Knowledge Graph-driven Tabular Data Discovery from Scientific Documents

Vijay S. Kumar[1], Varish Mulwad[2], Jenny Weisenberg Williams[1], Tim Finin[3], Sharad Dixit[1], Anupam Joshi[3]

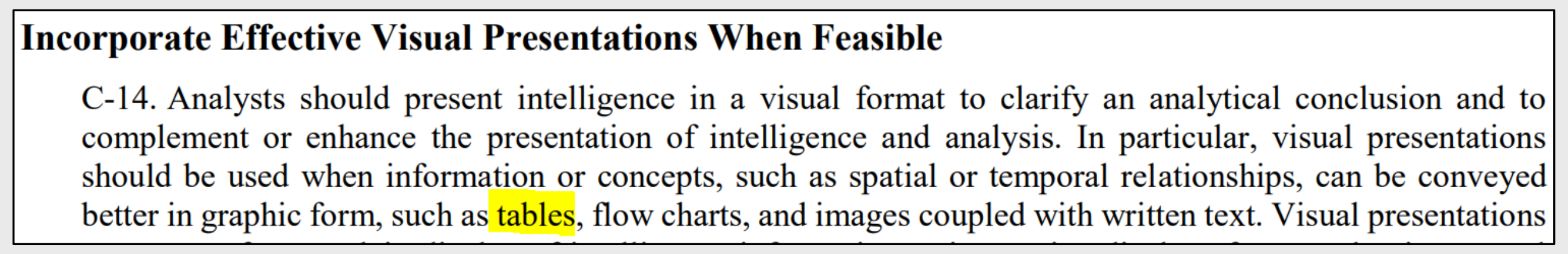
1: GE Research, Niskayuna, NY, USA, 2: GE Research, Bengaluru, KA, India, 3: University of Maryland, Baltimore County, Baltimore, MD, USA

Augmenting generative AI-driven search with valuable nuggets from tabular data sources is an emerging need

Intelligence Report Generation & Enhancement

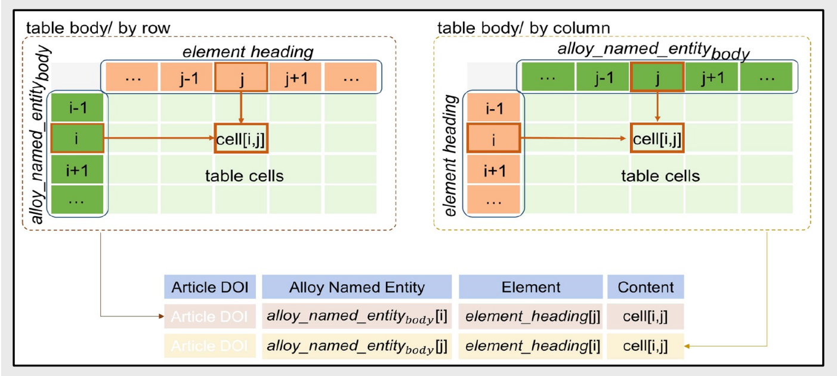
✓ Gather intelligence from more information sources

✓ Strengthen analysis reports with tabular data



Incorporate Effective Visual Presentations When Feasible

C-14. Analysts should present intelligence in a visual format to clarify an analytical conclusion and to complement or enhance the presentation of intelligence and analysis. In particular, visual presentations should be used when information or concepts, such as spatial or temporal relationships, can be conveyed better in graphic form, such as tables, flow charts, and images coupled with written text. Visual presentations

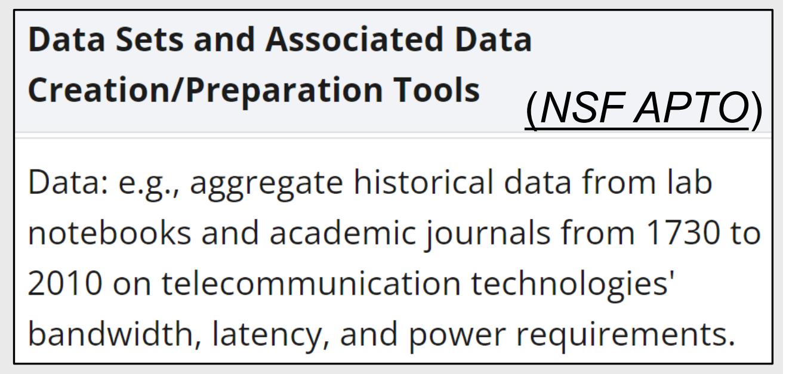


AI-assisted Scientific Research

✓ Augment understanding & discovery from literature

✓ Assemble datasets in low-data domains

(alloy materials discovery, technology forecasting)



Data Sets and Associated Data Creation/Preparation Tools (NSF APTO)

Data: e.g., aggregate historical data from lab notebooks and academic journals from 1730 to 2010 on telecommunication technologies' bandwidth, latency, and power requirements.

Despite prevalence of tables in technical documents, limited focus on tabular data discovery in scientific domains

Dataset	Document Type / Source	Domain	Corpus size	# tables
ChemTables	Patents / USPTO	Chemical	1,000	788
ArxivPapers	Preprints / arXiv	ML	104,723	277,996
ProCure (this work)	Papers & preprints / PubMed Central OA	Biomedical, clinical	62,777	120,417

Dataset	Downstream Task
PubTables-1M	Table detection, Table structure recognition
ChemTables	Table classification
ArxivPapers	Table extraction and segmentation
SciGen	Reasoning-aware table-to-text generation
TAT-QA	Question-answering over tables and text
S2abEL	Entity Linking for scientific tables

Scientific Tables are Hard!

