ChatGPT

Using Large Language Models to Simplify Real-World Evidence Research

Introduction

Real-world data (RWD) – health information collected outside of controlled trials (e.g. electronic health records, insurance claims, registries, wearables) – has become an invaluable resource for generating real-world evidence (RWE) on treatment effectiveness, safety, and outcomes [1]. However, extracting meaningful insights from RWD is challenging due to its volume and complexity. Advances in artificial intelligence (AI), especially **large language models (LLMs)**, offer powerful new tools to analyze these vast datasets. LLMs are generative AI systems (like GPT-4) trained on massive text corpora, enabling them to understand natural language and even generate human-like text. Crucially, LLMs can **process and curate enormous amounts of RWD at unprecedented speed and scale**, far beyond what manual analysis can achieve [2]. By doing so, they help researchers find patterns and relationships in healthcare data that might not be readily apparent to human analysts [3]. In turn, **LLM-driven analysis combined with human expertise can unlock critical insights**, providing a more comprehensive view of patient experiences and outcomes [4].

In this article, we explore how LLMs are enhancing RWE research and streamlining key tasks. We discuss practical applications – from literature reviews to clinical data mining – and examine how LLM-assisted workflows can make evidence generation more efficient. We also highlight important considerations for using AI responsibly in healthcare research, including data quality, ethical safeguards, and regulatory guidelines. Finally, we look ahead at future trends, such as multimodal AI integration and evolving standards, to understand how LLMs may further transform RWE generation.

Applications of LLMs in RWE Research

LLMs can augment many stages of the RWE research process. Key application areas include:

- **Literature Review and Evidence Synthesis:** Researchers can leverage LLMs to **rapidly review and summarize medical literature**, making the evidence synthesis process more efficient. For example, early studies of LLM-assisted systematic reviews indicate that LLMs can aid in screening studies and drafting summaries [5]. An LLM can quickly sift through hundreds of trial reports or real-world study publications, extract key findings, and even compile them into a coherent overview. This accelerates the traditionally labor-intensive process of literature review. (That said, human experts must still verify the accuracy of AI-generated summaries – a caution we will return to later.)

- **Clinical Data Extraction and Chart Review:** A large portion of RWD comes from unstructured clinical documentation, such as physician notes, discharge summaries, and pathology reports. **LLMs excel at natural language processing (NLP)** on these texts, helping to extract valuable information and insights. In practice, AI models have dramatically improved phenotyping – identifying patients with certain characteristics or outcomes – from EHRs. For instance, an NLP system matched with clinical expertise achieved about *94% accuracy in extracting key heart failure concepts from EHR notes, versus only 49% via manual review* [6]. This illustrates how LLMs (a form

of deep NLP) can efficiently comb through free-text records to find symptoms, comorbidities, or outcomes that would be tedious to detect by hand. Overall, LLMs can summarize patient histories, identify mention of adverse events, and flag relevant details in clinical narratives, thereby reducing the burden of manual chart review.

- **Patient Phenotyping and Cohort Identification:** Beyond individual records, LLMs can help researchers **identify cohorts of patients meeting specific criteria** across large databases. By parsing unstructured text fields alongside structured data, LLM-driven tools uncover patients who would be missed if one relied only on coded entries. For example, in an ophthalmology RWE study, relying solely on diagnosis codes identified about 330,000 patients with a certain condition, but by having AI analyze free-text ophthalmologist notes, researchers found an *additional 476,000 patients* – expanding the cohort to over 810,000 [7]. In other words, **LLMs can recognize condition mentions and clinical clues in narrative text (e.g. descriptions of symptoms or exam findings) that standard database queries overlook**, dramatically improving cohort capture. This capability is extremely valuable for real-world studies, where under-coding and variability in documentation often obscure important patient populations.

- **Data Interpretation and Insight Generation:** LLMs can also assist in **making sense of complex real-world data outputs**. After data are collected and analyzed, an LLM can be used to interpret statistical results or translate analytical findings into plain language. For instance, an LLM might summarize the outcomes of an RWE study (e.g. "In a dataset of 10,000 patients, treatment X was associated with a 20% lower hospitalization rate compared to treatment Y") and highlight key points for researchers or decision-makers. Similarly, LLMs can cross-reference multiple data sources and help generate hypotheses – such as suggesting why a certain patient subgroup responds differently to a therapy – by drawing on biomedical knowledge. This natural language generation ability allows LLMs to draft segments of study reports or create narratives explaining data patterns. By automating some of the interpretation and reporting tasks, LLMs free up researchers to focus on higher-level analysis and decision-making.

