Health Care Data Harmonization Using: Shiny, Clinical Expertise, and RDBMS

Seth Russell, MS
Peter E. DeWitt, Ph.D.

useR!
Aug 2025

Department of Biomedical Informatics
SCHOOL OF MEDICINE
UNIVERSITY OF COLORADO **ANSCHUTZ MEDICAL CAMPUS**

Disclosures

This work was supported by:

Eunice Kennedy Shriver National Institute of Child Health and Human Development grant R01HD105939 to Drs Sanchez-Pinto and Bennett.

Related Publications:

Development and Validation of the Phoenix Criteria for Pediatric Sepsis and Septic Shock

L. Nelson Sanchez-Pinto*, MD, MBI; Tellen D. Bennett*, MD, MS; Peter E. DeWitt**, PhD; Seth Russell**,MS; Margaret N. Rebull, MA; Blake Martin, MD; Samuel Akech, MBChB, MMED; David J. Albers, PhD;Elizabeth R. Alpern, MD, MSCE; Fran Balamuth, MD, PhD, MSCE; Melania Bembea, MD, MPH, PhD; Mohammod Jobayer Chisti, MBBS, MMed, PhD; Idris Evans, MD, MSc; Christopher M. Horvat, MD, MHA; Juan Camilo Jaramillo-Bustamante, MD; Niranjan Kissoon, MD; Kusum Menon, MD, MSc; Halden F. Scott, MD, MSCS; Scott L. Weiss, MD; Matthew O. Wiens, PharmD, PhD; Jerry J. Zimmerman, MD, PhD; Andrew C. Argent***, MD, MBBCh, MMeg; Lauren R. Sorce***, PhD, RN, CPNP-AC/PC; Luregn J. Schlapbach***, MD, PhD; R. Scott Watson***, MD, MPH; and the Society of Critical Care Medicine Pediatric Sepsis Definition Task Force JAMA. https://doi.org/10.1001/jama.2024.0196 Published online January 21, 2024.

International Consensus Criteria for Pediatric Sepsis and Septic Shock

Schlapbach LJ, Watson RS, Sorce LR, et al. *JAMA.* 2024;331(8):665–674. https://doi.org/10.1001/jama.2024.0179

https://github.com/magic-lantern/2025-useR

Disclosures

The study was approved with a waiver of consent by a central institutional review board at the University of Colorado, plus separate regulatory approvals at non-US sites.

https://github.com/magic-lantern/2025-useR

Motivation

Sepsis – infection with life-threatening organ dysfunction.

Our work focused on **Pediatric Sepsis**. Many pediatric survivors of sepsis have ongoing physical, cognitive, emotional, and psychological sequelae, which may have long-term effects on them and their families. Sepsis in children has important differences to adult sepsis, including:

- Age-specific variability of vital signs
- developmental age-dependent immune function,
- Pediatric-specific comorbidities, epidemiology, and outcomes

Sepsis is a leading cause of death among children worldwide. Prior to this work the most current pediatric-specific criteria for sepsis were published in 2005 based on expert opinion. In 2016, the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) defined sepsis as life-threatening organ dysfunction caused by a dysregulated host response to infection, but it excluded children.

https://github.com/magic-lantern/2025-useR

Study

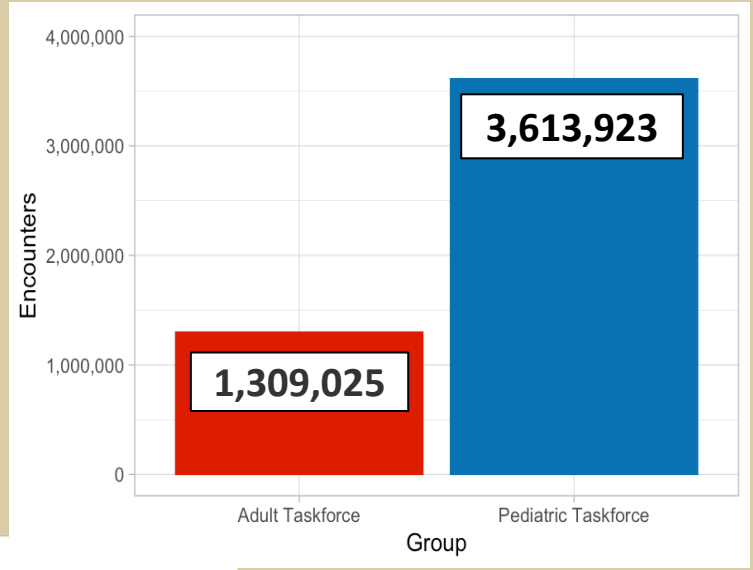
Aim 1: Determine the optimal clinical criteria for each pediatric organ dysfunction in differently resourced settings and care environments

Aim 2: Develop and validate novel pediatric sepsis criteria

Aim 3: Design, build, and evaluate prototype CDS tools to facilitate use of the new pediatric sepsis criteria

