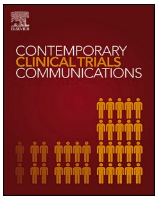
Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc



Linking clinical trial participants to their U.S. real-world data through tokenization: A practical guide

Michael J. Eckrote [a,1], Carrie M. Nielson [b,*,1], Mike Lu [b], Tyler Alexander [c], Rikisha Shah Gupta [b], Kim Wah Low [b], Zhiwei Zhang [b], Austin Eliazar [a], Reyna Klesh [a], Andrew Kress [a], Matt Bryant [b], Alex Asiimwe [b], Nicolle M. Gatto [d,e], Nancy A. Dreyer [f]

[a] *HealthVerity, Philadelphia, PA, USA*
[b] *Gilead Sciences, Foster City, CA, USA*
[c] *SimulStat Incorporated, Solana Beach, CA, USA*
[d] *Scientific Research & Strategy, Aetion Inc., New York, NY, USA*
[e] *Columbia Mailman School of Public Health, New York, NY, USA*
[f] *Dreyer Strategies LLC, Newton, MA, USA*

ARTICLE INFO

Keywords:
Real-world data
Tokenization
Data linkage
Privacy preservation

ABSTRACT

In drug development, the use of real-world data (RWD) has augmented our understanding of patients' health care experiences and the effects of treatments beyond clinical trials. Although electronic health record (EHR) data integration at clinical trial sites is a widely adopted practice, primarily for recruitment and data capture, a challenge to data utility is the fragmentation of health data across different sources.

Linking RWD sources to each other and to trial data – while preserving patient privacy through tokenization – aids in filling evidence gaps with outcome data and facilitates the generalization of effects from controlled trial environments to real-world settings. This paper describes the applications of RWD linkage and how they benefit both clinical development and real-world decision-making. Trial benefits include improving interpretability and generalizability (e.g., by remediating missing data or losses to follow-up), extending follow-up beyond trial closeout, and characterizing the applicability of trial results to under-represented groups.

The operational aspects of linking trial data to RWD are addressed, emphasizing the importance of using privacy-preserving record linking systems with established metrics of accuracy and precision, managing consent, and providing the necessary training and resources at trial sites to inform participants about providing access to their RWD through data linkage.

Advances in computing and the proliferation of patient data are accelerating access to diverse data sources. Analytic innovations and expanded uses of RWD for drug development have driven the use of novel clinical trial designs that integrate RWD sources with trial data. These efforts promote both more inclusive and efficient patient recruitment for clinical trials and more complete evidence of therapeutic effects than either clinical trials or RWD can provide on their own.

Site-based integration of electronic health record (EHR) data into trial operations is widely used to support recruitment and ascertain baseline and follow-up data, which has become more robust with EHR networks [1–3]. However, the fragmentation of patient health data makes it difficult to obtain a complete picture of a patient's health journey. The ability to link records from disparate sources to a single patient can overcome this limitation. Near-term opportunities arising through the incorporation of clinical trial participants' linked RWD include filling gaps in medication use, outcome data and health care utilization.

Drug development is increasingly centered around the patient [4], community, and healthcare provider (HCP). By thoughtfully fusing relevant data, research teams can broaden their view of patients' journeys across systems and gain a better understanding of treatment benefits and risks for groups of interest. From the patient perspective, the ability to collate clinical data linked across systems and share it with caregivers can facilitate the navigation of care options and enable them to identify relevant clinical trial options in an end-to-end ecosystem [5,6]. From the HCP and clinical trialist perspective, a disease-based

[*] Corresponding author.
E-mail address: carrie.nielson@gilead.com (C.M. Nielson).
[1] These authors contributed equally.

https://doi.org/10.1016/j.conctc.2024.101354
Received 5 January 2024; Received in revised form 13 August 2024; Accepted 16 August 2024
Available online 17 August 2024

Abbreviations

electronic master patient index	(eMPI)
external comparator arm	(ECA)
health technology assessment	(HTA)
loss to follow-up	(LTFU)
personally identifiable information	(PII)
privacy-preserving record linkage system	(PPRLS)
real-world data	(RWD)

ecosystem that uses health data linkage can help fill gaps in trial data caused by loss to follow-up, can extend observation time for long-term outcomes, can help contextualize trial findings for broader patient populations, and offers opportunities for real-world variable validation against trial standards. Here we describe how linkage of clinical trial data to those same patients RWD helps advance disease understanding and biomedical science. We then describe operational considerations required for trial participants to enable linkage to their RWD.

Applications of RWD-trial linkage

Using RWD to support treatment effect estimation and interpretation is one of the primary drivers of RWD-trial linkage. This section describes such applications, and other potential uses are discussed at the end of the manuscript.

