Leveraging Ontotext's Eligibility Design Assistant for Effective Patient Population Selection

Authors: Ivelina Nikolova, Katerina Serafimova, Georgi Iliev, Todor Primov

This is part of Ontotext's AI-in-Action initiative aimed at enabling data scientists and engineers to benefit from the AI capabilities of our products.

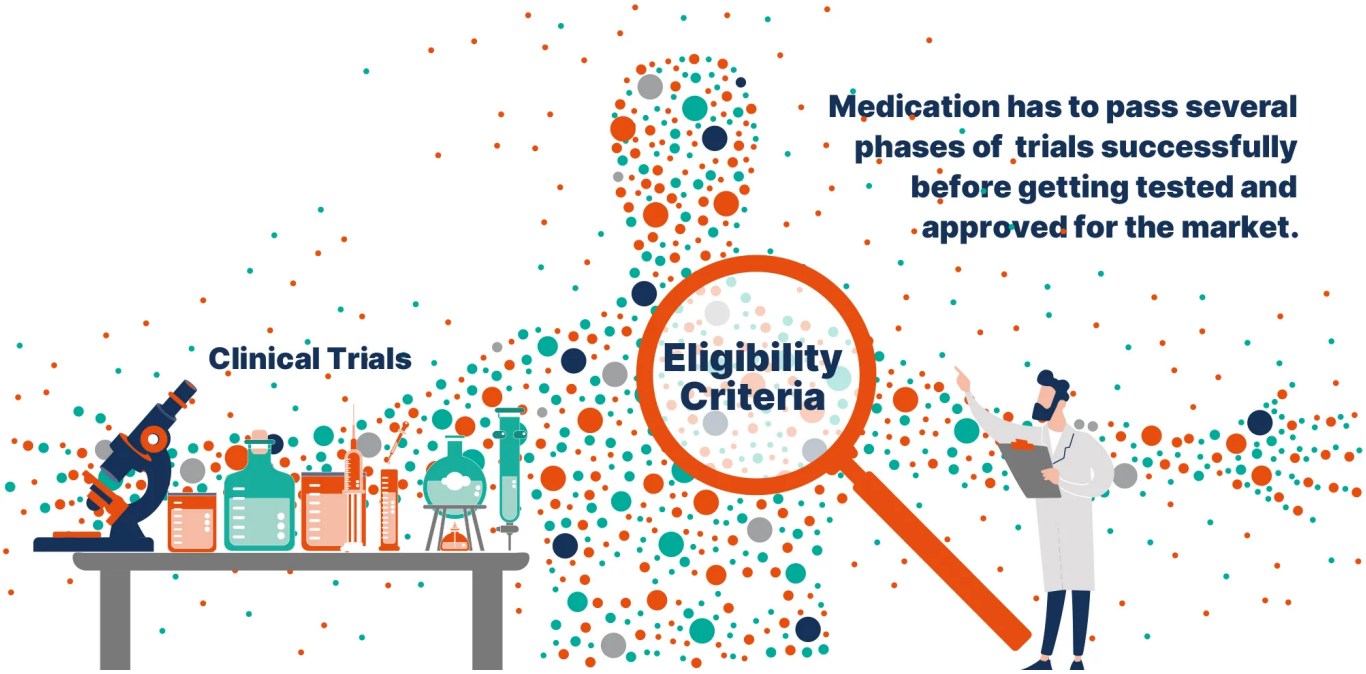
Drug discovery is an ever-growing field of research, which often involves different stakeholders, facing multiple challenges. Clinical trials play an important part in this process as any new medication has to pass several phases of such trials successfully before getting tested and approved for the market. That's why a good clinical trial design is crucial for accurately evaluating the various effects and safety of new drugs.

Ontotext's Clinical Trials Eligibility Design Assistant helps with one of the most challenging tasks in study design: selecting the proper patient population. To meet the expected study outcomes, researchers need to identify the most suitable set of eligibility criteria for participation. As they want to test for specific outcome measures (endpoints) during and after the trial, these criteria must allow them to measure the outcomes objectively and achieve the desired endpoints. In addition, they need to ensure that participants can undergo clinical trial interventions safely.

