Original Paper

Development and Evaluation of a Human-in-the-Loop Data Curation Training Program to Support a Digital Clinical Trial Platform: Descriptive Feasibility Study

Mirna Issa[1], BS; Olivia Yang[2], BS; Jeffrey Doeve[3], BS; Shreya Arcot[1], BS; Amanda Chang[2], BS; Shriya Amara[4], BS; Shiven Himanshu Bhakta[1], BS; Iman Baber[5], BS; Kawan Shali[3], BS; Qaiss Dweik[2], BS; Tony Kin Wai Hung[6], MD, MBA, MSCR, FACP, FAMIA

[1]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA, United States

[2]Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, CA, United States

[3]Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA, United States

[4]Department of Neuroscience, University of California, Los Angeles, Los Angeles, CA, United States

[5]Institute for Society and Genetics, University of California, Los Angeles, Los Angeles, CA, United States

[6]Hartford HealthCare Cancer Institute, Hartford, CT, United States

Corresponding Author:

Tony Kin Wai Hung, MD, MBA, MSCR, FACP, FAMIA
Hartford HealthCare Cancer Institute
85 Retreat Ave
Hartford, CT 06106
United States
Phone: 1 860-972-4183
Email: Tony.Hung@hhchealth.org

Abstract

Background: LookUpTrials is a clinician-facing digital platform designed to support point-of-care navigation of institution-specific oncology clinical trials, incorporating artificial intelligence–assisted summarization and search functionalities. While embedding curated trial knowledge tools in community oncology workflows is feasible, the human infrastructure required to sustain high-quality, up-to-date trial data over time remains underexplored.

Objective: This study aimed to evaluate the feasibility of a structured training program embedding trainees within a supervised human-in-the-loop workflow for oncology clinical trial data curation in support of the LookUpTrials digital platform.

Methods: We conducted a descriptive feasibility evaluation of a cohort-based training and workflow model. A total of 10 undergraduate trainees curated publicly available trial information from institutional portals and ClinicalTrials.gov across 5 academic medical centers. Trial throughput and participant experiences were summarized descriptively.

Results: Over 10 months, trainees curated 2503 oncology clinical trial entries across 5 institutions, with processing rates increasing over time. Participants reported that structured onboarding and peer support facilitated engagement with data curation workflows.

Conclusions: Standardized human-in-the-loop workflows for clinical trial data stewardship can be implemented with individuals without prior clinical trials or informatics experience when supported by structured training and quality assurance. This study complements prior feasibility work on embedding trial knowledge tools in community oncology settings by focusing on the sustainability of the underlying content pipeline. Further evaluation is needed to assess scalability across institutions and durability over longer periods of platform maturation.

Introduction

Access to timely, accurate, and institution-specific information about available oncology clinical trials remains a persistent operational challenge in routine cancer care [1-5]. Trial information is often fragmented across registries, institutional portals, and protocol documents, limiting clinicians' ability to efficiently surface relevant studies during time-constrained visits [2,3,5,6]. Digital clinical trial platforms have emerged to aggregate and present structured trial information at the point of care, but their real-world effectiveness depends not only on technical design and usability but also on the availability of high-quality, up-to-date trial data [7-12].

LookUpTrials is a clinician-facing digital platform designed to surface institution-specific oncology clinical trial information at the point of care. The platform integrates artificial intelligence (AI)–assisted protocol summarization and search functionalities to facilitate trial navigation in routine clinical workflows. Prior work has demonstrated the feasibility of embedding LookUpTrials within community oncology workflows [13,14]. However, sustaining such platforms requires ongoing data stewardship to ensure accuracy, completeness, and local operational relevance, including site-specific trial availability, investigator contacts, and recruitment status [2,15,16]. Scalable workforce models to support this data stewardship layer remain underexplored [2,16].