RWD can fill gaps in treatment experience or health outcomes due to trial loss to follow-up and can be used to extend the observation period after trial completion to augment information about treatment benefits and risks. It is rarely true that the same health outcomes are measured in RWD as in clinical trials unless the trials were designed to include RW outcomes [7]. However, RWD-trial linked data offers opportunities to evaluate real-world impact and can also be used to validate RWD-based variable definitions against trial standard definitions. RWD can be used to contextualize health outcomes and provide a direct comparison of the outcomes observed in the real-world setting to those observed in single-arm trials. For populations with different demographics and geographies than the trial, RWD can be used in population bridging or transportability studies to estimate effects relevant for policy makers, payers and prescribers. For each of these opportunities, the standards that have been developed for real-world evidence generation are relevant. For example, the RWD fit for purpose should be systematically evaluated, and the validity of key RWD measures should be known or determined during the trial-RWD study. Additional considerations for designing and analyzing a trial-RWD linked study intended for treatment effect estimation are described in this section.

Quantify and mitigate the impacts of trial losses to follow-up

Loss to follow-up (LTFU) can endanger effect-estimate power and precision and can introduce biases into trial results [8]. While strategies to engage participants in trials are essential for retaining participants, incorporating linked RWD presents an additional approach enabling prospective passive data collection. This method aids in evaluating the impact of selective losses to follow-up. Furthermore, in specific situations, linked RWD can address gaps in outcome ascertainment. Considerations of cost to recruit additional participants – in anticipation of LTFU – and the danger of producing a biased or underpowered estimate after LTFU should be weighed against the possibility of recovering information from RWD.

Facilitate long-term real-world outcome assessment after trial completion

For events that can be captured reliably in RWD after a trial ends, linkage offers an efficient approach to follow-up for long-term extension studies and trials that require long-term observation. This opportunity is especially relevant when surrogate trial endpoints are used as a substitute for rare events and for conditions that develop slowly [9]. For example, a trial with a disease progression endpoint could continue to follow patients through their linked RWD for worsening outcomes requiring hospitalization or even mortality in data sources with reliable capture of death, such as the Social Security Administration's Death Master File.

Contextualize and generalize trial findings

When it is not ethical or practically feasible to randomize participants to a control arm in a trial, an RWD-based external comparator arm (ECA) can sometimes be used. An ECA uses RWD from patients not treated with the study drug. ECAs can be used to contextualize the trial outcomes, to estimate treatment effects in the target population, or to support early stopping decisions if there are substantial differences in outcomes between the trial and ECA [10]. A common conundrum in RWD-based ECA design is that baseline comparability of trial and RWD patients is not fully known since baseline characteristics are measured differently in trial protocols and RWD derived from usual care. Access to a subset of patients with both trial and RWD data through linkage allows for direct comparison, validation, and bias analysis.

The development of generalizability (application of estimates in a study sample to its broader target population) and transportability (application of estimates in one to a different population) methods has allowed for RWD to extend trial results beyond trial participants [11] and to address trial selection bias [12]. These methods use trial-RWD linkage to provide the foundation for achieving external validity. This evidence is important to regulators and can be essential for health technology assessments (HTAs) and understanding treatment impact on a broader scale than is measurable in a trial.

A recent demonstration of generalizability methods involved a lung cancer screening trial for reducing mortality among smokers. Because screening efficacy was expected to differ by sex, trial sites were oversampled for female representation. To generalize the treatment effect estimate to the full population, several demographic and anthropometric variables were used to generate the probability of being in the trial and of being in the target population. Weighting by the inverse odds of being in the trial allowed for estimation of the generalized effect estimate. Ensuring key identifiability conditions held was indispensable for causal inference (e.g., exchangeability, positivity, consistency, lack of measurement error) [11].

Enable validation of RWD variables

Validation of key exposure, outcome, and potential confounding variables is a critical component of RWE study design [13], including RWD-trial linked designs. The use of RWD to remediate missing trial data or to extend follow-up for a trial outcome requires validation of the RWD measure against the trial standard. Availability of an RWD-trial linked patient set offers opportunities to conduct this validation in the patient population of interest. Methods for RWD algorithm validation studies have been well described [14] and can be built into trial-RWD linked study protocols to measure the extent of misclassification in RWD against the trial variables. Such validation studies can occur even before the trial is complete, using adaptive validation methods [15].

With results of validation studies available, bias analyses can be executed. Quantifying relevant biases is essential for causal inference. It is key to ensuring that uncertainty in effect estimates due to measurement or modeling errors is bounded by what is known – or reasonably assumed – about the magnitude of the errors. For example, knowing the sensitivity and specificity of an outcome measured in RWD allows for bias analyses to convey the robustness of a treatment effect estimate to misclassification [16–18].

