Responsible AI in UK Health Data

Setting Standards for Future Use of LLMs in Clinical Data Extraction

Authors: Arun Sujenthiran, Lucia Groizard, Melissa Estevez, Cornelius Thaiss, Natalia Viani, Kathi Seidl-Rathkopf

Contributors: Maria Alvarellos, Adam Manhi, Emma Salib, Amanda White

APRIL 2026

flatiron

flatiron

Executive Summary

Artificial intelligence (AI) is transforming the way researchers use electronic health records (EHRs) to generate real-world data (RWD) and real-world evidence (RWE) for clinical studies and scientific discovery. Large language models (LLMs) now make it possible to extract meaningful clinical information from unstructured EHR text at a scale and speed far beyond traditional manual abstraction. This capability has the potential to accelerate clinical research, support regulatory decisions, and improve patient care.

However, the data within EHRs are complex, inconsistently documented, and often ambiguous. Furthermore, LLMs can behave unpredictably, be sensitive to input variations, or reinforce biases present in source data. Without proper oversight, these challenges could undermine data quality, reduce the reliability of downstream analyses, and erode public and patient trust, especially when working with sensitive health data. While these risks are well recognised, the costs of limited adoption are often overlooked, as vast amounts of longitudinal patient data would otherwise remain inaccessible because manual curation does not scale. Therefore, responsible AI deployment represents not only a technical advance but an ethical imperative to maximise patient benefit, under-scoring the need for high-quality, transparent evaluation frameworks.

To address this gap, Flatiron Health has developed the Validation of Accuracy for LLM/ML-Extracted Information and Data (VALID) framework. This provides a structured, multi-dimensional approach to assessing the accuracy, reliability, and fitness-for-purpose of LLM-extracted clinical information. In the UK, Flatiron Health applies these principles within a robust governance approach built on standards that reflect a commitment to responsible innovation. By embedding structured, high-quality frameworks such as VALID in health research, the UK can set a global benchmark for trustworthy use of LLMs in healthcare to improve patient care and outcomes.

This white paper outlines how high-quality, well-governed health data is essential for safely developing and deploying LLMs in health research. It provides practical recommendations for ensuring that LLMs are adopted safely, responsibly, and to their full potential within the UK health data ecosystem.

Given the research potential of LLMs and the evolving regulatory requirements in the UK, our overriding recommendation is a more consistent and collaborative approach across UK system stakeholders, guiding the use of LLMs in clinical data extraction to maximise research benefits while maintaining patient and public trust.

flatiron

The Imperative for Responsible AI Adoption

Large-language Models and Artificial Intelligence

The growth in popularity of tools such as ChatGPT, have made the use of the terms “generative AI” and LLMs commonplace. But what is the relationship between something that can generate new text, and AI? The Central Digital and Data Office, provides a useful explanation for what these terms mean, and how they are all related (Central Digital and Data Office [CDDO], 2024).

Artificial intelligence (AI) is a broad term for a field of study which uses computers to emulate the products of human intelligence or to build capabilities which go beyond human intelligence. LLMs are a category of what is known as “generative AI”, which uses large quantities of data to train a model to learn the underlying patterns and structure of that data such that it is capable of generating new content. LLMs are trained on text and natural language data to predict the next word that sounds most plausible given prior text, based on large amounts of data (Figure 1) (Central Digital and Data Office [CDDO], 2024).

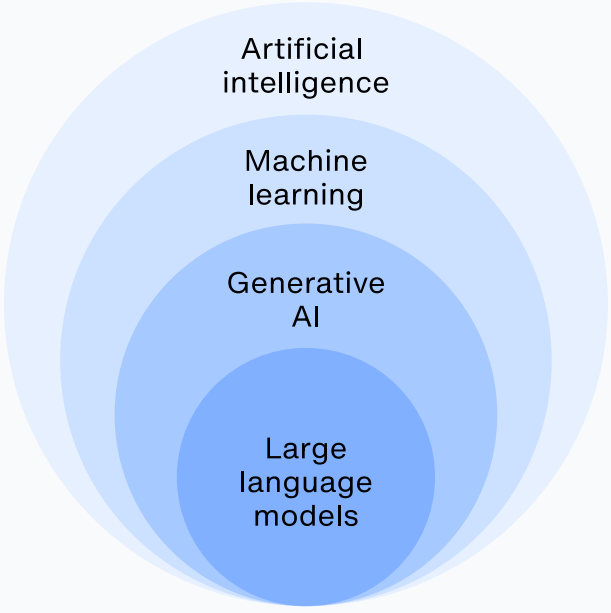

Figure 1. Generative AI and LLMs are sub-categories within the broader field of AI.