Usually, researchers select their eligibility criteria based on historical data. However, this data is often vast and heterogeneous, making analysis too complex. The publicly available datasets represent only a small fraction of the information. As companies tend to keep their proprietary datasets private, there's little visibility in the field, complicating the clinical trial design process even further. Playing around with a criteria selection without providing a proper rationale is not a viable strategy and it's where our application comes into play.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

Authors: Ivelina Nikolova, Katerina Serafimova, Georgi Iliev, Todor Primov



Ontotext's skills and how we can assist

Eligibility criteria and outcome measures are typically expressed in free text. That's why the usual method of analyzing them automatically involves traditional natural language processing techniques. However, these are limited to processing only a very narrow context and therefore can provide only a partial solution.

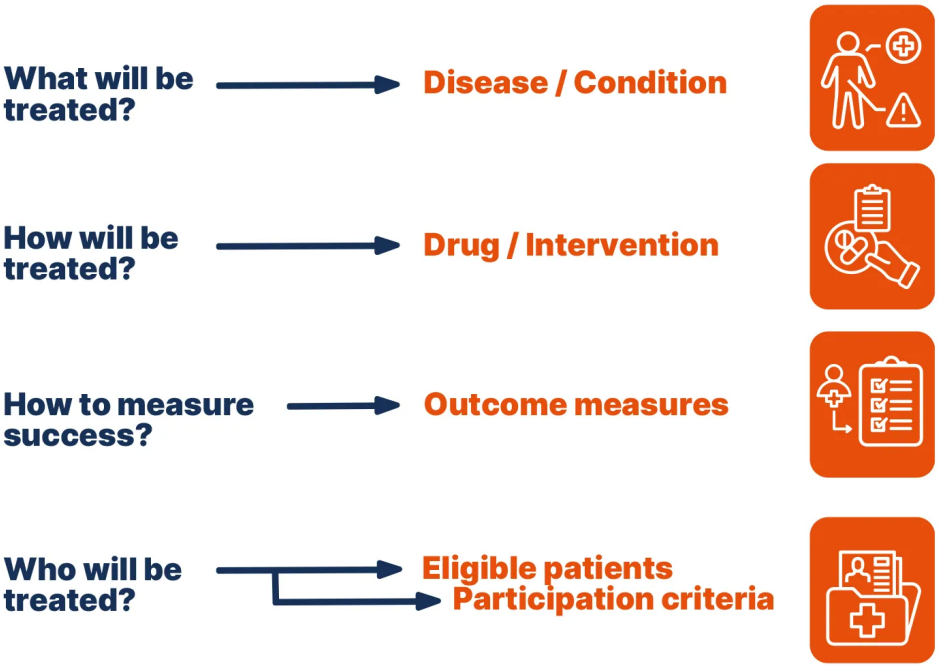
Instead, we propose a new approach: Retrieval-Augmented Generation (RAG), combining a highly normalized knowledge graph with large language models (LLMs). This method allows us to answer questions by searching within a very narrow context. This reduces LLM hallucinations and maintains provenance to the original data. At the same time, the knowledge graph provides data linking with well-known and well-structured datasets like UMLS and DrugCentral, and allows further generalization (for example over drug types).

What's a clinical trial about

The simplest way to explain what a clinical trial is about is shown in the visual below. In this

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

specific use case, we analyze only interventional clinical trials that aim to discover more about a particular intervention or treatment. When designing such a study, researchers have to answer a set of questions as shown in the visual.



Ontotext's Eligibility Design Assistant helps answer the last question: **Who will be treated?** This requires selecting the appropriate eligibility criteria to meet the chosen outcome measures.