Domain-specific Entities

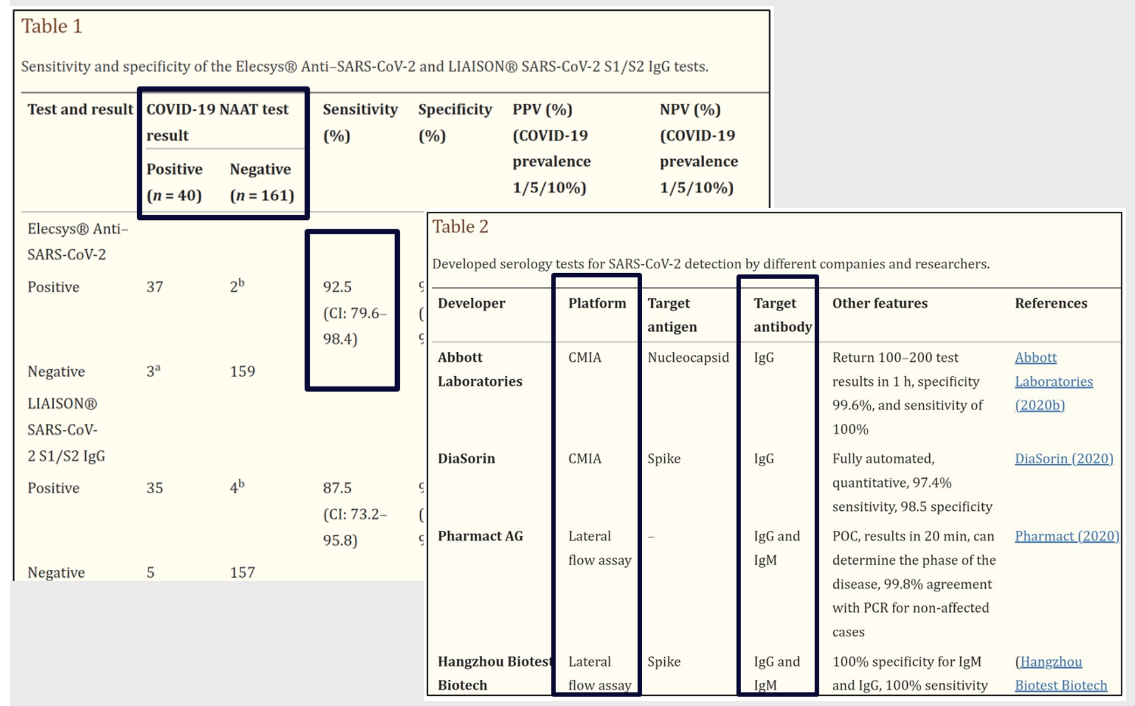


Table 1

Sensitivity and specificity of the Elecsys® Anti–SARS-CoV-2 and LIAISON® SARS-CoV-2 S1/S2 IgG tests.

Test and result	COVID-19 NAAT test result		Sensitivity (%)	Specificity (%)	PPV (%) (COVID-19 prevalence 1/5/10%)	NPV (%) (COVID-19 prevalence 1/5/10%)
	Positive (n = 40)	Negative (n = 161)				
Elecsys® Anti–SARS-CoV-2						
Positive	37	2[b]	92.5 (CI: 79.6–98.4)			
Negative	3[a]	159				
LIAISON® SARS-CoV-2 S1/S2 IgG						
Positive	35	4[b]	87.5 (CI: 73.2–95.8)			
Negative	5	157				

Table 2

Developed serology tests for SARS-CoV-2 detection by different companies and researchers.

Developer	Platform	Target antigen	Target antibody	Other features	References
Abbott Laboratories	CMIA	Nucleocapsid	IgG	Return 100–200 test results in 1 h, specificity 99.6%, and sensitivity of 100%	Abbott Laboratories (2020b)
DiaSorin	CMIA	Spike	IgG	Fully automated, quantitative, 97.4% sensitivity, 98.5 specificity	DiaSorin (2020)
Pharmact AG	Lateral flow assay	–	IgG and IgM	POC, results in 20 min, can determine the phase of the disease, 99.8% agreement with PCR for non-affected cases	Pharmact (2020)
Hangzhou Biotest Biotech	Lateral flow assay	Spike	IgG and IgM	100% specificity for IgM and IgG, 100% sensitivity	(Hangzhou Biotest Biotech

