Structuring Eligibility on Both Sides: The EXACT System for Precision Clinical Trial Matching

Adam BLUM[a,1]
[a] *HealthKey.ai*

Abstract. Fewer than one in twenty adults with cancer enrolls in a therapeutic clinical trial; the bottleneck is rarely a missing trial but the labor of determining who is truly eligible from criteria buried in prose and patient data scattered across systems. We present EXACT, a precision trial-matching system that decomposes matching into two structured tasks rather than reasoning end-to-end over free text. On the trial side, an attribute-specific Prompt Workbench, in which subject-matter experts author and iteratively refine prompts, converts eligibility prose into conjunctive-normal-form (CNF) logic. On the patient side, longitudinal records from many EHRs are harmonized via FHIR into PRomop, an oncology-extended OMOP patient record, and projected to a flat, query-ready feature layer. A stateless matcher (PATCH) evaluates each criterion and returns a tri-valued verdict — Eligible, Potentially eligible, or Ineligible — naming, for every "Potential," the minimal tests that would resolve it. Against expert-adjudicated ground truth, micro-averaged extraction F1 reached ≈82% for follicular lymphoma (100 trials, 84 attributes) and 89.9% for multiple myeloma (98 trials, 79 attributes). Because matching is a deterministic evaluation of extracted criteria against harmonized patient attributes, per-criterion extraction accuracy upper-bounds end-to-end matching accuracy under accurate patient data.

Keywords. clinical trial matching, eligibility criteria, information extraction, large language models, OMOP common data model, patient-centered informatics

1. Introduction

Fewer than one in twenty adults with cancer enrolls in a therapeutic clinical trial; participation rates of roughly 3–5% have persisted for decades [1]. The limiting factor is rarely the absence of a suitable trial — it is the labor of determining who is truly eligible. Eligibility criteria are written as narrative prose that combines diagnoses, biomarkers, laboratory thresholds, prior lines of therapy, and temporal windows, while the patient data needed to check them are scattered across multiple electronic health records (EHRs), formats, and institutions and are rarely assembled into a single usable picture. Manual screening by clinicians, coordinators, and patients does not scale: it is slow, costly, and quickly out of date.

Most automated matching tools apply coarse filters — disease, stage, and age — and return long lists of nominal matches, the majority of which fail when checked against the full criteria; the burden of verification falls back onto people. A second class of systems

[1]Corresponding Author: Adam Blum, HealthKey.ai; E-mail: adam@healthkey.ai.

uses large language models (LLMs) to reason over patient narrative and trial text end-to-end [2]. This is flexible, but its logic can drift, its verdicts are hard to reproduce, and it is difficult for a clinician to audit why a given patient was or was not matched.

We present EXACT (EXtracting Attributes from Clinical Trials) , a precision trial-matching system that takes a different stance: rather than reasoning over free text on both sides at once, it structures both sides first and then compares them. Trial criteria are extracted into explicit logic; patient records are harmonized into a standardized, query-ready representation; and a stateless engine evaluates each criterion, returning a three-valued verdict and, where a verdict is uncertain, the specific tests that would resolve it. This paper describes the architecture, the expert-in-the-loop extraction method and its evaluation, and — responding directly to reviewer questions — the relationship between the extraction accuracy we measure and end-to-end matching accuracy, together with limitations, cost, equity, and conflict-of-interest considerations.

2. Methods

2.1. System Overview

EXACT, available at https://github.com/healthkey-ai/exact, decomposes matching into four components arranged as one pipeline (Figure 1). First, the Attribute Criteria Extraction (ACE) engine turns trial prose into structured criteria using expert-authored prompts and LLMs. Second, a structured trial database stores the resulting inclusion and exclusion conditions with their thresholds, units, and categories. Third, the patient record is harmonized into a standardized store (PRomop). Fourth, the Patient Attribute–Trial Criteria Harmonizing (PATCH) engine evaluates each patient's attributes against trial logic. As Figure 1 shows, two human-facing loops surround this core: on the trial side, subject-matter experts curate extraction prompts through the Prompt Workbench (Section 2.2); on the patient side, patients and clinicians view matches, edit attributes, and onboard data through dedicated interfaces that write back to PRomop. Decomposing the problem into independently testable components — rather than delegating the whole task to a single model — is what makes each step measurable and each verdict traceable.