Human-in-the-loop approaches are increasingly recognized as essential complements to automated pipelines and registry-based feeds, which often lack the granularity and local context required for operational trial navigation [17-19]. In this study, undergraduate trainees were selected as a pragmatic test case to evaluate whether individuals without formal clinical or informatics training can be onboarded into standardized data curation workflows with appropriate supervision and quality assurance. This cohort provides a controlled environment to prototype training materials, workflow design, and quality control processes that could be adapted to more durable staffing models, such as research coordinators, centralized trial operations teams, or hybrid human-AI review pipelines.

We therefore evaluated the feasibility of a structured undergraduate training program embedding trainees within a supervised human-in-the-loop workflow for oncology clinical trial data curation supporting LookUpTrials. This work complements prior feasibility studies of embedding trial knowledge tools in community oncology by focusing on the sustainability of the underlying content pipeline rather than downstream clinical adoption or enrollment outcomes [20, 21].

Methods

Study Design and Setting

We conducted a feasibility evaluation of a structured training program embedding trainees within a supervised human-in-the-loop workflow for oncology clinical trial data curation in support of the LookUpTrials digital platform (Figure 1). The primary objective was to assess whether individuals without prior clinical trials or informatics experience could be onboarded into standardized data stewardship workflows with appropriate supervision and quality assurance. The evaluation focused on workflow feasibility and participant experience rather than clinical adoption or patient-level outcomes.

Figure 1. Overview of the undergraduate training and data curation workflow supporting the LookUpTrials platform. The figure illustrates the sequential processes through which undergraduate participants are recruited, trained, and engaged in oncology clinical trial data curation, including data abstraction, peer-led quality assurance, faculty review, and integration into the LookUpTrials platform.
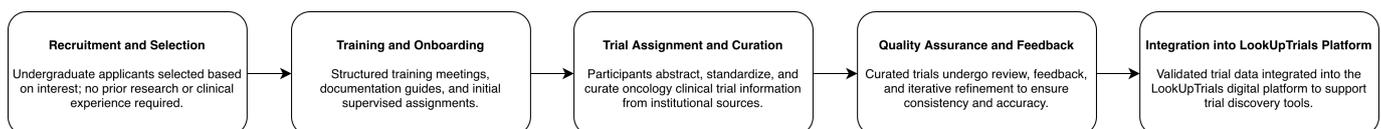
Participants

A total of 10 undergraduate trainees from a university-based student organization were recruited through a competitive application process based on interest in clinical research, availability for longitudinal participation, and commitment to a defined weekly time allocation. Participants had heterogeneous academic backgrounds and varying levels of prior research experience, reflecting the study's intent to assess the trainability of standardized workflows among individuals without formal clinical or informatics training.

Training and Workflow Design

Trainees underwent structured onboarding that included orientation to oncology clinical trial concepts, navigation of public trial registries and institutional trial portals, and use of standardized data abstraction templates. Training materials included written documentation, example entries, and step-by-step workflow guides. Trainees were embedded within a supervised, multitiered human-in-the-loop workflow that incorporated peer review and faculty oversight to ensure data accuracy, consistency, and completeness. Iterative updates to training materials and workflow documentation were made based on trainee feedback and observed sources of error.

Data Sources and Curation Process

Trainees curated publicly available oncology clinical trial information from institutional clinical trial portals and ClinicalTrials.gov across 5 academic medical centers. The sampling frame included all oncology clinical trials listed as active or recruiting on participating institutions' public

trial portals during the study period. Trials were selected systematically based on portal availability and recruitment status rather than through random sampling. Data collection was limited to 5 academic medical centers selected based on the availability of public institutional trial listings and existing program partnerships. The number of trials curated per institution was not intentionally balanced and varied according to institutional trial volume and portal structure.

Institutional clinical trial portals were used to capture site-specific information, such as local recruiting status and institution-level trial availability, which is not consistently available or timely in ClinicalTrials.gov. These data are necessary to support the institution-specific trial discovery functionality of the LookUpTrials platform. Extracted data elements included trial identifiers, disease site, eligibility descriptors, phase, recruiting status, and site-specific information. Data were entered into standardized templates designed to support downstream integration into the LookUpTrials platform. Quality assurance included peer review of entries and secondary verification by program leads to resolve discrepancies and clarify ambiguous protocol information.