Similar to open data … less text, more numbers … with ranges, multi-value cells

High Structural Heterogeneity

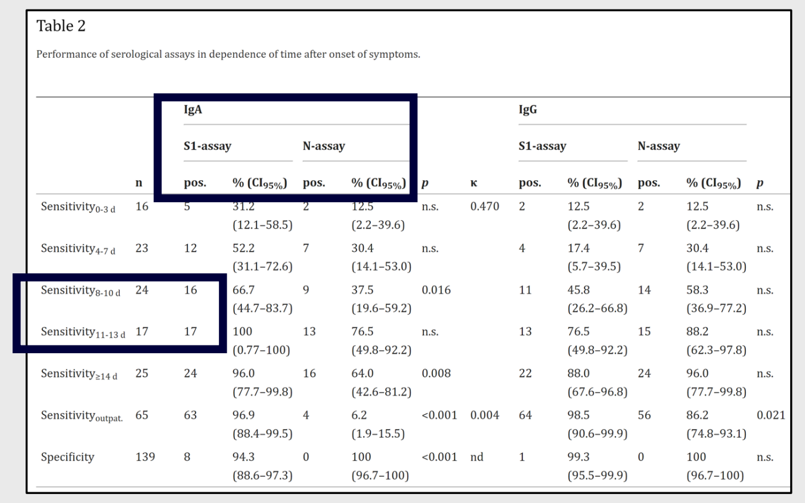


Table 2

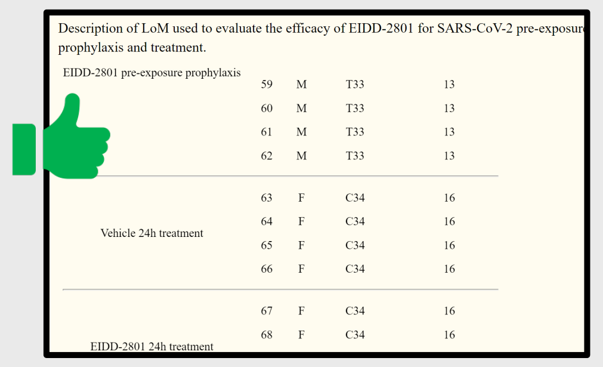
Performance of serological assays in dependence of time after onset of symptoms.

		IgA						IgG				
		S1-assay		N-assay				S1-assay		N-assay		
	n	pos.	% (CI$_{95\%}$)	pos.	% (CI$_{95\%}$)	p	κ	pos.	% (CI$_{95\%}$)	pos.	% (CI$_{95\%}$)	p
Sensitivity$_{0-3\ d}$	16	5	31.2 (12.1–58.5)	2	12.5 (2.2–39.6)	n.s.	0.470	2	12.5 (2.2–39.6)	2	12.5 (2.2–39.6)	n.s.
Sensitivity$_{4-7\ d}$	23	12	52.2 (31.1–72.6)	7	30.4 (14.1–53.0)	n.s.		4	17.4 (5.7–39.5)	7	30.4 (14.1–53.0)	n.s.
Sensitivity$_{8-10\ d}$	24	16	66.7 (44.7–83.7)	9	37.5 (19.6–59.2)	0.016		11	45.8 (26.2–66.8)	14	58.3 (36.9–77.2)	n.s.
Sensitivity$_{11-13\ d}$	17	17	100 (0.77–100)	13	76.5 (49.8–92.2)	n.s.		13	76.5 (49.8–92.2)	15	88.2 (62.3–97.8)	n.s.
Sensitivity$_{\geq 14\ d}$	25	24	96.0 (77.7–99.8)	16	64.0 (42.6–81.2)	0.008		22	88.0 (67.6–96.8)	24	96.0 (77.7–99.8)	n.s.
Sensitivity$_{outpat.}$	65	63	96.9 (88.4–99.5)	4	6.2 (1.9–15.5)	<0.001	0.004	64	98.5 (90.6–99.9)	56	86.2 (74.8–93.1)	0.021
Specificity	139	8	94.3 (88.6–97.3)	0	100 (96.7–100)	<0.001	nd	1	99.3 (95.5–99.9)	0	100 (96.7–100)	n.s.

Row and column headers … sub-columns … abridged header cells

Characterization	System Count	Precision	Recall
Tables with Header Rows	113,582	1.00	0.94
Tables with Header Columns	48,733	1.00	0.55
Tables with Concise Header Rows	36,182	0.84	0.94
Tables with Multi-level Header Rows	32,169	1.00	0.97
Tables with ONLY Numeric Data Cells	12,969	1.00	0.83
Tables with Concise Body	40,158	0.97	0.67
Horizontal Tables	21,863	0.95	0.50
Vertical Tables	7205	0.91	0.62

Lack of Information Reliability



Description of LoM used to evaluate the efficacy of EIDD-2801 for SARS-CoV-2 pre-exposure prophylaxis and treatment.

EIDD-2801 pre-exposure prophylaxis	59	M	T33	13
	60	M	T33	13
	61	M	T33	13
	62	M	T33	13
Vehicle 24h treatment	63	F	C34	16
	64	F	C34	16
	65	F	C34	16
	66	F	C34	16
EIDD-2801 24h treatment	67	F	C34	16
	68	F	C34	16

PMC7979515: SARS-CoV-2 Infection is Effectively Treated and Prevented by EIDD-2801