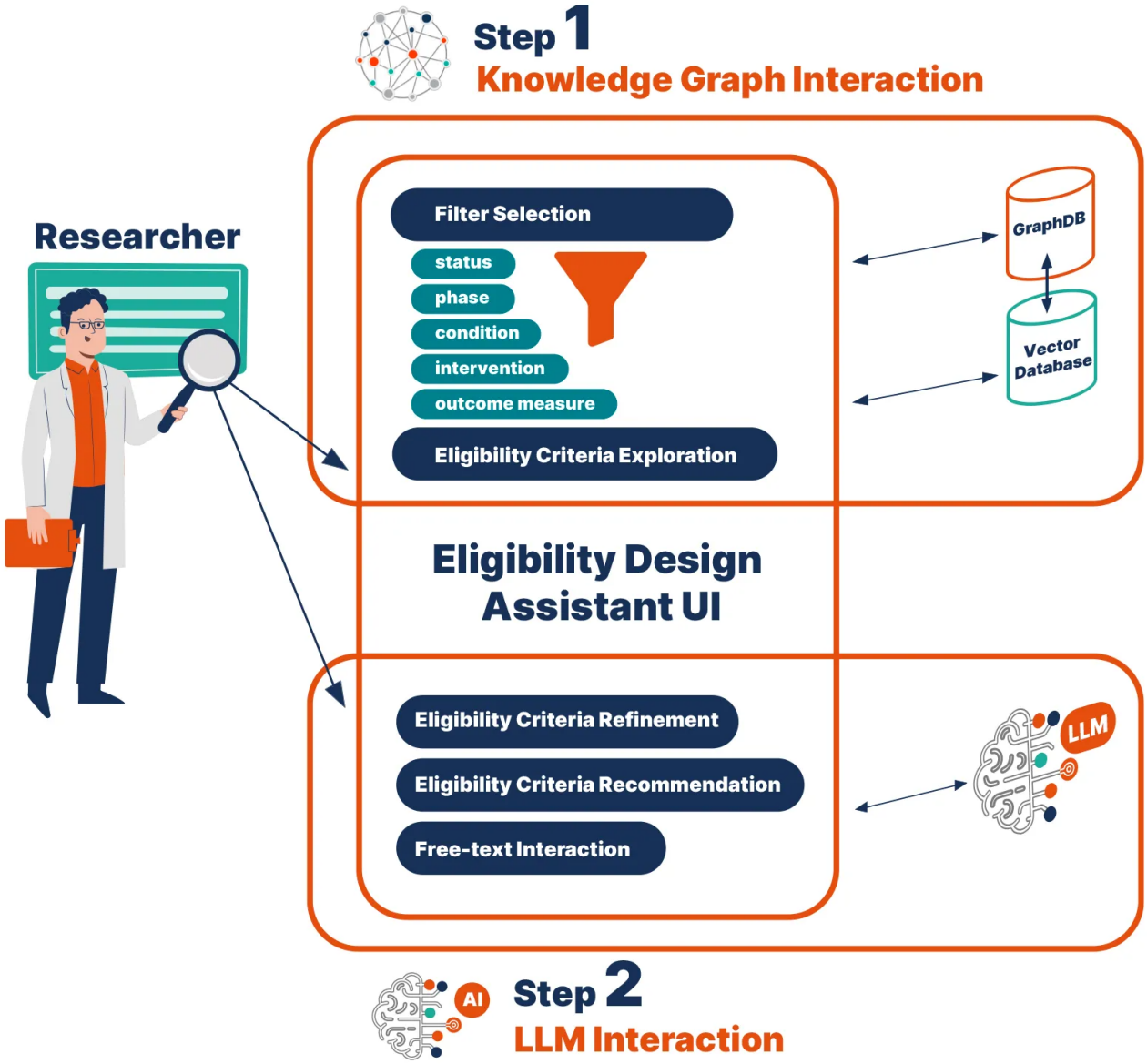
A sample record from ClinicalTrials.gov about the treatment of Asthma with Omalizumab, Placebo and Fluticasone can be seen here (Condition: Asthma; Intervention / Treatment: Drug:Omalizumab, Drug:Placebo, Drug:Fluticasone, Outcome measures, Participation criteria).

Introducing the Eligibility Design Assistant

The Eligibility Design Assistant is designed to assist researchers who have already defined the condition, intervention, and outcome measures of interest. It helps them select appropriate eligibility criteria to achieve the specified outcome measures and enables them

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

to analyze historical public data from ClinicalTrials.gov.



In the first stage Knowledge Graph Interaction, the Eligibility Design Assistant provides a variety of filters to subset the data, including:

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

- **Overall status of the study**: Currently limited to "Completed".
- **Phase**: The phase of the desired clinical trial can be selected.
- **Condition**: We have restricted the data to a subset relevant to only three conditions – Asthma, COPD, and Diabetes Mellitus.
- **Intervention**: Includes all interventions available in the dataset, listed in the relevant studies. They are normalized to *DrugCentral* resources in the knowledge graph and the names provided from *DrugCentral* are loaded in the filter.
- **Outcome measure**: An outcome measure can be selected from a list of all outcome measures found in clinical trials that match the previously selected filters. Free text can also be entered – it can represent just a common term from an outcome measure, which helps for a wider search (e.g. FEV1).