In summary, trial-RWD linkage presents important opportunities to reduce the impact of trial losses to follow-up and to extend observation periods for clinically relevant endpoints. In addition, linkage provides an opportunity to validate real-world outcomes and key covariates against trial reference standards in the linked subset. Incorporating the principles of RWE design improves the likelihood of valid causal inference. With careful consideration, researchers can ensure the quality of RWD, reduce bias, and improve the generalizability of treatment effect estimates from trial-RWD linked studies while minimizing the burden on patients and health systems.

Operationalization: the mechanics of linking trial data to RWD

The first decision must be to consider whether the RWD of interest are generally available and if so, in what types of data sources. For example, some EHR data have been developed to be rich for specific conditions whereas others have broader population coverage but less disease-specific depth. RWD sources are heterogeneous and can include commercial claims, electronic health records, patient registries, pharmacies, public insurance systems, death records and other sources [19]. They may be used individually or in combination, depending on data access agreements, and the quality [20] and completeness [21] of measures that are available for each patient will vary.

RWD-based study design must incorporate thorough evaluations, both to select the data sources that are fit for purpose and to assess the feasibility of using the fit-for-purpose data within the set of trial participants with sufficiently complete data. The opportunities and challenges of fusing heterogeneous RWD sources for causal inference has been described for big data applications in general and trial-RWD fusion specifically [12]. Refer to recent regulatory and industry standards for generation and use of real-world evidence [22–25].

While the concept of privacy-preserving data linkage has been used extensively in healthcare and other industries, linking a clinical trial participant's data to their RWD has only recently come into use. Here we outline the operational aspects for trial-RWD linkage.

1) Select a privacy-preserving record linking system (PPRLS) that can be used with most – if not all – RWD sources of interest
2) Develop consent processes
3) Enable sites to consent patients to tokenization and to collect and manage personally identifiable information (PII)
4) Identify fit-for-purpose RWD for linkage

PPRLS selection

Tokenization assigns an anonymized secure "token" to an individual's records through a salted one-way cryptographic hash function, which prevents reverse engineering of the token to reveal personal identifiers (Fig. 1). Generally, this function is performed by a trusted third party or is conducted behind the firewall where the RWD resides.

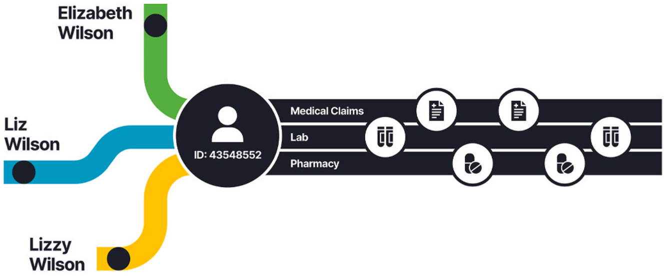


Fig. 1. Probabilistic privacy-preserving record linkage uses a set of personal identifiers to infer that records coming from disparate sources likely belong to the same patient. Masking such identifiers and replacing them with a token allows for the patient's deidentified data to be used for research.

While not all data sources contain the same personal identifiers, deterministic and probabilistic matching can be used to match individuals with varying degrees of accuracy. Understanding matching methods and performance metrics (Table 1) can help researchers select the appropriate PPRLS for the RWD being used.

Deterministic matching uses business rules for matching. For example, one rule might instruct the system to match two records based on matching Social Security number (SSN) and address fields [26]. Because SSNs are not generally recorded in RWD, most tokens are derived from a combination of first and last name, gender, date of birth, and zip code. Match accuracy can be improved by algorithms that can account for typographical errors or variations. Some systems offer many tokens that can be used singly or in combination [27–29]. In addition, the precision of a PPRL approach may improve when more patient attributes are included; however, sensitivity may decline if patients who lack a required identifier cannot be included. As sample sizes increase, the probability that identical tokens will be created for different patients due to chance (known as a hash collision or false-positive match) increases [30]. This problem can be minimized by conducting matching within patient subgroups of interest – e.g., by clinical criteria, geography, or other factors.

Accuracy of linkage systems: A clear description of the process of data linkage to RWD sources should be available, along with information about the provenance of the RWD sources and ideally the matching

Table 1
Metrics used to describe the performance of privacy-preserving record linkage systems relative to a gold standard for personal identity. Performance varies by the completeness of identifiers in RWD and the use of identifiers in the tokenization algorithms.