https://github.com/magic-lantern/2025-useR

Study



Ann & Robert H. Lurie Children's Hospital of Chicago

UPMC CHILDREN'S HOSPITAL OF PITTSBURGH

Seattle Children's HOSPITAL · RESEARCH · FOUNDATION

Children's Hospital of Philadelphia

JOHNS HOPKINS CHILDREN'S CENTER

浙江大学医学院附属儿童医院 THE CHILDREN'S HOSPITAL ZHEJIANG UNIVERSITY SCHOOL OF MEDICINE

Children's Hospital Colorado

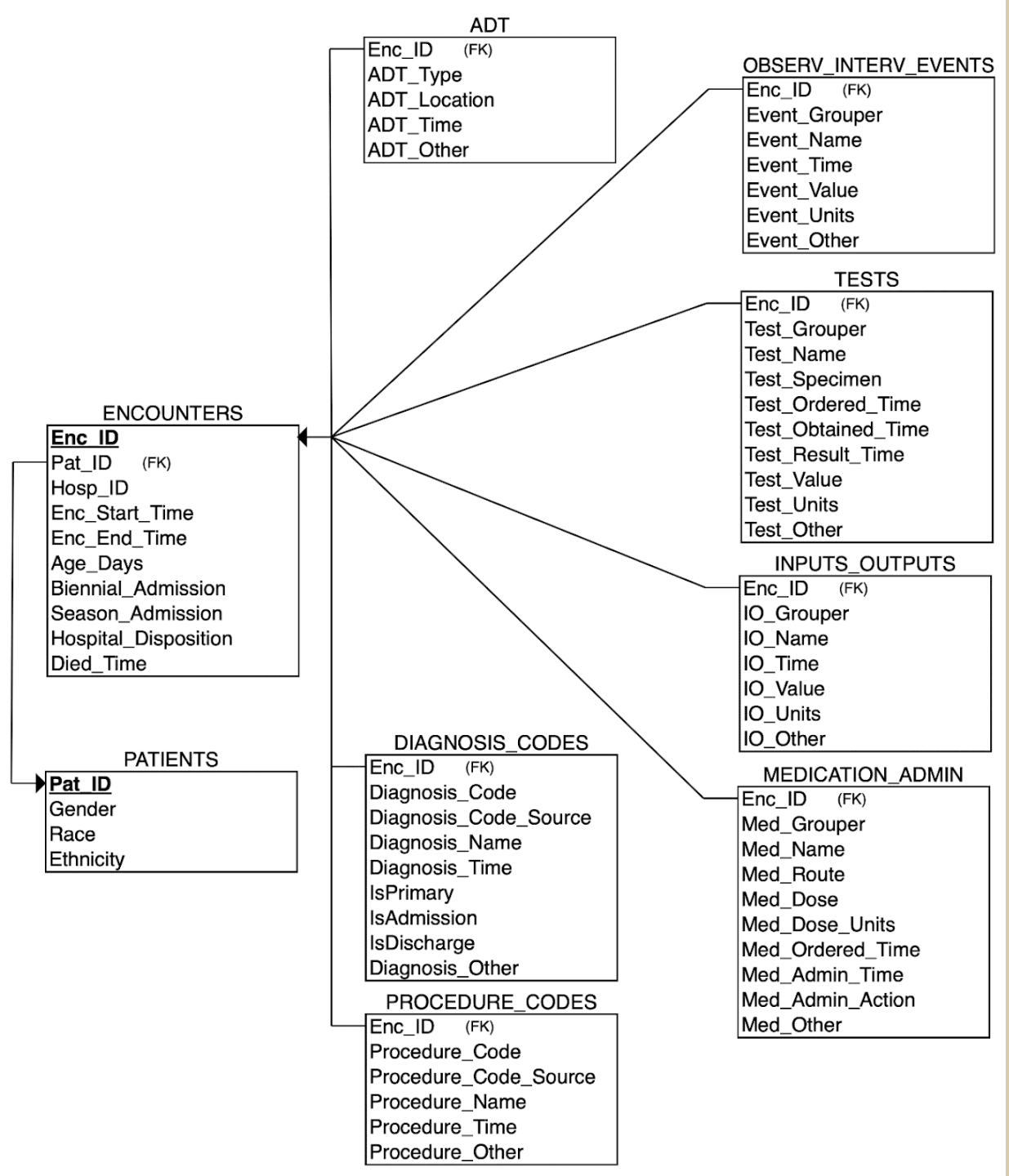
HGM Hospital General de Medellín Luz Castro de Gutiérrez - E.S.E.

KENYA MEDICAL RESEARCH INSTITUTE KEMRI

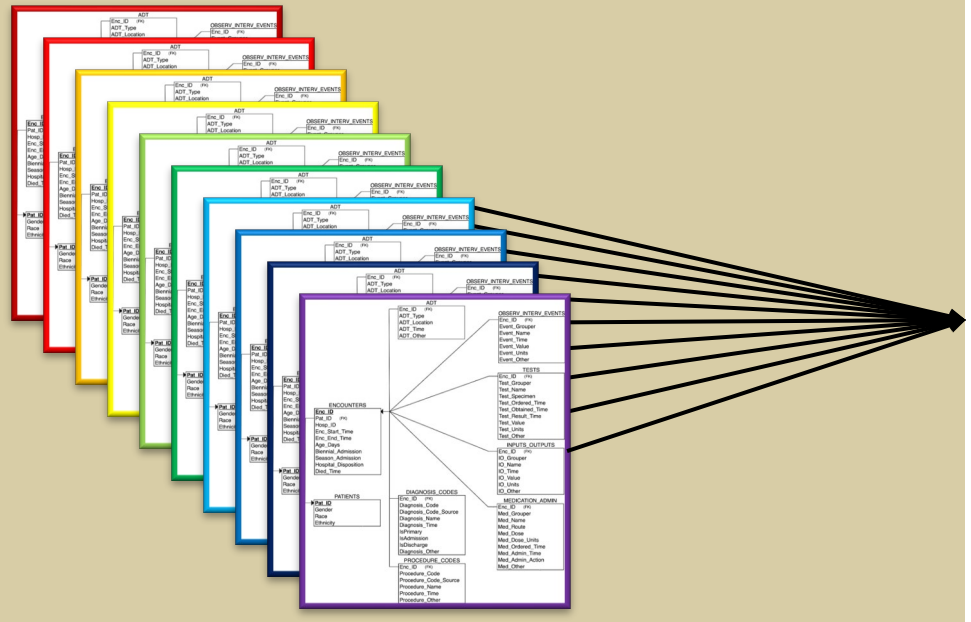
icddr,b

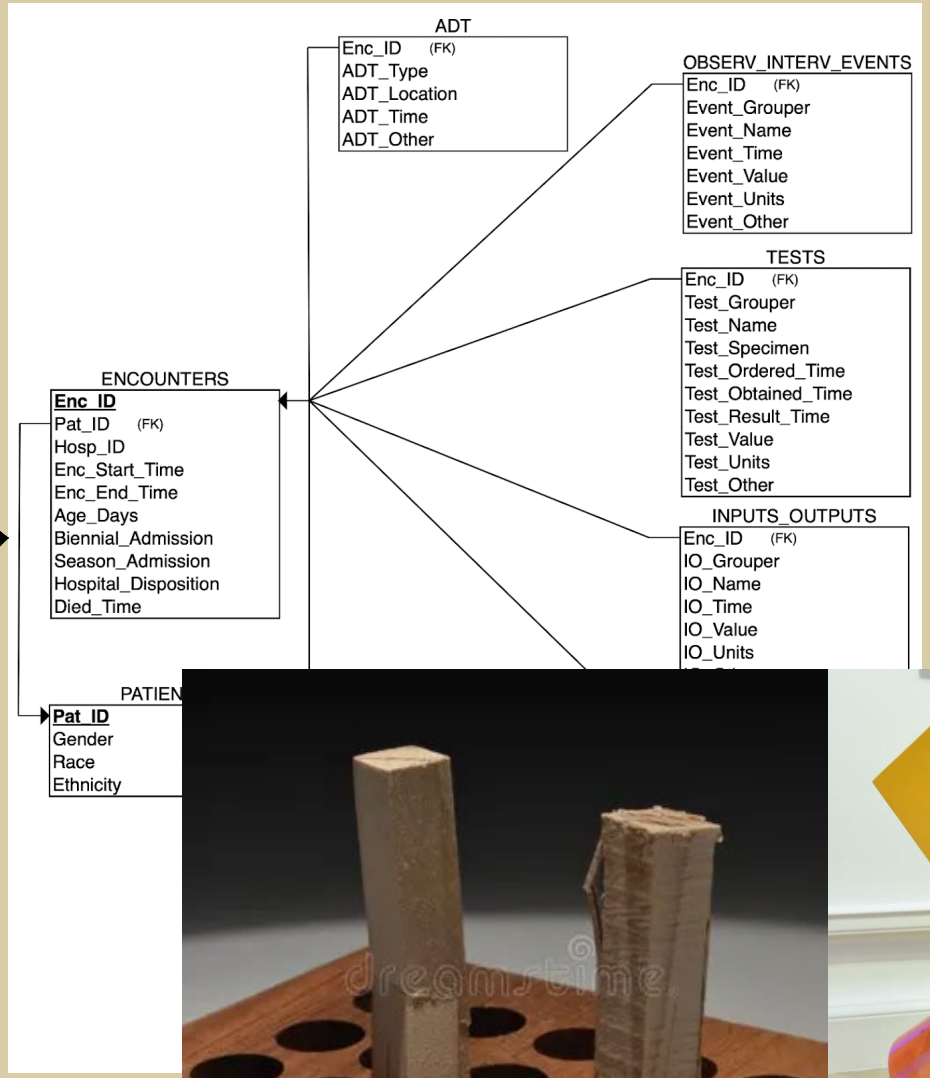
https://github.com/magic-lantern/2025-useR