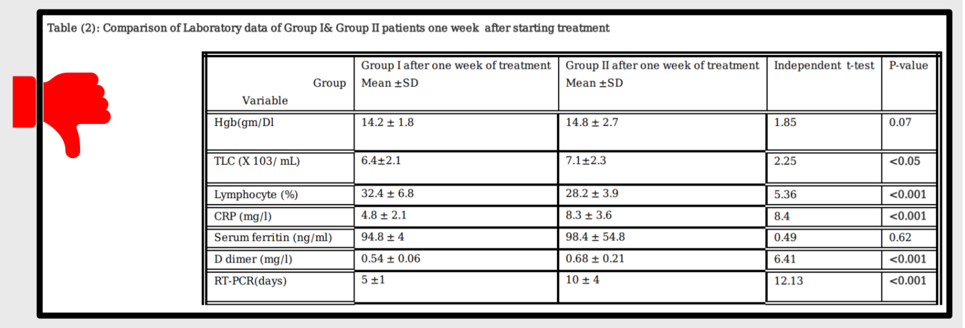


Table (2): Comparison of Laboratory data of Group I& Group II patients one week after starting treatment

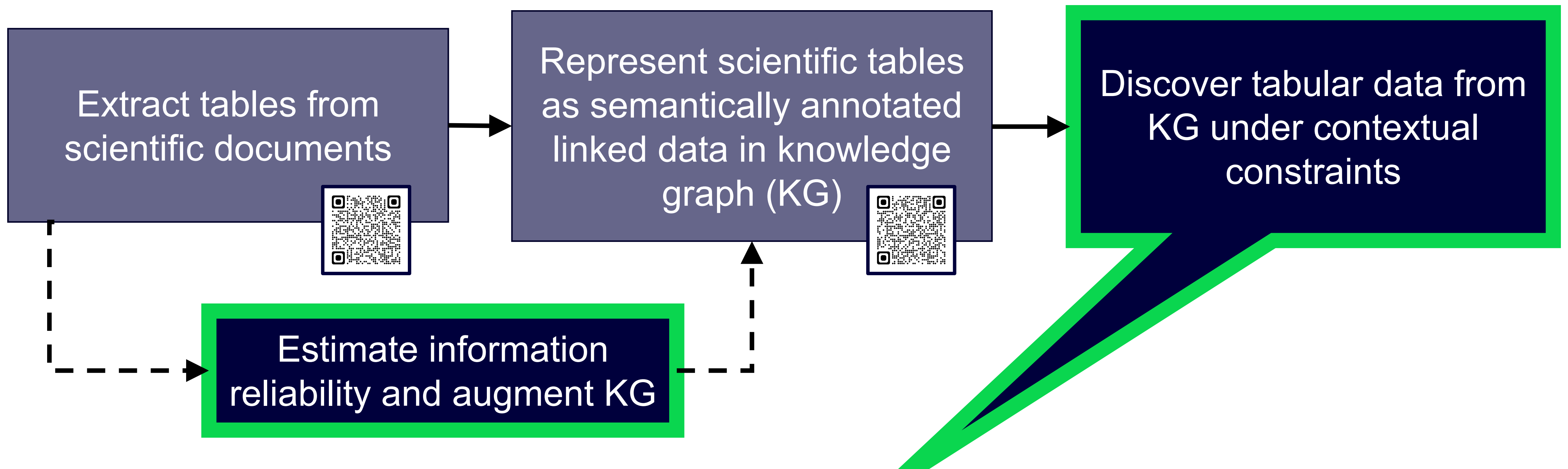
Group / Variable	Group I after one week of treatment Mean ±SD	Group II after one week of treatment Mean ±SD	Independent t-test	P-value
Hgb(gm/Dl	14.2 ± 1.8	14.8 ± 2.7	1.85	0.07
TLC (X 103/ mL)	6.4±2.1	7.1±2.3	2.25	<0.05
Lymphocyte (%)	32.4 ± 6.8	28.2 ± 3.9	5.36	<0.001
CRP (mg/l)	4.8 ± 2.1	8.3 ± 3.6	8.4	<0.001
Serum ferritin (ng/ml)	94.8 ± 4	98.4 ± 54.8	0.49	0.62
D dimer (mg/l)	0.54 ± 0.06	0.68 ± 0.21	6.41	<0.001
RT-PCR(days)	5 ±1	10 ± 4	12.13	<0.001

PPR230896: Efficacy and Safety of Ivermectin for Treatment and Prophylaxis of COVID-19 Pandemic

Our automated rule-based structural characterization of 120,000+ tables shows high variability amongst scientific tables