Term	Definition	Synonym or related term
True positive (TP)	Matching records identified through linkage that truly belong to the same person	
False positive (FP)	Matching records identified through linkage that truly do not belong to the same person	Hash collision
True negative (TN)	Records not identified as matches that truly do not belong to the same person	
False negative (FN)	Records not identified as matches that truly belong to the same person	
Terms that summarize TP, FP, TN, and FN		
Precision	TP/(TP + FP) Of all matching records identified, the proportion that truly belong to the same person	Positive predictive value
Recall	TP/(TP + FN) Of all true matches in the sample, the proportion that the matching algorithm identified	Sensitivity, true positive rate
F-Score	(2 x Precision x Recall)/(Precision + Recall) A summary score of precision and recall	
False discovery rate	FP/(FP + TP) Of all matching records identified, the proportion that do not truly belong the same person	
Accuracy	(TP + TN)/(TP + TN + FP + FN) A summary of the true matches that were correctly identified as matches and true non-matches that were correctly not identified as matches	
Specificity	TN/(TN + FP) Of all truly non-matching records, the proportion identified as not belonging to the same person	True negative rate
Additional terms used in tokenization performance evaluation		
Hash collision	Records from two different people mistakenly assigned the same token (through a hash function), resulting in a false positive match	False positive
Fill rate	1 – missingness rate Proportion of records with sufficiently complete identifiers to produce a token	

performance. The performance of a PPRLS can be quantified using metrics such as recall (a.k.a. sensitivity), false discovery rate, and precision (a.k.a. positive predictive value, Table 1). An evaluation of multiple tokens showed high precision (≥99 %) for tokens that used combinations of names (with or without soundex algorithms to account for phonetically similar matches), gender, DOB, zip code, and SSN; however, recall was low (23–65 %) when precise names, SSN, or address were required [27]. Relying on a token with low recall would reduce the number of matched samples and could introduce bias if matches are achieved differentially by a characteristic relevant to the study. As with the evaluation of diagnostic tests, choice of a PPRLS depends on the tolerance for inaccurate matches (false positives and false negatives). Industry and government agencies have reported on simulated linkage under various scenarios (e.g., high-quality identifiers like SSNs, moderate quality with missing SSNs, and low-quality identifiers, for example with spelling errors and transposed numbers) [31]. This evaluation demonstrated a wide range in accuracy across PPRL techniques, even with high-quality identifiers (e.g., F1 model accuracy scores between 0.56 and 0.96, where 1 is perfect accuracy). In practice, the use of various combinations of identifiers to create tokens with high accuracy has been a focus of PPRLS research on RWD [27,28,31]. For example, matches on multiple tokens have been shown to result in a false positive rate as low as single tokens that included SSN [27].

Researchers should consider both the average PPRLS performance and the completeness of the identifiers (or "fill rate") available for matching, recognizing that due to privacy concerns SSNs are not widely available in data sources [31]. Because most RWD require combinations of non-SSN identifiers, it is advisable to evaluate tokenization systems whose models are sophisticated enough to support such combinations. Regardless of overall PPRLS performance, the impact of linkage on selection into analytic cohorts should be evaluated. Some biases have been identified in linkage performance – for example, due to variations in document completeness and naming practices by ethnicity or country of origin [27,32]. Researchers should continue to evaluate whether linked samples differ on relevant characteristics from the underlying patient populations they are intended to represent.

8. *Consent process considerations*

Stakeholders should devise a strategy for RWD linkage patient consent similar to trial consent, including complying with the HIPAA compound authorization rule (45 CFR 164.501, 164.508, 164.512(i)). This means that if sponsors decide to combine the patient consent for RWD linkage with the broader clinical trial participant consent, sponsors will be required to incorporate a mechanism to allow subjects to opt-out (e.g., a separate checkbox). Consideration must then be given to whether potential participants who decline linkage would be allowed to enroll in the trial. If instead the RWD linkage consent is in a separate vehicle from the main trial consent, sponsors should consider positioning the "sub-consent" in a patient accessible way.

Within this process, stakeholders may also seek additional patient authorizations for the release of identifiable health-related information (for example, for regulatory requirements). Identifiable records can be differentiated from de-identified RWD in the sense that patients have consented for trial stakeholders to now review the actual source medical records, which often might expose the patient's identity. These records can augment any de-identified RWD and trial data, enabling covered entities to make more informed decisions for study activities, like screening and enrollment, and intra-trial medical monitoring of participants. While these applications remain exploratory, they may create opportunities for patients to be notified of future clinical studies, as seen in recent announcements by commercial pharmacy and laboratories regarding expansion into clinical trial services [33].

In our experience, including an option for the patient to consent to RWD linkage at the time of study enrollment yields better response than soliciting consent to linkage later in the study, perhaps reactively due to factors such as data missingness, losses to follow-up, evidence gaps, and evolving treatment landscape. By implementing tokenization at the onset of the trial, sponsors preserve the opportunity for future RWD analytics, all while reducing the operational burden of an additional consent later.

