Data Linkage in Practice: A Living Systematic Review of Clinical Trials in The United States (US) Utilizing Linkage to Real World Data

Evelyn Rizzo[1], Kevin Kallmes[2], Thomas Dougherty[3]

[1]Mobility HEOR Akron, Ohio, [2]Nested Knowledge, St Paul, Minnesota; [3]Clinical Data Science and Evidence, Novo Nordisk Inc., Plainsboro, New Jersey

Plain Language Summary

Why does it matter? Data linkage connects clinical trial results to real-world data like insurance claims and medical records, helping researchers answer new questions about medicines.

What did we do? We reviewed published clinical trials in the US that linked trial data with real-world information, like insurance claims and electronic health records. We analyzed how and why these studies used linked data in their research.

What did we find? 31 trials used data linkage, mostly linking to insurance claims. Studies covered a range of diseases, and on average, 65% of participants' data could be linked. The main reasons were to study treatment effectiveness, costs, safety, and survival.

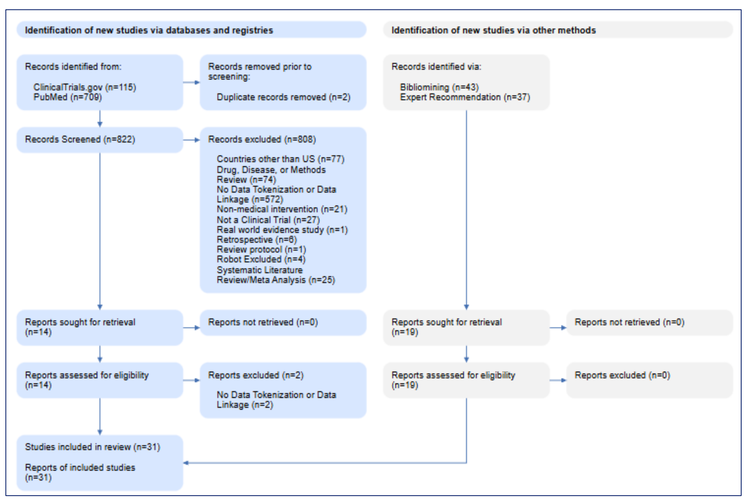
Objective

- Data linkage and tokenization are increasingly being adopted to address current limitations in clinical trials; however, research on topics and implementation of linkage/tokenization in practice is limited.
- The objectives of this systematic review were to describe and quantify examples of published clinical trials in the US that used data linkage and evaluate the analytical goals and uses of linked data.

Methods

- Relevant articles were identified through PubMed and ClinicalTrials.gov searches implemented on an artificial intelligence-assisted systematic literature review platform (AutoLit, Nested Knowledge), for publications between 2014-2025.
- Articles were included if they reported a pharmacological intervention and a US-based study population. **(Figure 1)**
- Study background, objective, patient disease state, type of linked data, linked data elements, and linkage methods were extracted from each study.