Read more about this work here

A semantics-driven approach to tabular data discovery

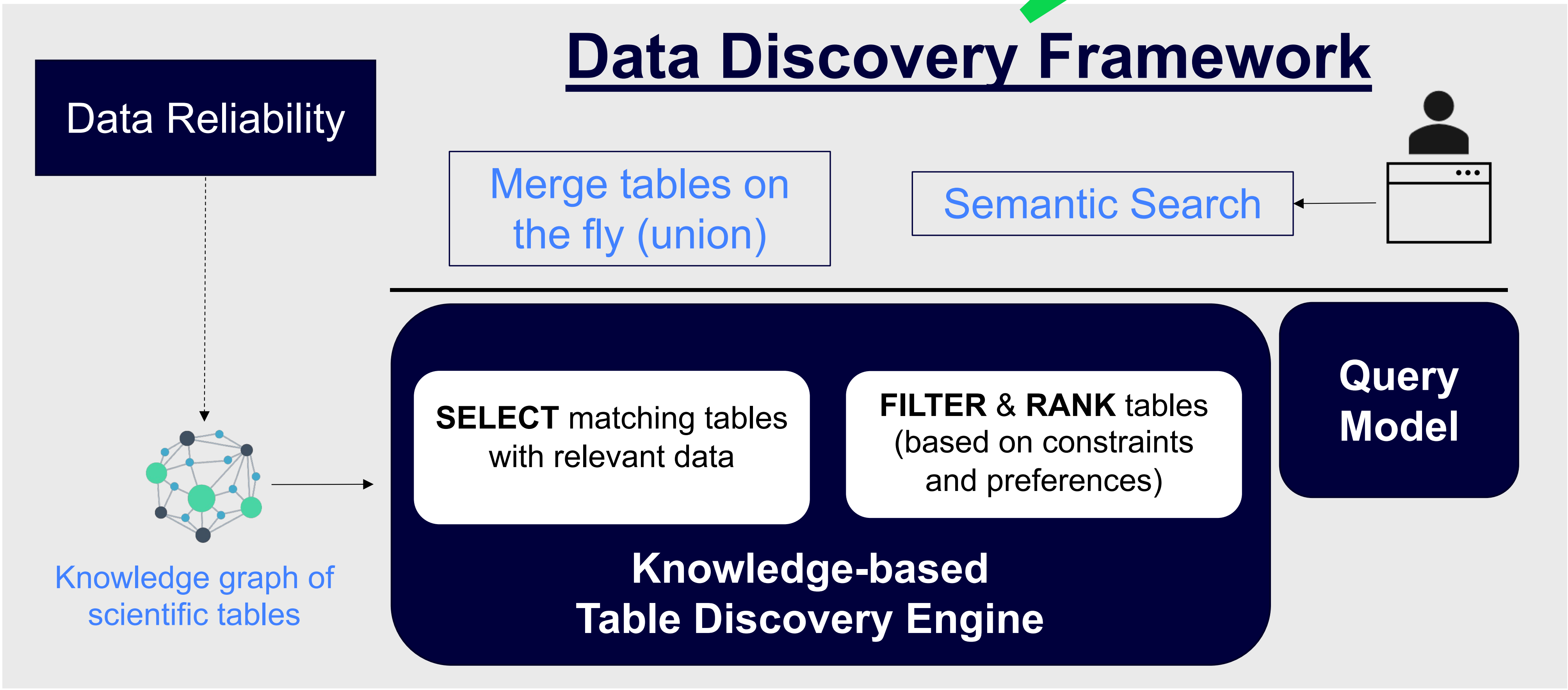


Extract tables from scientific documents

Represent scientific tables as semantically annotated linked data in knowledge graph (KG)

Discover tabular data from KG under contextual constraints

Estimate information reliability and augment KG



Data Discovery Framework

Data Reliability

Merge tables on the fly (union)

Semantic Search

Knowledge graph of scientific tables

SELECT matching tables with relevant data

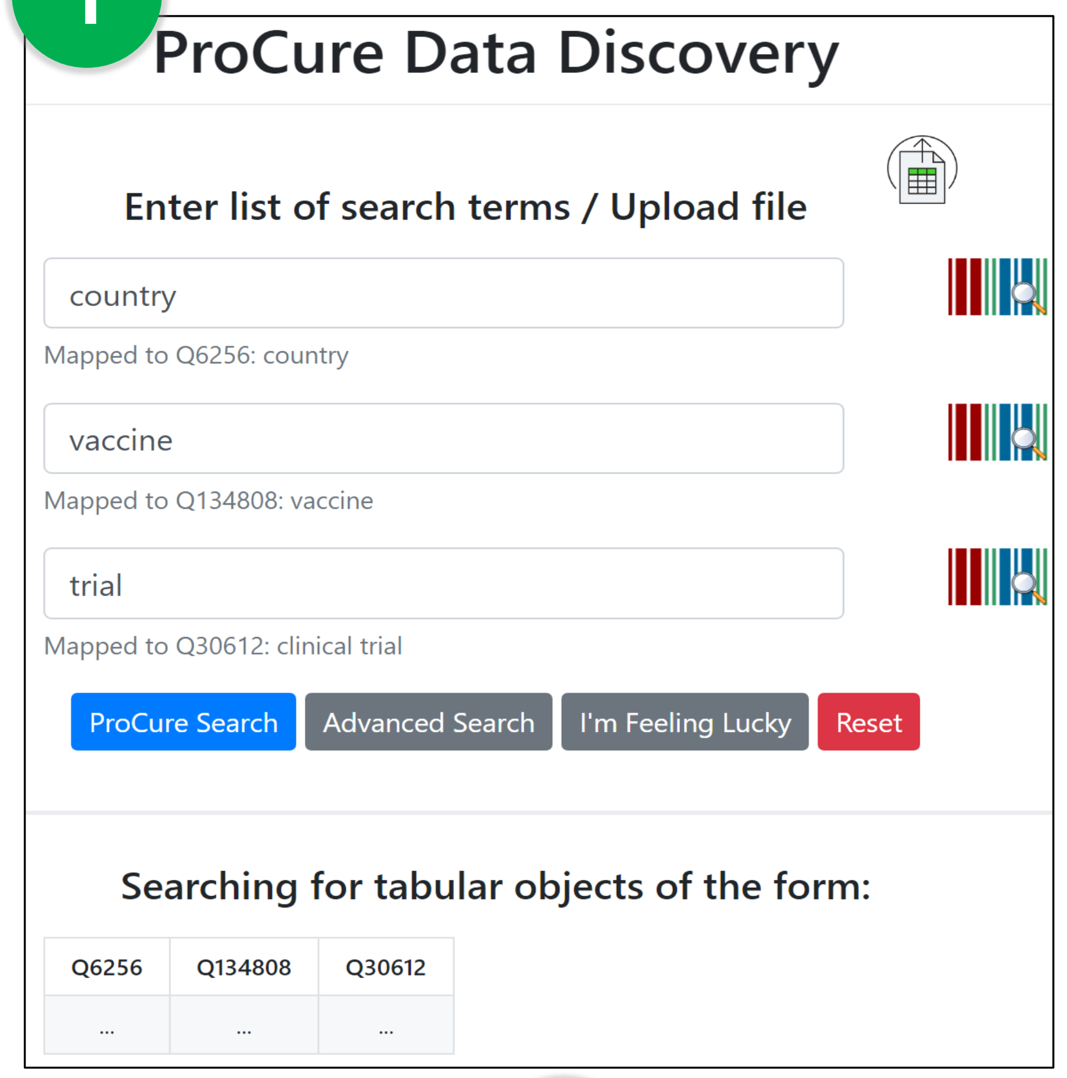
FILTER & **RANK** tables (based on constraints and preferences)