The consent revocation process can be accomplished through a dedicated third-party, perhaps a CRO, functional service provider or other technology vendor. Sponsors should also consider a long-term strategy for digitally managing such consent preferences to ensure only patients who have consented to have their RWD linked are included in future analysis datasets. This becomes especially crucial when other trial-related consent management solutions are decommissioned after study closeout.

Support for consent and PII collection and management

To ensure implementation success, sponsors should have a robust plan for enabling trial sites to gather the requisite PII for a trusted third party to create linkable patient tokens. Sites require materials to support the participant through the consent process, including training on the intricacies of the informed consent language related to RWD. This is especially important considering that privacy preservation, sub-study objectives, and supporting tokenization and answering participants' questions about tokenization might be atypical for trial operations. This not only means providing hands on training during regularly scheduled touchpoints like investigator meetings, site initiation visits, and site "office-hours," but also documentation and "leave-behinds" or tip sheets on RWD and tokenization.

In addition to managing the consenting process, research sites play a pivotal role in enabling the capability of RWD linkage, through tokenization, based on their role as a covered entity with access to patients' PII. As previously mentioned, tokens are assigned to the same individual's healthcare records, regardless of the source of the data, using demographic information about the individual (such as name, date of birth, gender, zip code) at the data source, to create the one-way hash. This means that sites must install a batch engine behind the site's firewall, or in most cases in clinical trials, trial coordinators must transcribe the information into a hosted and secure web-platform.

This activity is one reason for proactively collecting PII and consent: retroactive implementation can create numerous operational hurdles, especially if trial sites have already closed out. This introduces an opportunity to systematically manage the governance of the PII and patient consent collection process over multiple studies. Referring to the strategy of standardizing the prospective tokenization of all subjects at the start of the trial outlined earlier, sponsors might consider making PII capture and centralization a part of many of their portfolio's trials, preserving the opportunity to link to RWD when future evidence generation needs arise. Using an electronic master patient index (eMPI), sponsors can outsource the management of such protected data to a trusted third party to preserve privacy while enabling utility where applicable. eMPI solutions also help create a consistent source of truth for patient data while streamlining the process for managing identity and patient privacy preferences.

Selected RWD sources for linkage

Once patient PII is transformed into a de-identified token, the process of further defining which RWD will be linked to trial patients can continue.

Data availability is an important consideration, especially where linked RWD plays a crucial role in a sponsor's evidence generation plan. Several other tactics can be used to provide real-time insights on the likelihood of linking a trial patient to relevant fit-for-purpose RWD. One of these tactics is overlap analysis: once tokens are assigned to a patient in linkable datasets (trial and RWD), an overlap report can determine which RWD sources include the patient's records during which time periods. This can provide a preliminary indication whether the necessary RWD will be available for the patient. Periodic re-linkage of trial participants to accessible RWD can extend the length of observation period for patients who have changed jobs or geographies or aged out of parents' insurance. Current data restrictions make it unlikely that a patient only found in ex-U.S. or only in U.S. Medicare/Medicaid RWD will have linkable data for US trial sites. To the extent data are available for research, re-linkage efforts can help maintain continuity in data collection and improve the completeness of follow-up.

While patient privacy preservation remains an important consideration during the entire RWD linkage lifecycle, the final step of the process usually concludes with some level of advanced privacy review, ensuring HIPAA compliance is achieved and there will be little risk of re-identifying patients. There may be opportunities to negotiate with the privacy review team in terms of what potential PII is most important and which may be omitted or grouped into large categories, e.g., Northeast USA, to still achieve reasonably strong privacy protection. Depending on the use case, sponsors may require a HIPAA certification through expert third-party determination.

