



A QUANTITATIVE AND METRICS DRIVEN FRAMEWORK FOR MAXIMIZING DATA UTILITY WHILE BALANCING RE-IDENTIFICATION RISK

PHUSE 2021

June 14-18, 2021



Agenda

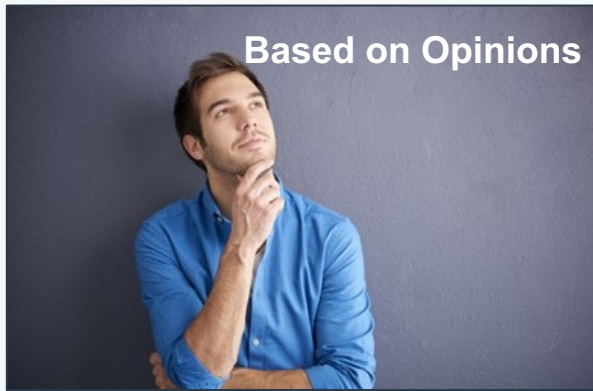
- **Background on Risk Modeling**
- **Background on Anonymization**
- **RLS' Metrics Driven Data Utility Optimization (MD-DUO) Framework**
- **Key Takeaways**

“The difficult part of anonymization isn’t only mitigating the risk – it’s mitigating the risk while *maximizing data utility*”

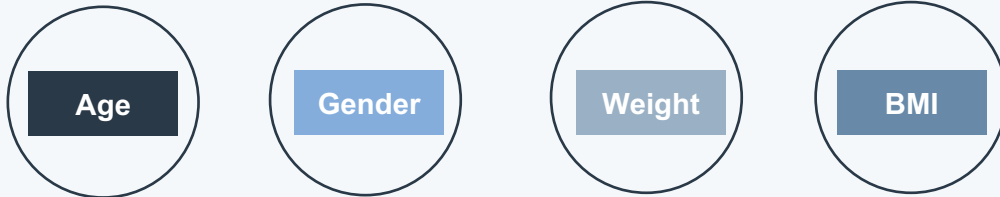
Risk Modeling: Qualitative vs. Quantitative Approaches

- **QUALitative Risk Modeling**

- *Where we're at / we've come from...*
- **Anonymization Rules are typically...**

