

Themed Section: Artificial Intelligence in Health Economics and Outcomes Research

Use of Large Language Models to Extract Cost-Effectiveness Analysis Data: A Case Study

Xujun Gu, MSPH, Hanwen Zhang, MS, Divya Patil, MS, Zafar Zafari, PhD, Julia Slejko, PhD, Eberechukwu Onukwugha, PhD

ABSTRACT

Objectives: Cost-effectiveness analyses (CEA) generate extensive data that can support much health economic research. However, manual data collection is time-consuming and prone to errors. Development in artificial intelligence (AI) and large language models (LLMs) offers a solution for automating this process. This study aims to evaluate the accuracy of LLM-based data extraction and assess its feasibility for supporting CEA data collection.

Methods: We evaluated the performance of the custom ChatGPT model (GPT), the Tufts CEA Registry (TCRD), and the researcher-validated data (RVE) in extracting 36 predetermined variables from 34 selected structured articles. Concordance rates between GPT and RVE, TCRD and RVE, and GPT and TCRD were calculated and compared. Paired student's *t* tests assessed differences in accuracy, and concordance rates across 36 variables were provided.

Results: The accuracy of GPT (GPT & RVE) was comparable to the accuracy of TCRD (TCRD & RVE) (mean 0.88, SD 0.06 vs mean 0.90, SD 0.06, $P = .71$). The performance of GPT varied across variables. GPT outperformed TCRD in capturing "Population and Intervention Details" but struggled with complex variables like "Utility."

Conclusions: This study demonstrated that LLMs, such as GPT, can be a promising tool for automating CEA data extraction, offering comparable accuracy to established registries. However, human supervision and expertise is essential to address challenges in complex variables.

Keywords: artificial intelligence (AI), cost-effectiveness analysis (CEA), data extraction, large language models (LLMs).

VALUE HEALTH. 2025; 28(11):1637–1645

Highlights

- To our knowledge, this is the first study using large language models, specifically GPT-4o, to extract cost-effectiveness analyses data (GPT) and compare it with (1) an existing database, ie, Tufts Cost-Effectiveness Analyses Registry and (2) researcher-validated extraction.
- The analysis found that GPT exhibited overall accuracy similar to Tufts Cost-Effectiveness Analyses Registry, with no statistically significant difference. However, GPT particularly underperformed in 4 variables: types of utilities, number of utilities, ratio quadrant, and number of ratios.
- Currently, large language models can assist researchers by automating simple variable extraction, filling missing data, and acting as a reference tool during manual data extraction to enhance efficiency under human supervision. Future research should explore batch data extraction methods and strategies for handling complex or subjective variables.

Introduction

With the advent of generative artificial intelligence (AI) tools, especially large language models (LLMs), such as Chat Generative Pre-Trained Transformer (ChatGPT) developed by Open AI, the applications of generative AI in various aspects of health economics and outcomes research (HEOR) are increasingly recognized and implemented.¹ LLMs process text by analyzing the probabilistic relationships between words and predicting the most probable next word based on learned patterns to generate responses.² For example, if a user inputs "The color of the rose is," the LLM calculates the probabilities of possible next words based on learned patterns from training data, which can be "red" with a probability of 0.6, "white" with 0.25, or "yellow" with 0.15. The model then selects the most probable next word based on the specific context to generate a response. The effectiveness of an LLM depends not only on the algorithm but also on the scale, diversity, and quality of its training data. For example, ChatGPT-4o is trained using self-supervised learning on diverse data sets, including publicly available books, research articles, and web-based documents.² This broad training corpus allows it to develop a structured understanding of various fields and synthesize information.

HEOR, particularly, can benefit from LLMs due to the availability of open-access publications, government reports, and academic websites. LLMs can leverage these sources to learn domain-specific knowledge to enhance research. One emerging application is the use of LLMs in conducting health economic modeling, but the reliability and accuracy of the outcomes of this approach remain uncertain.^{3,4} Another application is in evidence synthesis for systematic literature reviews. Researchers have utilized LLMs to automate various stages of the review process. These applications include streamlining the initial article screening^{5,6} and extracting data for analysis.⁵⁻¹⁰ These advancements highlight the power of generative AI in addressing traditional challenges in HEOR, though future investigation and improvement are necessary to understand the mechanisms and limitations.

However, to our best knowledge, we are not aware of any prior work that has specifically explored the use of LLMs for extracting cost-effectiveness analysis (CEA) data, particularly involving a large and comprehensive set of variables. This exploration of LLMs

for extracting CEA data is worthwhile for several reasons. First, the number of CEA papers has increased significantly since the first health technology assessment article was published in 1977,¹¹ generating extensive yet underutilized data. Second, the process of collecting CEA data is labor-intensive and time-consuming, involving extensive data extraction, validation, and documentation. The data extraction may also focus on adding new fields to records that had been previously extracted, for example, when new variables are introduced to an existing database. Finally, despite multiple procedures to ensure data quality, inaccurate information may still exist in the dataset. In several cases,^{5,7-9} the extraction and consensus processes were conducted internally without strict external monitoring and thus rely heavily on the integrity and consistency of the readers. As such, even with rigorous, bespoke protocols, potential bias may persist because of variability across individual readers, and changes in readers over time.