11. Other opportunities for RWD to enhance clinical trials

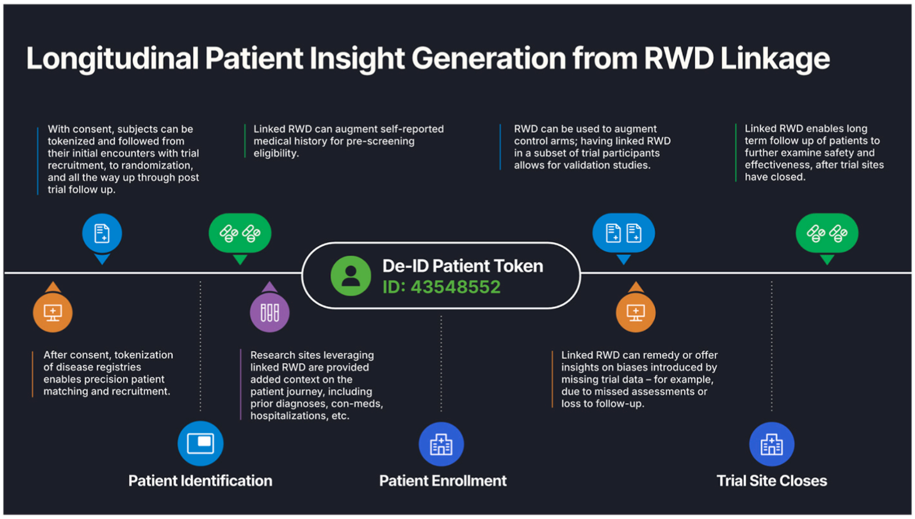
We have described some examples of how RWD linkage can enhance clinical trials by improving trial completeness, reliability, and generalizability. Implementation of data linkage both between RWD and trial data and across various RWD sources will allow for other advances in trial design and execution. For example, the ability to link patients (through their health data) to trial opportunities allows for targeted recruitment. Sponsors can leverage new trial delivery approaches designed to integrate with the healthcare system to meet patient needs and preferences. Targeted media can be used to direct a patient toward a web-based pre-screener, which allows them to link to their RWD and connect with a clinical site (Fig. 2). Sponsors can also promote patient engagement, which may be anonymous, before, during and after a clinical trial to exchange health and trial data. For example, a patient's pharmacy encounter can trigger a notification of their potential eligibility for a relevant trial, facilitated further if electronic health data and/or personal health record data can be consented for access.

RWD is being used for trial feasibility analyses (or "protocol optimization"). Combining multiple linked-RWD sources can yield sufficiently complete and large datasets to build trial emulation models, accelerating the implementation of feasible trial designs. These linked RWD sources can support advances in systematic, computationally complex emulations such as Trial Pathfinder [34].

In addition to opportunities afforded by linkage of deidentified RWD, patient-enabled linkage can empower disease-specific communities to promote patient advocacy to inform clinical trial design. This model has proven feasibility: for example, in GO2 for Lung Cancer, patients volunteer for an online registry, consent to provide their electronic health data, and can then be notified of relevant trials (go2.org).

Beyond these initial opportunities for linked RWD, future innovations include:

1. Using advanced techniques like secure multiparty computing and cipher encryption to combine RWD and consumer information. This enables **precision patient identification and trial promotion via digital media** for hard-to-reach patient communities. Patients who are interested can consent to share their medical history for web-based pre-screening. Linked RWD can augment self-reported medical history, potentially improving the completeness and efficiency of screening.
2. Using linked RWD allows sponsors and their partners to find a pool of **patients who fit certain criteria for a clinical trial.** For example, a lab service provider can identify patients with a specific diagnosis and lab results. With RWD that doesn't reveal personal information, partners can create groups of anonymous patients who match the trial requirements. These partners and healthcare organizations can then use lists of patients who agreed in advance to join clinical trials. By linking this information with the initially evaluated RWD, they can make recruitment efforts more focused on a group of patients closely aligned with the trial's criteria.
3. Extending the look-back period for medical records available to trial sites can **improve accuracy of eligibility screening.** When data are limited to what is available in an individual health system, sites and trial coordinators may screen-fail patients due to a lack of sufficient medical history to determine trial eligibility. Linked RWD provides a more complete view of the patient's journey, surfacing medical history information like prior diagnoses, prior or concomitant medications, hospitalizations, and lab values. Incorporating such information into pre-screening algorithms can make patient outreach more efficient.



Longitudinal Patient Insight Generation from RWD Linkage

With consent, subjects can be tokenized and followed from their initial encounters with trial recruitment, to randomization, and all the way up through post trial follow up.

Linked RWD can augment self-reported medical history for pre-screening eligibility.

RWD can be used to augment control arms; having linked RWD in a subset of trial participants allows for validation studies.

Linked RWD enables long term follow up of patients to further examine safety and effectiveness, after trial sites have closed.

De-ID Patient Token
ID: 43548552

After consent, tokenization of disease registries enables precision patient matching and recruitment.

Research sites leveraging linked RWD are provided added context on the patient journey, including prior diagnoses, con-meds, hospitalizations, etc.

Linked RWD can remedy or offer insights on biases introduced by missing trial data – for example, due to missed assessments or loss to follow-up.

Fig. 2. Tokenization is a component of privacy-preserving record linkage systems (PPRLS) that allows linkage of disparate data sources, including RWD and trial data, to construct a complete patient journey. The process uses personally identifiable information (PII) and assigns an anonymized secure "token" to an individual's records through a one-way cryptographic hash function. The same token is assigned to the individual's records, regardless of the source of the data, using personal identifiers (e.g., date of birth, gender, zip code) at each data source.

Conclusion

Integrating fragmented health data can result in a more complete representation of the patient's experience than any one RWD source or clinical trial can offer on its own. Linkage across RWD sources or between RWD and trial data can confer gains in scientific understanding, as well as improving recruitment efficiency and data quality. Technologic advances in privacy-preserving record linkage systems and benchmarking have improved linkage accuracy, although record and variable validation studies remain an essential component of reliable evidence generation. For trial-RWD linkage, patient consent and linkage operational hurdles are surmountable with appropriate planning and supporting materials for sites to inform study participants.