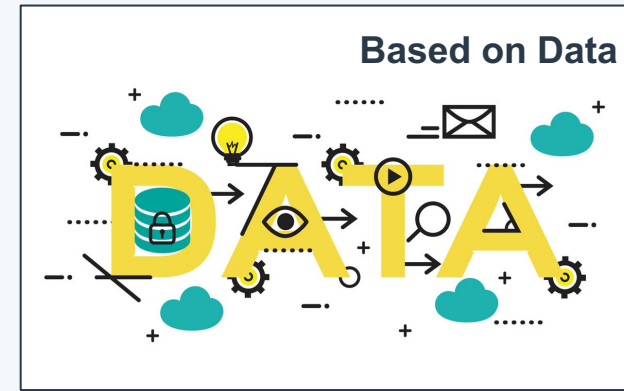


- Developed **identifier by identifier** and **independent from one another...**

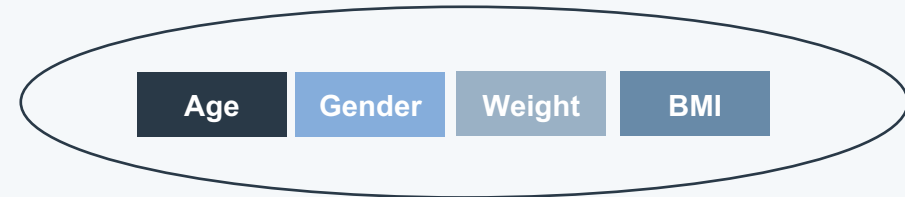


- **QUANTitative Risk Modeling**

- *Where we're going / need to be...*
- **Anonymization Rules are...**



- Developed **by combining all identifiers together...**

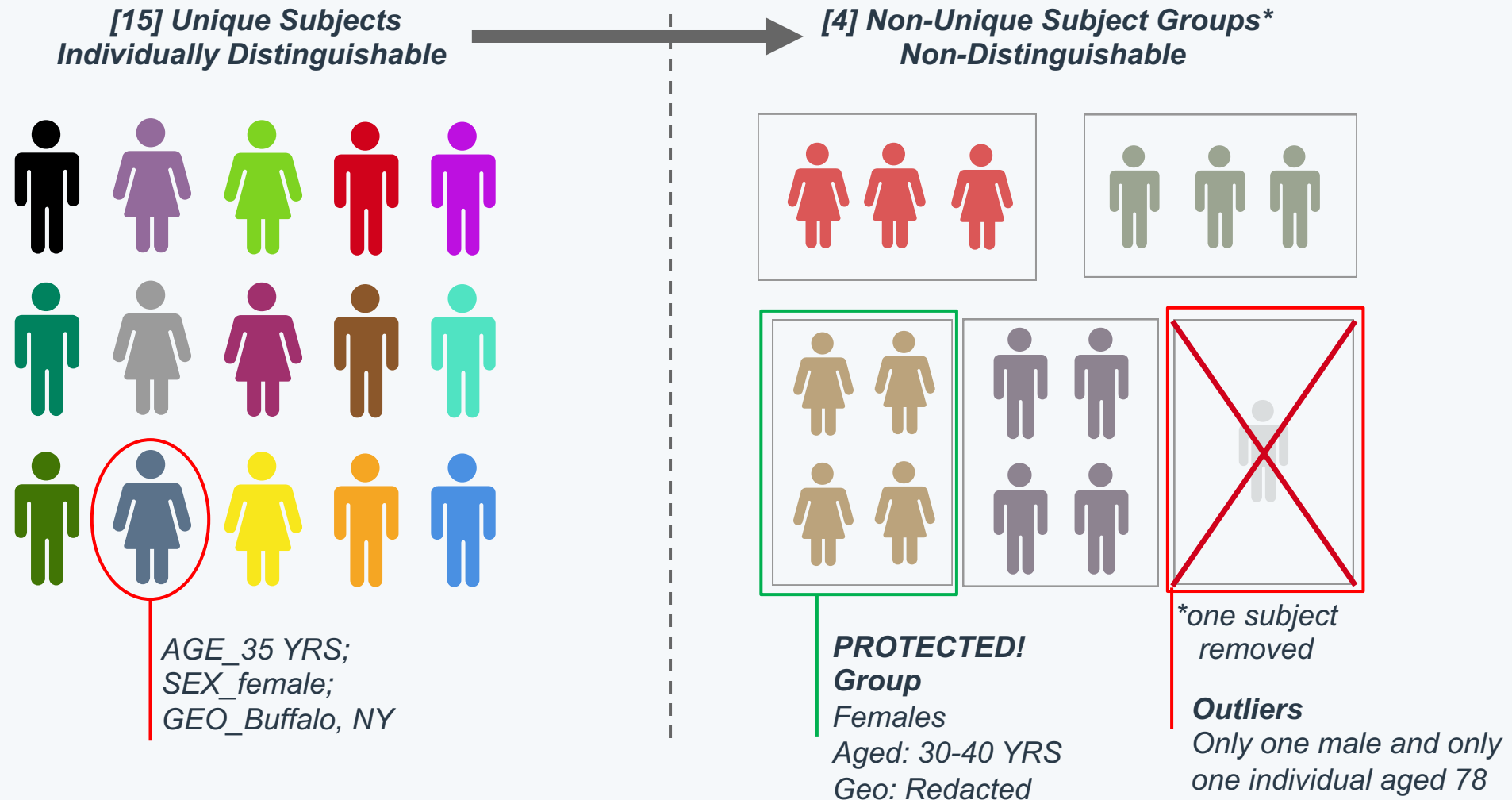


**Quantitative Risk Assessments¹ help
you generate all possible
anonymization options across all
patients *and* all combinations of
identifiers²**

¹ you also need the right [automation] tools/software

² this is incredibly difficult (or time consuming) to do with a qualitative approach

Quantitative Risk Modeling and Anonymization is About Grouping Individuals

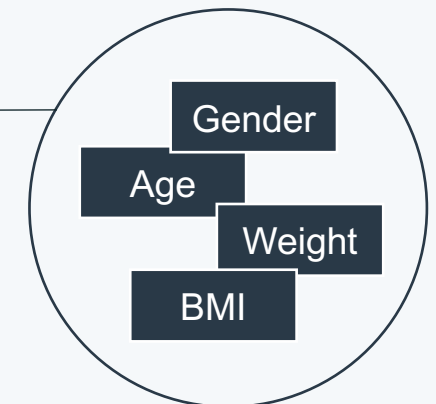


The recipients of disclosures can help determine your 'Group Size'...