The Tufts Center for the Evaluation of Value and Risk in Health's CEA Registry (Tufts CEA Registry)¹² is one exemplar of a resource-intensive data collection process. It has collected studies from 1976 to date, giving researchers a valuable opportunity to study topics such as novel health measurement assessment,¹³ industry sponsorship bias in HEOR,¹⁴ meta-analysis for CEA,¹⁵ the development of CEA by geographic region,¹⁶ etc. Staffs conduct systematic literature searches first to identify studies that meet inclusion criteria, specifically (1) original CEA and (2) measures of health benefits reported in quality-adjusted life years (QALYs) or disability-adjusted life years. After this, 2 trained readers, usually a master or doctoral student in relevant fields, extract data from selected articles, after which a consensus meeting is held to ensure the quality and consistency of the extracted data. It is estimated that, as of 2022, over 40 000 person-hours have been invested to build the registry.¹⁷

This study aims to evaluate the possibility and accuracy of LLMs, specifically GPT, in extracting CEA data from studies. By comparing GPT's extracted data to the Tufts CEA Registry and researcher-validated data, we assess GPT's concordance rates: the degree to which outputs match established datasets. Unlike previous studies comparing LLM-extracted data with individual human extractions,^{5,7-10} our study benchmarks GPT against validated and widely used Tufts CEA Registry, ensuring a more reliable assessment of its accuracy. To our knowledge, no prior research has evaluated an LLM's ability to extract variables essential for replicating cost-effectiveness models. This research will focus on GPT's ability to accurately identify, classify, and structure variables specific to HEOR, such as incremental cost-effectiveness ratios. As transparency and reproducibility gain importance in HEOR, this study marks a key step in exploring LLMs' accuracy, scalability, and potential in automated CEA data extraction and prior model replication.

Methods

Data Source and Prompts for GPT

Three data sources were used for comparison. The first one is Tufts CEA Registry data (TCRD), the largest CEA dataset in the world, including over 200 variables for over 13 000 CEA articles.¹² The second source of data is a researcher-validated extraction (RVE), in which researchers read the articles and extracted the variables manually. Three researchers were responsible for the RVE data collection. All of them have contributed at least 320 hours of data collection for at least 100 articles at Tufts Medical Center. The RVE data source is viewed as the gold standard in this study. The third data source is based on the LLM. Generative Pre-

trained Transformer 4o (GPT-4o) was selected as the base model. Based on that, we built a customized GPT called "Cost-effectiveness Analysis Data Extractor" (<https://chatgpt.com/g/g-A1jiDLrgl-cost-effectiveness-analysis-data-extractor>), which can automatically capture all the predetermined variables in an uploaded article and present the result in a structured table. This custom GPT is designed exclusively for CEA data extraction and related questions. For the sake of simplicity, "CEA Registry Data," "researcher-validated extraction," and "GPT-extracted data" were noted as "TCRD," "RVE," and "GPT," respectively, in the remaining sections of this study.

GPT-4o has been available for paid users since May 2024.¹⁸ As of November 2024, the customized GPT, which is based on GPT-4o, is not free. We used this paid version for 3 reasons. First, the custom feature allows users to tailor GPT for specific topics by combining prompts and knowledge, which is essential for focusing on specialized domains, such as HEOR. Second, custom GPT can be published and shared, allowing consistent use across multiple users for collaborative tasks. Free users can access shared GPT to a limited extent but cannot create or share their own. Third, GPT-4o was one of the top-performing models at the time this study was initiated (November 2024), newer models, such as ChatGPT-o3 and Deepseek-R1, had not yet been released. GPT-4o achieved an 88% score on the Massive Multitask Language Understanding benchmark,¹⁹ which tests reasoning across 57 subjects, making it the most advanced model available. Of note, other models lack the ability to customize and share GPTs for collaborative use.

The prompt for GPT was designed based on prompt engineering and adhered to 4 principles to optimize efficiency and accuracy^{20,21}: (1) Explicit Task Instructions: the prompt specified the exact task, directing GPT to read the provided article and extract specified data; (2) Contextual Background: to enhance relevance, the prompt situated the task within HEOR by incorporating domain-specific terminology and referencing economic evaluation frameworks, ensuring GPT understood the focus of the extraction; (3) Input and Output Specifications: the expected input was a PDF file, and the output was a structured table displaying 36 extracted variables in a predetermined order; and (4) Supplemental Reference Material: 2 additional files (see Appendix A and B in Supplemental Materials found at <https://doi.org/10.1016/j.jval.2025.05.008>), a variable list and a codebook, were uploaded alongside the prompt. These files contained definitions, data types, and potential values for each variable, clarifying the extraction criteria. A zero-shot prompting strategy was applied, wherein no examples of ideal output or context were provided.^{9,22-24} An iterative prompt refinement approach was used to improve the quality and structure of the output: the GPT-generated results were reviewed, and the prompts were adjusted accordingly before retesting the performance. This process was repeated 6 times in our study until the extracted data aligned with the specified format. By systematically refining the prompts, this approach improved both the reliability and reproducibility of the final extracted data set. The final iteration of the prompt, presented in Appendix C in Supplemental Materials found at <https://doi.org/10.1016/j.jval.2025.05.008>, was used to generate the "Cost-effectiveness Analysis Data Extractor" in conjunction with Appendix A and B in Supplemental Materials found at <https://doi.org/10.1016/j.jval.2025.05.008>.

