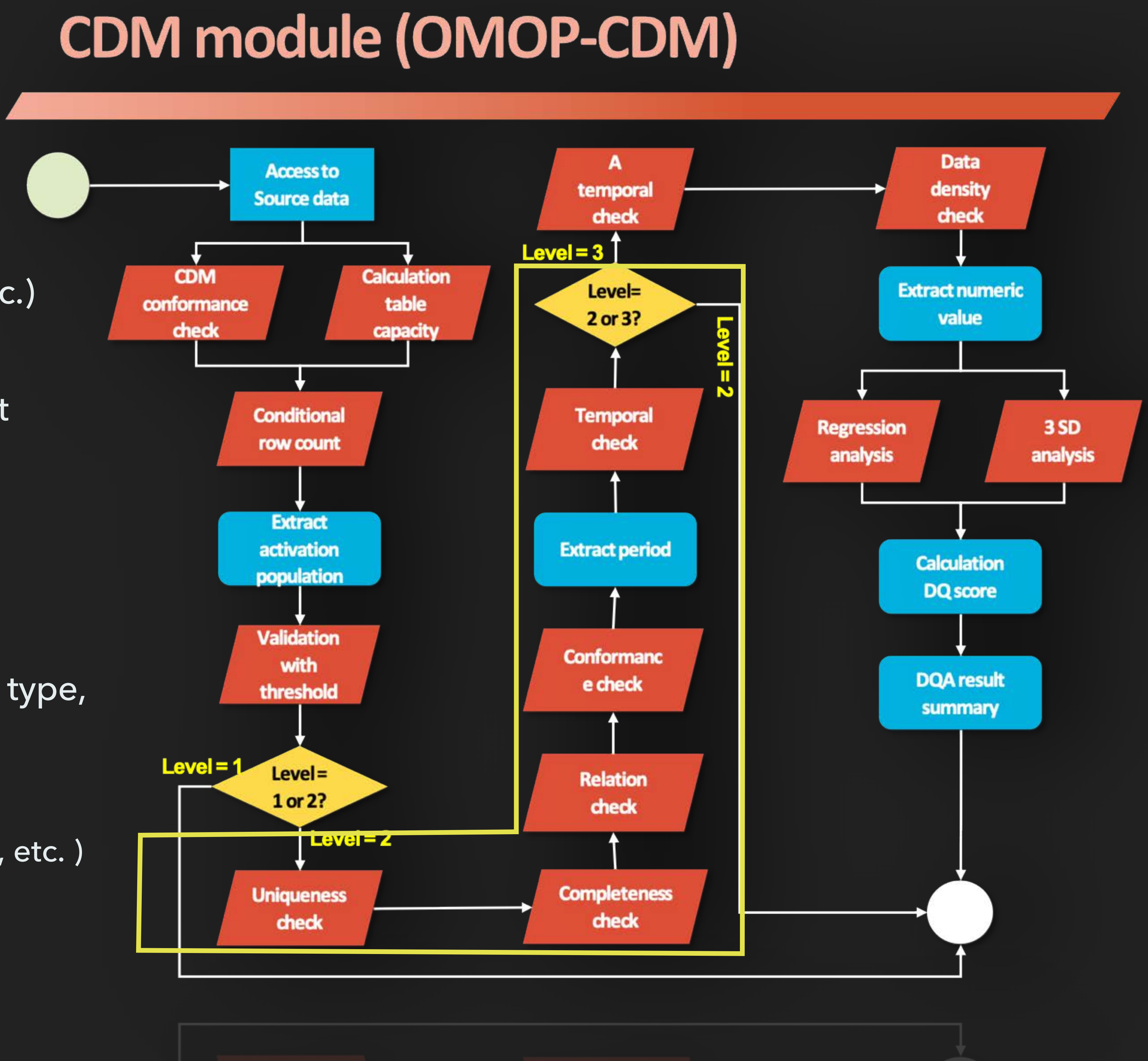
Data Quality Assessment process (CDM part)-1