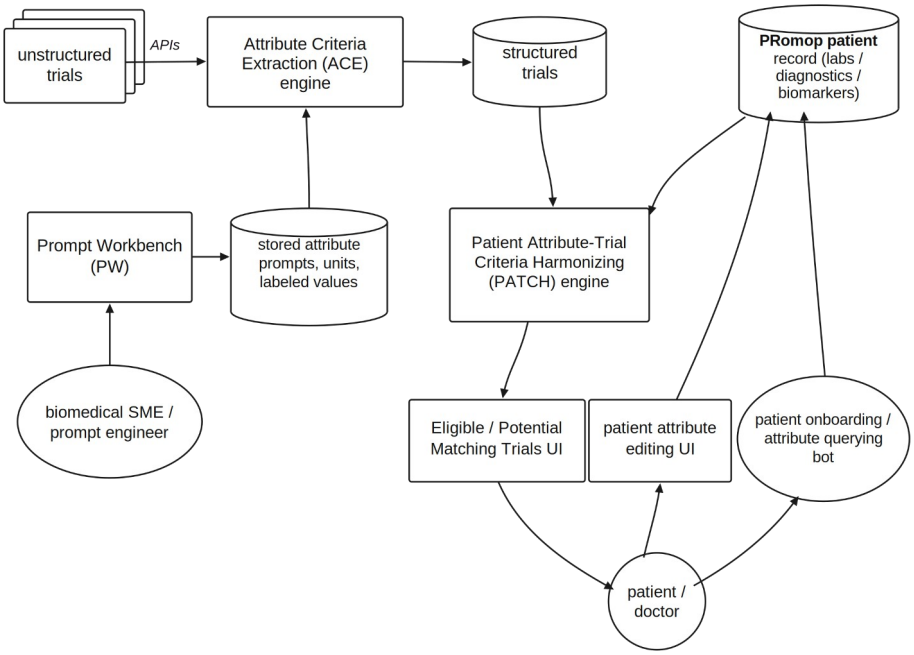

Figure 1. EXACT system architecture. On the trial side, the Attribute Criteria Extraction (ACE) engine converts unstructured trials into structured trials, driven by attribute prompts that biomedical subject-matter experts author in the Prompt Workbench. The Patient Attribute–Trial Criteria Harmonizing (PATCH) engine evaluates structured trials against the PRomop patient record (labs, diagnostics, biomarkers) and produces eligible/potential matches. Patients and clinicians view results, edit attributes, and onboard data through dedicated interfaces that write back to PRomop.

2.2. Trial-Side Extraction: the Prompt Workbench

Extraction is performed one attribute at a time. Rather than a single generic prompt, biomedical subject-matter experts (SMEs) author an attribute-specific prompt for each eligibility concept — laboratory thresholds, biomarker status, prior therapy, staging, performance status, temporal windows. The Prompt Workbench is the interactive environment in which they do so: an SME labels ground truth on representative trials, runs the prompt, inspects disagreements, and edits the prompt, iterating until accuracy plateaus. Edits are small and inspectable rather than model retraining. As a concrete example, adding a single instruction — "interpret Karnofsky performance status on its own scale, not as ECOG" — raised accuracy on that attribute by more than ten points. Domain knowledge enters the prompts as structured, reusable variables (for example, $therapy_type for therapy hierarchies and components, and $molecular_marker for known mutation lists) and as curated concept sets and composite rules (for example, the CRAB and SLiM criteria for myeloma [6]). Because a prompt targets an attribute, not a trial, the resulting library is shared across the whole corpus.

2.3. Trial Representation

Extracted criteria are expressed in conjunctive normal form (CNF): a conjunction (AND) of clauses, each a disjunction (OR) of conditions. CNF was chosen for two reasons. It is efficient — clauses evaluate quickly and a single failed required clause short-circuits the verdict — and it is interpretable: every verdict traces back to a specific attribute and to the passage in the trial text from which it was extracted, so a clinician can verify the system's reading rather than trust it.