Benefits of AI and LLMs for EHR Data Extraction

With the capacity to rapidly extract insights from massive datasets, AI is enabling research at a scale and speed impossible with traditional data processing approaches. EHRs are an especially valuable source of such data because they contain rich, longitudinal clinical information. When transformed into real-world evidence (RWE) they can accelerate the generation of clinically actionable insights, potentially impacting the lives of millions of patients and healthcare providers (Estevez *et al*., 2025, Goldacre, 2022).

Many EHR fields contain structured data (e.g. diagnoses codes, demographic information), facilitating automated processes for data minimisation, cleaning, extraction and analysis. On the other hand, unstructured EHR data, such as clinician or pathology notes or imaging reports, by their nature are more difficult to extract and standardise at scale, historically requiring manual curation by experts and significantly decelerating the process of RWD transformation (Estevez *et al*., 2025).

Thanks to recent advances in machine learning (ML) and LLMs, researchers are increasingly turning to these technologies to extract clinical relevant data points from complex, unstructured documents in the EHR to support efforts to generate RWE rapidly and at scale (Adamson *et al*., 2023; Cohen *et al*., 2025; HDR UK, 2023). As an example, Flatiron Health researchers have successfully used ML models to extract key clinical information from unstructured EHR documents for cancer patients in the Flatiron Health anonymised database:

- **Scale**: Castleman Disease is a rare disease involving the lymph nodes with poor outcomes and limited treatment options, and can later develop into lymphoma. Building a sufficiently large cohort of patients to understand real-world treatment is a challenge, due to the limited number of patients as well as the effort required for manual chart abstraction (e.g. no ICD-10 code). Flatiron Health researchers utilised a natural language processing (NLP)-based ML model to allow for the largest analysis of real-world patients with Castleman's Disease to date. This model successfully identified ultra-rare needles in haystack populations like patients with Castleman's disease (e.g. clinical characteristics, treatment trends, and real-world overall survival). These data could ultimately inform and improve treatment of patients with this rare disease (Cohen *et al*., 2023).

- **Speed and depth**: There is growing evidence to support the prognostic and predictive relevance of circulating tumour (ct) DNA in patients with early stage colorectal cancer (CRC). To learn more about the real-world utilisation of tumour-informed, personalised ctDNA assays in patients diagnosed with stage I-III CRC and a commercial ctDNA test, Flatiron Health researchers used ML models to extract data on the uptake of ctDNA testing over time as well as how those results correlated with other biomarkers and the development of specific sites of metastasis. This is the largest study to date of early stage CRC ctDNA testing in routine practice with the potential to predict risk for liver metastasis and other outcomes. Future work could translate these findings into clinically actionable insights and treatment strategies (Fidyk *et al*., 2024).

Whilst both of these use cases applied traditional ML models to extract data from EHRs, useful learnings can be taken for LLM extraction approaches. The same features that make LLMs so useful (e.g. that they can process vast volumes of unstructured clinical text and generate structured variables at unprecedented speed) simultaneously introduce new complexities and risks (Estevez *et al*., 2025).

flatiron

Risks of AI and LLMs for EHR Data Extraction

LLMs are trained to predict the next word that sounds most plausible given prior text, based on large amounts of data, so while they excel at producing fluent, confident-sounding answers, they do not actually understand facts and are often lacking clinical domain knowledge. **In healthcare, this matters because sounding right is not the same as being right.** An LLM can generate medical language that looks authoritative while missing key context, misinterpreting data, or fabricating details (Central Digital and Data Office [CDDO], 2024). The risks of this could include poor data quality, impacting the reliability of downstream analyses and public and patient trust. Within the context described in this paper, LLMs are deployed as tools to assist in structuring unstructured electronic health record data for research use. They do not replace clinician judgement, are not embedded in live care pathways, and are not used to generate patient-facing outputs.