- DQA Level 1
  - CDM Conformance check (E.g. Table name, column name, Data type, etc.)
  - Calculation of data table volume size
  - Conditional row count (E.g. Duplicate row check, Missing data , Special char)
  - Validation with threshold (E.g. Row count result > number of data population)



Data Quality Assessment process (CDM part)-2

- DQA Level 2
  - Data uniqueness check
    (E.g. Primary key duplicated check, etc.)
  - Data completeness check
    (E.g. Missing data, measurement result value, etc.)
  - Data relation check
    (E.g. Primary key, Foreign key, etc.)
  - Data conformance check
    (E.g. Valid with Data type and column type, etc.)
  - Atemporal error data
    (E.g. Visit_start_date > visit_start_datetime, etc.)



CDM module (OMOP-CDM)

Access to Source data

CDM conformance check

Calculation table capacity

Conditional row count

Extract activation population

Validation with threshold

Level = 1

Level = 1 or 2?

Level = 2

Uniqueness check

Completeness check

Relation check

Conformance check

Extract period

Temporal check

Level = 2 or 3?

Level = 2

Level = 3

A temporal check

Data density check

Extract numeric value

Regression analysis

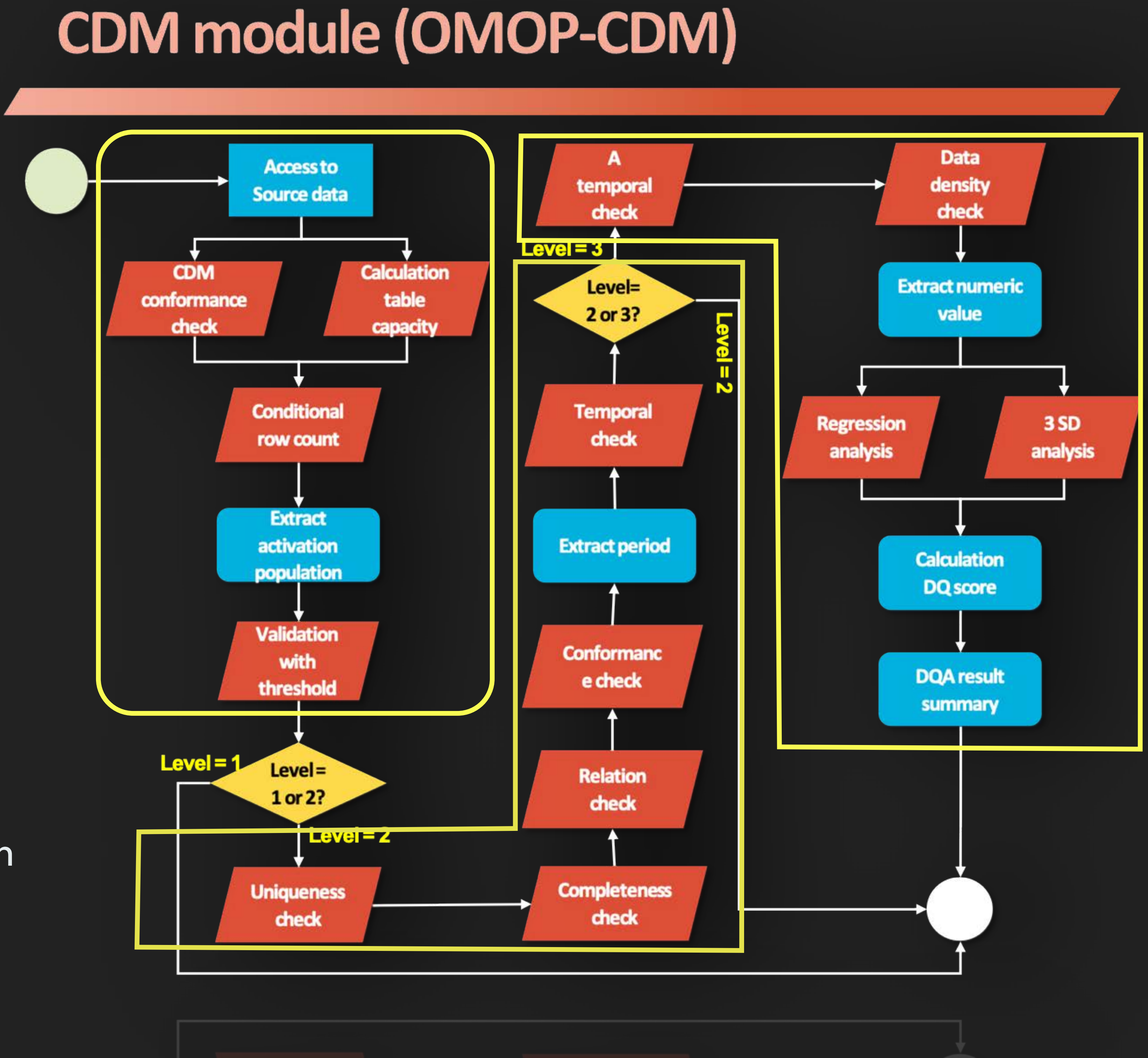
3 SD analysis

Calculation DQ score

DQA result summary

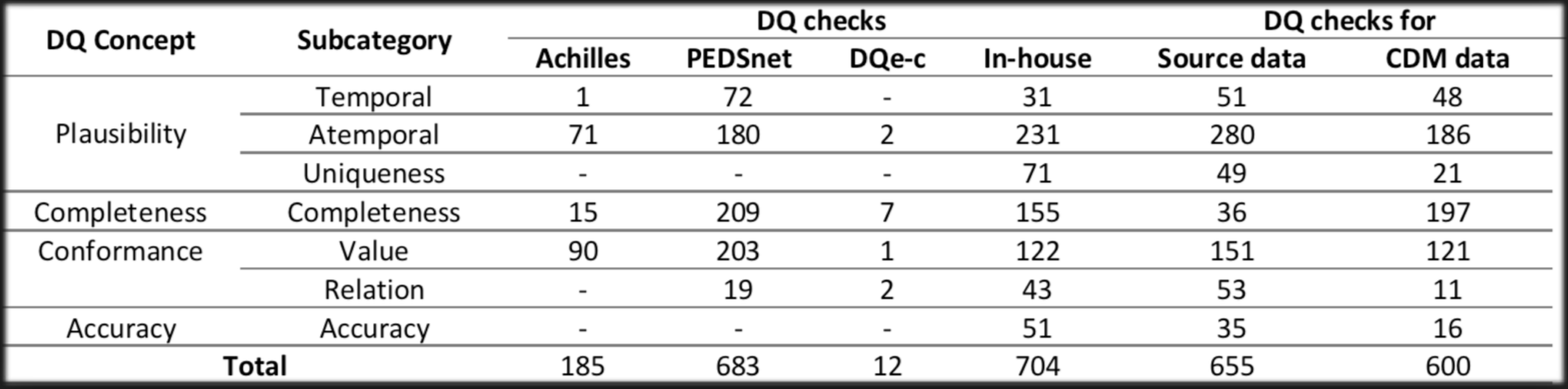
Data Quality Assessment process (CDM part)-3

- DQA Level 3
  - Temporal data error check (E.g. Drug quantity cannot be more than 600 , etc.)
  - Data density check (E.g. Note or Note nlp data density, etc.)
  - Data check for measurement value outlier
    - Linear regression with continuous variable
    - 3 Standard deviation with categorical variables
  - Data quality score calculation with DQA result