Outcomes and Measures

The primary feasibility outcomes included trial throughput (number of trials curated over time) and descriptive changes in individual processing rates across the study period. The secondary outcomes included participant-reported experiences with onboarding, workflow clarity, peer support, and perceived learning, assessed via postprogram surveys with Likert-scale and open-ended items. Given the exploratory nature of the study and small sample size, analyses were descriptive and focused on characterizing workflow performance and participant engagement rather than formal hypothesis testing.

Ethical Considerations

This study was conducted as an educational program evaluation and did not meet the definition of human subjects research. The project evaluated a training workflow and analyzed program-level feasibility outcomes using aggregated data and deidentified participant feedback. No identifiable private information was collected, and no patient-level data or protected health information were accessed. All clinical trial information was obtained from publicly available registries and institutional trial portals.

In accordance with US Department of Health and Human Services regulations for the protection of human subjects (45 Code of Federal Regulations 46), this work did not require institutional review board review or approval [22]. As the study did not involve human subjects research as defined by these regulations, informed consent was not required. Participant responses were collected voluntarily and analyzed in aggregate, and no financial compensation was provided.

Results

Participant Characteristics

Participant demographic characteristics are summarized in Table 1. In total, 10 undergraduate trainees participated in the program across the study period. Participants represented diverse academic majors within the life sciences and had heterogeneous prior research experience. Several participants reported no prior exposure to oncology or clinical trials at baseline. This heterogeneity aligned with the study objective of evaluating the feasibility of onboarding individuals without formal clinical or informatics training into standardized data curation workflows.

Table 1. Demographic characteristics of undergraduate trainees participating in the clinical trial data curation workflow (n=10).

Characteristics	Values, n (%)
Sex	
Male	4 (40)
Female	6 (60)
Academic standing	
Sophomore	4 (40)
Junior	6 (60)
Academic major	
Biology	3 (30)
Biochemistry	2 (20)
Human Biology and Society	1 (10)
Microbiology, Immunology, and Molecular Genetics	3 (30)
Neuroscience	1 (10)
Cohort	
Cohort 1	6 (60)
Cohort 2	4 (40)
Prior research experience (years)	
None	5 (50)

Characteristics	Values, n (%)
<1	2 (20)
1-2	2 (20)
>2	1 (10)

Trial Curation Output and Coverage

The number of curated oncology trials by institution is summarized in Table 2. Over a 10-month period, trainees curated a total of 2503 oncology clinical trial entries sourced from publicly available institutional trial portals and ClinicalTrials.gov across 5 academic medical centers.

The volume of curated trials varied by institution, reflecting differences in the size and scope of publicly listed oncology trial portfolios. Curated trials spanned multiple disease sites and trial phases, providing broad coverage of oncology clinical research activity represented within the source portals (Table 3).

Table 2. Processed oncology clinical trials by institution across 5 academic medical centers (n=2503).

Institutions	Processed trials, n (%)
UCLA[a]	311 (12.4)
Cedars-Sinai Medical Center	92 (3.7)
University of California, Irvine	584 (23.3)
University of California, San Francisco	1004 (40.1)
University of California, San Diego	512 (20.5)

[a]UCLA: University of California, Los Angeles.

Table 3. Processed oncology clinical trial entries by cancer type across all participating institutions (n=1790).[a]

Cancer types	Processed trials, n (%)
Head and neck	85 (4.7)
Gastrointestinal	211 (11.8)
Gynecologic	85 (4.7)
Breast	113 (6.3)
Sarcoma	73 (4.1)
Lung	156 (8.7)
Central nervous system	155 (8.7)
Skin	175 (9.8)
Genitourinary	178 (9.9)
Hematologic	267 (14.9)
Miscellaneous	292 (16.3)

[a]Trials classified under nonspecific disease categories (eg, "solid tumors") were not assigned to a single cancer type and therefore were not included in this distribution.