Article and Variable Selection

The articles to be reviewed were selected based on 5 criteria:

1. Only articles included in TCRD were considered, enabling consistency and ease of comparison with GPT and RVE.

- Articles must include the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) Checklist (either the 2013 or the 2022 version).^{25,26} We applied this strict inclusion criterion, prioritizing studies that adhered to standardized reporting frameworks (CHEERS), which provided complete data for key CEA variables. This approach minimized variability in reporting practices, allowing a more controlled evaluation of GPT's extraction ability. Although this resulted in a smaller sample size, it ensured that the extracted data could be systematically compared, strengthening the internal validity of our findings.
- Only articles published in or after 2013 were selected because the CHEERS Checklist was first introduced in 2013.²⁵
- In this study, we only included articles with a focus on the US market, ensuring comparability across the selected articles.
- Only studies reporting cost-effectiveness in terms of \$/QALY were included.

Additionally, articles with missing PubMed ID or marked as partial reviews by Tufts were excluded. According to Tufts CEA Registry, partial reviews involved extracting only a limited set of variables and typically applied to publications in journals with an impact factor < 2 or articles not published in predefined priority journals.

The variables selected for comparison were based on both the available variables in TCRD and the contents of the CHEERS Checklist 2022. A total of 36 variables were chosen, covering 6 areas: (1) Study Identification & Publication Details; (2) Study Design & Methodology; (3) Population & Intervention Details; (4) Incremental Cost-Effectiveness Ratios; (5) Utility; and (6) Currency & Year Details. The details of these variables were provided in Appendix A, and their codebook was included in Appendix B in Supplemental Materials found at <https://doi.org/10.1016/j.jval.2025.05.008>.

Extraction Process

RVE was conducted first before the other 2 data extractions, to minimize bias in the researchers' data extraction after reviewing the results from the other 2 methods. Then, for GPT, each article was uploaded to the custom GPT, which generated a structured table of variables. Lastly, we downloaded the TCRD data set on

November 19, 2024, and used R Studio²⁷ to select the specific 36 variables for 34 articles.

Comparison

We constructed 3 comparison groups to evaluate data accuracy across sources, with RVE serving as the gold standard (right answer) to benchmark the accuracy of other data sources. The groups are defined as follows:

- GPT & TCRD: This group assesses the extent to which GPT replicates data collection tasks performed for TCRD. The concordance rate reflects the matching rate between GPT and TCRD.
- GPT & RVE: This group evaluates the accuracy of GPT in extracting data consistent with standards set by trained researchers. The concordance rate reflects the accuracy of GPT.
- TCRD & RVE: This group examines the reliability and completeness of TCRD by comparing it with researcher-validated data, identifying discrepancies or missing information. The concordance rate reflects the accuracy of TCRD.

For all comparisons, an exact match between data sources is coded as "1" (fully concordant), and a nonmatch is coded as "0" (nonconcordant). Partial matches are assigned a concordance score based on the percentage alignment between values, ranging from 0 to 1. The detailed information on comparison criteria for each variable is in Appendix A in Supplemental Materials found at <https://doi.org/10.1016/j.jval.2025.05.008>. The final concordance rate for each article is calculated by dividing the total concordance score by the total number of variables. For example, an article achieving a score of 30/36 would have a concordance rate of 0.833. The mean concordance rate for each method (ie, GPT, TCRD, and RVE) is calculated as the total score divided by the total number of articles. Variable-specific concordance rates are reported to identify underperformed variables.

Sensitivity Analysis

Different wording or structure of prompts may yield different outputs, which is an inherent characteristic of LLMs. To assess the robustness of the GPT, which is how much the results depend on the prompt phrasing, we conducted a sensitivity analysis using

Figure 1. Flow diagram of articles included in the study.

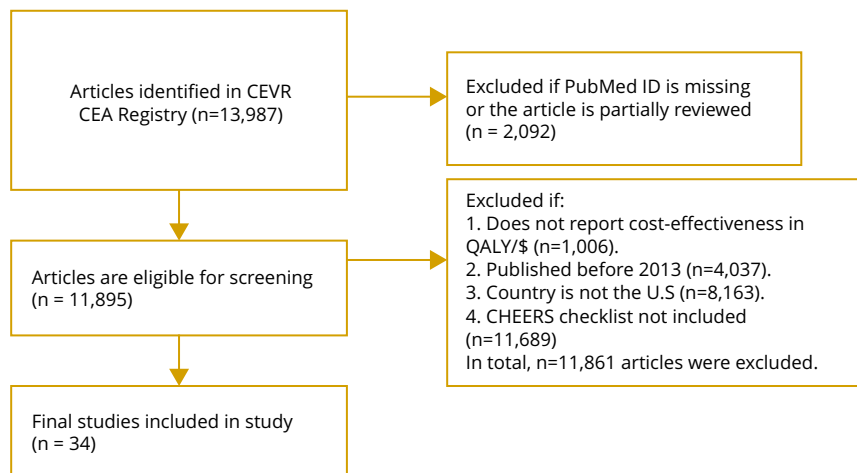
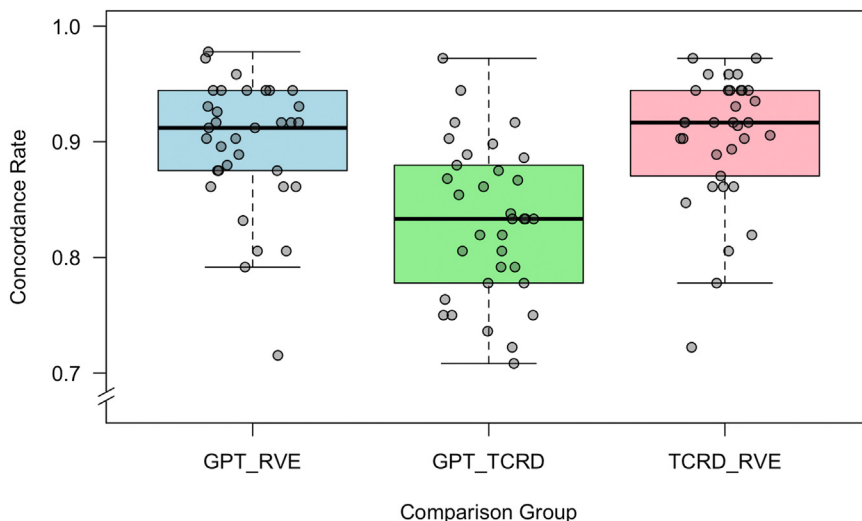