Data Model


ADT
Enc_ID (FK)
ADT_Type
ADT_Location
ADT_Time
ADT_Other
OBSERV_INTERV_EVENTS
Enc_ID (FK)
Event_Grouper
Event_Name
Event_Time
Event_Value
Event_Units
Event_Other
TESTS
Enc_ID (FK)
Test_Grouper
Test_Name
Test_Specimen
Test_Ordered_Time
Test_Obtained_Time
Test_Result_Time
Test_Value
Test_Units
Test_Other
INPUTS_OUTPUTS
Enc_ID (FK)
IO_Grouper
IO_Name
IO_Time
IO_Value
IO_Units
IO_Other
ENCOUNTERS
Enc_ID
Pat_ID (FK)
Hosp_ID
Enc_Start_Time
Enc_End_Time
Age_Days
Biennial_Admission
Season_Admission
Hospital_Disposition
Died_Time
PATIENTS
Pat_ID
Gender
Race
Ethnicity
DIAGNOSIS_CODES
Enc_ID (FK)
Diagnosis_Code
Diagnosis_Code_Source
Diagnosis_Name
Diagnosis_Time
IsPrimary
IsAdmission
IsDischarge
Diagnosis_Other
MEDICATION_ADMIN
Enc_ID (FK)
Med_Grouper
Med_Name
Med_Route
Med_Dose
Med_Dose_Units
Med_Ordered_Time
Med_Admin_Time
Med_Admin_Action
Med_Other
PROCEDURE_CODES
Enc_ID (FK)
Procedure_Code
Procedure_Code_Source
Procedure_Name
Procedure_Time
Procedure_Other

Provided Data



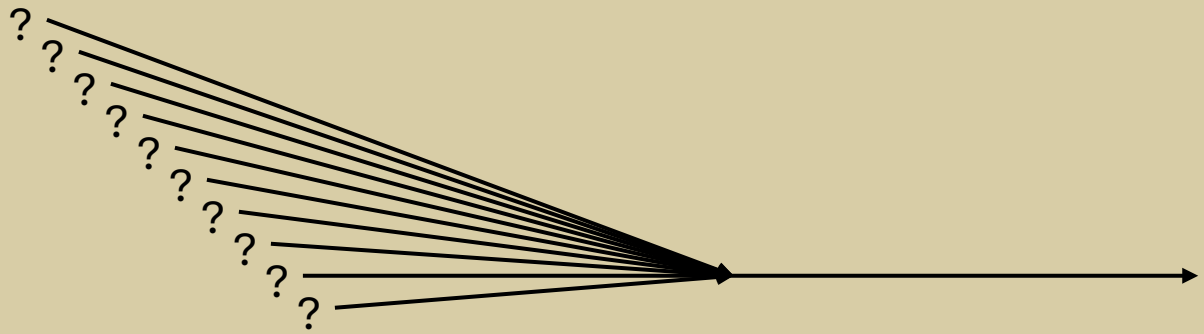


https://www.dreamstime.com/square-peg-round-hole-having-been-forced-image245578184

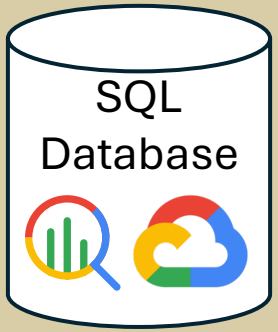
https://www.royalscottishacademy.org/artworks/9375-dylan-esposito-round-peg-in-a-square-hole-2022/

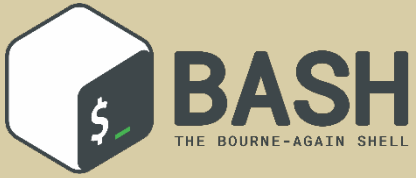
2,276,521,018 rows of data

Data Harmonization



1. Map multiple source row types to one harmonized type
2. Site and source specific unit harmonization
3. Site and source specific value transformations
4. Review
5. Repeat





https://github.com/magic-lantern/2025-useR

	A	B	C	D	E	F	G	H	I
1	site	test_name_mapped	col_1	col_1_value	col_2	col_2_value	match_type	col_3	col_3_value
2	site	ADENOVIRUS	test_name	ADENOVIRUS PCR QUALITATIVE			inclusion		
3	site	ADENOVIRUS	test_name	ADENOVIRUS PCR QUANT			inclusion		
4	site	ADENOVIRUS	test_name	ADENOVIRUS PCR QUANTITATIVE			inclusion		
5	site	ADENOVIRUS	test_name	ADENOVIRUS	test_grouper	RESPIRATORY PATHOGEN PCR PANEL	inclusion		
6	site	ADENOVIRUS	test_name	RAPID ADENOVIRUS CULTURE			inclusion		
7	site	ALBUMIN	test_name	ALBUMIN, SERUM			inclusion		
8	site	ALBUMIN	test_name	ALBUMIN (SERUM) - OUTSIDE LAB			inclusion		
9	site	ALBUMIN	test_name	ALBUMIN (SERUM)			inclusion		
10	site	ALBUMIN	test_name	ALBUMIN			inclusion		
11	site	ALC	test_name	LYMPH COUNT ABSOLUTE			inclusion		
12	site	ALC	test_name	LYMPH COUNT ABSOLUTE	test_grouper	ANTI-CD20 THERAPY & B-CELL PAN	exclusion		
13	site	ALC	test_name	LYMPH COUNT ABSOLUTE	test_grouper	SURFACE MARKER T&B PANEL	exclusion		
14	site	ALC	test_name	LYMPHOCYTES (ABS #) - OUTSIDE LAB			inclusion		
15	site	ALC	test_name	LYMPHOCYTES (ABS #) - OUTSIDE LAB	test_grouper	OUTSIDE LAB-CBC	exclusion	test_units	%
16	site	ALC	test_name	LYMPHOCYTES (ABS #)			inclusion		
17	site	ALK_PHOS	test_name	ALK PHOS, SERUM			inclusion		
18	site	ALK_PHOS	test_name	ALKALINE PHOSPHATASE, S			inclusion		
19	site	ALK_PHOS	test_name	ALKALINE PHOSPHATASE (S) - OUTSIDE LAB			inclusion		
20	site	ALK_PHOS	test_name	ALKALINE PHOSPHATASE (S)			inclusion		
21	site	ALT	test_name	ALT (SGPT)			inclusion		
22	site	ALT	test_name	GPT/ALT (SERUM) - OUTSIDE LAB			inclusion		
23	site	ALT	test_name	GPT/ALT (SERUM)			inclusion		
24	site	ANC	test_name	ABSOLUTE NEUTROPHIL COUNT - OUTSIDE LAB			inclusion		
25	site	ANC	test_name	ABSOLUTE NEUTROPHIL COUNT - OUTSIDE LAB	test_grouper	OUTSIDE LAB-CBC	exclusion	test_units	%
26	site	ANC	test_name	ABSOLUTE NEUTROPHIL COUNT			inclusion		
27	site	ANC	test_name	ANC# FINAL			inclusion		
28	site	AST	test_name	AST (SGOT)			inclusion		
29	site	AST	test_name	GOT/AST (SERUM) - OUTSIDE LAB			inclusion		