Query Model

Knowledge-based Table Discovery Engine

- Synthesize (relational) tabular response to semantic search requests
- Specify diverse set of contextual constraints
- Auto-generation of SPARQL query/ies corresponding to search request
- Extend with preliminary on-the-fly table generation capability

Table Discovery Prototype System



ProCure Data Discovery

Enter list of search terms / Upload file

country

Mapped to Q6256: country

vaccine

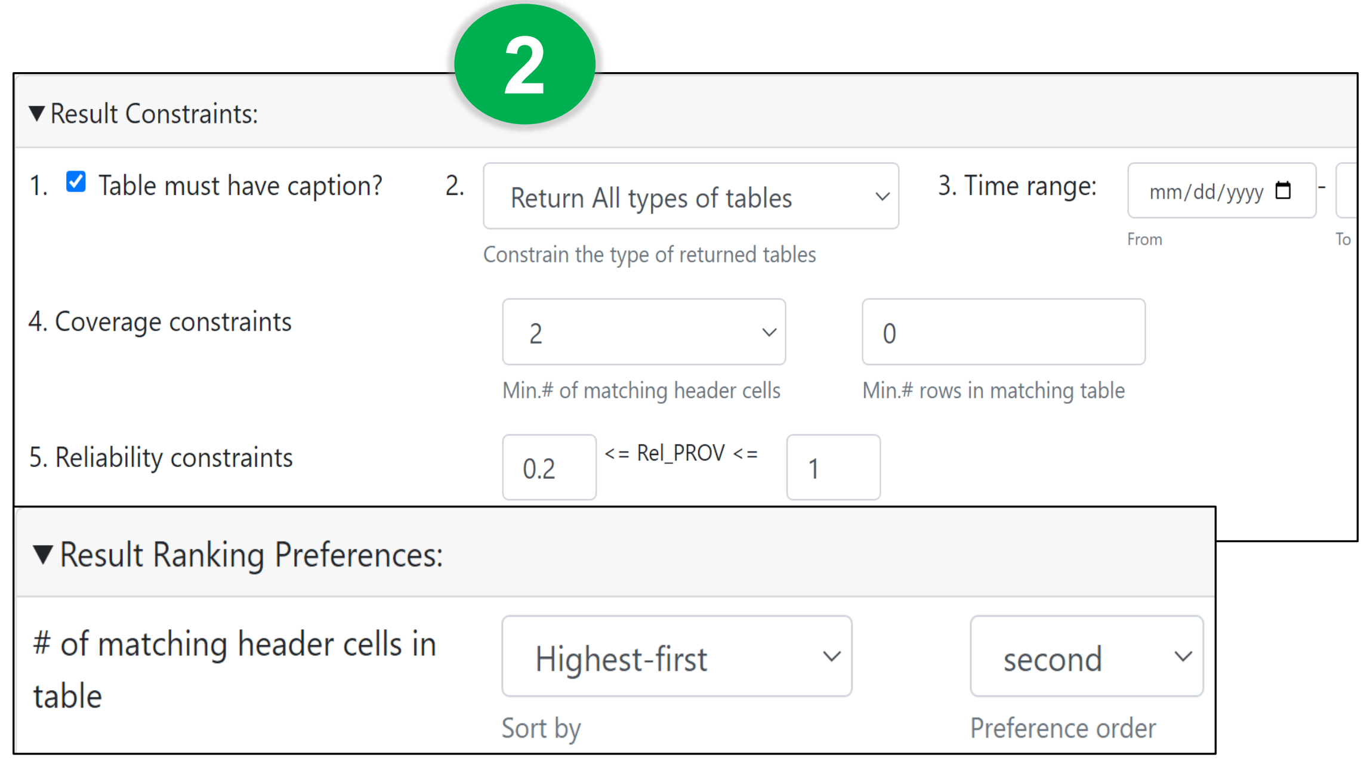
Mapped to Q134808: vaccine

trial

Mapped to Q30612: clinical trial

Searching for tabular objects of the form:

Q6256	Q134808	Q30612
...	...	...