Workflow Performance Over Time

A descriptive analysis of individual trial processing rates indicated increases in throughput over time for most trainees (Figure 2). These changes were consistent with increasing familiarity with trial structures, data abstraction conventions, and workflow expectations rather than formal performance benchmarking. Variability in processing rates across trainees was observed, reflecting differences in baseline experience, availability, and learning trajectories.

Figure 2. Individual changes in trial processing rates over time (n=10). Each line represents a single participant, illustrating trial processing rates measured during their first assigned trial abstraction and their most recent completed trial abstraction within the study period. Line style and symbol indicate prior research experience (dashed lines with triangles=no prior research experience; solid lines with circles=any prior research experience). The trial processing rate was defined as the number of oncology clinical trials fully abstracted and standardized per hour. Data are presented descriptively to illustrate within-participant variation.
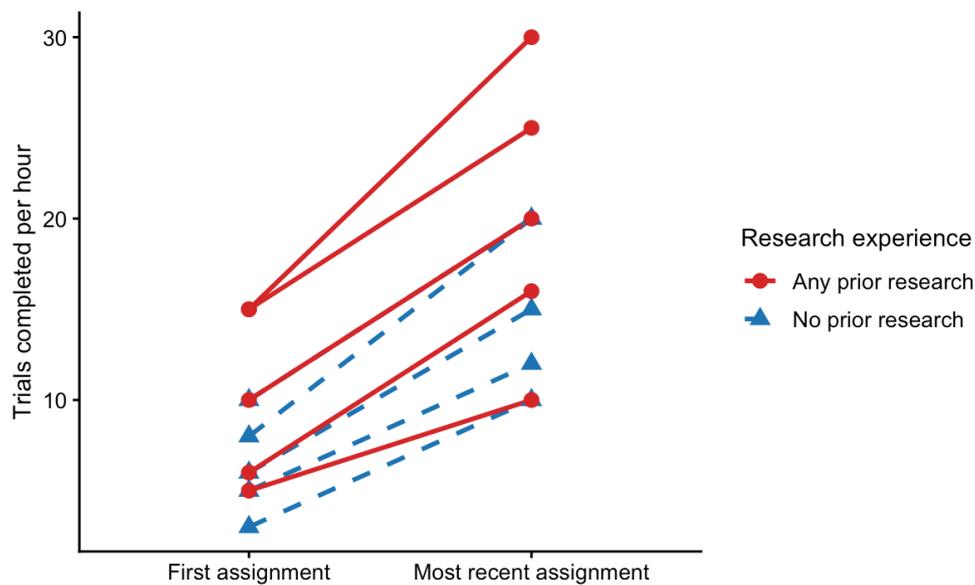
Quality Assurance and Workflow Iteration

Throughout the study period, peer review and secondary verification by program leads identified common sources of ambiguity in trial protocols, including inconsistent terminology across institutional portals and variability in how eligibility criteria and recruitment status were reported. Iterative refinements to training materials and workflow documentation were made in response to these challenges. Over time, fewer clarifications were required per curated trial, suggesting increasing alignment with standardized abstraction conventions and quality assurance processes.

Participant-Reported Experience

Postprogram surveys indicated that participants perceived structured onboarding, shared documentation, and peer support as helpful components of the workflow. Participant ratings of select program components are shown in Figure 3. Trainees reported increased familiarity with clinical trial structure, eligibility terminology, and the role of structured trial information in supporting digital clinical research tools. Participants also noted challenges related to interpreting heterogeneous trial descriptions across institutions and managing time commitments alongside academic responsibilities (Textbox 1). Qualitative feedback suggested that the supervised, iterative workflow design supported learning and engagement despite these challenges.