Data Quality Challenges and LLMs

Because data within EHRs are collected primarily for clinical care delivery and administrative purposes, they often contain missing or poorly formatted information, variable coding, free-text notes, and inconsistencies that complicate secondary analysis and the generation of reliable real-world data (RWD) (Goldacre, 2022). Furthermore, LLMs that are not specifically fit-for-purpose may also struggle with subjectivity and context-dependent findings in clinical documentation and, if not properly trained on clinical data, will be limited in their ability to navigate complex medical terminology, abbreviations, and diverse documentation styles (Estevez *et al*., 2025).

In the UK, the 2022 Goldacre Review emphasised that the safe, efficient, and transparent use of NHS health data is essential to maintaining public trust and ensuring research impact.

Given the sensitivity of health data and the potential impact on patient care if not used correctly, public trust and the reliability of subsequent research findings could be undermined if the use of LLMs in healthcare is not appropriately managed or governed. A robust evaluation framework is therefore essential to ensure LLM use with health data is accurate, ethical, and trustworthy (Goldacre, 2022; Adamson *et al*., 2023).

The Current Landscape of Frameworks for the Use of AI/ML/LLMs in Health Data Research

As AI systems become increasingly integrated into the curation and analysis of health data (Silberling, 2026), both researchers and regulatory bodies have recognised the need for rigorous, globally recognised frameworks to ensure their safe and responsible use, particularly when applied in a clinical care setting. In 2025, the U.S. Food and Drug Administration (FDA) issued a framework emphasising the importance of data provenance, model versioning, and continuous monitoring. Similarly, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) has published good practice reports and checklists, including the SUITABILITY Checklist, which provides recommendations for assessing the quality of RWD derived from EHRs (Estevez *et al*., 2025).

The European Union Artificial Intelligence Act provides comprehensive requirements for high-risk medical AI systems, including robust risk-mitigation processes, high-quality training datasets, transparent documentation, and human oversight (European Commission, 2025). In the UK, the Medicines and Healthcare products Regulatory Agency (MHRA) has outlined a data strategy aimed at embedding modern data technologies into regulatory pathways to ensure timely access to real-world data (RWD) and evidence (RWE) while maintaining standards strong enough to support decisions across the product lifecycle (Medicines and Healthcare Products Regulatory Agency [MHRA], 2024). In Japan, the Ministry of Health, Labour and Welfare (MHLW) has addressed these technological shifts by issuing the "Guidelines for the Utilization of Medical Digital Data for AI Research and Development" (2024). These guidelines primarily focus on identifying high-level ethical, legal, and social considerations, and aim to serve as a cornerstone for the appropriate and effective utilisation of health data in the face of the rapid evolution of AI technologies. Although welcome and necessary, these initiatives remain high-level, and none specifically address the unique challenges posed by LLMs, whose behaviour can be unstable, opaque, and highly sensitive to input variation (Estevez *et al*., 2025). **As a result, despite the growing use of LLMs to analyse unstructured EHR text, there is still no consensus or agreed-upon standard for validating their safety, reliability, or fitness for purpose in extracting fit-for-purpose RWD from clinical records.**

Building on a foundation of responsible data governance and research-ready oncology datasets, Flatiron Health has developed the Validation of Accuracy for LLM/ML-Extracted Information and Data (VALID) Framework. This framework provides a structured methodology to evaluate the accuracy of LLM-extracted clinical information, ensuring that the insights derived from unstructured EHR data meet rigorous quality standards (Estevez *et al*., 2025).

Flatiron Health's VALID Framework

Flatiron Health's VALID Framework offers a holistic evaluation strategy. The framework is built on three foundational pillars: Variable-level Performance Metrics, Verification Checks, and Replication and Benchmarking Analyses (Figure 2) (Estevez *et al.*, 2025).