▼Result Constraints:

1. ☑ Table must have caption?

2. Return All types of tables — Constrain the type of returned tables

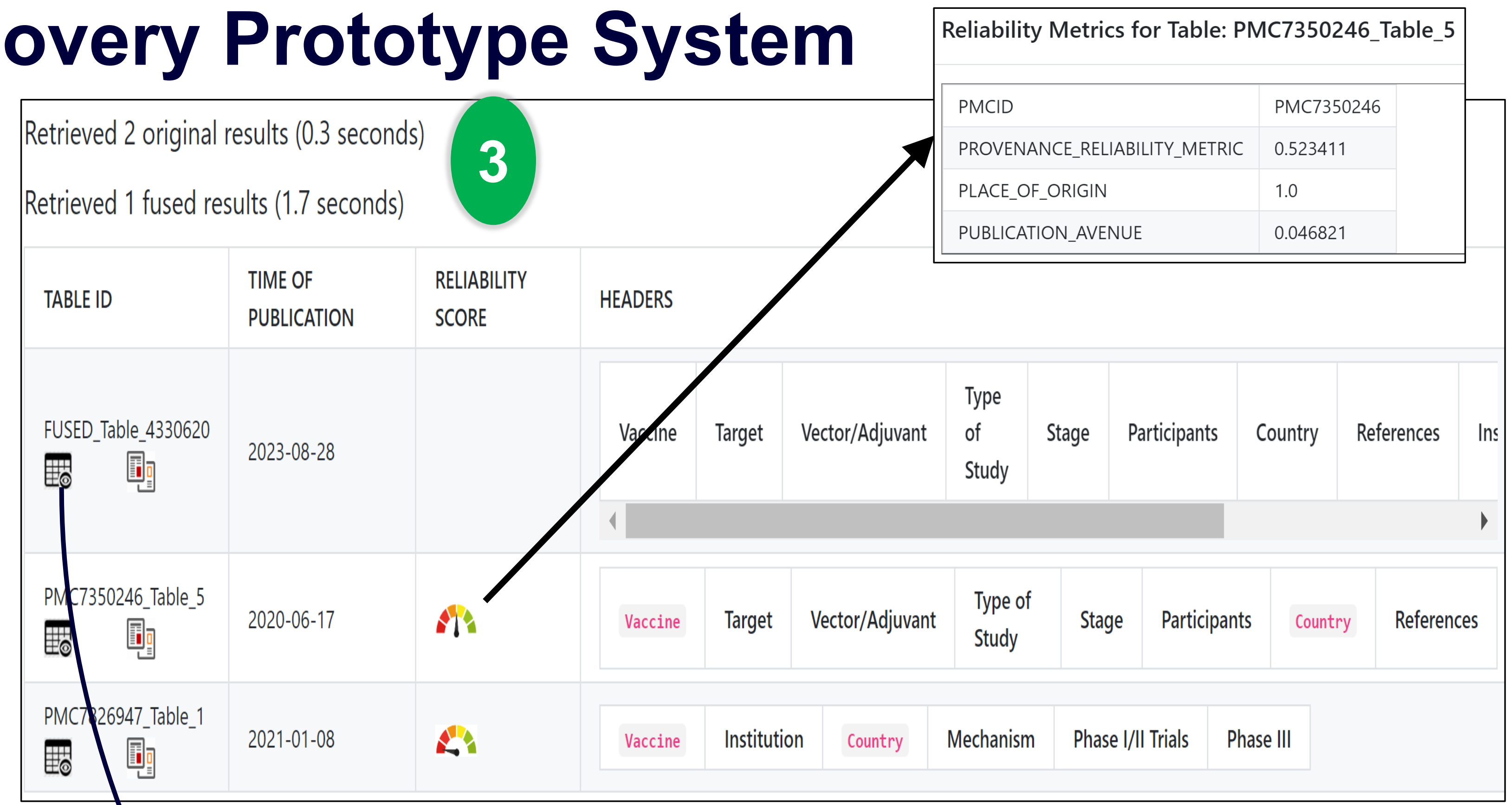
3. Time range: mm/dd/yyyy - (From, To)

4. Coverage constraints: 2 (Min.# of matching header cells); 0 (Min.# rows in matching table)

5. Reliability constraints: 0.2 <= Rel_PROV <= 1

▼Result Ranking Preferences:

# of matching header cells in table: Highest-first (Sort by); second (Preference order)



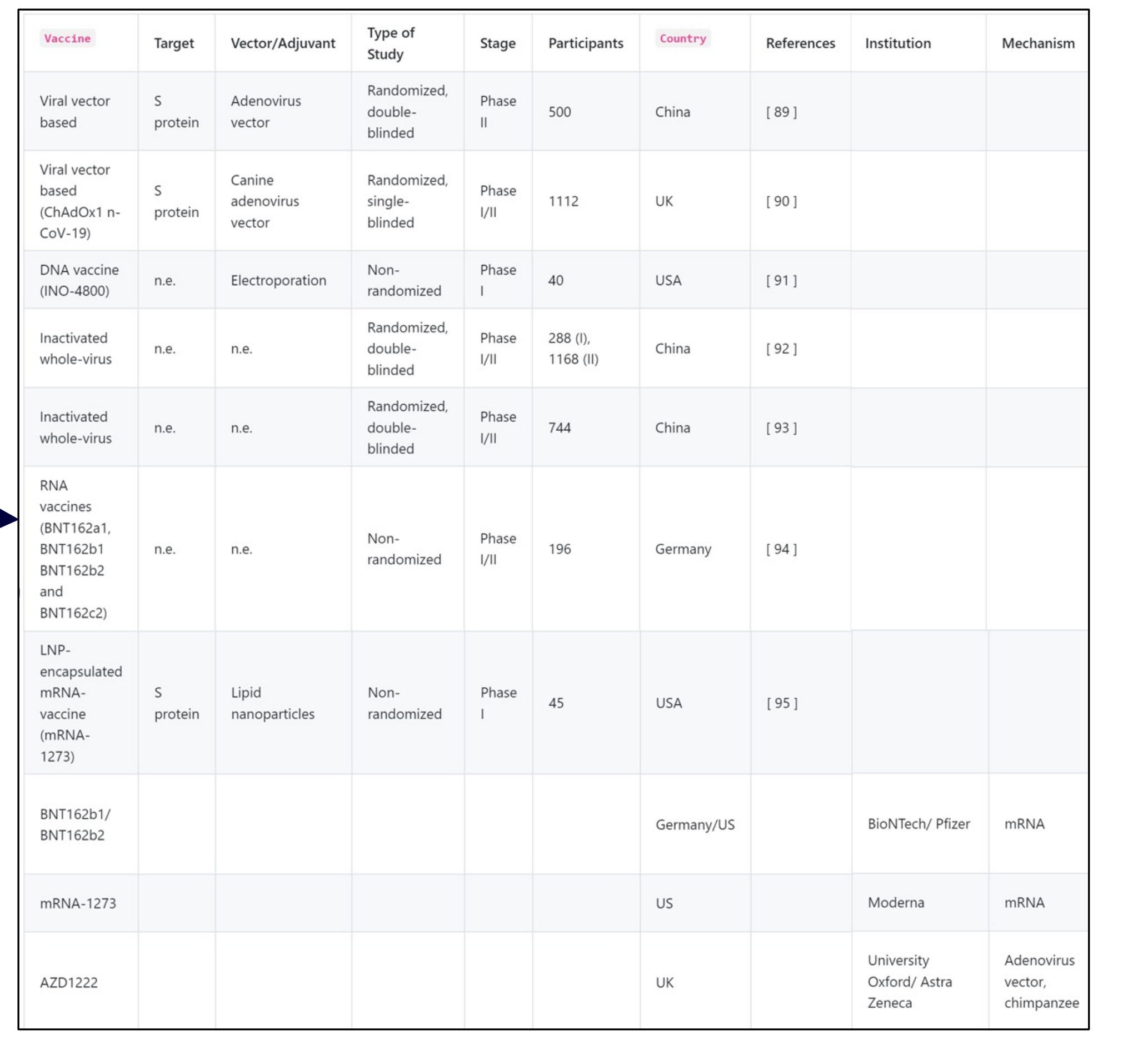
Retrieved 2 original results (0.3 seconds)

Retrieved 1 fused results (1.7 seconds)

TABLE ID	TIME OF PUBLICATION	RELIABILITY SCORE	HEADERS
FUSED_Table_4330620	2023-08-28		Vaccine, Target, Vector/Adjuvant, Type of Study, Stage, Participants, Country, References, Ins
PMC7350246_Table_5	2020-06-17		Vaccine, Target, Vector/Adjuvant, Type of Study, Stage, Participants, Country, References
PMC7826947_Table_1	2021-01-08		Vaccine, Institution, Country, Mechanism, Phase I/II Trials, Phase III

Reliability Metrics for Table: PMC7350246_Table_5

PMCID	PMC7350246
PROVENANCE_RELIABILITY_METRIC	0.523411
PLACE_OF_ORIGIN	1.0
PUBLICATION_AVENUE	0.046821



Vaccine	Target	Vector/Adjuvant	Type of Study	Stage	Participants	Country	References	Institution	Mechanism
Viral vector based	S protein	Adenovirus vector	Randomized, double-blinded	Phase II	500	China	[89]		
Viral vector based (ChAdOx1 n-CoV-19)	S protein	Canine adenovirus vector	Randomized, single-blinded	Phase I/II	1112	UK	[90]		
DNA vaccine (INO-4800)	n.e.	Electroporation	Non-randomized	Phase I	40	USA	[91]		
Inactivated whole-virus	n.e.	n.e.	Randomized, double-blinded	Phase I/II	288 (I), 1168 (II)	China	[92]		
Inactivated whole-virus	n.e.	n.e.	Randomized, double-blinded	Phase I/II	744	China	[93]		
RNA vaccines (BNT162a1, BNT162b1 BNT162b2 and BNT162c2)	n.e.	n.e.	Non-randomized	Phase I/II	196	Germany	[94]		
LNP-encapsulated mRNA-vaccine (mRNA-1273)	S protein	Lipid nanoparticles	Non-randomized	Phase I	45	USA	[95]		
BNT162b1/ BNT162b2						Germany/US		BioNTech/ Pfizer	mRNA
mRNA-1273						US		Moderna	mRNA
AZD1222						UK		University Oxford/ Astra Zeneca	Adenovirus vector, chimpanzee

This research was based on work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2021-21022600004].
The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.