Figure 2. Boxplot for the concordance rates for 3 comparison groups. TCRD_RVE: comparison between the registry and researcher-validated data. GPT_TCRD: comparison between GPT-extracted data and the registry. GPT_RVE: comparison between GPT-extracted data and researcher-validated data. The box is the IQR, and bold line in the middle of the box is median. The double slash on the y-axis indicates that the value is not start from origin.



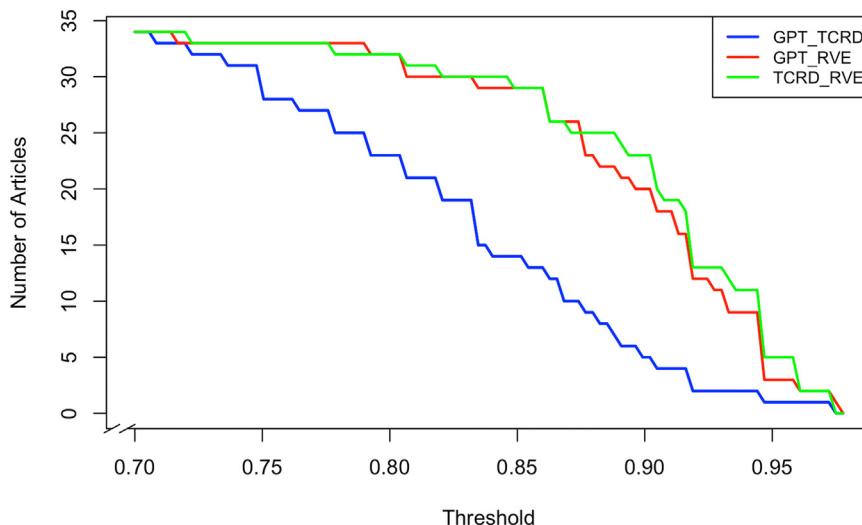
GPT indicates ChatGPT model.

the base-case prompts, along with 4 alternative versions with the same semantic content but different expression styles (see Appendix C in Supplemental Materials found at <https://doi.org/10.1016/j.jval.2025.05.008>). We used the prompt sensitivity score (PSS) to quantify the sensitivity.²⁸ This analysis helps evaluate whether results are replicable and consistent under minor prompt changes. Details can be found in Appendix D in Supplemental Materials found at <https://doi.org/10.1016/j.jval.2025.05.008>.

Results

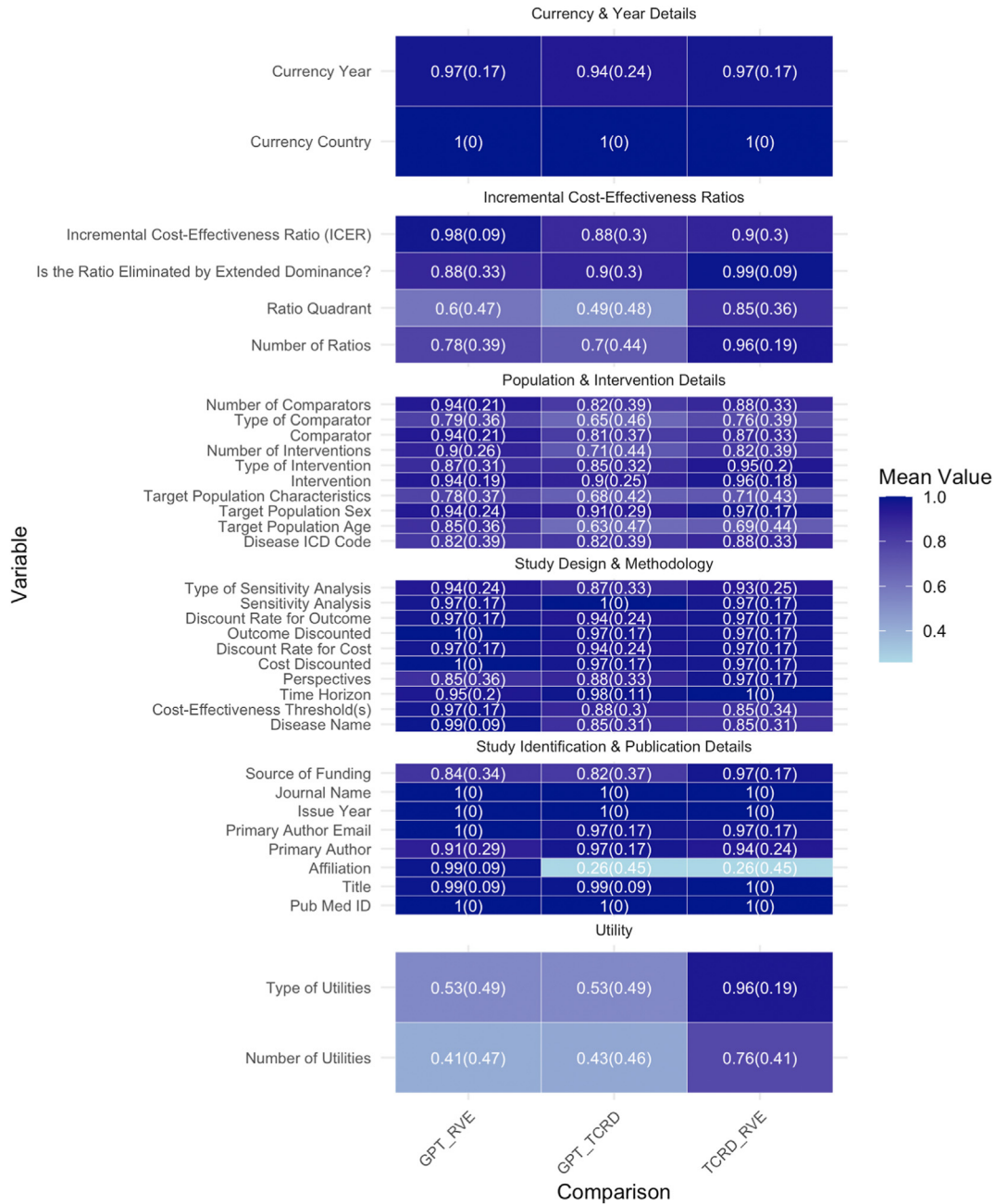
From the initial 13 987 articles in TCRD, 11 895 remained after excluding articles with missing identifiers or articles categorized in the TCRD database as “partially reviewed.” A total of 11 861 articles were excluded based on the inclusion criteria. Based on absolute totals, 1006 did not report \$/QALY, 4037 were published before 2013, 8163 did not focus on the US setting, and 11 689 did not include the