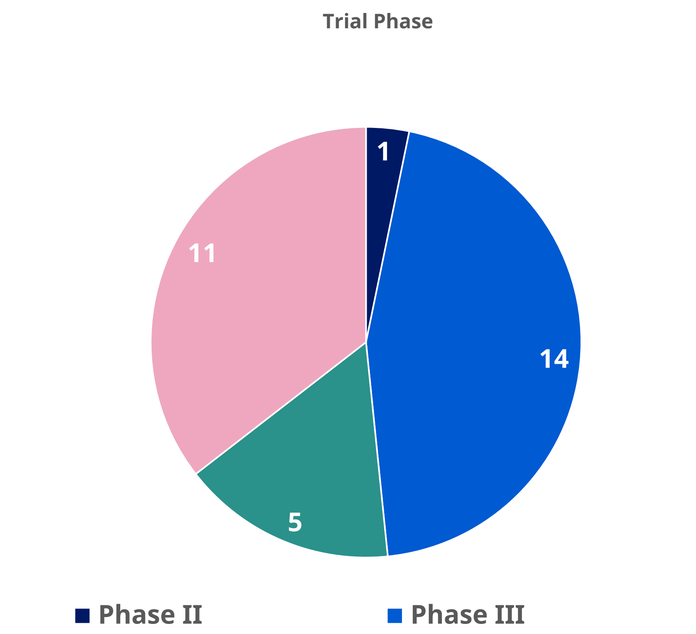
Figure 1: PRISMA diagram



Results

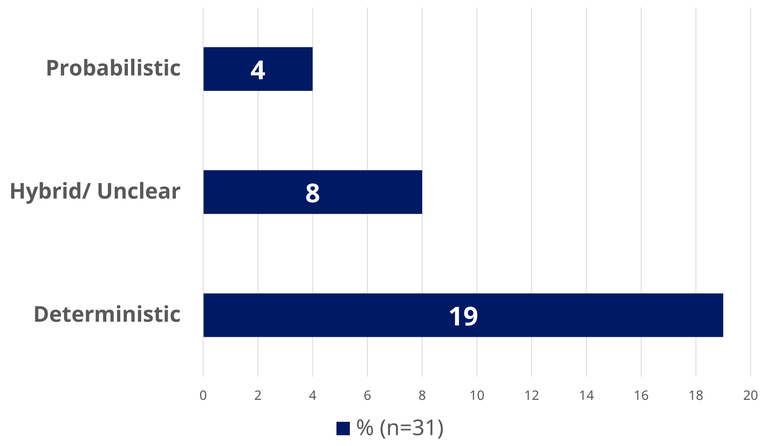
- Out of 902 abstracts screened, 31 publications reporting trials with linkage were included in this review. **(Figure 1)**
- There were 11 interventional trials, 1 phase II, 14 phase III, and 5 phase IV trials. **(Figure 2)**

Figure 2: Distribution of Trial Phases Among Included Studies



- Most studies used deterministic linkage (61.3%), followed by methods which were hybrid or unclear (25.8%), and probabilistic linkage (12.9%). (**Figure 3/Table 1)**

Figure 3: Distribution of Linkage Methods



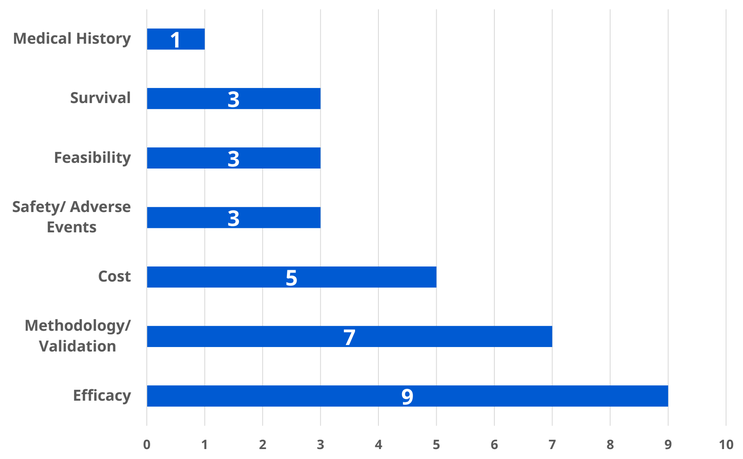
- Trial data were linked with real-world datasets,
- The studies were sponsored by industry (8), academic (6) and government institutions (17). **(Table 1)**
- including claims data (74.2%), registries (16%), and electronic health records (10%). **(Table 1)**
- The disease states were: Cardiovascular Risk(10), Cancer/Tumors (5), Aortic Stenosis (2), Kidney Disease (3), Women's Health (3), and Other (7). **(Table 1)**
- Of the 28 studies that reported the percentage of the population that was successfully linked, the range was 11.6%-100% and average of 64.7%. (**Table 1)**