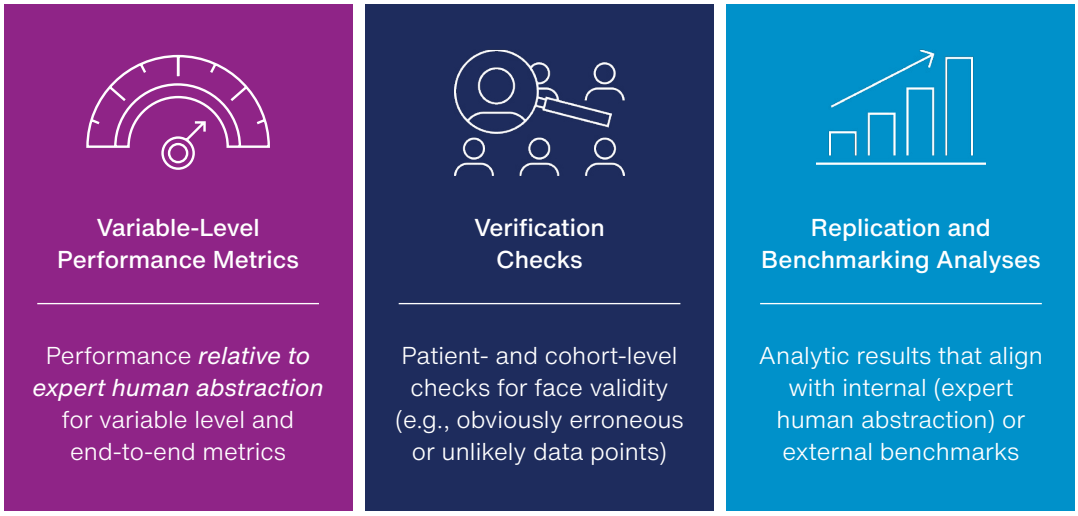

Figure 2.

- **Variable-Level Performance Metrics**: Variable-level performance metrics give the accuracy and completeness of a data curation approach by comparing it against a source of truth label. Standard measures of accuracy include recall (measures how good a model is at finding all the cases it's supposed to find), precision (measures how often the cases identified by the model are actually correct), F1 score (combines recall and precision into a single measure of a model's accuracy) and completeness (rates how much of the expected data is actually present in the dataset). Performance metrics are calculated using a testing dataset that the model has never encountered, is representative of the target population, and is large enough to ensure reliable results.

  Even with expert human review, some variation between reviewers is inevitable due to the inherent clinical complexity and ambiguity of documentation in EHRs. By evaluating differences in human abstraction, the metrics can tell whether low model performance is actually due to ambiguous data, rather than a problem with the model quality itself. Even if an LLM demonstrates good performance on individual variables, when multiple raw data elements are combined to derive

a variable, compounding errors can significantly reduce end-to-end accuracy. End-to-end evaluation metrics are therefore essential for complex derived variables such as line of therapy or biomarker status at key timepoints. The level of benchmarking required could vary by use case: high-risk applications may warrant a rigorously adjudicated reference standard, whereas lower-risk cases can adopt a risk-based approach. However, when high accuracy against reliable labels indicates strong human abstraction quality, single-abstracted reference data may be sufficient for variables with strong performance metrics, and in some cases human abstraction benchmarks may not be necessary at all.

Finally, extending the variable-level performance metrics and relative performance difference between LLMs and human abstractors against a common benchmark to stratified sub-groups can indicate whether there is model bias towards or against a sub-group and whether it is due to errors in the model or within the dataset itself. Model bias can be evaluated by testing whether AI-generated data reproduces well-established findings. For example, known survival differences between demographic groups should be reflected in the model's outputs when compared against reliable reference data, with results examined across relevant subgroups where sample sizes allow.

- **Verification Checks**: Verification checks examine patient-level data for conflicting or incorrect entries and ensure that overall cohort distributions match clinical expectations. These checks act as a proxy for accuracy, helping to confirm the dataset's reliability and usability. Even with high LLM extraction performance, small errors can affect usability and erode trust in the data. Verification checks identify likely model-generated inconsistencies, and addressing them improves the dataset's overall quality and confidence in its use. Stratified verification checks can also identify model bias and ensure that certain subgroups are not at risk for being disproportionately removed from analyses as well as for variables with rare classes and those with less prevalent demographic subgroups.