Figure 3. Step plot for the number of articles meeting dynamic concordance rate thresholds for 3 comparison groups. TCRD_RVE: comparison between the registry and researcher-validated data. GPT_TCRD: comparison between GPT-extracted data and the registry. GPT_RVE: comparison between GPT-extracted data and researcher-validated data. Each line represents the accumulate count of articles in each comparison group meets the threshold. The double slash on the x-axis indicates that the value is not start from origin.



GPT indicates ChatGPT model.

Figure 4. Heatmap for the mean concordance rates for various variables across 3 comparison groups. TCRD_RVE: comparison between the registry and researcher-validated data. GPT_TCRD: comparison between GPT-extracted data and the registry. GPT_RVE: comparison between GPT-extracted data and researcher-validated data. The name on the side is the name for specific variable, grouped by several categories: (1) Study Identification and Publication Details; (2) Study Design and Methodology; (3) Population and Intervention Details; (4) Incremental Cost-Effectiveness Ratios; and (5) Utility. The color gradient reflects the mean concordance rate for 34 articles in each group, with darker shades meaning higher value. The number in the brackets is the SD for the corresponding variable in each group.



GPT indicates ChatGPT model; ICD, International Classification of Diseases.

CHEERS checklist. Consequently, 34 articles met the criteria and were included in the analysis (Fig. 1).

The accuracy of GPT (GPT & RVE) and TCRD (TCRD & RVE) is 0.88 (SD: 0.06) and 0.90 (SD: 0.06) respectively, with a paired student's *t* test *P* value .7128. The matching rate of GPT and TCRD is 0.82 (SD: 0.07). The boxplot in Figure 2 shows the distribution of concordance rates across 3 comparison groups. TCRD & RVE (median: 0.91, IQR:

0.07) and GPT & RVE (median: 0.89, IQR: 0.07) show the high medians and the narrowest IQRs, but GPT & TCRD (median: 0.83, IQR: 0.11) shows some variability and more outliers compared with others. Figure 3 evaluates the number of articles meeting dynamic concordance thresholds across 3 comparison groups. At lower thresholds (eg, 0.7), all groups exhibit high concordance. As the threshold increases (eg, 0.8), differences emerge, with GPT & RVE

Table 1. Prompt sensitivity scores by extracted variable across 5 prompt versions.

Variable	PSS
PubMed ID	0.00
Title	0.03
Affiliation	0.00
Primary author	0.02
Primary author email	0.00
Issue year	0.00
Journal name	0.00
Source of funding	0.27
Disease name	0.01
Disease ICD code	0.04
Cost-effectiveness threshold(s)	0.04
Time horizon	0.00
Perspective	0.03
Cost discounted	0.00
Discount rate for cost	0.01
Outcome discounted	0.00
Discount rate for outcome	0.01
Currency country	0.00
Currency year	0.06
Sensitivity analysis	0.00
Type of sensitivity analysis	0.04
Target population age	0.05
Target population sex	0.24
Target population characteristics	0.05
Intervention	0.04
Type of intervention	0.16
Number of interventions	0.09
Comparator	0.02
Type of comparator	0.17
Number of comparators	0.08
Number of ratios	0.22
Ratio quadrant	0.14
Is the ratio eliminated by extended dominance?	0.09
Incremental cost-effectiveness ratio	0.04
Number of utilities	0.18
Type of utilities	0.31
Average PSS	0.07

ICD indicates International Classification of Diseases; PSS, prompt sensitivity score.

maintaining the highest number of articles above the threshold, followed by TCRD & RVE and GPT & TCRD.