CRediT authorship contribution statement

Michael J. Eckrote: Writing – review & editing, Writing – original draft, Project administration, Methodology, Conceptualization. **Carrie M. Nielson:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Conceptualization. **Mike Lu:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Tyler Alexander:** Writing – review & editing, Writing – original draft, Project administration. **Rikisha Shah Gupta:** Writing – review & editing, Methodology. **Kim Wah Low:** Writing – review & editing, Methodology. **Zhiwei Zhang:** Writing – review & editing, Methodology. **Austin Eliazar:** Writing – review & editing, Methodology. **Reyna Klesh:** Writing – review & editing. **Andrew Kress:** Writing – review & editing. **Matt Bryant:** Writing – review & editing, Conceptualization. **Alex Asiimwe:** Writing – review & editing, Conceptualization. **Nicolle M. Gatto:** Writing – review & editing, Methodology, Conceptualization. **Nancy A. Dreyer:** Writing – review & editing, Methodology.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

MJE, AE, RK, and AK are employees of and hold equity in HealthVerity. AA, CMN, KWL, MB, ML, RGS, and ZZ are employees of and hold stock in Gilead Sciences Inc. TA is contracted to Gilead Sciences Inc. NMG is an employee of Aetion Inc; she holds stock options in Aetion Inc and Pfizer Inc. NAD is affiliated with Dreyer Strategies LLC and PicnicHealth.

References

[1] M. Dugas, M. Lange, C. Muller-Tidow, P. Kirchhof, H.U. Prokosch, Routine data from hospital information systems can support patient recruitment for clinical studies, Clin. Trials 7 (2010) 183–189.

[2] E.C. O'Brien, S.R. Raman, A. Ellis, B.G. Hammill, L.G. Berdan, T. Rorick, et al., The use of electronic health records for recruitment in clinical trials: a mixed methods analysis of the Harmony Outcomes Electronic Health Record Ancillary Study, Trials 22 (2021) 465.

[3] S.R. Raman, L.G. Qualls, B.G. Hammill, A.J. Nelson, E.K. Nilles, K. Marsolo, et al., Optimizing data integration in trials that use EHR data: lessons learned from a multi-center randomized clinical trial, Trials 24 (2023) 566.

[4] M. Algorri, N.S. Cauchon, T. Christian, C. O'Connell, P. Vaidya, Patient-centric product development: a summary of select regulatory CMC and device considerations, J. Pharmaceut. Sci. 112 (2023) 922–936.

[5] V.A. Rudrapatna, A.J. Butte, Opportunities and challenges in using real-world data for health care, J. Clin. Invest. 130 (2020) 565–574.

[6] E.S. Russell, E. Aubrun, D.C. Moga, S. Guedes, W. Camelo Castillo, J.R. Hardy, et al., FDA draft guidance to improve clinical trial diversity: opportunities for pharmacoepidemiology, J Clin Transl Sci. 7 (2023) e101.

[7] R.J. LoCasale, C.L. Pashos, B. Gutierrez, N.A. Dreyer, T. Collins, A. Calleja, et al., Bridging the gap between RCTs and RWE through endpoint selection, Ther Innov Regul Sci. 55 (2021) 90–96.

[8] E.A. Akl, M. Briel, J.J. You, X. Sun, B.C. Johnston, J.W. Busse, et al., Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review, BMJ 344 (2012) e2809.

[9] FDA-NIH Biomarker Working Group, BEST (Biomarkers, EndpointS, and Other Tools) Resource, Silver Spring, MD, 2016.

[10] S. Ventz, L. Comment, B. Louv, R. Rahman, P.Y. Wen, B.M. Alexander, et al., The use of external control data for predictions and futility interim analyses in clinical trials, Neuro Oncol. 24 (2022) 247–256.

[11] K. Inoue, W. Hsu, O.A. Arah, A.E. Prosper, D.R. Aberle, A.A.T. Bui, Generalizability and transportability of the national lung screening trial data: extending trial results to different populations, Cancer Epidemiol. Biomarkers Prev. 30 (2021) 2227–2234.

[12] E. Bareinboim, J. Pearl, Causal inference and the data-fusion problem, Proc. Natl. Acad. Sci. U.S.A. 113 (2016) 7345–7352.

[13] U.S. Department of Health and Human Services, Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products: guidance for industry. https://www.fda.gov/media/152503/download, 2024.

[14] E.J. Weinstein, M.E. Ritchey, V. Lo Re 3rd, Core concepts in pharmacoepidemiology: validation of health outcomes of interest within real-world healthcare databases, Pharmacoepidemiol. Drug Saf. 32 (2023) 1–8.