- 'Group Sizes' and Risk Thresholds are inextricably linked...



Groups are based on combinations of identifiers



Group Size = 'Equivalence Class' =
The number of individuals in each group

Anonymization helps reduce suppression of identifiers by grouping

Example: Age Transformation Options in Datasets

5 YR, 10 YR, 20 YR, 40 YR and Suppress

Original Values	Transformation Options						Suppress
AGE	Level-0	Level-1	Level-2	Level-3	Level-4	Level-5	
85	18	[15, 20[	[10, 20[	[0, 20[	[0, 40[	*	
80	19	[15, 20[	[10, 20[	[0, 20[	[0, 40[	*	
85	20	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*	
83	21	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*	
84	22	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*	
82	23	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*	
80	24	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*	
83	25	[25, 30[	[20, 30[	[20, 40[	[0, 40[	*	
83	26	[25, 30[	[20, 30[	[20, 40[	[0, 40[	*	
82	27	[25, 30[	[20, 30[	[20, 40[	[0, 40[	*	
80	28	[25, 30[	[20, 30[	[20, 40[	[0, 40[	*	
85	29	[25, 30[	[20, 30[	[20, 40[	[0, 40[	*	
81	30	[30, 35[	[30, 40[	[20, 40[	[0, 40[	*	
84	31	[30, 35[	[30, 40[	[20, 40[	[0, 40[	*	
85	32	[30, 35[	[30, 40[	[20, 40[	[0, 40[	*	
81	33	[30, 35[	[30, 40[	[20, 40[	[0, 40[	*	
85	34	[30, 35[	[30, 40[	[20, 40[	[0, 40[	*	
80	35	[35, 40[	[30, 40[	[20, 40[	[0, 40[	*	
81	36	[35, 40[	[30, 40[	[20, 40[	[0, 40[	*	
80	37	[35, 40[	[30, 40[	[20, 40[	[0, 40[	*	
81	38	[35, 40[	[30, 40[	[20, 40[	[0, 40[	*	
23	39	[35, 40[	[30, 40[	[20, 40[	[0, 40[	*	
36	40	[40, 45[	[40, 50[	[40, 60[	[40, 80[	*	
24	41	[40, 45[	[40, 50[	[40, 60[	[40, 80[	*	
38							

Example: Transformed Age Values in Documents

40 YR Age Band Transformation Option	
Patient 17359385 was a 41-80 yo fe with transient symptoms of aphasia; of neurologic symptoms.	Original Age
Patient 12395438 was a 41-80 yo m difficulty in getting out of bed for three autopsy was performed, and the last stable disease.	Transformed to Age Group
Patient 14836739 was a 41-80 yo m hypertension and diabetes mellitus p nausea, vomiting, diarrhea and posit hospitalization. An autopsy revealed and the wall of the large intestine, w	

RLS' Metrics Driven Data Utility Optimization (MD-DUO) Framework...

1 Generate Possible Transformations	2 Risk of Reidentification (ROR)	3 Clinical Utility (CU)	4 Information Loss (IL)	5 Optimal Approach Based on Tradeoffs
...across all QIs Transformations must meet the minimum risk threshold. Transformations which allow for higher granularity are preferred over outright redaction/suppression.	...across all possible transformation scenarios The ROR metric is used to filter transformation scenarios that meet the risk threshold.	...for independent review of transformation options across QIs Internal sponsor teams prioritize QIs and evaluate the clinical utility (CU) of anonymization. Clinical and Operational Complexity may also be taken into account.	...across all possible transformation scenarios IL metric was used to rank the different transformation scenarios in terms of data utility / data loss.	...using optimal ROR, IL, and CU tradeoffs If new information is found that was previously missed, the process may be updated to incorporate new measurements if necessary.