The heatmap in Figure 4 evaluates 6 categories of variables, showing concordance rates across 3 comparison groups. High concordance rates (mean ≥ 0.8) are observed for sections “Study Design & Methodology,” “Currency & Year Details,” and “Study Identification & Publication Details.” However, variable “Affiliation” performs poorly in GPT & TCRD and TCRD & RVE (mean: 0.26, SD: 0.45), driven by

missing values in the TCRD. “Population & Intervention Details” sees high performance in GPT & RVE but lower concordance and higher variability in GPT & TCRD and TCRD & RVE for variables such as “Number of Comparators” and “Target Population Age.” “Utility” shows the largest discrepancies, with GPT-based methods underperforming significantly (mean: 0.41-0.53) compared with TCRD & RVE (mean: 0.96-0.76).

For the sensitivity analysis, we calculated PSS across 36 variables from 34 articles, using 5 different prompt versions. Table 1 shows that the overall average PSS was 0.07, indicating that the model generally produced consistent outputs regardless of prompt wording. This low level of variation suggests strong generalization and reproducibility of the model’s responses overall. However, a few variables demonstrate substantially higher PSS, indicating greater variability and reduced reliability in those specific areas: type of utilities (0.31), target population sex (0.24), source of funding (0.27), and number of ratios (0.22). In contrast, many simpler or more explicitly stated variables, such as PubMed ID, Affiliation, Issue Year, Journal Name, and Cost Discounted, exhibit a 0 PSS, suggesting perfect agreement across prompts.

Discussion

This study assessed the feasibility and accuracy of GPT-extracted data compared with RVE and TCRD. Overall, GPT demonstrated a high concordance rate compared with TCRD and RVE, although it varied on specific variables. The non-significant *P* value .7128 indicated that no statistically significant difference in accuracy between GPT and TCRD. The distribution of TCRD & RVE (accuracy of TCRD) indicated that TCRD generally provides reliable and accurate information, whereas some outliers occurred. The distribution of GPT & RVE (accuracy of GPT) showed that GPT aligned well with RVE. Although the distribution of GPT & TCRD (matching rate between GPT & TCRD) showed variability and more outliers compared with others, it still had ability to replicate the data collection tasks. For specific variables, the high concordance rate in Study Design & Methodology, Currency & Year Details, and Study Identification & Publication Details across 3 groups is likely due to their standardized nature and lower complexity. In contrast, the low concordance rate for “Affiliation” highlighted the influence of missing data in the TCRD, in which only 6 out of 34 “Affiliation” can be found in TCRD. Similarly, variability in “ratio quadrant” suggested that interpreting cost-effectiveness data, particularly its classification based on knowledge and data, can be a challenge for both GPT and RVE. Also, nonstandardized or complex variables, such as “Type of Comparator” and “Number of Interventions” can be difficult to extract. The result for “Utility” showed the most significant gaps, with GPT failing to extract consistently. This may be because some utility names are complex. Additionally, CEA models often incorporate multiple utilities, which can be numerous and detailed, making it difficult for GPT to extract and accurately classify them. In summary, this study demonstrated the promising role of generative AI, specifically utilizing a custom GPT in automating extraction of data from CEA literature. By successfully retrieving complex variables with a high degree of concordant rate compared with the other 2 sources, this approach demonstrated several opportunities for refining and expanding CEA data collection. Through these results, we can draw meaningful insights into both the advantages and current limitations of using generative AI in the field of HEOR.

Opportunities for CEA Data Collection

A key opportunity identified is the potential for LLMs to standardize TCRD by identifying and resolving missing or

Figure 5. Example of GPT's step-by-step instructions to calculate the incremental cost-effectiveness ratio. This is a screenshot from the custom GPT: Cost-effectiveness Analysis Data Extractor. The user's query is in the top right corner and the GPT's answer appears after the user's query.

How do I calculate incremental cost-effectiveness ratios (ICERs) in here?

To calculate Incremental Cost-Effectiveness Ratios (ICERs), the formula is as follows:

$$\text{ICER} = \frac{\text{Cost}_{\text{Intervention}} - \text{Cost}_{\text{Comparator}}}{\text{Effectiveness}_{\text{Intervention}} - \text{Effectiveness}_{\text{Comparator}}}$$

Where:

- **Cost** is the total cost (e.g., discounted lifetime cost per patient).
- **Effectiveness** is the health benefit (e.g., QALYs or life-years gained).
- The numerator represents the incremental cost of the intervention compared to the comparator.
- The denominator represents the incremental effectiveness of the intervention compared to the comparator.

Application to the Provided Study

From the clinical trial model for intensive GDMT (HIC):

- **Cost (HIC)** = \$67,300
- **Cost (UC)** = \$61,100
- **QALYs (HIC)** = 5.97
- **QALYs (UC)** = 5.16

$$\text{ICER}_{\text{HIC}} = \frac{67,300 - 61,100}{5.97 - 5.16} = \frac{6,200}{0.81} = \$7,654 \text{ per QALY (approximately } \$8,000/\text{QALY as reported).}$$

From the clinical trial model for high-intensity care plus (HIC-P):

- **Cost (HIC-P)** = \$158,000

Can you elaborate on sensitivity analysis? What influences ICER variations the most?

Message Cost-effectiveness Analysis Data Extractor

ChatGPT can make mistakes. Check important info.

GPT indicates ChatGPT model; QALY, Quality-adjusted life-years.

inconsistent data. Given the TCRD's long history and the periodic introduction of new variables, older entries often lack recently introduced variables or exhibit inconsistencies because of evolving definitions. For instance, in cases in which the first author differs from the corresponding author, TCRD occasionally records the first author as the "Primary Author" but links the "Primary Author's Email" to the corresponding author. Similarly, a recently added, but not included in this study, variable "Source Code," tracking whether authors share code for transparency, is available only for 25 out of 13 987 articles. All 25 articles were published after 2023. LLMs could efficiently identify and correct these discrepancies, ensuring more comprehensive and accurate data sets.