Interactive, LLM-assisted workflows are emerging across these applications. For example, a researcher might query an LLM with a prompt like "Summarize the key efficacy outcomes of these five real-world studies on Drug A" and get a coherent summary in seconds. Or an analyst could feed the LLM a de-identified patient record and ask "Does this patient meet the criteria for our registry cohort?" to quickly check eligibility. By intelligently handling language and context, LLMs serve as powerful assistants throughout the RWE research cycle – from gathering evidence to analyzing data to communicating findings.

Real-World Examples of LLM-Enhanced RWE

LLM-driven curation of unstructured clinical notes can capture implicit evidence that traditional methods miss. For example, analyzing free-text physician notes for clues of metastasis (e.g. "spots on bone scan") enabled one study to identify five times more metastatic cancer cases than were recorded through structured staging codes alone [8]. The AI model recognized references to tumor spread that doctors documented in narrative form but didn't formally code, greatly increasing the detected incidence of metastasis. In addition to such outcomes, LLM/NLP techniques can extract other key clinical variables (like tumor grades, lab values, or risk factors) buried in text, enriching the data available for RWE analyses [9]. This deeper capture of information from notes improves risk stratification and tracking of disease progression across patient populations. [8]

Similarly, tapping into unstructured EHR data can dramatically expand patient cohorts for real-world studies. In one case, researchers initially identified about 330,000 patients with a certain eye disease using standard

diagnosis codes. After deploying NLP/LLM methods to scour ophthalmologist notes for mentions of that disease, an additional 476,000 patients were found – patients who had the condition documented in free text but not coded formally [7]. This brought the total cohort to over 810,000 patients. The figure above illustrates how combining structured data (dark portion) with unstructured note mining (light portion) uncovered a much larger patient population. By leveraging LLMs to read clinical notes, RWE studies can capture the "hidden" patients and ensure analyses reflect the full real-world patient spectrum. In this example, the AI also utilized imaging data (scans and photos) alongside text, demonstrating a multimodal approach to identifying disease progression [10]. [7] [11]

These examples highlight the tangible impact of LLMs in RWE generation. In both cases, **human experts played a crucial role** alongside the AI: domain specialists helped define what textual cues to look for (e.g. phrases indicating metastasis or diagnostic hints in notes), and they validated the AI's outputs for accuracy [12]. The end result is a more robust evidence base – greater capture of relevant events and patients – than either humans or algorithms could achieve alone. By integrating LLM-driven automation with expert oversight, researchers can significantly enhance the completeness and depth of real-world evidence.

Ensuring Responsible Use of LLMs in RWE (Quality, Ethics & Compliance)

While LLMs offer powerful advantages, their use in healthcare research **must be approached cautiously and thoughtfully**. RWD itself can be messy – it is often noisy, sparse, and inconsistently collected [13]. Naively applying AI to such data can **amplify biases or errors** present in the data [13]. Therefore, **data quality and preparation are paramount**. Researchers should carefully preprocess RWD (for example, handling missing values, standardizing terminologies, and removing identifiers) before feeding it to LLMs [14]. It's also essential to incorporate medical domain knowledge into the AI workflow: expert input helps the model focus on meaningful signals and avoid "learning" spurious patterns. In fact, expert consensus holds that *special precautions are needed when using AI on RWD – including rigorous validation on real-world cohorts, algorithmic transparency, and integration of clinical knowledge – to avoid potentially harmful conclusions* [15]. In practice, this means any insights an LLM produces should be cross-checked against clinical reasoning and established evidence. **Human-in-the-loop validation is critical**: LLM outputs (whether a literature summary or a cohort identification) must be reviewed by researchers or clinicians for accuracy and plausibility before being trusted. Remember that LLMs, for all their fluency, can sometimes **"hallucinate" incorrect facts or misinterpret context**, especially in zero-shot settings. For example, one evaluation found that LLM-generated summaries of medical evidence sometimes contained misinterpretations or omissions that **could lead to medical harm if unchecked** [16]. Thus, maintaining a **strong quality control process with human oversight** is non-negotiable when using LLMs for RWE.