Table 1. Characteristics of Included Studies

Sponsor / Study	Objective	Disease / Condition	Data Linked	Method	% Pop Linked
Edwards Lifesciences	Cost-effectiveness	Severe aortic stenosis	Medicare claims	Probabilistic	77.50%
Natl Cancer Inst. NCORP	Safety/Adverse Events	Metastatic prostate cancer	Medicare claims	Deterministic	56%
Genzyme/Sanofi	Efficacy/Safety	Kidney transplant	OPTN registry	Exact/Deterministic	89%
Univ. Rochester NCORP	Feasibility/Validation	Advanced cancer	NDI, Obituaries	Probabilistic	72%
BC Centre for Cardiovascular Health	Cost	Aortic stenosis	Medicare claims	Probabilistic	76.50%
Eli Lilly and Company	Feasibility	Rheumatoid arthritis	Claims	Hybrid (scoring)	88%
Funding PE Drawz: NIH	Methodology	Hypertension	EHR/Labs	Hash/link	63%
N/A (Academic, Propofol GA-CARES)	Survival	Cancer (surgical)	EMR, Cancer Registry	Manual/EHR	100%
NIH R01AG058971 (ALLHAT)	Safety	Hypertension	Medicare claims	Deterministic	50%
SAFE-PCI Trial (Academic)	Methodology	CAD, PCI in women	CathPCI registry	Hierarchical	82%
VA NEPHRON-D	Validation	Diabetic kidney disease	EMR, Trial Report	Deterministic	71.10%
Sunnybrook Health (CLEANJoint)	Safety	Joint replacement	OHIP (claims, admin)	Deterministic	N/A
NIH R01 CA165277	Methodology/Cost	Pediatric leukemia	PHIS billing	Deterministic	96%
Indiana IMPACT Study	Feasibility/Validation	Depression, CVD	EMR, Medicare, Medicaid	Deterministic	13%
Janssen/Johnson & Johnson	Medical history	COVID-19 (vaccine)	EHR, claims, labs	Tokenization	N/A
Genentech (AD-LINE)	Efficacy/Safety	Early AD, dementia	Medicare claims	Deterministic	55%
Genentech (WeSMA)	Feasibility	SMA	Open Claims, RWD	Tokenization	96%
NIH R01AG067498	Survival	Hypertension	NDI, CMS, SSA records	Deterministic	69.30%
Sanofi, AHRQ, PCORI	Effectiveness/Safety	Influenza, Alzheimerâ€™s, NH residents	Medicare/NH assessmt	Deterministic	N/A
NHLBI R01HL136708	Clinical outcomes	Post-stent, DAPT	Registry, claims	Deterministic	65% registry, 34% claims
AHRQ, KL2TR001870, R01HL136679	Concordance	Hypertension (SPRINT)	EHR	Deterministic	35%
Astellas	External validity	Kidney transplant	OPTN registry	Deterministic	97%
NHLBI, ACC NCDR	Ischemic/Bleeding events	PCI, elderly	Claims	Deterministic	11.60%
NHLBI (ALLHAT)	Gout outcome	Hypertension	CMS, VA claims	Deterministic	71.80%
FDA, Burroughs Wellcome	Method comparison	CVD	Medicare claims	Deterministic	55.60%
NCI, HOPE Foundation	Long-term sequelae	Colorectal cancer	Medicare claims	Deterministic	37.80%
NIH/NEI K23EY022949, R01 EY015473	OAG risk	Hysterectomy, OAG	Medicare claims	Deterministic	93.40%
NHLBI	Economic outcome	Post-menopause	Medicare claims	Deterministic	62.50%
NHLBI	CTS outcome	Postmenopausal women	Medicare claims	Deterministic	79.70%
NIH/NHLBI Contracts NO1-HC-35130, etc.	Fracture risk	Hypertension	Medicare, VA	Deterministic	74.90%
NCI/NCORP / CCOP / UM1CA182883-03	Adverse consequences	Prostate cancer, finasteride	Medicare claims	Deterministic	73.80%

- The key objectives for using linkage were efficacy (9), cost (5), methodology/validation (7), safety/adverse events (3), feasibility (3), survival (3), and medical history (1). (**Figure 4/Table 1))**

Figure 4: Primary Objectives for Data Linkage



Conclusions

- This review demonstrates the increased use of data linkage by US-based government, industry and academic centers in clinical trials for drugs for a broad range of therapeutic areas and objectives.
- These findings show a burgeoning role for linkage in expanding outcome collection and analysis across diverse disease areas.

This study was sponsored by Novo Nordisk Inc.
Presented at the ISPOR EU conference in Glasgow, Scottland 2025