Paper ML23

Using Custom GPT to Write Edit Checks

Andrew Burd, Formation Bio, New York, NY, USA

ABSTRACT

Study builders across the industry must master various EDC systems, each with its own programming language for edit checks—Medidata uses C#, Viedoc uses JavaScript, and Veeva and REDCap rely on formulas similar to Excel. Often, the focus on perfecting the eCRFs and Visit Schedule leaves automated edit checks, the first line of defense against bad data, as an afterthought. Robust edit checks ensure data integrity and patient safety while also providing cost savings by reducing time on rework and manual data review. To address this, we developed a custom GPT, an AI model tailored for specific tasks, which leverages the study builder user guide to automatically generate edit checks directly from a study design specification. Automating this process significantly reduces the time and effort required to program edit checks during study start-up and ensures that edit checks are no longer overlooked.

INTRODUCTION

Clinical study data collection relies heavily on Electronic Data Capture (EDC) systems to ensure accurate and reliable data. However, the creation and maintenance of edit checks—automated rules that detect data entry errors or inconsistencies—are often treated as secondary tasks. This oversight can lead to inadequate validation of crucial patient data and increased manual review, ultimately impacting data quality, timelines, and cost.

Study builders face the challenge of working with multiple EDC platforms, each requiring proficiency in different scripting languages or formula systems. Medidata Rave uses C#, Viedoc uses JavaScript, and Veeva and REDCap often rely on Excel-like formulas to define edit checks. Managing these diverse languages and approaches can be time-consuming and prone to human error.

To streamline the process, our team sought to leverage advancements in large language models. We developed a custom GPT that interprets study design specifications and user guides to automatically generate platform-specific edit checks. This paper details the motivation behind this approach, explains how our custom GPT solution was built, and outlines the benefits observed in early adoption.

THE CHALLENGE OF MULTIPLE EDC SYSTEMS

Clinical data management teams must balance the intricacies of multiple EDC systems. Each system's unique programming language or formula syntax necessitates specialized expertise. When pivoting between Medidata, Viedoc, Veeva, and REDCap, developers must recall or re-learn the nuances of each platform.

The Cost of Neglecting Edit Checks

By focusing primarily on eCRF layouts and visit schedules, study builders may not allocate sufficient time and resources to writing robust edit checks. Missing or underdeveloped checks can lead to:

- Data inconsistencies that go unnoticed until later in the study, resulting in rework.
- Increased manual data review, consuming more time and resources.
- Potential risks to patient safety if critical alerts or data validations are absent.

When edits are overlooked, downstream correction efforts can be costly and time-consuming, causing delays in study timelines.

DEVELOPING A CUSTOM GPT SOLUTION

Recognizing the limitations of manual edit check development, we created a custom GPT model to automate this process. This model is tailored to interpret both the study design specification and the EDC user guides, generating formatted, system-specific code for each edit check.

Leveraging Large Language Models

GPT-based models are known for their ability to understand and generate text. By training our custom GPT on documentation from various EDC platforms, we enabled it to:

1. Parse structured inputs from study design specifications (e.g., field names, allowable values).
2. Identify the relevant validation logic (e.g., range checks, cross-field checks).
3. Output scripts in the correct language or syntax for the specified EDC platform.

IMPLEMENTATION AND WORKFLOW

Our automated workflow begins with the study builder uploading the study design specification. The custom GPT then references an internal library of user guides and best practice templates to produce code snippets or formula-based checks that can be directly copied into the EDC system.

Key Steps in the Process

1. **Specification Upload**: The user provides a structured specification containing field definitions, visit schedules, and data rules.
2. **GPT Generation**: The custom GPT parses the specification, identifies logic for each edit check, and produces a code snippet in the appropriate language (C#, JavaScript, or an Excel-like formula).
3. **Verification and Customization**: Study builders review the generated edit checks, making minor adjustments if needed to align with study-specific nuances.
4. **Deployment**: Code snippets are copied into the EDC system or integrated through an automated upload feature.

BENEFITS AND EARLY RESULTS

Implementing this custom GPT approach has led to several immediate benefits:

Time Savings

Automating edit check creation significantly reduces development time, allowing study builders to focus on more complex tasks or final review.

Enhanced Data Quality

With an automated, standardized process, fewer checks go unnoticed or under-prioritized. This leads to more robust data validation from the outset.

Reduced Costs

Early detection of data inconsistencies lowers the amount of rework and manual cleaning required later in the study, resulting in cost savings for sponsors and CROs.

CONCLUSION

Overlooking edit checks can have serious implications for data quality, patient safety, and study costs. By leveraging a custom GPT solution, we have streamlined the edit check creation process, significantly reduced development time, and improved overall data integrity. As AI technologies continue to advance, further refining these models and integrating them seamlessly into EDC platforms will only bolster their benefits. We encourage clinical data professionals to explore AI-driven solutions for minimizing human error and maximizing efficiency in their workflow.

REFERENCES

(References go here if needed, following your preferred citation style.)

ACKNOWLEDGMENTS

(If you wish to acknowledge any contributors, list them here.)

RECOMMENDED READING

(If you have any recommended reading for further exploration, list it here.)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Andrew Burd
Company: Formation Bio
Address: 16 E 34th St floor 10, New York, NY 10016
Work Phone: 301-712-5387
Email: andrewburd2@gmail.com
Website: https://www.formation.bio/