In the second stage *LLM interaction*, the Eligibility Design Assistant enables users to refine the criteria and obtain criteria recommendations. These criteria can be refined into four categories and, currently, the following are provided: **diagnosis, treatment, procedure,** and **vital parameters**.

Each category corresponds to mentions of the same type in the respective criteria (see examples in the table below).

Criterion	Type
"Severe coronary artery disease"	Diagnosis
"Patients with pulmonary hypertension due to COPD, undergoing routine invasive measurement of hemodynamic parameters."	Procedure
"Have had treatment with a stable regimen of high-dose inhaled corticosteroids (ICS) for at least 8 weeks prior to screening"	Treatment
"Heart rate < 55 bpm or >105 bpm"	Vital parameters

Once the criteria have been refined, they are visualized by category, additionally split into inclusion and exclusion criteria. Researchers can then submit pre-defined questions about these criteria to the application or write a free-text question. As a result, they can iteratively retrieve the eligibility criteria that are most appropriate for achieving the pre-selected outcome measure within the given settings (condition, phase, intervention).

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

Behind the scene

Data

For this application and specific use case, we have used a complete snapshot of *ClinicalTrials.gov* from Q1'24, covering over 490,000 studies. The original eligibility criteria from the studies have been parsed into individual semi-structured string literals to facilitate downstream processing by the LLM. We have also included normalizations of conditions, covering over 1.1 million instances of *UMLS* concepts as well as normalizations of interventions, comprising over 2.5 million instances of *ChEMBL* or *DrugCentral* concepts.

LLM's role in the Eligibility Design Assistant

The LLM helps process the criteria resulting from the pre-selected overall status, phase, condition, intervention, and outcome measure. At this stage, the number of criteria can range from tens to hundreds. This makes it necessary to have additional automatic processing to help researchers prioritize criteria matching their preferences for the study design.

To achieve this, the ***LLM is first used as a classifier*** that further refines and classifies the criteria into four categories (as detailed in the Introducing the Eligibility Design Assistant section [Link to section]). Then, the ***LLM also functions as a recommender***. It analyzes the subsets of eligibility criteria (such as inclusion/exclusion, diagnosis, treatment, procedure, and vital parameters) to recommend the most important criteria.

Sample demo

The following presentation provides a quick walkthrough of the application.

LLDI and the knowledge graph contribution

Ontotext's LinkedLifeData Inventory (LLDI) provided significant support in developing the Eligibility Design Assistant by supplying all required source datasets. LLDI provides a ready-

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

to-use RDF version of the ClinicalTrials.gov dataset, which otherwise isn't publicly available in this format. This version comes with a semantic model that facilitates efficient data querying and subsetting. LLDI also provides key data points normalized to reference datasets such as UMLS, ChEMBL, and DrugCentral. It ensures the information is up-to-date and its data loader simplifies loading the necessary datasets.

Using a knowledge graph and linked data makes it possible to extend the application's functionality even further. For example, we can expand queries from a single drug to groups of drugs with similar mechanisms of action (addressing the same target – protein, pathway, and so on). This helps researchers enhance the scope and depth of their research.

To sum it up

Ontotext's Eligibility Design Assistant functions as a lightweight UI, integrated with standard GraphDB and datasets from LLDI (including Clinical Trials, DrugCentral, ChEMBL, and UMLS). It helps researchers design clinical trials by automatically processing a large number of trials and their corresponding eligibility criteria. The application generates well-informed recommendations and provenance data for specific exclusion or inclusion criteria necessary for achieving particular outcome measures.

To minimize errors, such as hallucinations, we have restricted the set of possible values and criteria used to query the LLM, leveraging a highly normalized knowledge graph. Each step is designed to ensure high precision, with consistency tests performed throughout the whole process. The data is well-structured, even in such a domain, which is typically characterized by textual data.

The Eligibility Design Assistant demonstrates a technical use case for RAG, where the LLM is used together with a focused knowledge graph. It iteratively derives knowledge and, in this way, facilitates critical research tasks in the clinical research domain.

Do you want to learn more about our Eligibility Design Assistant?

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

Leveraging Ontotext's Eligibility Design Assistant for Effective Patient Population Selection

Authors: Ivelina Nikolova, Katerina Serafimova, Georgi Iliev, Todor Primov

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com