1. Generate All Possible Transformations Across all Quasi-Identifiers

Source/Input Data

SUBJID	SITEID	AGE	GENDER	ETHNICITY	WEIGHT	BMI	CC
14231020	G106	84	FEMALE	Hispanic or Latino	60	25.77785520	FRA
13271017	R024	83	MALE	Black	60	25.77785520	FRA
10551019	F087	48	FEMALE	Hispanic or Latino	60	25.77785520	FRA
12251001	A062	63	FEMALE	Hispanic or Latino	60	25.77785520	FRA
10571011	F087	41	MALE	Black	60	25.77785520	FRA
12091002	G106	70	MALE	Black	60	25.77785520	FRA
12421019	A062	73	MALE	Black	60	25.77785520	FRA
15211011	G106	56	MALE	Hispanic or Latino	60	25.77785520	FRA
11071009	R024	66	MALE	Hispanic or Latino	60	25.77785520	FRA
11481025	F087	69	MALE	Hispanic or Latino	60	25.77785520	FRA
10421025	R024	83	MALE	Black	60	29.99801349	FRA
10431021	F087	81	MALE	Hispanic or Latino	60	29.99801349	FRA
13281025	A062	59	FEMALE	Hispanic or Latino	60	29.99801349	FRA
14481022	R024	53	MALE	Black	60	29.99801349	FRA
14411006	G106	40	MALE	Black	60	29.99801349	FRA
13231016	A062	47	MALE	Black	60	29.99801349	FRA
12101014	G106	63	MALE	Black	60	29.99801349	FRA
13211016	A062	79	MALE	Black	60	29.99801349	FRA
11411014	F087	60	MALE	Black	60	29.99801349	FRA
13071025	G106	50	MALE	Hispanic or Latino	60	29.99801349	FRA
11261018	F087	44	MALE	Hispanic or Latino	60	29.99801349	FRA
14501022	R024	65	MALE	Hispanic or Latino	60	29.99801349	FRA
15501018	R024	39	FEMALE	Black	62	30.71901334	FRA
14311002	F087	62	FEMALE	Black	62	30.71901334	FRA
11191025	G106	69	MALE	Black	62	30.71901334	FRA
14481020	A062	78	MALE	Black	62	30.71901334	FRA
12491017	R024	40	MALE	Hispanic or Latino	62	30.71901334	FRA
11551002	G106	59	MALE	Hispanic or Latino	62	30.71901334	FRA
13031021	F087	43	MALE	Hispanic or Latino	62	30.71901334	FRA
10021015	A062	74	MALE	Hispanic or Latino	62	30.71901334	FRA
14031007	F087	81	FEMALE	Black	63	24.06682986	FRA
10231024	G106	85	MALE	Black	63	24.06682986	FRA
10431019	F087	83	MALE	Hispanic or Latino	63	24.06682986	FRA
14021001	R024	79	FEMALE	Hispanic or Latino	63	24.06682986	FRA
13211018	A062	55	MALE	Black	63	24.06682986	FRA
15411006	R024	74	MALE	Black	63	24.06682986	FRA
11291015	R024	52	MALE	Hispanic or Latino	63	24.06682986	FRA
15491017	A062	42	MALE	Hispanic or Latino	63	24.06682986	FRA
11591024	F087	58	MALE	Hispanic or Latino	63	24.06682986	FRA
12111004	G106	61	MALE	Hispanic or Latino	63	24.06682986	FRA

AGE Transformation Options

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
18	[15, 20[	[10, 20[	[0, 20[	[0, 40[	*
19	[15, 20[	[10, 20[	[0, 20[	[0, 40[	*
20	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*
21	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*
22	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*
23	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*
24	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*
25	[25, 30[	[20, 30[	[20, 40[	[0, 40[	*
26	[25, 30[	[20, 30[	[20, 40[	[0, 40[	*

ETHNICITY Transformation Options

Level-0	Level-1	Level-2
Black	Non-Caucasian	*
Hispanic or La...	Non-Caucasian	*
White	Caucasian	*

COUNTRY Transformation Options

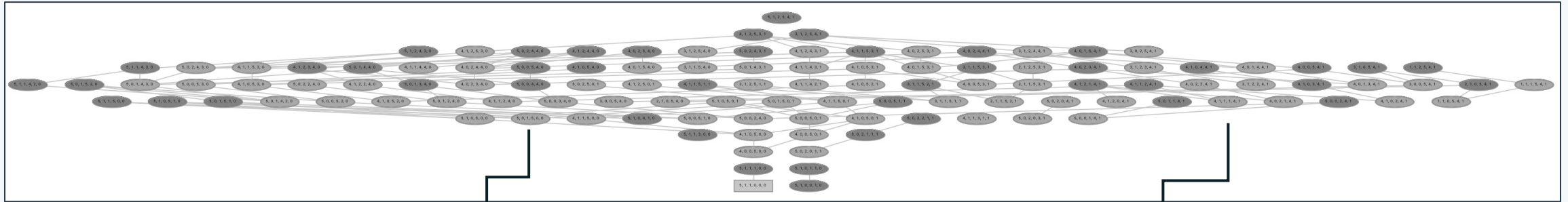
Level-0	Level-1
FRA	EUR
GER	EUR
POL	EUR
USA	NA