- **Replication and Benchmarking Analyses**: Replication and benchmarking analyses check whether results from LLM-extracted data match those from trusted reference datasets. References can include expert-curated internal data or well-established external datasets, like the Surveillance, Epidemiology, and End Results (SEER) programme and the National Cancer Registration and Analysis Service (NCRAS). Unlike checks that focus on a handful of data points, these analyses look at the whole dataset to answer research questions, helping to reveal how errors in multiple variables might interact and affect findings. This makes replication analyses important for ensuring that RWD are fit-for-purpose for research and regulatory use. Analyses can be done on broad patient groups (e.g. early vs. advanced disease) or on specific subgroups, such as those defined by a particular biomarker or treatment. Replication and benchmarking analyses can also assess model bias if they can replicate well-established stratifications in sub-groups (e.g. overall survival being lower in certain demographics).

With these three pillars, the VALID framework offers several key advantages. It promotes transparency by clearly documenting how performance is measured and how results are derived. It ensures reproducibility through standardised metrics, verification checks, and replication analyses, allowing assessments to be consistently repeated across datasets, models, or research teams. It evaluates a model's fitness-for-purpose by determining whether the extracted data are sufficiently reliable and accurate for their intended research or regulatory application, while identifying areas where human review or additional quality checks may be needed. Model development and evaluation are supported by secure development lifecycle practices, including access controls, audit logging, privacy-preserving error analysis techniques, and safeguards to mitigate risks such as data leakage, prompt manipulation, or unintended memorisation of sensitive information.

By providing transparent, reproducible, and context-aware evaluation of LLM-extracted data, the VALID framework ensures that insights derived from EHRs are accurate, reliable, and trustworthy. This combination of scientific assessment and ethical oversight builds confidence in the use of LLMs, ensuring that data-driven insights are both actionable and responsibly generated.

flatiron

Flatiron Health's Approach in the UK to Responsible Use of AI and Patient Data

Flatiron Health's mission to improve patient lives by learning from every person with cancer underpins a commitment to ethical, equitable use of health data, including ensuring that AI systems operate without causing harmful bias or discrimination (Flatiron Health, 2025; Ryals, 2025). In the UK, Flatiron Health partners with stakeholders across the healthcare ecosystem (e.g. NHS Trusts and Health Boards, the research community and patients) to convert routinely collected clinical data into research-ready insights in full compliance with standards set by the Health Research Authority (HRA)'s Confidentiality Advisory Group (CAG) and Research Ethics Committee (REC). This ensures that all access to NHS patient data is ethically approved, legally authorised, and conducted under strict governance frameworks that protect patient confidentiality and privacy in line with the UK's unique regulatory requirements. All processing activities are conducted in accordance with the principles of data minimisation and storage limitation under Article 5 UK GDPR. Extraction processes are designed to capture only data necessary for clearly defined research purposes. Intermediate datasets used for validation and quality assurance are subject to technical and organisational safeguards, including access controls and audit mechanisms.

Each NHS data partner receives:

- highly-curated data on their own patients on a regular basis to unlock a range of research, clinical and commercial benefits locally
- access to Flatiron's global anonymised research-ready datasets for research
- a proportionate share in the revenue Flatiron generates through its research partnerships across the life sciences sector.

In the UK, the use of NHS patient data for research purposes is conducted in accordance with the UK General Data Protection Regulation (UK GDPR), the Data Protection Act 2018, and applicable approvals from the Health Research Authority (HRA), including the Confidentiality Advisory Group (CAG) where relevant. Processing of special category health data is undertaken for scientific research purposes under Article 9(2)(j) UK GDPR, subject to appropriate safeguards, and in line with recognised ethical and governance frameworks.

Importantly, the LLM-enabled extraction processes described in this paper are used solely to support research data curation and are not deployed to make automated clinical decisions or solely automated decisions affecting individual patients under Article 22 UK GDPR.

Because patient involvement is a cornerstone of responsible health data research, Flatiron Health UK incorporates input from people with lived experience (including patients, survivors and carers) to guide data governance, research priorities, and policies around data access and use through its Patient Voices Panel (Flatiron Health UK, 2025a) and Research Transparency Panel (Flatiron Health UK, 2025b). Engaging people with lived experience ensures that research priorities, policies and governance frameworks reflect the needs, values, and preferences of those whose information is being used. This participation enhances trust and transparency, giving patients confidence that their data are handled ethically and securely, while also ensuring that research using Flatiron's RWD is always in the public and patient interest. Flatiron Health UK honours national opt-out mechanisms and embeds additional opt-out mechanisms with each NHS partner, ensuring patients can make an informed choice about the use of their health data for research. Flatiron Health UK also maintains a publicly accessible data access register which documents the organisations and use cases underpinning each access request for data.