2.4. Patient-Side Harmonization

On the patient side, longitudinal data are aggregated from many EHR systems via HL7 FHIR [4] into a single stream and normalized into PRomop, an OMOP CDM 5.4-based patient record with oncology and genomics extensions [3,5] that holds the patient's labs, diagnostics, and biomarkers (available at https://github.com/healthkey-ai/promop). From this store, PRomop maintains a flat, denormalized, query-ready projection (PatientRecord) of the attributes that matching, decision support, and predictive models all consume.

We chose the OMOP Common Data Model rather than a proprietary schema deliberately. A standard CDM with shared vocabularies (accessed through the OHDSI Athena service) provides interoperability across sites, a mature analytics tool stack, and reusable cohort definitions, and it avoids proprietary lock-in for both patient data and trial criteria [3]. The flat PatientRecord projection is the keystone: the same denormalized row simultaneously powers eligibility screening, standard-of-care decision-support rules, and machine-learning features, so a criterion can be evaluated without expensive joins across the normalized model.

2.5. Therapy Representation with HemOnc Concepts

Therapy is one of the most frequent bases for eligibility and, correspondingly, one of the hardest attribute categories to structure (Section 3.2). A large share of inclusion and exclusion criteria turn on prior treatment: whether a patient has received a particular agent, a component of a regimen, or any member of a therapeutic category, and how many prior lines of therapy have been given. Any of these can determine either eligibility or ineligibility — one trial may require prior exposure to a drug class while another excludes patients who have received it.

To represent therapy consistently on both sides, EXACT uses the HemOnc oncology vocabulary [7], available through OHDSI Athena, as the shared reference for agents, regimen components, and their categories. On the trial side, extraction prompts resolve therapy mentions to HemOnc concepts, so a criterion expressed as a specific drug, a component, or a class is normalized to one controlled hierarchy. On the patient side, the PRomop record is converted to express the therapies a patient has received as ordered lines — first-line, second-line, and later — each line stated in HemOnc concepts. Because the criterion and the patient's treatment history are expressed in the same vocabulary and its category hierarchy, PATCH can evaluate a therapy criterion — for example, receipt of a proteasome inhibitor in an earlier line — as a structured lookup against the line-organized record, traversing the HemOnc hierarchy to decide whether a specific agent satisfies a class-level requirement rather than reasoning over free text.

2.6. Matching: the PATCH Engine

PATCH is stateless: it takes a patient's harmonized attributes and a trial's CNF criteria and returns one of three verdicts. A trial is Eligible when all required criteria are satisfied by the patient's known attributes; Ineligible when a criterion is contradicted by the patient's data; and Potentially eligible when one or more required attributes are missing or unknown — an explicit "not yet a no" rather than a forced decision. The three-valued design is deliberate: real records are incomplete, and by reading the logical structure PATCH computes the minimal set of tests or questions that would convert a Potential into a definite verdict, turning a data gap into a concrete next step for the patient or navigator.

2.7. Evaluation Design and Safeguards Against Overfitting

Extraction accuracy is measured against expert-adjudicated ground truth. For each disease, the most recent trials in the corpus are annotated independently by two clinical annotators who are blind to the system's output; disagreements are adjudicated to a frozen gold standard, and inter-annotator agreement is reported. Ambiguous or vague ("fuzzy") criteria are handled with a written annotation guide; genuinely unresolved cases are adjudicated rather than guessed, and the edge rules are documented so that annotation is reproducible.

Two design choices guard against overfitting to the trials seen during prompt development. First, there is no model training: prompts are authored on a development subset, frozen, and then scored only on held-out annotated trials, so reported accuracy reflects performance on unseen trials rather than on the examples used to write the prompts. This also clarifies the analysis structure — the annotated evaluation set (for follicular lymphoma, 100 trials) is scored as a held-out test set; it is not a training set from which the remainder of the corpus is separately validated. Second, prompts target attributes (for example, "bilirubin upper limit") rather than individual trials, and are versioned and re-evaluated as the corpus grows, which limits both overfitting and drift. Maintenance therefore scales sub-linearly with the number of trials: a small SME team curates a shared, reusable prompt library that each new trial reuses, rather than authoring per-trial logic.