BMI Transformation Options

Level-0	Level-1	Level-2	Level-3	Level-4
21.01948691	[20, 25[	[20, 30[	[10, 30[	*
21.07208942	[20, 25[	[20, 30[	[10, 30[	*
21.08639800	[20, 25[	[20, 30[	[10, 30[	*
21.12680354	[20, 25[	[20, 30[	[10, 30[	*
21.17533564	[20, 25[	[20, 30[	[10, 30[	*
21.38113387	[20, 25[	[20, 30[	[10, 30[	*
21.49368841	[20, 25[	[20, 30[	[10, 30[	*

2a. Generate Transformation Options Based on Combinations of Identifiers

All Combinations of Transformations (160+)



Example Group 1

4, 0, 2, 4, 4, 1

Identifier	Level	Transformation
Age	4	40 YR BANDS
Gender	0	RETAIN
Ethnicity	2	SUPPRESS
Weight	4	40 KG BANDS
BMI	4	SUPPRESS
Country	1	SUPPRESS

ANONYMOUS

Example Group 2

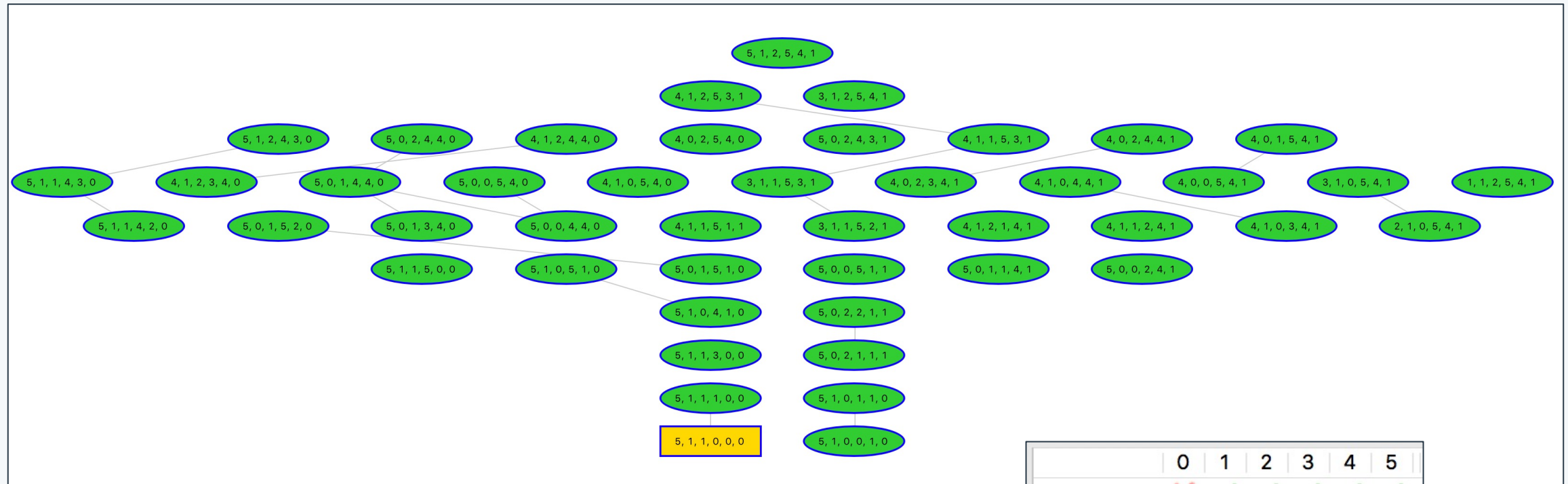
3, 0, 0, 5, 4, 0

Identifier	Level	Transformation
Age	3	20 YR BANDS
Gender	0	RETAIN
Ethnicity	0	RETAIN
Weight	5	SUPPRESS
BMI	4	SUPPRESS
Country	0	RETAIN