As described, Flatiron has made significant investments in ML/AI in the US to improve the quality of data product offerings and enhance the data curation process. In the UK, Flatiron is leveraging and applying these learnings in line with its HRA approval and applicable UK law to expedite and enhance its data curation process, and increase the value delivered to research partners. Flatiron's investment and research is focused on a number of areas including pan-tumour extraction of oral therapies from unstructured clinical data and the development of an approach to enhancing the accuracy and reliability of ML-extracted data.

This commitment to the ethical use of health data and AI is backed by findings from the UK-based, independent charitable organisation, the Health Foundation. In a large survey, they reported broad-based support by the public and NHS staff for the responsible use of AI in healthcare and that "decision-making accuracy and transparency are critical enablers for public confidence in AI, and these issues should be important focuses for design, regulation and training – as well as engaging the public on the resulting trade-offs," (Thornton *et al.*, 2024).

Recommendations and Best Practices: LLMs and UK Health Data

Developing robust quality frameworks for the use of LLMs in UK health data extraction will require coordinated, cross-sector collaboration. Stakeholders across the health data ecosystem, including data custodians, researchers, industry innovators, policymakers, and patient and public representatives, must work together to implement frameworks that can be adapted to different contexts while maintaining consistent standards.

Ongoing collaboration will ensure these frameworks evolve alongside advances in technology and changes in policy, enabling them to remain relevant and effective. Cross-sector research partnerships will also accelerate methodological maturity and help ensure that AI-extracted real-world data are reliable for clinical, epidemiological, and regulatory use.

For example, for NHS organisations and those managing Trusted Research Environments or Secure Data Environments, implementing robust, multi-dimensional quality assessment frameworks will help ensure accuracy, protect data integrity, and maintain public trust. For industry, a multi-layered “safety by design” approach, which is embedded throughout the model development lifecycle, should include expert human-in-the-loop review, validation processes, and fail-safe mechanisms. This will ensure that AI innovation in UK healthcare is not only technically advanced, but clinically safe, ethically grounded, and worthy of public trust.

Our overriding recommendation is therefore to strengthen consistency and collaboration across the UK health data ecosystem to support the safe, responsible, and effective adoption of LLMs in clinical data extraction.

Conclusion

High-quality data is the foundation of trustworthy real-world evidence, and this remains true as the field enters the era of large language models. While LLMs offer transformative potential for scaling and accelerating the curation of clinical data, their benefits can only be realised when paired with rigorous safeguards that ensure accuracy, reproducibility, and fairness. **Flatiron Health's VALID framework provides a practical, transparent model for what "good" looks like in the quality assessment of LLM-extracted health data, bridging scientific, operational, and ethical expectations.**

Flatiron Health UK is committed to advancing UK standards while ensuring that patient and public trust remains central to the responsible use of health data. But no single organisation can define or uphold these expectations alone. Progress requires collaboration across critical stakeholders, including government, academia, industry and patient advocacy groups. Given its pioneering health system, strong governance structures, and established models of patient and public involvement, the UK has the opportunity to lead the world in setting shared, robust standards for the safe and responsible adoption of LLMs in healthcare, ensuring their promise translates into meaningful, equitable impact for patients.

Flatiron Health UK stands ready to work with all key stakeholders across the UK healthcare ecosystem to advance the UK's global position as a leader in responsible deployment of AI solutions in healthcare. Contact us to get involved or for further information.

References

1. Central Digital and Data Office. (2024, January). *Generative AI framework for HM Government. UK Government*. https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government

2. Japanese Ministry of Health, Labour and Welfare. (2024). *Guidelines for the Utilization of Medical Digital Data for AI Research and Development*. https://www.mhlw.go.jp/content/001310044.pdf

3. Estevez, M., Singh, N., Dyson, L., *et al*. (2025). *Ensuring reliability of curated EHR-derived data: The validation of accuracy for LLM/ML-extracted information and data (VALID) framework.* arXiv. https://ascopubs.org/doi/10.1200/CCI-25-00215

4. Goldacre, B. (2022). *Better, broader, safer: Using health data for research and analysis (A review commissioned by the Secretary of State for Health and Social Care)*. Department of Health and Social Care.