2.8. Extending to New Diseases

EXACT currently supports five diseases: follicular lymphoma, multiple myeloma, breast cancer, chronic lymphocytic leukemia, and mantle cell lymphoma. Extending it to a new disease follows a fixed procedure. All current trials for the disease are analyzed to enumerate the attribute criteria they use to determine eligibility or ineligibility. Attributes that appear only once across the disease's trials are excluded, so the attribute set captures recurring, generalizable criteria rather than idiosyncratic one-offs. The prompt-authoring process (Section 2.2) is then carried out only for attributes that are new to the disease and not already covered by the existing shared library.

This is where reuse pays off across diseases as well as across trials. Because much of the eligibility vocabulary — laboratory thresholds, performance status, common biomarkers, and prior-therapy structure — recurs across cancers, each disease we add introduces fewer genuinely new attributes than the last, and the SME prompt work

required to onboard it has correspondingly diminished. The marginal cost of covering an additional disease falls as the shared attribute library approaches saturation.

3. Results

3.1. Extraction Accuracy

For follicular lymphoma, the corpus comprised 1,518 trials drawn from ClinicalTrials.gov, ISRCTN, and the EUCTR; the 100 most recent trials were annotated across 84 patient attributes, yielding 533 individual attribute criteria. Micro-averaged extraction accuracy on the held-out annotated set was F1 ≈ 82% (precision 80%, recall 85%). The method replicated in a second disease with a larger active trial base: of 3,447 actively recruiting multiple myeloma trials, 98 were labeled across 79 eligibility attributes, yielding a micro-averaged extraction F1 of 89.9%. (Disease-specific derived attributes such as MeetsCRAB and MeetsSLiM are computed for myeloma trials but are not the basis of this figure, which spans all 79 attributes.) Table 1 summarizes these results; work extending the evaluation to additional diseases, including chronic lymphocytic leukemia, mantle cell lymphoma, and breast cancer, is ongoing.

Table 1. Extraction evaluation summary across diseases and venues. Accuracy is micro-averaged on held-out, expert-adjudicated trials.

Disease	Venue	Corpus / annotation	Extraction accuracy
Follicular lymphoma	ASH 2025	1,518 trials; 100 SME-annotated; 84 attributes; 533 criteria	F1 ≈ 82% (P 80%, R 85%)
Multiple myeloma	BSH 2026	3,447 actively recruiting trials; 98 labeled; 79 attributes	F1 89.9%

3.2. Where Extraction Is Hardest

Accuracy is not uniform across attribute categories. Simple, well-coded attributes (diagnoses, laboratory thresholds with explicit units, biomarker status) extract most reliably. Therapy-related constraints are the consistent weak point: precision on therapy criteria sits at roughly 77%, because they require reasoning over therapy hierarchies, drug classes and components, line-of-therapy semantics, and temporal relationships that are frequently implicit in the trial text. This is a property of the attribute category rather than of a particular disease, and it defines where additional domain resources are most valuable (Section 4.6).

3.3. Extraction Accuracy as a Proxy for End-to-End Matching Accuracy

A recurring and fair question is how per-criterion extraction accuracy relates to the metric that ultimately matters — whether the right patient reaches the right trial. EXACT's architecture makes this relationship explicit. Once trial criteria are extracted into CNF and patient attributes are harmonized, the match itself is not a learned step: PATCH deterministically evaluates each clause of the criteria against the patient's attributes. Under the assumption that patient information is accurate and complete, the matcher introduces no error of its own — a verdict is correct if and only if the criteria that

determine it were extracted correctly. Extraction is therefore the sole fallible component, and per-criterion extraction accuracy places a ceiling on achievable matching accuracy: the system cannot match better than it reads. This is why we report per-criterion F1 rather than only an end-to-end ranking metric. Per-criterion measurement is more granular and more rigorous — it localizes error to specific attributes (revealing, for example, that therapy criteria are the weak point), whereas an aggregate ranking score conflates extraction error, patient-data gaps, and scoring heuristics and is more easily inflated.