https://github.com/magic-lantern/2025-useR

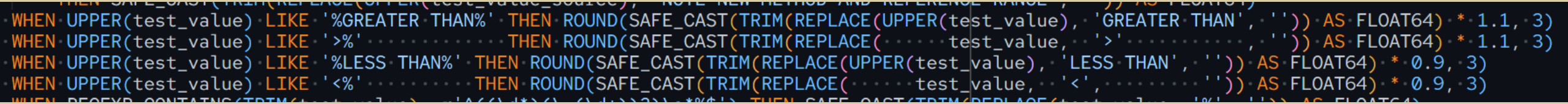
	A	B	C	D	E	F	G	H	I
1	site	test_name_mapped	col_1	col_1_value	col_2	col_2_value	match_type	col_3	col_3_value
2	site	ADENOVIRUS	test_name	ADENOVIRUS PCR QUALITATIVE			inclusion		
3	site	ADENOVIRUS	test_name	ADENOVIRUS PCR QUANT			inclusion		
4	site	ADENOVIRUS	test_name	ADENOVIRUS PCR QUANTITATIVE			inclusion		
5	site	ADENOVIRUS	test_name	ADENOVIRUS	test_grouper	RESPIRATORY PATHOGEN PCR PANEL	inclusion		
6	site	ADENOVIRUS	test_name	RAPID ADENOVIRUS CULTURE			inclusion		
7	site	ALBUMIN	test_name	ALBUMIN, SERUM			inclusion		
8	site	ALBUMIN	test_name	ALBUMIN (SERUM) - OUTSIDE LAB			inclusion		
9	site	ALBUMIN	test_name	ALBUMIN (SERUM)			inclusion		
10	site	ALBUMIN	test_name	ALBUMIN			inclusion		
11	site	ALC	test_name	LYMPH COUNT ABSOLUTE			inclusion		
12	site	ALC	test_name	LYMPH COUNT ABSOLUTE	test_grouper	ANTI-CD20 THERAPY & B-CELL PAN	exclusion		
13	site	ALC	test_name	LYMPH COUNT ABSOLUTE	test_grouper	SURFACE MARKER T&B PANEL	exclusion		
14	site	ALC	test_name	LYMPHOCYTES (ABS #) - OUTSIDE LAB			inclusion		
15	site	ALC	test_name	LYMPHOCYTES (ABS #) - OUTSIDE LAB	test_grouper	OUTSIDE LAB-CBC	exclusion	test_units	%
16	site	ALC	test_name	LYMPHOCYTES (ABS #)			inclusion		
17	site	ALK_PHOS	test_name	ALK PHOS, SERUM			inclusion		
18	site	ALK_PHOS	test_name	ALKALINE PHOSPHATASE, S			inclusion		
19	site	ALK_PHOS	test_name	ALKALINE PHOSPHATASE (S) - OUTSIDE LAB			inclusion		
20	site	ALK_PHOS	test_name	ALKALINE PHOSPHATASE (S)			inclusion		
21	site	ALT	test_name	ALT (SGPT)			inclusion		
22	site	ALT	test_name	GPT/ALT (SERUM) - OUTSIDE LAB			inclusion		
23	site	ALT	test_name	GPT/ALT (SERUM)			inclusion		
24	site	ANC	test_name	ABSOLUTE NEUTROPHIL COUNT - OUTSIDE LAB			inclusion		
25	site	ANC	test_name	ABSOLUTE NEUTROPHIL COUNT - OUTSIDE LAB	test_grouper	OUTSIDE LAB-CBC	exclusion	test_units	%
26	site	ANC	test_name	ABSOLUTE NEUTROPHIL COUNT			inclusion		
27	site	ANC	test_name	ANC# FINAL			inclusion		
28	site	AST	test_name	AST (SGOT)			inclusion		
29	site	AST	test_name	GOT/AST (SERUM) - OUTSIDE LAB			inclusion		

https://github.com/magic-lantern/2025-useR

B	C	D	E	F	G	H	I	J
test_grouper	test_name	test_units	test_specimen	test_other	test_count	min_string	max_strin	min
GLUCOSE, BLOOD, POINT OF CARE	NOVANAME		BLOOD		71556	51	see below	51
GLUCOSE, BLOOD, POINT OF CARE	GLUPOC	mg/dL	BLOOD		71554	10	see below	10
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	NRBC	K/cu mm	BLOOD		57482	0	see below	0
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	RBC	M/cu mm	BLOOD		57419	0.44	see below	0.44
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	MCV	fL	BLOOD		57292	100	see below	46.8
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	MCH	pg	BLOOD		57251	10.5	see below	10.5
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	MCHC	g/dL	BLOOD		57228	21.9	see below	21.9
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	RDW	%	BLOOD		57215	10.4	see below	10.4
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	HCT	%	BLOOD		56742	10	see below	6.5
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	HGB	g/dL	BLOOD		56244	1.6	see below	1.6
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	MPV	fL	BLOOD		52330	10	see below	7
COMPREHENSIVE METABOLIC PANEL	GFRNA	mL/min/1.73 sqm	BLOOD		50231	108	Unable to cal	20
COMPREHENSIVE METABOLIC PANEL	ASTALT		BLOOD		50224	0.1	Unable to cal	0.1
COMPREHENSIVE METABOLIC PANEL	GFRAA	mL/min/1.73 sqm	BLOOD		50224	108	Unable to cal	23
COMPREHENSIVE METABOLIC PANEL	GLU	mg/dL	BLOOD		50224	100	see below	3
COMPREHENSIVE METABOLIC PANEL	BCR		BLOOD		50222	10	Unable to cal	2
COMPREHENSIVE METABOLIC PANEL	CREATININE	mg/dL	BLOOD		49446	0.1	see below	0.1
COMPREHENSIVE METABOLIC PANEL	ALBUMIN	g/dL	BLOOD		48794	0.9	see below	0.9
COMPREHENSIVE METABOLIC PANEL	BUN	mg/dL	BLOOD		48792	1	see below	1
COMPREHENSIVE METABOLIC PANEL	PROT	g/dL	BLOOD		48557	1.5	see below	1.5
COMPREHENSIVE METABOLIC PANEL	CALCIUM	mg/dL	BLOOD		47718	10	see below	4.7
COMPREHENSIVE METABOLIC PANEL	ANIONGAP	mmol/L	BLOOD		47295	0	Unable to cal	0