Ethical and privacy considerations are equally important. RWE often involves sensitive patient information from real-world settings, so **protecting patient privacy and data security** is a core requirement. Any use of LLMs on clinical data must comply with regulations like HIPAA in the US, GDPR in Europe, and other applicable data protection laws. Typically, this involves de-identifying patient data before analysis (e.g. removing names, addresses, and any direct identifiers) [17]. In situations where LLMs or AI tools need to learn from multiple data sources, organizations are exploring privacy-preserving techniques such as **federated learning** (where the AI model is trained across decentralized data without raw data leaving secure servers) and synthetic data generation (using simulated patient data that mirrors real data statistics without exposing real individuals) [18]. **Transparency** is another ethical imperative. Stakeholders (from researchers to regulators) should be informed when an analysis or piece of evidence was produced with AI assistance. This transparency builds trust and allows proper

scrutiny of the methods used. To that end, documentation of the LLM's role, parameters, and performance should be part of reporting an AI-assisted RWE study. Recent initiatives are beginning to formalize this; for example, the *AI-in-RWE Transparency (AIRT) Checklist* defines standards for reporting **model training, validation, explainability, and bias mitigation** in publications involving AI-generated RWE [19] [20]. Such frameworks guide researchers to clearly describe how an LLM was applied, how its outputs were verified, and what steps were taken to ensure fairness and reliability.

Crucially, regulators and oversight bodies are actively developing guidelines for AI in evidence generation. Health agencies want to ensure that when AI-informed RWE is submitted (for drug approvals, device safety monitoring, health technology assessments, etc.), it meets high standards of credibility. For example, **NICE (the UK's health authority)** has stated that any AI methods used in evidence generation should come with *demonstrable value and be balanced against risks like algorithmic bias or lack of transparency* [21]. They emphasize that AI **should augment, not replace, human judgment**, insisting on an informed human-in-the-loop for critical decision points [22]. Likewise, the FDA and EMA are examining how to integrate AI-derived insights while upholding scientific rigor and patient safety. The overarching theme is that **responsible use of LLMs in RWE requires rigorous validation, ethical safeguards, and compliance with evolving standards**. By adhering to these principles – data quality control, bias mitigation, privacy protection, transparency, and oversight – researchers can harness LLMs productively without compromising on trust or patient welfare.

Future Outlook: LLMs, Multimodal Integration, and Evolving Standards

The intersection of LLMs and RWE research is a fast-moving frontier. Looking ahead, we can expect several trends that will shape how these AI tools are integrated into healthcare studies:

- **Multimodal Data Integration:** Future AI models will not be limited to text; we are already seeing the rise of **multimodal LLMs** that can process and combine various data types (text, images, signals, etc.) for a more holistic analysis. In healthcare, this means an LLM might take in clinical notes *and* medical images, or genomic data, to provide richer insights. For example, researchers have begun integrating radiology images with text reports so that an AI can both describe imaging findings and correlate them with clinical history [23]. Similarly, as noted earlier, combining ophthalmic images with text-based patient records has enhanced detection of disease progression [11]. This multimodal approach can improve accuracy and context – e.g. an AI could flag a finding on an X-ray and simultaneously summarize the patient's symptoms and lab results that relate to that finding [24]. Such capabilities are in early stages but **hold great promise for more comprehensive real-world evidence**, as many health questions span multiple data sources. We may soon see RWE studies where an LLM-based system analyzes everything from patient interview transcripts to wearable sensor readings in one unified framework.

- **Domain-Specialized LLMs:** Another likely trend is the development of **LLMs tailored to biomedical and real-world data domains**. Current off-the-shelf LLMs (like general GPT models) are trained on broad internet text, which may not include the nuances of clinical language or the specifics of healthcare data. Companies and research groups are beginning to train or fine-tune LLMs on medical text (such as clinical notes corpora, biomedical literature, and RWD datasets) to create models that speak the language of healthcare more fluently. By incorporating domain-specific knowledge and terminology, these specialized models can produce more relevant and accurate results for RWE tasks [25]. For instance, a pharma-focused LLM might better understand drug names, trial endpoints, or adverse event terminology, thus reducing misinterpretations. We

can expect the gap between general AI and domain-focused AI to narrow as *"medicine-aware"* LLMs become available, making AI assistance even more effective for real-world evidence generation.