WARNING: NON-ANONYMOUS

2b. Reduce Transformation Options to Only Anonymous Transformations

All Combinations of **Anonymous** Transformations (39)



Note: Cannot retain original AGE values – anonymous options require transformation

	0	1	2	3	4	5
AGE	✗	✓	✓	✓	✓	✓
GENDER	✓	✓				
ETHNICITY	✓	✓	✓			
WEIGHT	✓	✓	✓	✓	✓	✓

3. Refine Anonymous Transformations Based on Clinical Utility (CU)

All Combinations of **Anonymous** Transformations (5) that allow Age Transformation (vs Suppression) and Retaining Gender



	0	1	2	3	4	5
AGE	✗	✓	✓	✓	✓	✗
GENDER	✓	✗				
ETHNICITY	✓	✓	✓			
WEIGHT	✓	✓	✓	✓	✓	✓

Why this can be helpful/necessary:

- Secondary research into Gender discrepancies
- Studying Elderly vs. Younger Subpopulations

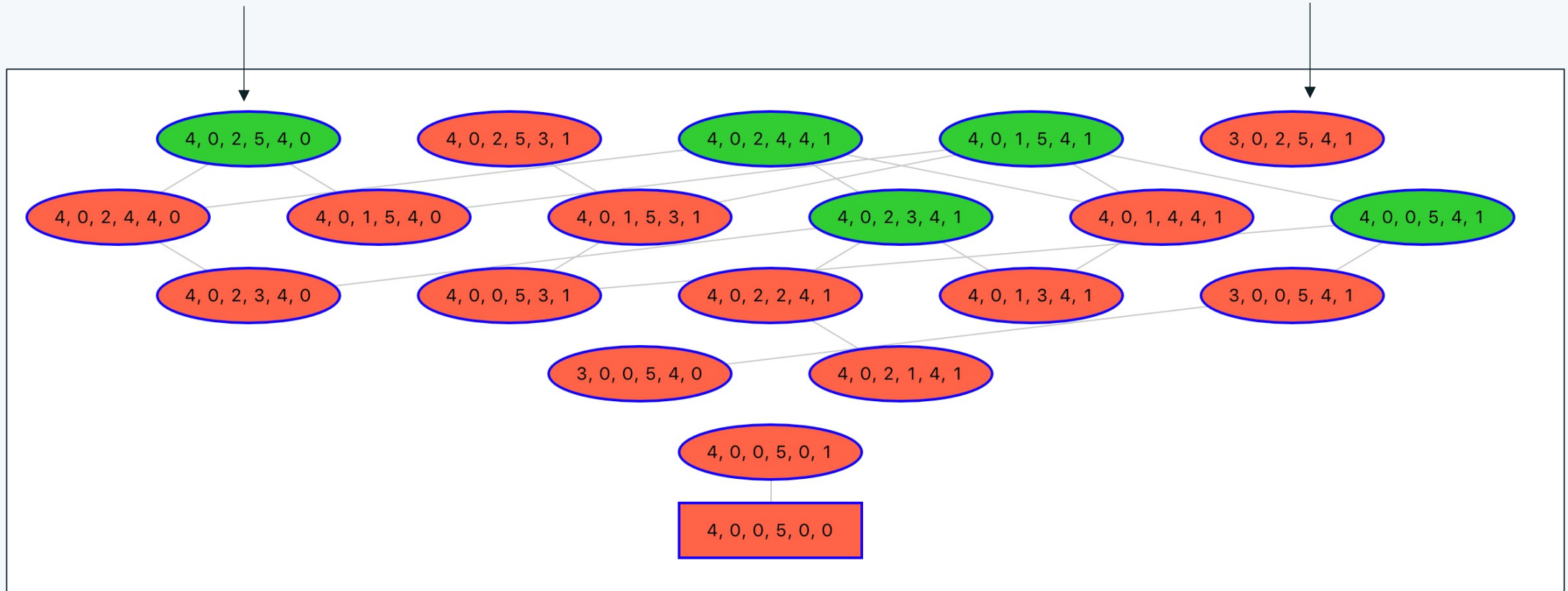
NOTE: Researchers often request rules with NON-ANONYMOUS groupings!

ANONYMOUS Transformations (5)

The only options you can provide

NON-ANONYMOUS Transformations (15)

Options the requester might want/request...