https://github.com/magic-lantern/2025-useR

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
test_grouper	test_name	test_units	test_othe	test_count	min_strin	max_strin	min	max	mean	sd	median	quart25	quart75	mode	test_name_mapped
GLUCOSE, BLOOD, POINT OF CARE	NOVANAME			71556	51	see below	51	51	51						
GLUCOSE, BLOOD, POINT OF CARE	GLUPOC	mg/dL		71554	10	see below	10	600	137.37	81.51					GLUCOSE
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	NRBC	K/cu mm		57482	0	see below	0	388	0.09	2.55					
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	RBC	M/cu mm		57419	0.44	see below	0.44	10.69	3.74	0.87					
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	MCV	fL		57292	100	see below	46.8	126.5	84.41	7.27					MCV
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	MCH	pg		57251	10.5	see below	10.5	46.6	28.4	2.81					MCH
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	MCHC	g/dL		57228	21.9	see below	21.9	38	33.63	1.62					
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	RDW	%		57215	10.4	see below	10.4	54.4	15.08	2.92					RDW
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	HCT	%		56742	10	see below	6.5	62.5	31.49	6.81					
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	HGB	g/dL		56244	1.6	see below	1.6	22	10.59	2.32					HGB
COMPLETE BLOOD COUNT (CBC) + AUTO DIFF	MPV	fL		52330	10	see below	7	15.1	10.27	1.11					MPV
COMPREHENSIVE METABOLIC PANEL	GFRNA	mL/min/1.73 sqm		50231	108	Unable to cal	20	173	75.62	46.74					
COMPREHENSIVE METABOLIC PANEL	ASTALT			50224	0.1	Unable to cal	0.1	30.4	1.7	1.27					
COMPREHENSIVE METABOLIC PANEL	GFRAA	mL/min/1.73 sqm		50224	108	Unable to cal	23	200	88.75	54.96					
COMPREHENSIVE METABOLIC PANEL	GLU	mg/dL		50224	100	see below	3	1845	114.67	54.96					GLUCOSE
COMPREHENSIVE METABOLIC PANEL	BCR			50222	10	Unable to cal	2	490	33.5	28.03					
COMPREHENSIVE METABOLIC PANEL	CREATININE	mg/dL		49446	0.1	see below	0.1	12.5	0.48	0.48					CREATININE
COMPREHENSIVE METABOLIC PANEL	ALBUMIN	g/dL		48794	0.9	see below	0.9	5.9	3.74	0.73					ALBUMIN
COMPREHENSIVE METABOLIC PANEL	BUN	mg/dL		48792	1	see below	1	128	12.48	10.47					BUN
COMPREHENSIVE METABOLIC PANEL	PROT	g/dL		48557	1.5	see below	1.5	11.6	6.01	1.1					



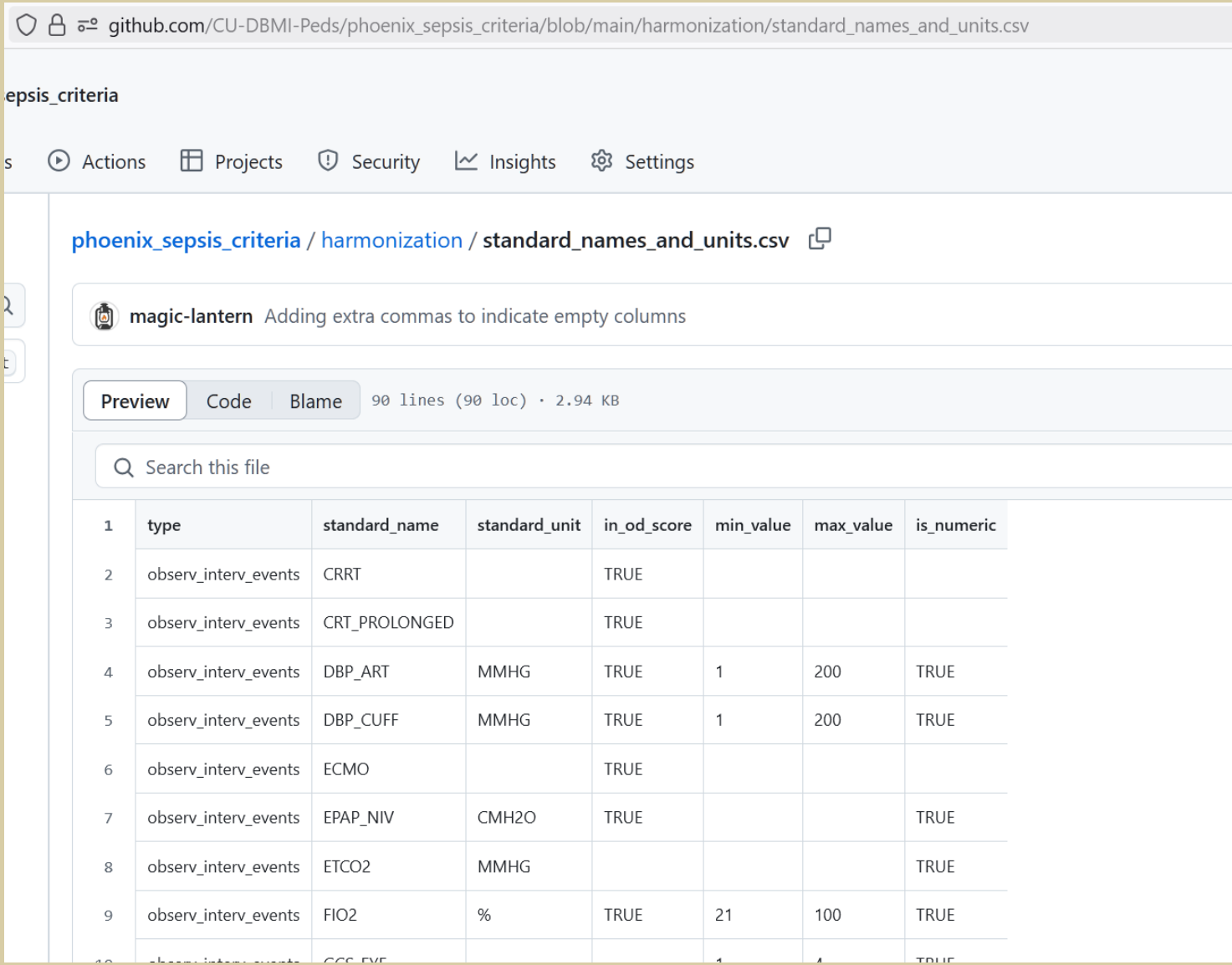
https://github.com/magic-lantern/2025-useR

Unit harmonization

	A	B	C	D	E	F
1	**site**	**test_name**	**unit_match**	**unit_dest**	**conversion_operation**	**conversion_factor**
2	site	ALBUMIN	MG/DL	G/DL	/	1000
3	site	ALC	UL	10E3/UL	/	1000
4	site	AST	MG/DL	U/L	*	10000
5	site	CREATININE	NMOL/ML	MG/DL	/	10000
6	site	CRP	MG/L	MG/DL	/	10
7	site	CRP	MG/LITER	MG/DL	/	10
8	site	CRP	MG/LT	MG/DL	/	10
9	site	CRP	ML/L	MG/DL	/	10
10	site	D_DIMER	NG/ML (FEU)	MG/L FEU	/	1000

https://github.com/magic-lantern/2025-useR

Value Thresholding



1	type	standard_name	standard_unit	in_od_score	min_value	max_value	is_numeric
2	observ_interv_events	CRRT		TRUE			
3	observ_interv_events	CRT_PROLONGED		TRUE			
4	observ_interv_events	DBP_ART	MMHG	TRUE	1	200	TRUE
5	observ_interv_events	DBP_CUFF	MMHG	TRUE	1	200	TRUE
6	observ_interv_events	ECMO		TRUE			
7	observ_interv_events	EPAP_NIV	CMH2O	TRUE			TRUE
8	observ_interv_events	ETCO2	MMHG				TRUE
9	observ_interv_events	FIO2	%	TRUE	21	100	TRUE

https://github.com/CU-DBMI-Peds/phoenix_sepsis_criteria/blob/main/harmonization/standard_names_and_units.csv