Figure 3. Participant ratings of selected program components (n=10). Points represent individual participant ratings on a 5-point Likert scale. Diamonds indicate median values, and horizontal bars represent the IQR. Components correspond to modifiable elements of the undergraduate training and data curation workflow (eg, training resources, peer support, and workload) and are presented descriptively to illustrate variability in participant perceptions.
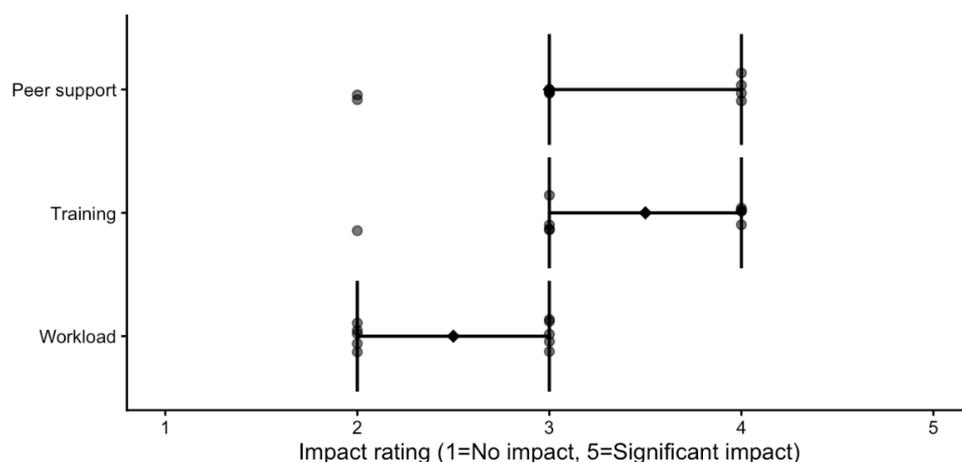
Textbox 1. Principal challenges and solutions encountered by participants.

Challenges

- Lack of uniformity between clinical trial sites and dynamic statuses of clinical trial progress
- Difficulty in the production of accurate and cohesive data compilations among multiple program members
- Lack of research experience among program members and background knowledge on oncology broadly
- Ensuring consistent satisfaction and adequate use of time among program members

Solutions

- Emphasis on quality control and consistent collaborations to discuss arising issues and uncertainties among program members
- Use of standardized training materials, shared data abstraction templates, and hands-on onboarding to promote consistency
- Use of a shared communication platform and open access to members' previous work for review
- Open dialogue among program members and communicating feasible deadlines to support sustained participation

---

Discussion

In this feasibility evaluation, we found that a standardized, supervised human-in-the-loop workflow for oncology clinical trial data curation can be implemented with individuals without prior clinical trials or informatics training. Over a 10-month period, trainees curated more than 2500 trial entries across 5 academic medical centers, with descriptive increases in processing rates over time and favorable participant-reported experiences with structured onboarding and peer support. These findings suggest that with appropriate workflow design, training materials, and quality assurance processes, nonclinician trainees can contribute meaningfully to the data stewardship layer required to support real-world digital clinical trial platforms.

Digital clinical trial platforms increasingly integrate automated data pipelines and AI-assisted features to support point-of-care trial navigation. However, our findings reinforce that human-in-the-loop data stewardship remains a critical complement to automation, particularly for maintaining institution-specific trial information that is operationally relevant but inconsistently represented in registries. The observed learning curves and workflow stabilization over time highlight the importance of standardized abstraction conventions, shared documentation, and multitiered quality assurance in sustaining high-quality trial data. From an informatics perspective, this work underscores that the success of AI-enabled clinical research tools depends not only on algorithmic performance but also on the design of scalable sociotechnical workflows that maintain the underlying data infrastructure.

This study complements prior feasibility evaluations of embedding LookUpTrials within community oncology workflows by focusing on the sustainability of the content pipeline rather than downstream clinical adoption or enrollment outcomes [13,14]. Together, these findings support a layered model of digital clinical trial platforms in which clinician-facing tools can be embedded within routine workflows, and the underlying data stewardship processes required to maintain those tools can be supported through structured human-in-the-loop workflows. Framing digital trial platforms as sociotechnical systems with interdependent technical and human components may inform more durable implementation strategies as such tools scale across institutions.