Additionally, the study highlights a less-than-perfect concordance rate between the TCRD and RVE, suggesting potential biases in human data collection. Variations often arise with

subjective variables (eg, "target population characteristics") as opposed to more objective ones (eg, "type of sensitivity analysis"). This variability is potentially exacerbated in the TCRD, in which numerous readers contribute, each potentially interpreting variables differently. An LLM-driven approach offers a solution by applying a single, consistent perspective across all articles and enhancing consistency in variable classification. Beyond standardization, LLMs can also analyze and provide insights beyond variable classification. The model can interpret and generate narrative explanations, contextual insights, and summaries useful for exploratory analyses. For instance, GPT could provide steps in calculating specific incremental cost-effectiveness ratios (Fig. 5), summarizing key points of each study or identifying related research, and supporting deeper, more flexible exploration of data sets.

Limitation of This Study

However, this study has some limitations. First, it focused on only a subset of variables within TCRD, excluding certain complex variables, such as “Global Burden of Disease Tier.” Including more variables may yield different outcomes, and future research should consider these to fully evaluate the model’s capabilities.²⁹ The small sample size (34 out of 13 987) reflects the strict inclusion criteria necessary to ensure consistent and structured reporting for evaluating LLM’s performance. Although this practice improved internal validity, it also limited external validity to other CEA studies with incomplete or inconsistent variables reporting, potentially challenging GPT’s ability to interpret non-standardized variable names and concepts. Additionally, it could also exclude studies with comprehensive variables that did not cite or report the CHEERS checklist. Future research should use other criteria or quality benchmarks to expand the study sample while maintaining methodological rigor.

This study opted not to report traditional agreement statistics, such as Cohen’s kappa or the concordance correlation coefficient^{5,29} because many variables in our analysis have values on a continuum between 0 and 1, making these measures less appropriate. Additionally, the small sample size of 36 variables across 34 articles limited the utility and interpretability of such statistical measures in this context.

Another limitation of this study is the use of a custom GPT interface instead of an API. An API offers greater customization for the result, such as adjusting parameters, for example, “temperature” (influences the balance between creativity and accuracy) and “frequency penalty” (reduces repetition). These features could streamline large-scale screening and storage of thousands of articles in a database. However, our custom GPT approach addressed key API drawbacks: API requests are stateless, meaning that they lack memory of prior interactions, making follow-up questions challenging with significant coding adjustments. This limits dynamic, multistep interactions necessary for extracting complex CEA data.

Furthermore, in our GPT data extraction process, the GPT occasionally returned general response like “Multiple” for the “number of utilities.” However, when prompted further, it was able to provide the specific number for each specific measure (see example in Appendix E in Supplemental Materials found at <https://doi.org/10.1016/j.jval.2025.05.008>). Therefore, how to design the prompt strategy to make the output more accurate, reliable, and reproducible will be an important area for future research.

Challenge and Consideration for the Future

Despite the phenomenal progress of LLMs in recent years, human feedback remains critical,³⁰ especially from experts in specific fields. This can be structured into 3 pillars of LLM: input, algorithms, and output. First, the quality of input data influences its performance. Even the best model cannot generate reliable output based on poor sources. Utilizing high-quality inputs, such as peer-reviewed research or experts’ responses, ensures a robust knowledge foundation. Human screening and selection help filter irrelevant or unreliable data before training. Second, experts can improve model algorithms by offering structured feedback and correcting errors, refining the model iteratively and reducing hallucinations in specialized contexts. Finally, the model-generated responses require continuous validation by experts to assess accuracy and relevance to domain-specific contents. In HEOR, human review ensures alignment with theory or standards, such as the CHEERS checklist.

Another future consideration is whether to build on an existing model or create a domain-specific one. Established models offer fast deployment and broad knowledge but risk misinterpreting domain-specific terms or hallucinating because of a lack of targeted training. In contrast, a “HEOR-AI” could deliver higher precision with curated inputs, although it demands substantial data and computing power and may lack flexibility across HEOR’s interdisciplinary scope. A balanced approach, fine-tuning existing models with continuous expert input, offers the best of both: broad capabilities with domain-specific accuracy. Ultimately, expert involvement across the input-algorithm-output chain remains vital.

Conclusions

To our best knowledge, this study is the first to utilize an LLM for extracting data from CEA articles and compare its concordance rates with TCRD and RVE. The findings are promising but not conclusive, showing that GPT can achieve high concordance rates in data extraction. The performance of GPT, however, depends on article structure, specific variables, and prompt engineering. Notably, GPT outperformed manual data extraction in some cases, offering comparable results with lower resource investment. Although this approach has the potential to expedite data collection for institutions, researchers’ oversight and expertise remain critical to ensure data quality. In summary, our findings highlight GPT’s potential to improve the efficiency and scalability of data collection for HEOR applications.

Author Disclosures

Author disclosure forms can be accessed below in the Supplemental Material section. Dr Slejko is an editor for *Value in Health* and had no role in the peer-review process of this article.

Supplemental Material

Supplementary data associated with this article can be found in the online version at <https://doi.org/10.1016/j.jval.2025.05.008>.