- **Regulatory Evolution and Best Practices:** As LLM applications mature, **regulatory and governance frameworks will continue to evolve**. Regulators are already collaborating with academia and industry to formulate guidelines for AI in healthcare research. We will likely see more consensus on questions like: What documentation is required if an analysis used an LLM? How should AI outputs be validated for regulatory submissions? What level of transparency about the model and data is necessary for an RWE study to be deemed "regulatory-grade" [26]? Initiatives like the European Union's proposed AI Act are aiming to set boundaries on high-risk AI systems (which would include many healthcare applications) [26]. In parallel, industry standards and checklists (such as AIRT) will gain traction in ensuring AI-enhanced studies are reported with clarity and rigor [20]. All these efforts point toward an environment where **LLM-assisted evidence generation is accepted, even expected, provided it adheres to agreed-upon quality and transparency norms**. Researchers who stay abreast of these guidelines and embed them in their workflows will be well-positioned to leverage AI while meeting all necessary compliance.

- **Continuous Learning and Collaboration:** Finally, the future will see RWE researchers, clinicians, data scientists, and AI systems working in closer partnership. The most successful applications of LLMs so far have come when AI is used to augment human expertise – not replace it. This trend will continue, with LLMs handling the heavy lifting of data processing and initial analysis, and humans providing direction, interpretation, and critical judgement. As LLM capabilities improve, researchers will need to continually update their own skills in using these tools (for example, mastering how to craft effective prompts, or how to fine-tune models on their data). There is also growing interest in **interactive AI systems** that can explain their reasoning or uncertainty, which could further build user trust. In the coming years, we may interact with LLMs almost like colleagues – asking them "why do you think this outcome happened?" and getting an evidence-backed explanation. Such developments would make the research process more iterative and insightful. The ultimate goal is that **LLMs become an integral, trusted part of the RWE research toolbox**, enabling teams to derive insights faster and more comprehensively than ever, while always under the guiding hand of human expertise.

In conclusion, using LLMs to simplify and enhance real-world evidence research is a game-changer for healthcare and drug development. These models can **dramatically streamline literature reviews, extract hidden gems from clinical data, and help interpret complex findings**, thus accelerating the journey from raw data to actionable evidence. By harnessing LLMs' strengths – and doing so responsibly – researchers can focus on asking the right questions and making informed decisions, rather than getting bogged down in data drudgery. Early successes in applying LLMs to RWE show improved efficiency and depth of analysis, from finding more patients and outcomes to generating deeper insights into patient journeys [27]. At the same time, we must heed the lessons learned: ensure data quality, maintain transparency, involve human experts, and uphold ethical standards. With appropriate guardrails in place, LLMs are poised to become indispensable allies in real-world evidence generation. As one industry expert observed, *by harnessing the power of LLMs together with human expertise, we can unlock more comprehensive insights into treatment outcomes and patient experiences, ultimately improving care* [27]. The coming years will undoubtedly bring even more innovation at this intersection of AI and RWE – and those prepared to integrate these advancements into their research strategies will lead the way in translating real-world data into real-world improvements in health.

Sources: The information above is drawn from a range of current publications and expert analyses. Key references include industry case studies of LLMs in healthcare [28] [7], research on AI-driven RWD analytics [1] [14], and emerging guidelines on ethical AI use in evidence generation [21] [20], among others. Each citation in the text corresponds to the source material for verification and further reading.

---

[1] [6] [13] [14] [15] [18] AI in RWE Studies: Applications, Challenges & Impact | IntuitionLabs
https://intuitionlabs.ai/articles/ai-in-rwe-rwd-studies

[2] [4] [7] [8] [9] [10] [11] [12] [17] [23] [24] [25] [27] [28] What is the Role of Large Language Models in Real-World Evidence Generation? | veranahealth.com
https://veranahealth.com/what-is-the-role-of-large-language-models-in-real-world-evidence-generation/

[3] [21] [22] Use of AI in evidence generation: NICE position statement | NICE
https://www.nice.org.uk/position-statements/use-of-ai-in-evidence-generation-nice-position-statement

[5] The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review - PubMed
https://pubmed.ncbi.nlm.nih.gov/40332983/

[16] Evaluating Large Language Models on Medical Evidence Summarization | Population Health Sciences
https://phs.weill.cornell.edu/news/evaluating-large-language-models-medical-evidence-summarization

[19] [20] [26] ISPOR - The AI-in-RWE Transparency (AIRT) Checklist: Essential and Desirable Standards for AI-Enhanced Real-World Evidence
https://www.ispor.org/heor-resources/presentations-database/presentation-cti/ispor-europe-2025/poster-session-3-2/the-ai-in-rwe-transparency-airt-checklist-essential-and-desirable-standards-for-ai-enhanced-real-world-evidence