Several limitations warrant consideration. First, the sample size was small and drawn from a single undergraduate program, which may limit generalizability to other trainee populations or institutional contexts. Second, while peer review and secondary verification were incorporated into the workflow, formal benchmarking of data accuracy, interrater reliability, or error rates against gold-standard sources was not performed in this feasibility evaluation. Third, the evaluation focused on workflow feasibility and participant experience rather than clinical outcomes, platform usability, or downstream effects on trial referral or enrollment. Finally, this study reflects an early-stage implementation context, and workflow performance may differ in more mature operational settings or under sustained production demands.

Future work should evaluate the scalability and durability of standardized human-in-the-loop workflows across diverse workforce models, including funded research coordinators, centralized trial operations teams, and hybrid human-AI review pipelines. Incorporating formal data quality metrics, automation-assisted abstraction, and continuous quality monitoring may further strengthen the sustainability of trial data stewardship processes. As digital clinical trial platforms mature, integrating these workflow designs into enterprise research operations could help ensure that AI-enabled tools remain supported by reliable, up-to-date, and operationally relevant trial data over time.

Data Availability

The clinical trial data extracted and processed by undergraduate participants were sourced from publicly available institutional clinical trial portals and ClinicalTrials.gov. The final curated dataset is publicly available through the LookUpTrials mobile app. Additional information about data processing methods is available upon request from the corresponding author.

Authors' Contributions

MI, OY, JD, S Arcot, AC, S Amara, SHB, IB, KS, and QD contributed to data interpretation and manuscript drafting. TKWH supervised the project and provided final revisions. All authors reviewed and approved the final version.

Conflicts of Interest

TKWH serves as an advisor to the TeamX Health chapter at University of California, Los Angeles, and is a developer of the LookUpTrials app. TKWH also helped design and implement the undergraduate training and onboarding workflows described in this study. MI, OY, JD, SA, AC, SHB, IB, KS, and QD participated as undergraduate trainees involved in the data curation workflows described. The authors declare no other personal financial conflicts of interest.

References

1. Rimel BJ. Clinical trial accrual: obstacles and opportunities. Front Oncol. 2016;6:103. [doi: 10.3389/fonc.2016.00103] [Medline: 27200292]
2. Fleury ME. Consensus recommendations for improving the cancer clinical trial matching environment. Cancer. Jan 1, 2024;130(1):11-15. [doi: 10.1002/cncr.35034] [Medline: 37851508]
3. Castillo BS, Boehmer L, Schrag J, et al. Oncologist-reported barriers and facilitators to offering cancer clinical trials to their patients. Curr Oncol. May 28, 2024;31(6):3017-3029. [doi: 10.3390/curroncol31060230] [Medline: 38920714]
4. Nipp RD, Hong K, Paskett ED. Overcoming barriers to clinical trial enrollment. Am Soc Clin Oncol Educ Book. Jan 2019;39:105-114. [doi: 10.1200/EDBK_243729] [Medline: 31099636]
5. Wong C, Zhang S, Gu Y, et al. Scaling clinical trial matching using large language models: a case study in oncology. Presented at: Proceedings of the 8th Machine Learning for Healthcare Conference, PMLR; Apr 11-12, 2023; New York. URL: https://proceedings.mlr.press/v219/wong23a.html [Accessed 2026-03-12]
6. Mathes G, Bochum S, Werner P, Fegeler C, Sigle S. Design and implementation of a community driven clinical trial information management system for precision oncology settings. Stud Health Technol Inform. Aug 30, 2024;317:105-114. [doi: 10.3233/SHTI240844] [Medline: 39234712]
7. Hu YH, Cheng YY, Lan CC, Su YH, Sung SF. An intelligent trial eligibility screening tool using natural language processing with a block-based visual programming interface: development and usability study. JMIR Med Inform. Dec 11, 2025;13:e80072. [doi: 10.2196/80072] [Medline: 41380085]
8. Shriver S, Semy S, Lister Z, Arafat W, Fleury M. Blue-button: a tool for institution-agnostic, EHR-integrated regional clinical trial matching. JCO Oncol Pract. Oct 2024;20(10_suppl):65-65. [doi: 10.1200/OP.2024.20.10_suppl.65]
9. Inan OT, Tenaerts P, Prindiville SA, et al. Digitizing clinical trials. NPJ Digit Med. 2020;3:101. [doi: 10.1038/s41746-020-0302-y] [Medline: 32821856]
10. Köpcke F, Prokosch HU. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res. Jul 1, 2014;16(7):e161. [doi: 10.2196/jmir.3446] [Medline: 24985568]
11. Viergever RF, Karam G, Reis A, Ghersi D. The quality of registration of clinical trials: still a problem. PLoS ONE. 2014;9(1):e84727. [doi: 10.1371/journal.pone.0084727] [Medline: 24427293]
12. Huić M, Marušić M, Marušić A. Completeness and changes in registered data and reporting bias of randomized controlled trials in ICMJE journals after trial registration policy. PLoS ONE. 2011;6(9):e25258. [doi: 10.1371/journal.pone.0025258] [Medline: 21957485]
13. Sharma D, Hamm CM, Hung TKW, et al. Modernizing clinical trial accessibility: Integrating the AI-powered LookUpTrials app into the CTN program. JCO. Jun 2025;43(16_suppl):43. [doi: 10.1200/JCO.2025.43.16_suppl.e13558]