4. Ranking Information Loss (Quantitatively)

List View: Transformation Groups

- Apply **Quantitative Data Utility models** such as **Information Loss (IL)**
- Shows **how much information is lost** per each Grouping
- Helps you select options that remove the **least amount data**

Transformation	Anonymity	Min. score
[5, 1, 1, 0, 0, 0]	ANONYMOUS	0.32192397162915776 [32.1924%]
[5, 0, 0, 4, 4, 0]	ANONYMOUS	0.32932769255608063 [32.93277%]
[5, 1, 1, 1, 0, 0]	ANONYMOUS	0.33664262512125154 [33.66426%]
[5, 1, 0, 0, 1, 0]	ANONYMOUS	0.33679202033414257 [33.6792%]
[5, 1, 0, 1, 1, 0]	ANONYMOUS	0.35167621863946974 [35.16762%]
[5, 0, 0, 2, 4, 1]	ANONYMOUS	0.358153536604926 [35.81535%]
[5, 1, 1, 3, 0, 0]	ANONYMOUS	0.3657199418971466 [36.57199%]
[5, 0, 1, 3, 4, 0]	ANONYMOUS	0.36571994189714685 [36.57199%]
[5, 0, 1, 4, 4, 0]	ANONYMOUS	0.3947462367479162 [39.47462%]
[5, 0, 1, 5, 1, 0]	ANONYMOUS	0.4025778971704015 [40.25779%]
[5, 1, 0, 4, 1, 0]	ANONYMOUS	0.4104333378400504 [41.04333%]
[5, 0, 1, 1, 4, 1]	ANONYMOUS	0.41163601092794244 [41.1636%]
[5, 0, 0, 5, 1, 1]	ANONYMOUS	0.41179378807675837 [41.17938%]
[5, 0, 0, 5, 4, 0]	ANONYMOUS	0.41421356237309515 [41.42136%]
[4, 0, 0, 5, 4, 1]	ANONYMOUS	0.4245510125004821 [42.4551%]
[5, 0, 2, 1, 1, 1]	ANONYMOUS	0.4275130760350385 [42.75131%]
[5, 0, 2, 2, 1, 1]	ANONYMOUS	0.4410179195542048 [44.10179%]
[5, 0, 1, 5, 2, 0]	ANONYMOUS	0.45872806359159823 [45.87281%]
[4, 1, 0, 3, 4, 1]	ANONYMOUS	0.47174706547152767 [47.17471%]
[4, 0, 2, 3, 4, 1]	ANONYMOUS	0.47174706547152767 [47.17471%]
[5, 1, 1, 5, 0, 0]	ANONYMOUS	0.483809488904126 [48.38095%]
[5, 0, 2, 4, 4, 0]	ANONYMOUS	0.49211988466087053 [49.21199%]
[4, 0, 1, 5, 4, 1]	ANONYMOUS	0.49465566305928754 [49.46557%]
[5, 1, 0, 5, 1, 0]	ANONYMOUS	0.5004983093078867 [50.04983%]
[4, 1, 0, 4, 4, 1]	ANONYMOUS	0.5030268051587061 [50.30268%]
[4, 0, 2, 4, 4, 1]	ANONYMOUS	0.5030268051587061 [50.30268%]
[4, 1, 0, 5, 4, 0]	ANONYMOUS	0.5140570838547955 [51.40571%]
[4, 0, 2, 5, 4, 0]	ANONYMOUS	0.5140570838547955 [51.40571%]
[4, 1, 1, 2, 4, 1]	ANONYMOUS	0.5255950353526435 [52.5595%]
[2, 1, 0, 5, 4, 1]	ANONYMOUS	0.5372351738214316 [53.72352%]
[5, 1, 1, 4, 2, 0]	ANONYMOUS	0.5390866046748071 [53.90866%]
[5, 1, 1, 4, 3, 0]	ANONYMOUS	0.5390866046748071 [53.90866%]
[5, 0, 2, 4, 3, 1]	ANONYMOUS	0.5491994506513005 [54.91995%]
[3, 1, 0, 5, 4, 1]	ANONYMOUS	0.551514993913153 [55.1515%]
[4, 1, 2, 3, 4, 0]	ANONYMOUS	0.5642185155646857 [56.42185%]
[4, 1, 1, 5, 1, 1]	ANONYMOUS	0.5858483860489572 [58.58484%]
[4, 1, 2, 4, 4, 0]	ANONYMOUS	0.5974635949187606 [59.74636%]
[3, 1, 1, 5, 2, 1]	ANONYMOUS	0.6003512644204019 [60.03513%]