Two consequences follow. First, the accuracy we report is conservative with respect to real matching: many trial verdicts are decided by a single disqualifying criterion (a contradicted exclusion short-circuits to Ineligible), so an individual verdict rarely depends on every extracted attribute being simultaneously correct. Second, where extraction is uncertain or a required attribute is missing, the tri-valued design converts that uncertainty into a Potentially eligible verdict with named resolving tests, rather than forcing a false Eligible or false Ineligible. Residual extraction error is thus biased toward a safe, actionable outcome rather than a silent mismatch. The corollary is that end-to-end matching accuracy is bounded above by extraction accuracy and, in deployment, further protected by the conservative verdict semantics — but it also depends on the accuracy of the patient record (Section 2.4), which remains the subject of ongoing prospective study (Section 4.4).

4. Discussion

4.1. Relation to Other AI Matchers

The contrast with end-to-end LLM matchers such as TrialGPT [2], TrialMatchAI, and OncoLLM is not stylistic but definitional. These systems are typically posed and evaluated as ranking problems: given a patient vignette — often a partial, synthesized description drawn from a benchmark dataset — they order the trials in that dataset by predicted suitability or relevance. They do not render a concrete eligibility determination, do not eliminate ineligible trials from consideration, and do not incorporate the patient's own priorities. EXACT is built on the opposite commitment. For each trial it decides, criterion by criterion, whether the patient is fully eligible or a potential match — and, for a potential, which specific tests would resolve it — and it removes ineligible trials from consideration outright. Only then does it rank the surviving eligible trials, using multi-criteria decision analysis (MCDA) over factors that matter to the individual (Section 4.3). Eligibility is thus a determination rather than a score, and ranking operates on genuinely eligible trials weighted by patient preference rather than on an undifferentiated list.

Structuring both sides also yields three properties that end-to-end matching cannot easily guarantee. EXACT is auditable: every verdict traces to a specific attribute and a passage in the trial, so clinicians can verify rather than trust. It is robust to missing data: potentially-eligible verdicts with minimal-test guidance keep gaps from becoming silent dead ends. And it is expert-steerable: domain knowledge enters through SME-refined prompts and concept sets, with no model retraining, so a clinician can correct a systematic error directly and observe the effect on held-out accuracy.

4.2. Maintenance and Overfitting in Practice

Two practical questions follow from an interactive, expert-in-the-loop method: how many experts are needed as trials accumulate, and does interactive prompt development overfit to the trials being viewed? Because prompts are attribute-scoped and reused, maintenance scales sub-linearly — a small SME team, rather than one expert per trial, curates the shared library, and new trials draw on existing prompts. The same effect holds across diseases: each new disease introduces fewer novel attributes than the last (Section 2.8). Overfitting is bounded structurally by the held-out evaluation of Section 2.7: prompts are frozen before scoring on unseen annotated trials, and versioning with re-evaluation on the growing corpus surfaces drift. The Karnofsky example illustrates the intended dynamic — an inspectable edit that generalizes to an attribute, not a patch fitted to one protocol.

4.3. Beyond Eligibility: Ranking Around the Patient

Eligibility narrows the field; it does not, by itself, tell a patient which trial is right for them. When several trials remain eligible, EXACT ranks them with multi-criteria decision analysis (MCDA) over factors that matter to the individual — expected therapeutic benefit, treatment burden (visits, procedures, demands on time), anticipated toxicity, and travel distance to study sites — weighted by the patient's own priorities. The intent is to let patients and clinicians ask not only “can I join?” but “is this the right one for me?”

4.4. Cost, Integration, and Equity

Integration cost is lowered by building on standards. FHIR ingestion, an OMOP store, and TEFCA-based exchange mean that connecting a new site or trial source reuses existing connectors and vocabularies rather than bespoke integration, and the reusable prompt library amortizes extraction cost across the corpus. Equity and generalizability remain open questions. Multi-source FHIR ingestion reduces dependence on any single EHR, and therefore on any single institution's population, but performance in community (non-academic) settings and among under-represented groups has not yet been established prospectively and should not be assumed from the disease-level accuracy reported here.