Review

A	B	D	E	F	G	H
test_name	in_od_score	num_encounters	num_patients	num_obs	obs_per_encounter	obs_per_patient
LACTATE	TRUE	15095	11522	50645	0.298	0.228
LACTATE	TRUE	18192	17093	157112	0.116	0.109
LACTATE	TRUE	2	2	2	1	1
LACTATE	TRUE	119	114	211	0.564	0.54
LACTATE	TRUE	294	273	362	0.812	0.754
LACTATE	TRUE	6973	5552	47940	0.145	0.116
LACTATE	TRUE	12462	11934	194170	0.064	0.061
LACTATE	TRUE	8215	8214	107131	0.077	0.077
LACTATE	TRUE	11505	9045	94784	0.121	0.095

https://github.com/CU-DBMI-Peds/phoenix_sepsis_criteria/tree/main/harmonization

https://github.com/magic-lantern/2025-useR

Process on RDBMs

```sql
/*
Views and tables to curate tests
*/
-- view to join together default/standard labs with mappings
CREATE OR REPLACE VIEW `**REDACTED**.full.test_name_mapping` AS
-- allow for both default names for tests and custom mappings
SELECT DISTINCT * FROM (
    -- default names for tests
    SELECT * FROM (
      SELECT DISTINCT
        site,
        standard_name AS test_name_mapped,
        'test_name' AS col_1,
        standard_name AS col_1_value,
        SAFE_CAST(NULL AS STRING) AS col_2,
        SAFE_CAST(NULL AS STRING) AS col_2_value,
        SAFE_CAST(NULL AS STRING) AS col_3,
        SAFE_CAST(NULL AS STRING) AS col_3_value,
        'inclusion' AS match_type,
        standard_unit
      -- cross join to map all sites to standard tests
      FROM `**REDACTED**.full.test_name_mapping_configuration` c,
        `**REDACTED**.full.standard_names_and_units` s
      WHERE s.type = 'tests'
    ) std
    WHERE NOT EXISTS (
      SELECT col_1_value
      FROM  `**REDACTED**.full.test_name_mapping_configuration` i
      WHERE i.col_1 = 'test_name'
      AND i.col_1_value = std.col_1_value
      AND i.site = std.site
    )
    UNION ALL
    -- custom mappings
    SELECT DISTINCT
        c.site,
        c.test_name_mapped,
        c.col_1,
        TRIM(UPPER(c.col_1_value)) AS col_1_value,
        c.col_2,
        TRIM(UPPER(c.col_2_value)) AS col_2_value,
        c.col_3,
        TRIM(UPPER(c.col_3_value)) AS col_3_value,
        c.match_type,
        s.standard_unit
    FROM `**REDACTED**.full.test_name_mapping_configuration` c
    LEFT JOIN `**REDACTED**.full.standard_names_and_units` s
        ON c.test_name_mapped = s.standard_name
        AND s.type = 'tests'
)
ORDER BY site, test_name_mapped
;
```

https://github.com/CU-DBMI-Peds/phoenix_sepsis_criteria/blob/main/harmonization/harmonize_tests.sql

Process on RDBMs

```sql
CREATE OR REPLACE TABLE `**REDACTED**.full.tests_phase1` AS

SELECT DISTINCT
    base.* EXCEPT(test_units, test_value),
    test_units AS test_units_source,
    COALESCE(unit_dest, test_units, standard_unit) AS test_units,
    CASE
        WHEN conversion_operation = '/' THEN ROUND(SAFE_CAST(test_value AS FLOAT64) / SAFE_CAST(conversion_factor AS FLOAT64), 3)
        WHEN conversion_operation = '*' THEN ROUND(SAFE_CAST(test_value AS FLOAT64) * SAFE_CAST(conversion_factor AS FLOAT64), 3)
        ELSE SAFE_CAST(test_value AS FLOAT64)
    END AS test_value
FROM (
    SELECT DISTINCT
        t.* EXCEPT(test_value, test_value_source, test_units, test_name, test_name_source),
        --  if a test_value_source already exists, preserve it
        COALESCE(test_value_source, test_value) AS test_value_source,
        -- fixing some values that can be converted to numeric with basic rules
        CASE
            WHEN UPPER(test_value_source) LIKE '%NOTE NEW METHOD AND REFERENCE RANGE%'
                AND test_value IS NULL
                THEN SAFE_CAST(TRIM(REPLACE(UPPER(test_value_source), 'NOTE NEW METHOD AND REFERENCE RANGE', '')) AS FLOAT64)
            WHEN UPPER(test_value) LIKE '%GREATER THAN%' THEN ROUND(SAFE_CAST(TRIM(REPLACE(UPPER(test_value), 'GREATER THAN', '')) AS FLOAT64) * 1.1, 3)
            WHEN UPPER(test_value) LIKE '>%'             THEN ROUND(SAFE_CAST(TRIM(REPLACE(      test_value,  '>'           , '')) AS FLOAT64) * 1.1, 3)
            WHEN UPPER(test_value) LIKE '%LESS THAN%' THEN ROUND(SAFE_CAST(TRIM(REPLACE(UPPER(test_value), 'LESS THAN', '')) AS FLOAT64) * 0.9, 3)
            WHEN UPPER(test_value) LIKE '<%'          THEN ROUND(SAFE_CAST(TRIM(REPLACE(      test_value,  '<',          '')) AS FLOAT64) * 0.9, 3)
            WHEN REGEXP_CONTAINS(TRIM(test_value), r'^((\d*)(\.(\d+))?)\s*%$') THEN SAFE_CAST(TRIM(REPLACE(test_value, '%', '')) AS FLOAT64)
        ELSE SAFE_CAST(TRIM(test_value) AS FLOAT64)
        END AS test_value,
        CASE
            WHEN REGEXP_CONTAINS(TRIM(test_value), r'^((\d*)(\.(\d+))?)\s*%$') THEN '%'
```

https://github.com/CU-DBMI-Peds/phoenix_sepsis_criteria/blob/main/harmonization/harmonize_tests.sql

Process on RDBMs

```sql
/*
The above includes all the tests we want, but also includes a few we don't want. Drop those
using 'exclusion' rows from the configuration table
*/
CREATE OR REPLACE TABLE `**REDACTED**.full.tests_phase2` AS

SELECT
  *
FROM `**REDACTED**.full.tests_phase1`

EXCEPT DISTINCT

SELECT
  t.*
FROM `**REDACTED**.full.tests_phase1` t
INNER JOIN `**REDACTED**.full.test_name_mapping` m
  ON
    m.match_type = 'exclusion'
    AND m.col_1 = 'test_name'
    AND TRIM(UPPER(t.test_name_source)) = m.col_1_value
    AND TRIM(UPPER(t.site)) = TRIM(UPPER(m.site))
    AND (
      (
        m.col_2 IS NULL
        AND m.col_3 IS NULL)
      OR (
        m.col_2 = 'test_grouper'
        AND TRIM(UPPER(t.test_grouper)) = m.col_2_value
        AND m.col_3 IS NULL)
      OR (
        m.col_2 = 'test_grouper'
        AND m.col_3 = 'test_units'
        AND TRIM(UPPER(t.test_grouper)) = m.col_2_value
        AND TRIM(UPPER(t.test_units_source)) = m.col_3_value)
      OR (
        m.col_2 = 'test_specimen'
        AND TRIM(UPPER(t.test_specimen)) = m.col_2_value)
      OR (
        m.col_2 = 'test_units'
        AND TRIM(UPPER(t.test_units)) = m.col_2_value)
    )
;
```

https://github.com/CU-DBMI-Peds/phoenix_sepsis_criteria/blob/main/harmonization/harmonize_tests.sql