14.   Hung KT, Dunn L, Sherman EJ, et al. LookUpTrials: assessment of an artificial intelligence-powered mobile application to engage oncology providers in clinical trials. JCO Global Oncology. Aug 2023;9(Supplement_1):111-111. [doi: 10.1200/GO.2023.9.Supplement_1.111]

15.   Jain N, Mittendorf KF, Holt M, et al. The My Cancer Genome clinical trial data model and trial curation workflow. J Am Med Inform Assoc. Jul 1, 2020;27(7):1057-1066. [doi: 10.1093/jamia/ocaa066] [Medline: 32483629]

16.   Venkatesan A, Karamanis N, Ide-Smith M, Hickford J, McEntyre J. Understanding life sciences data curation practices via user research. F1000Res. 2019;8(ELIXIR):1622. [doi: 10.12688/f1000research.19427.1]

17.   Ni Y, Wright J, Perentesis J, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak. Apr 14, 2015;15:28. [doi: 10.1186/s12911-015-0149-3] [Medline: 25881112]

18.   Doan A, Ardalan A, Ballard J, et al. Human-in-the-loop challenges for entity matching. Presented at: SIGMOD/PODS'17; May 14-19, 2017; Chicago, IL, USA. [doi: 10.1145/3077257.3077268]

19.   Szostak J, Ansari S, Madan S, et al. Construction of biological networks from unstructured information based on a semi-automated curation workflow. Database (Oxford). Jun 17, 2015;2015:bav057. [doi: 10.1093/database/bav057] [Medline: 26200752]

20.   Beck JT, Vinegra M, Dankwa-Mullan I, et al. Cognitive technology addressing optimal cancer clinical trial matching and protocol feasibility in a community cancer practice. JCO. May 20, 2017;35(15_suppl):6501-6501. [doi: 10.1200/JCO.2017.35.15_suppl.6501]

21.   Kostka J, Zerillo JA, Kruse A, et al. Clinical trial enrollment expansion to the community. JCO. Mar 1, 2016;34(7_suppl):112-112. [doi: 10.1200/jco.2016.34.7_suppl.112]

22.   Protection of Human Subjects, Code of Federal Regulations, title 45 (2024): part 46. Code of Federal Regulations. 2024. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46 [Accessed 2026-03-17]

Abbreviations

AI: artificial intelligence