4.5. Limitations, Responsible Deployment, and Conflict of Interest

The clearest technical limitation is therapy reasoning (Section 3.2). We do not regard human-in-the-loop review as its long-term fix; the intended remedy is architectural — a richer therapy-ontology knowledge layer that supplies drug-class, component, and line-of-therapy structure to the extraction prompts — combined with continued SME refinement. Human review is a safeguard, not the mechanism by which the gap closes.

Conflicting criteria are handled explicitly. Cross-source disagreements in the patient record are reconciled by a dedicated reconciliation component; genuinely contradictory or vague trial criteria surface as Potential and are never silently forced to a verdict, following the same conservative principle that governs missing data.

The author holds equity in HealthKey.ai, which develops EXACT; this is a commercial interest that could bias evaluation. We mitigate it in four concrete ways: core components are released as open source; matching logic is transparent CNF that any party can inspect; ground-truth annotation is performed independently and blind to the system, with adjudication; and independent third-party validation is planned. We regard external replication on independently held data — and eventual prospective enrollment outcomes — as the decisive evidence, and report current results with that standard in mind.

5. Conclusion

EXACT reframes trial matching as two structured tasks — reading trial criteria into explicit logic and harmonizing patient records into a query-ready representation — connected by a stateless, auditable matcher that prefers an honest "Potential, and here is how to resolve it" to a confident guess. Because the match is a deterministic evaluation of extracted criteria against accurate patient attributes, the per-criterion extraction accuracy we report (≈82% for follicular lymphoma and 89.9% for multiple myeloma) bounds and stands in for end-to-end matching accuracy, while the tri-valued design keeps residual error safe and actionable. The remaining work is prospective: usability, enrollment outcomes, community-setting generalizability, and independent validation.

Acknowledgements

The author thanks his developer colleagues at HealthKey who helped developed EXACT including Leonid Morozov, Vladimir Tarasov and Nikita Shpilevoy, and the clinical annotators, led by Samar Elkassas who developed and adjudicated the ground-truth standard. HealthKey advisors Jude Fitzgibbon and Steven Labkoff on provided essential guidance on assessing accuracy and disease criteria. Paul Ahlstrom and Jennifer Ahlstrom of HealthTree foundation gave important feedback on patient needs in deciding on appropriate clinical trials and practices on garnering a comprehensive longitudinal patient record.

Disclosure

A. Blum holds equity in HealthKey.ai. Mitigations of the resulting conflict of interest are described in Section 4.6.

References

[1] Unger JM, Cook E, Tai E, Bleyer A. The role of clinical trial participation in cancer research: barriers, evidence, and strategies. Am Soc Clin Oncol Educ Book. 2016;35:185-198, doi:10.1200/EDBK_156686.

[2] Jin Q, Wang Z, Floudas CS, Chen F, Gong C, Bracken-Clarke D, et al. Matching patients to clinical trials with large language models. Nat Commun. 2024;15:9074, doi:10.1038/s41467-024-53081-z.

[3] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574-578.

[4] Bender D, Sartipi K. HL7 FHIR: an agile and RESTful approach to healthcare information exchange. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS); 2013 Jun 20-22; Porto, Portugal. p. 326-331, doi:10.1109/CBMS.2013.6627810.
[5] Osterman TJ, Terry M, Miller RS. Improving cancer data interoperability: the promise of the Minimal Common Oncology Data Elements (mCODE) initiative. JCO Clin Cancer Inform. 2020;4:993-1001, doi:10.1200/CCI.20.00059.
[6] Rajkumar SV, Dimopoulos MA, Palumbo A, Blade J, Merlini G, Mateos MV, et al. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. Lancet Oncol. 2014;15(12):e538-e548, doi:10.1016/S1470-2045(14)70442-5.
[7] Warner JL, Dymshyts D, Reich CG, Gurley MJ, Hochheiser H, Moldwin ZH, et al. HemOnc: a new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. J Biomed Inform. 2019;96:103239, doi:10.1016/j.jbi.2019.103239.