[15] L.J. Collin, R.F. MacLehose, T.P. Ahern, R. Nash, D. Getahun, D. Roblin, et al., Adaptive validation design: a bayesian approach to validation substudy design with prospective data collection, Epidemiology 31 (2020) 509–516.

[16] T.L. Lash, M.P. Fox, R.F. MacLehose, G. Maldonado, L.C. McCandless, S. Greenland, Good practices for quantitative bias analysis, Int. J. Epidemiol. 43 (2014) 1969–1985.

[17] S.R. Newcomer, S. Xu, M. Kulldorff, M.F. Daley, B. Fireman, J.M. Glanz, A primer on quantitative bias analysis with positive predictive values in research using electronic health data, J. Am. Med. Inf. Assoc. 26 (2019) 1664–1674.

[18] L.H. Smith, M.B. Mathur, T.J. VanderWeele, Multiple-bias sensitivity analysis using bounds, Epidemiology 32 (2021) 625–634.

[19] J.M. Franklin, K.L. Liaw, S. Iyasu, C.W. Critchlow, N.A. Dreyer, Real-world evidence to support regulatory decision making: new or expanded medical product indications, Pharmacoepidemiol. Drug Saf. 30 (2021) 685–693.

[20] S.T. Liaw, J.G.N. Guo, S. Ansari, J. Jonnagaddala, M.A. Godinho, A.J. Borelli, et al., Quality assessment of real-world data repositories across the data life cycle: a literature review, J. Am. Med. Inf. Assoc. 28 (2021) 1591–1599.

[21] N.G. Weiskopf, G. Hripcsak, S. Swaminathan, C. Weng, Defining and measuring completeness of electronic health records for secondary use, J. Biomed. Inf. 46 (2013) 830–836.

[22] National Institute for Health and Care Excellence, NICE real-world evidence framework. https://www.nice.org.uk/corporate/ecd9, 2022.

[23] N.M. Gatto, S.E. Vititoe, E. Rubinstein, R.F. Reynolds, U.B. Campbell, A structured process to identify fit-for-purpose study design and data to generate valid and transparent real-world evidence for regulatory uses, Clin. Pharmacol. Ther. 113 (2023) 1235–1239.

[24] U.S. Department of Health and Human Services, Considerations for the use of real-world data and real-world evidence to support regulatory decision-making for drug and biological products: guidance for industry. https://www.fda.gov/media/171667/download, 2023.

[25] S.V. Wang, S. Pinheiro, W. Hua, P. Arlett, Y. Uyama, J.A. Berlin, et al., STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies, BMJ 372 (2021) m4856.

[26] J. Nagels, S. Wu, V. Gorokhova, Deterministic vs. Probabilistic: best practices for patient matching based on a comparison of two implementations, J. Digit. Imag. 32 (2019) 919–924.

[27] E.V. Bernstam, R.J. Applegate, A. Yu, D. Chaudhari, T. Liu, A. Coda, et al., Real-world matching performance of deidentified record-linking tokens, Appl. Clin. Inf. 13 (2022) 865–873.

[28] L.B. Mirel, D.M. Resnick, J. Aram, C.S. Cox, A methodological assessment of privacy preserving record linkage using survey and administrative data, Stat. J. IAOS 38 (2022) 413–421.

[29] U. Tachinardi, S.J. Grannis, S.G. Michael, L. Misquitta, J. Dahlin, U. Sheikh, et al., Privacy-preserving record linkage across disparate institutions and datasets to enable a learning health system: the national COVID cohort collaborative (N3C) experience, Learn Health Syst 8 (2024) e10404.

[30] S.L. Huynh T, J. Leshin, T. Haskell, Assessment of the relationship between collision rate and sample size using a large US mortality dataset, Value Health 25 (2022) S206.

[31] Frederick National Laboratory for Cancer Research, Evaluating the performance of privacy preserving record linkage systems (PPRLS). https://surveillance.cancer.gov/reports/TO-P2-PPRLS-Evaluation-Report.pdf, 2023.

[32] J.T. Lariscy, Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox, J. Aging Health 23 (2011) 1263–1284.

[33] LabCorp Launches New COVID-19 Clinical Trial Site to Connect Patients with U.S. Research Trials, 2020. https://www.businesswire.com/news/home/20200601005419/en/LabCorp-Launches-New-COVID-19-Clinical-Trial-Site-to-Connect-Patients-With-U.S.-Research-Trials.

[34] R. Liu, S. Rizzo, S. Whipple, N. Pal, A.L. Pineda, M. Lu, et al., Evaluating eligibility criteria of oncology trials using real-world data and AI, Nature 592 (2021) 629–633.