R01 Sepsis

Tests

local file modified (UTC):
2025-06-05 15:15:23
nrow: 59584400

Selected Mapped Test

LACTATE

Subset

All options are selected by default

Selected the sites to summarize

chco chop dhaka hgm jhh kenya lurie picdb pittsburgh seattle

Selected the years to summarize

2010-2011 2010-2018 2012-2013 2014-2015 2016-2017 2018-2019

Summary Table Columns

Year column is not shown by default

Selected columns to show in the summary table

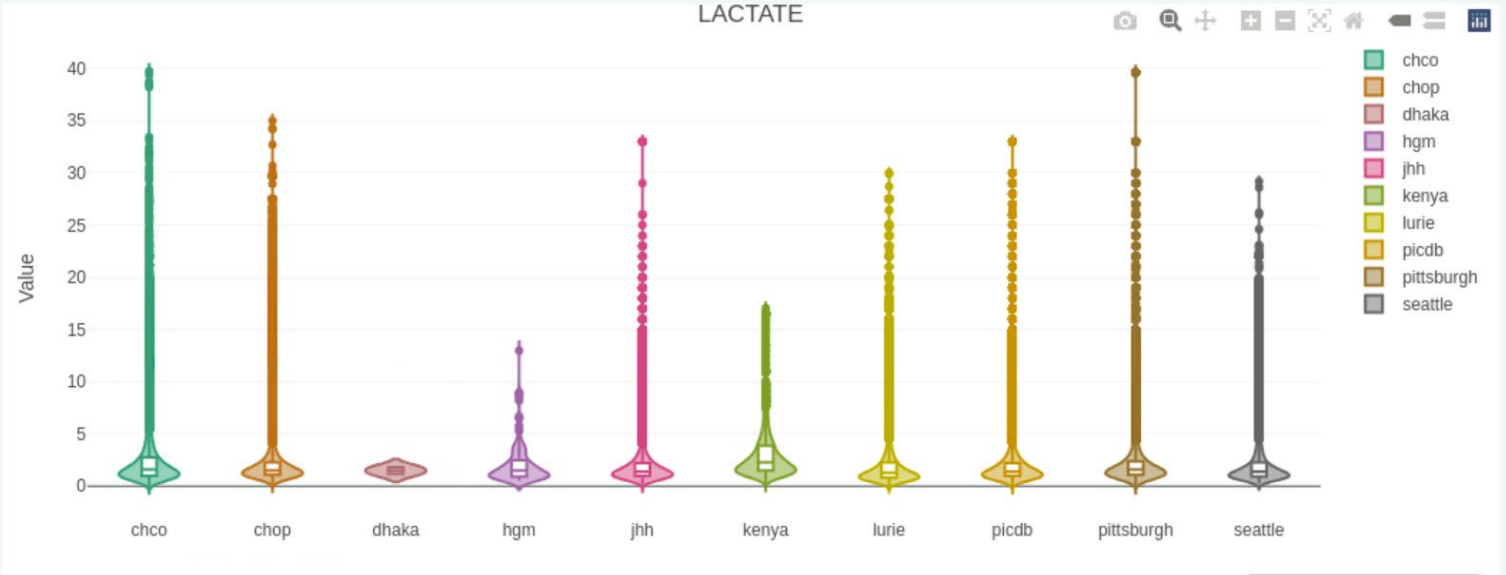
site test_name_source test_units test_units_source

Split

site

Color

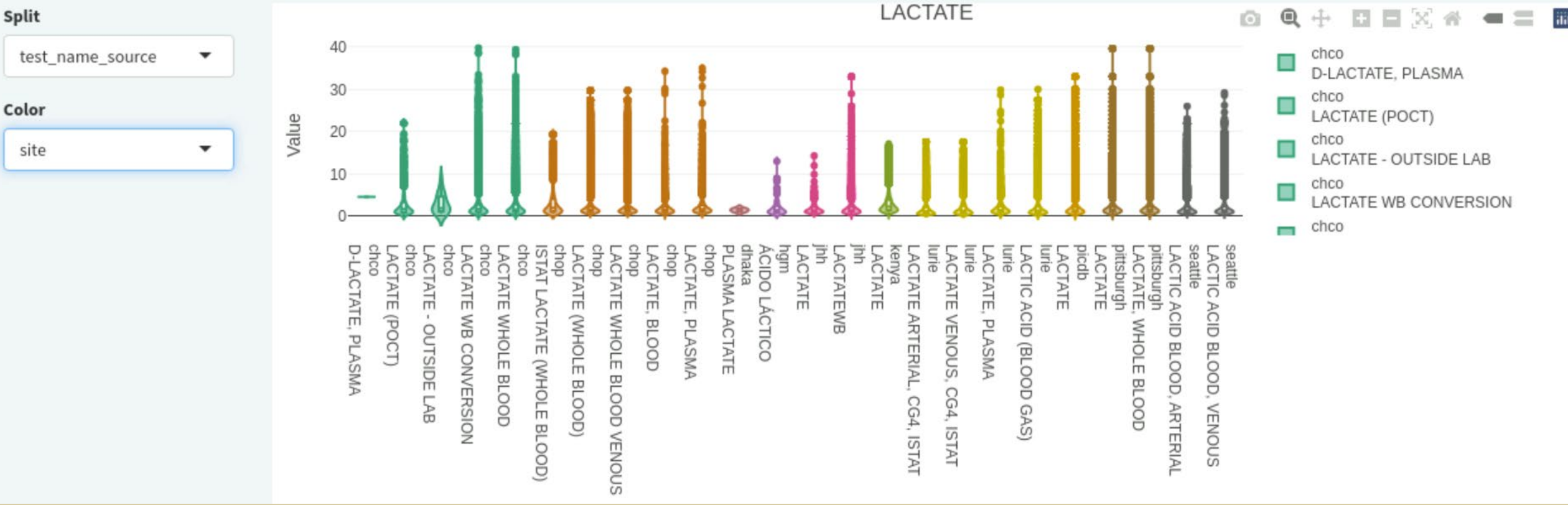
site



Show 25 entries

Search:

	site	test_name_source	test_units	test_units_source	rows	missing value	min_value	q25_value	q50_value	q75_value	max_value
1	picdb	LACTATE	MMOL/L	MMOL/L	194065	0	0.000	1.000	1.400	2.200	33.000
2	pittsburgh	LACTATE	MMOL/L	MMOL/L	52731	0	0.100	1.100	1.600	2.400	39.600
3	seattle	LACTIC ACID BLOOD, ARTERIAL	MMOL/L	MMOL/L	20984	0	0.270	0.900	1.400	2.400	23.000
4	seattle	LACTIC ACID BLOOD, ARTERIAL	MMOL/L	MMOL/L.	55350	0	0.270	0.900	1.400	2.200	26.000
5	chop	LACTATE (WHOLE BLOOD)	MMOL/L	MMOL/L	103966	0	0.450	1.100	1.600	2.400	29.700
6	lurie	LACTIC ACID (BLOOD GAS)	MMOL/L	MEQ/L	21172	0	0.000	0.900	1.300	2.400	30.000
7	chop	LACTATE WHOLE BLOOD VENOUS	MMOL/L	MMOL/L	46602	0	0.450	1.100	1.500	2.100	29.700