5. Adamson, B., Waskom, M., Blarre, A., *et al*. (2023). Approach to machine learning for extraction of real-world data variables from electronic health records. *Frontiers in Pharmacology*, 14, 1180962. https://doi.org/10.3389/fphar.2023.1180962

6. Cohen, A. B., Adamson, B., Larch, J. K., *et al*. (2025). Large language model extraction of PD-L1 biomarker testing details from electronic health records. *AI in Precision Oncology*, 2(2), 57–64.

7. HDR UK. (2023). *CogStack information retrieval and extraction platform gives access to underused data.* https://www.hdruk.ac.uk/case-studies/cogstack-information-retrieval-and-extraction-platform-2/

8. Cohen, A.B., Estevez, M, Kelly, J.G., *et al*. *Clinical Characteristics, Treatment Trends, and Outcomes of Patients with HHV-8-Negative/Idiopathic Multicentric Castleman Disease Treated with Siltuximab in a Machine Learning-Selected Real-World Cohort*. Blood. 142 (2023) 907-909. American Society of Hematology. https://doi.org/10.1182/blood-2023-500179

9. Fidyk, E., Kalesinskas, L., Krismer, K., *et al*. (2024). Real-world ctDNA testing patterns, associated biomarkers, and sites of metastasis in early stage colorectal cancer [Abstract 3610]. Journal of Clinical Oncology, 42(16_suppl), 3610. American Society of Clinical Oncology. https://doi.org/10.1200/JCO.2024.42.16_suppl.3610

10. Silberling, A. (January 2026) *Anthropic announces Claude for Healthcare following OpenAI's ChatGPT Health reveal*. TechCrunch. https://techcrunch.com/2026/01/12/anthropic-announces-claude-for-healthcare-following-openais-chatgpt-health-reveal/

11. Estevez, M. (September 2025) *Can You Trust Real-World Data Extracted by a Large Language Model (LLM)?* Flatiron Health. https://resources.flatiron.com/real-world-evidence/can-you-trust-real-world-data-extracted-by-a-large-language-model-llm

References cont.

12. Ryals, CA. (July 2025). *Responsible AI in action: Flatiron's approach to AI fairness and bias assessment*. Flatiron Health. https://resources.flatiron.com/real-world-evidence/responsible-ai-in-action-flatirons-approach-to-ai-fairness-and-bias-assessment

13. Flatiron Health. (2025). *About us*. https://flatiron.com/about-us

14. European Commission. (2025). *Artificial intelligence in healthcare*. https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en

15. Medicines and Healthcare Products Regulatory Agency. (2024). *MHRA data strategy 2024–2027*. https://www.gov.uk/government/publications/mhra-data-strategy-2024-2027/mhra-data-strategy-2024-2027

16. Thornton, N., Binesmael, A., Horton, T., & Hardie, T. (July 2024). *AI in health care: What do the public and NHS staff think? The Health Foundation*. https://www.health.org.uk/reports-and-analysis/analysis/ai-in-health-care-what-do-the-public-and-nhs-staff-think

17. Seidl-Rathkopf, K., Schwarz, A., Viani, N., *et al*. (2025). *A framework for evaluating performance of LLM-based extraction from the electronic health record across different healthcare systems*. ESMO AI & Digital Oncology.

18. Groizard L, Dolezalova N, Kushnir M, *et al*. *Privacy-preserving error analysis loop For ML-based extraction of oncology EHR data*. https://resources.flatiron.com/publications/privacy-preserving-error-analysis-loop-for-ml-based-extraction-of-oncology-ehr-data

19. Flatiron Health UK. (2025a). *Putting patients' voices at the heart of everything we do*. https://flatironhealth.co.uk/patients-and-public

20. Flatiron Health UK. (2025b). *Robust health data governance*. https://flatironhealth.co.uk/data-governance

21. Flatiron Health UK. (2025c). *Patient data privacy notice*. https://flatironhealth.co.uk/legal-privacy/patient-data-privacy-notice/


flatiron