DQUEEN's DQ check

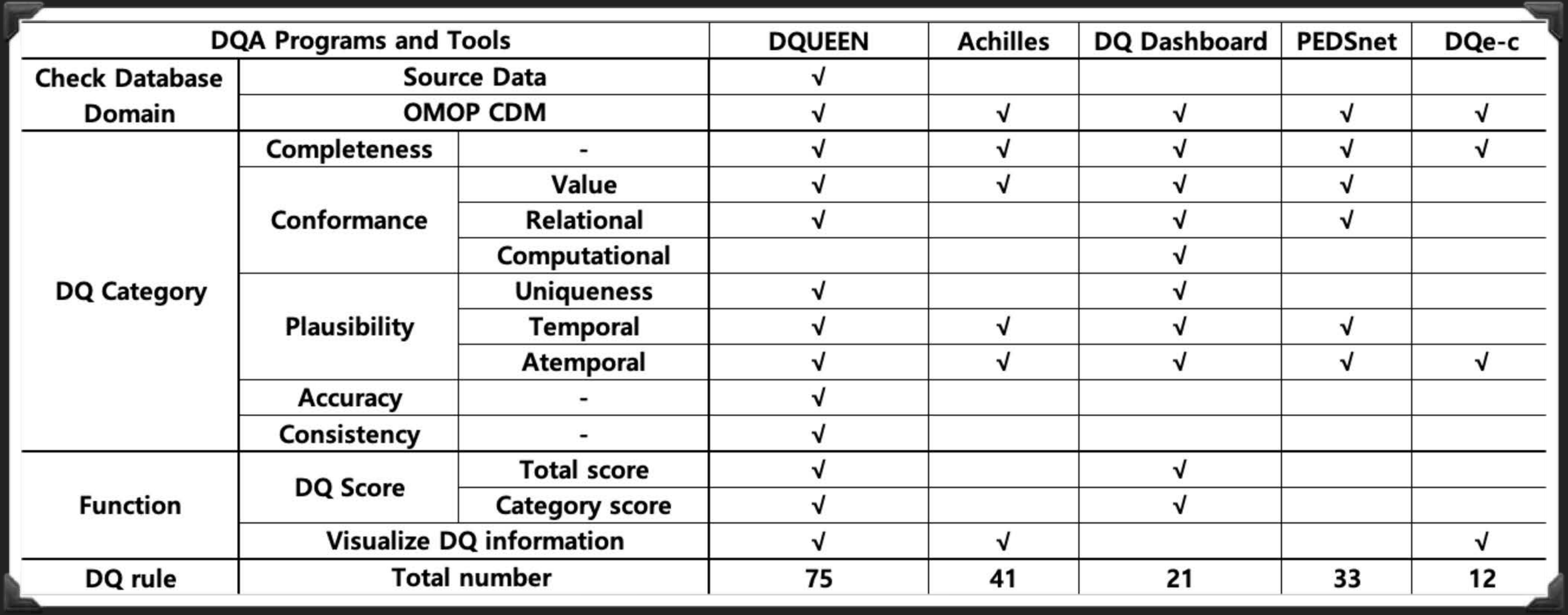
- The total number of DQ checks is 1,255 (742 checks by the three DQ tools and 513 In- house checks).
  - 321 duplicate DQ checks were excluded.
  - We adopt '**Accuracy**' for additional DQ concept
- Table 1 shows the DQ checks applied to each field for DQA.



DQ Concept	Subcategory	DQ checks				DQ checks for	
		Achilles	PEDSnet	DQe-c	In-house	Source data	CDM data
Plausibility	Temporal	1	72	-	31	51	48
	Atemporal	71	180	2	231	280	186
	Uniqueness	-	-	-	71	49	21
Completeness	Completeness	15	209	7	155	36	197
Conformance	Value	90	203	1	122	151	121
	Relation	-	19	2	43	53	11
Accuracy	Accuracy	-	-	-	51	35	16
Total		185	683	12	704	655	600

Table 1. Integrated Data Quality Concept and check- in OHDSI, PEDSnet and DQe-c DQA Process of DQUEEN

Compare with Other DQ tools



DQA Programs and Tools			DQUEEN	Achilles	DQ Dashboard	PEDSnet	DQe-c
Check Database Domain	Source Data		√				
	OMOP CDM		√	√	√	√	√
DQ Category	Completeness	-	√	√	√	√	√
	Conformance	Value	√	√	√	√	
		Relational	√		√	√	
		Computational			√		
	Plausibility	Uniqueness	√		√		
		Temporal	√	√	√	√	
		Atemporal	√	√	√	√	√
	Accuracy	-	√				
	Consistency	-	√				
Function	DQ Score	Total score	√		√		
		Category score	√		√		
	Visualize DQ information		√	√			√
DQ rule	Total number		75	41	21	33	12

Thank you!