Lower % =
Lower IL

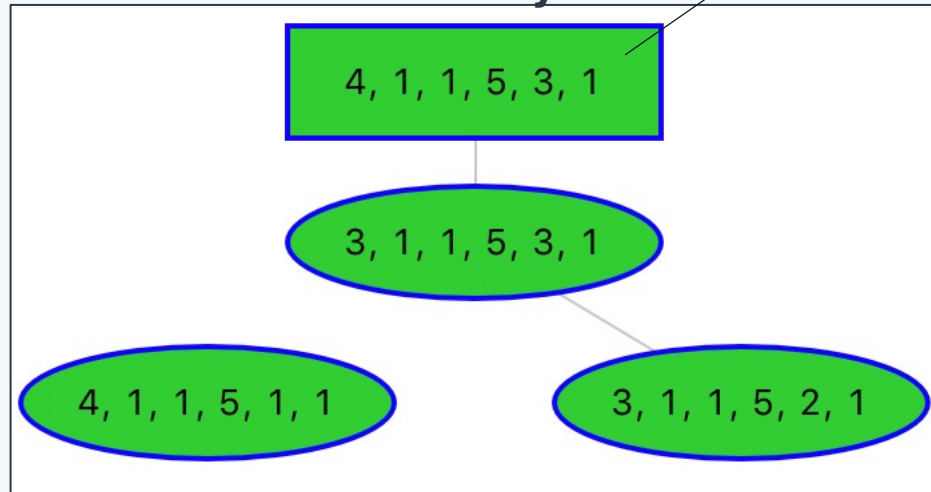
Highest % =
Higher IL

5. Making the Trade Off and Bringing it All Together

1. Filter Transformation Levels

	0	1	2	3	4	5
AGE	✓	✓	✓	✓	✓	✗
GENDER	✓	✓				
ETHNICITY	✓	✓	✗			
WEIGHT	✓	✓	✓	✓	✓	✓
BMI	✓	✓	✓	✓	✗	

2a. Select Desired Option Based on Clinical Utility



3. Generate Final Transformation Rules

Identifier	Level	Transformation
Age	4	40 YR BANDS
Gender	1	SUPPRESS
Ethnicity	1	CAUCASIAN / NON-CAUCASIAN
Weight	5	SUPPRESS
BMI	4	20 UNIT BANDS
Country	1	SUPPRESS

2b. Select Desired Option Considering Information Loss (IL)

Transformation	Anonymity	Min. score
[4, 1, 1, 5, 1, 1]	ANONYMOUS	0.5858483860489572 [58.58484%]
[3, 1, 1, 5, 2, 1]	ANONYMOUS	0.6003512644204019 [60.03513%]
[3, 1, 1, 5, 3, 1]	ANONYMOUS	0.6003512644204019 [60.03513%]
[4, 1, 1, 5, 3, 1]	ANONYMOUS	0.6493355199721982 [64.93355%]

Summary Remarks and Key Takeaways

- **Quantitative Risk Modeling is critical to generating anonymous transformation options that can meet data utility objectives**
- **RLS' Metrics Driven Data Utility Optimization (MD-DUO) Framework provides a structured, high efficiency way to incorporate quantitative methodologies**
- **The quantitative approaches can take subjectivity out of the equation and provide a definitive set of guidelines from which to base disclosures**
- **Some important considerations...**
 - Vast majority of data sharing today uses the qualitative approach
 - Quantitative risk assessments often demonstrate non-anonymous results when evaluating qualitative rules
 - Data sharing agreements can give you a false sense of security and there is a tendency to reduce conservatism
 - Qualitative approaches are not wrong but quantitative approaches underpinning anonymization efforts serves as a more robust, repeatable and defensible approach



**A QUANTITATIVE AND METRICS DRIVEN FRAMEWORK FOR MAXIMIZING
DATA UTILITY WHILE BALANCING RE-IDENTIFICATION RISK**

FOR FOLLOW UP: JLYONS@RLSCIENCES.COM

PHUSE 2021

June 14-18, 2021