Article and Author Information

Accepted for Publication: May 9, 2025

Published Online: June 27, 2025

doi: <https://doi.org/10.1016/j.jval.2025.05.008>

Author Affiliations: Department of Practice, Sciences, and Health Outcomes Research, University of Maryland Baltimore, Baltimore, MD, USA (Gu, Zafari, Slejko, Onukwugha); University of Texas at Austin College of Pharmacy, Austin, TX, USA (Zhang); Lumanity, Boston, MA, USA (Patil); The University of Maryland-Institute for Health Computing (UM-IHC), North Bethesda, MD, USA (Zafari).

Correspondence: Xujun Gu, MSPH, Department of Practice, Sciences, and Health Outcomes Research, University of Maryland Baltimore, Baltimore, MD 21201, USA. Email: xgu@umaryland.edu

Authorship Confirmation: All authors certify that they meet the ICMJE criteria for authorship.

Funding/Support: The authors received no financial support for this research.

Acknowledgment: The authors thank The Center for the Evaluation of Value and Risk in Health at Tufts Medical Center for generously granting student access to the CEA Registry.

REFERENCES

1. Fleurence RL, Bian J, Wang X, et al. Generative AI for health technology assessment: opportunities, challenges, and policy considerations - an ISPOR working group report. *Value Health*. 2025;28(2):175–183.
2. Open AI. How ChatGPT and our foundation models are developed. <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed>. Accessed March 20, 2025.
3. Reason T, Rawlinson W, Langham J, Gimblett A, Malcolm B, Klijn S. Artificial intelligence to automate health economic modelling: a case study to evaluate the potential application of large language models. *Pharmacoecon Open*. 2024;8(2):191–203.
4. Chhatwal J, Yildirim I, Balta D, et al. Can large language models generate conceptual health economic models? *Value Health*. 2024;27(6):S123.
5. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. *J Med Internet Res*. 2024;26(1):e48996.
6. Khraisha Q, Put S, Kappenberg J, Warritch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res Synth Methods*. 2024;15(4):616–626.
7. Schopow N, Osterhoff G, Baur D. Applications of the natural language processing tool ChatGPT in clinical practice: comparative study and augmented systematic review. *JMIR Med Inform*. 2023;11(1):e48933.
8. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*. 2017;7(2):e012545.
9. Reason T, Langham J, Gimblett A. Automated mass extraction of over 680,000 PICO from clinical study abstracts using generative AI: a proof-of-concept study. *Pharm Med*. 2024;38(5):365–372.
10. Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. *Syst Rev*. 2024;13(1):158.
11. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med*. 1977;296(13):716–721.
12. Center for the Evaluation of Value and Risk in Health. The cost-effectiveness analysis registry. Institute for Clinical Research and Health Policy Studies, Tufts Medical Center. Center for the Evaluations of Value and Risk in Health. www.cearegistry.org. Accessed November 11, 2024.
13. Basu A, Carlson J, Veenstra D. Health years in total: a new health objective function for cost-effectiveness analysis. *Value Health*. 2020;23(1):96–103.
14. Xie F, Zhou T. Industry sponsorship bias in cost effectiveness analysis: registry based analysis. *BMJ*. 2022;377:e069573.
15. Rosettie KL, Joffe JN, Sparks GW, et al. Cost-effectiveness of HPV vaccination in 195 countries: a meta-regression analysis. *PLoS One*. 2021;16(12):e0260808.
16. Butt T, Liu GG, Kim DD, Neumann PJ. Taking stock of cost-effectiveness analysis of healthcare in China. *BMJ Glob Health*. 2019;4(3):e001418.
17. A New Era for the CEA registry - May 2022 webinar. YouTube. <https://www.youtube.com/watch?v=vw1nIWAS-PY>; Published 2022. Accessed November 3, 2024.
18. Open AI. Introducing GPT-4o and more tools to ChatGPT free users. <https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/>. Accessed March 20, 2025.
19. Open AI. Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>. Accessed March 20, 2025.
20. Wang J, Shi E, Yu S, et al. Prompt engineering for healthcare: methodologies and applications. Cornell University. <https://arxiv.org/abs/2304.14670>; Published March 23, 2024. Accessed November 10, 2024.
21. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. *Ann Biomed Eng*. 2023;51(12):2629–2633.
22. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Med Inform*. 2024;12(1):e55318.
23. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc*. 2024;31(9):1812–1820.
24. Yim RP, Rudrapatna VA. Zero-shot prompting is the most accurate and scalable strategy for abstracting the mayo endoscopic subscore from colonoscopy reports using GPT-4. <https://www.medrxiv.org/content/10.1101/2024.03.22.24304745v1>; Published March 24, 2024. Accessed November 10, 2024.
25. Husereau D, Drummond M, Petrou S, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *BMJ*. 2013;346:f1049.
26. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. *BMC Med*. 2022;20(1):23.
27. RStudio Team. RStudio: Integrated Development for R. <https://www.posit.co/>; Published 2024. Accessed November 28, 2024.
28. Zhuo J, Zhang S, Fang X, Duan H, Lin D, Chen K. ProSA: assessing and understanding the prompt sensitivity of LLMs. Cornell University. <https://arxiv.org/abs/2410.12405>; Published October 16, 2024.
29. Hsueh JY, Nethala D, Singh S, et al. Exploring the feasibility of GPT-4 as a data extraction tool for renal surgery operative notes. *Urol Pract*. 2024;11(5):782–789.
30. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Cornell University. <https://arxiv.org/abs/2203.02155>; Published March 4, 2022. Accessed March 15, 2022.