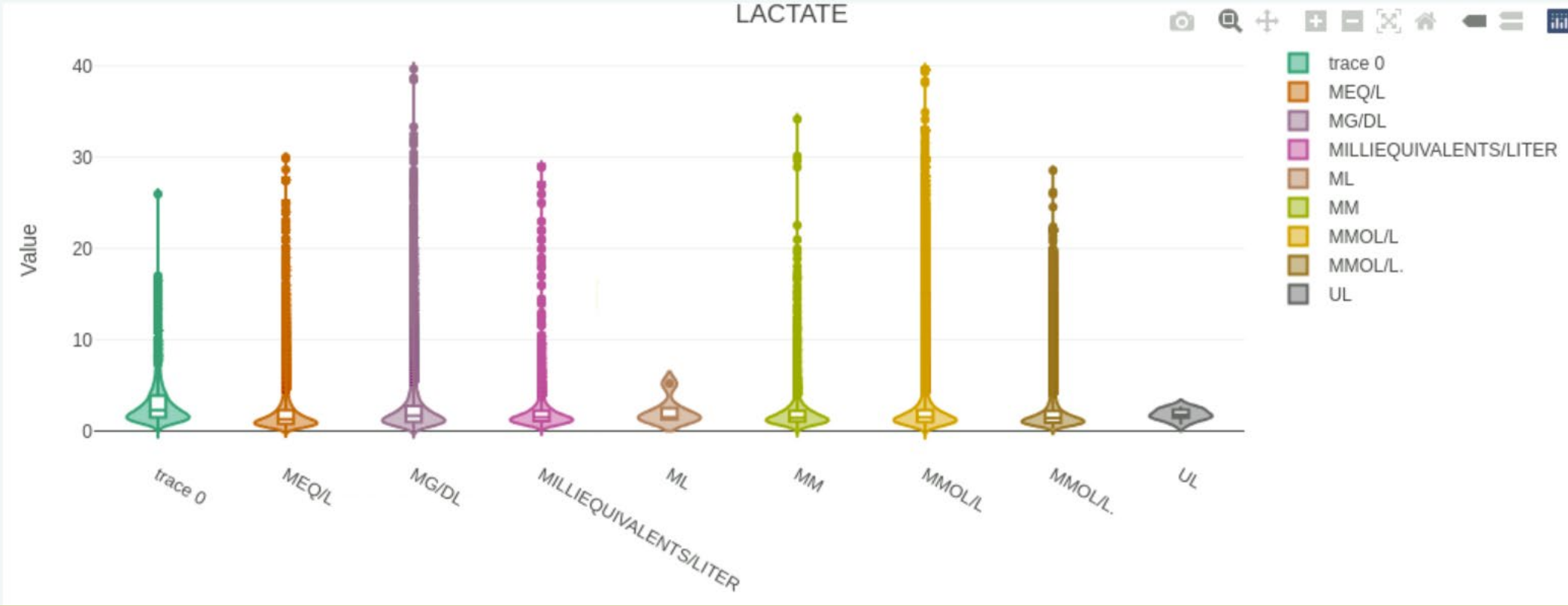
https://github.com/magic-lantern/2025-useR

Split

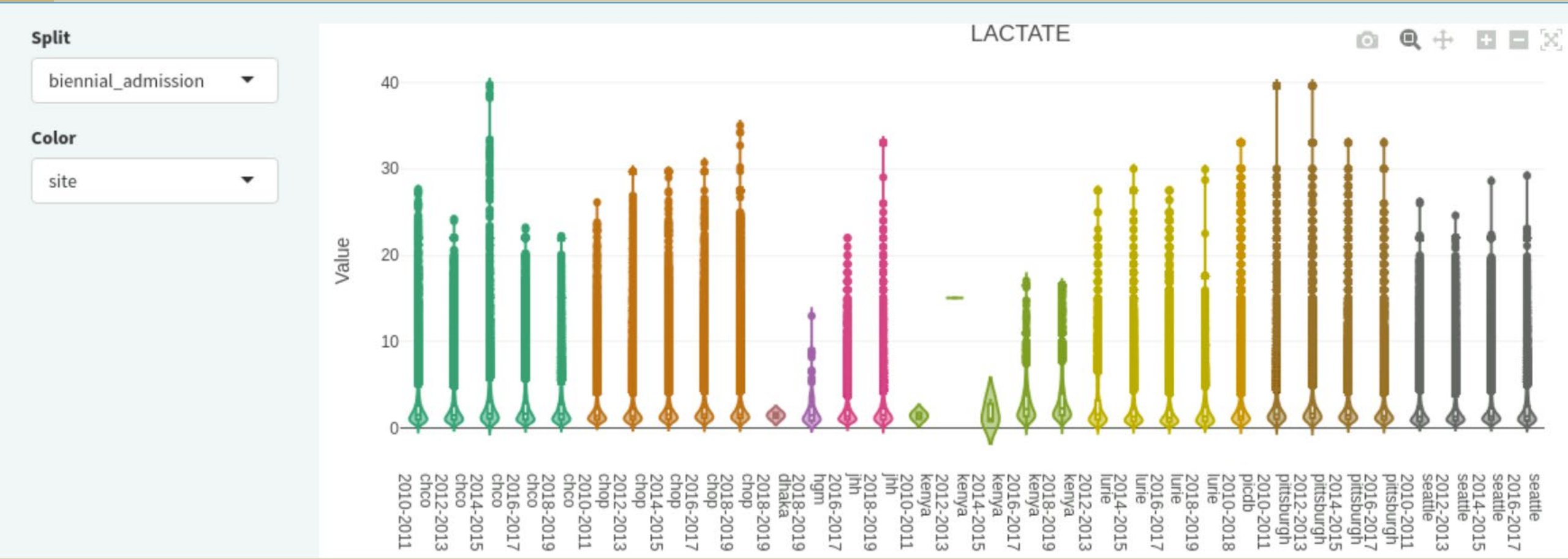
test_units_source

Color

test_units_source

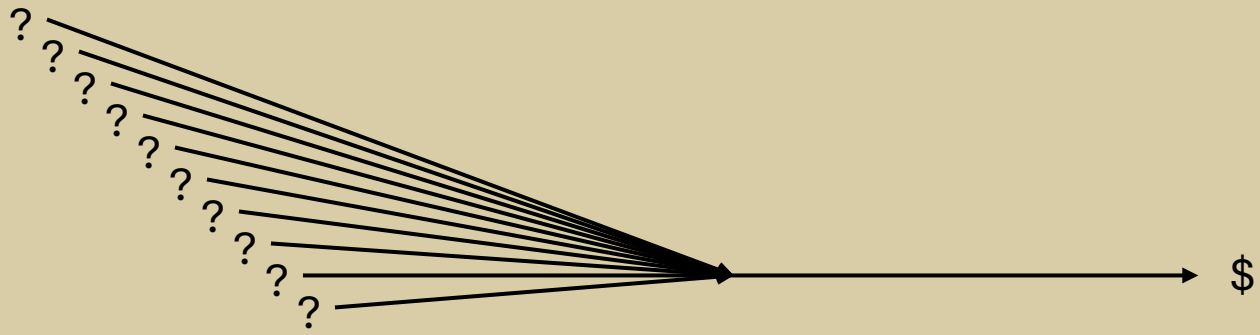


https://github.com/magic-lantern/2025-useR



https://github.com/magic-lantern/2025-useR

Harmonization: the Unsung Hero



Slides available at:

https://github.com/magic-lantern/2025-useR

Please reach out to me with questions!