PharmacoEconomics - Open (2024) 8:205–220
https://doi.org/10.1007/s41669-024-00476-9

ORIGINAL RESEARCH ARTICLE

# Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models

Tim Reason[1] · Emma Benbow[1] · Julia Langham[1] · Andy Gimblett[1] · Sven L. Klijn[2] · Bill Malcolm[3]

Accepted: 1 February 2024 / Published online: 10 February 2024

## Abstract

**Background** The emergence of artificial intelligence, capable of human-level performance on some tasks, presents an opportunity to revolutionise development of systematic reviews and network meta-analyses (NMAs). In this pilot study, we aim to assess use of a large-language model (LLM, Generative Pre-trained Transformer 4 [GPT-4]) to automatically extract data from publications, write an R script to conduct an NMA and interpret the results.

**Methods** We considered four case studies involving binary and time-to-event outcomes in two disease areas, for which an NMA had previously been conducted manually. For each case study, a Python script was developed that communicated with the LLM via application programming interface (API) calls. The LLM was prompted to extract relevant data from publications, to create an R script to be used to run the NMA and then to produce a small report describing the analysis.

**Results** The LLM had a > 99% success rate of accurately extracting data across 20 runs for each case study and could generate R scripts that could be run end-to-end without human input. It also produced good quality reports describing the disease area, analysis conducted, results obtained and a correct interpretation of the results.

**Conclusions** This study provides a promising indication of the feasibility of using current generation LLMs to automate data extraction, code generation and NMA result interpretation, which could result in significant time savings and reduce human error. This is provided that routine technical checks are performed, as recommend for human-conducted analyses. Whilst not currently 100% consistent, LLMs are likely to improve with time.

**Key Points for Decision Makers**

This is a promising first assessment of the feasibility of using LLMs to automate data extraction, analysis and result interpretation, which could result in significant time savings and reduce human error in the NMA process.

This study has shown that GPT-4 can successfully replicate the results of four NMAs in two disease areas for two outcome types (binary and time-to-event).

There is a need for further research to develop and test LLM-based processes.

## 1 Introduction

Many countries have health technology assessment (HTA) agencies to systematically evaluate the efficacy, safety and value of new interventions to inform price negotiations and reimbursement decisions [1, 2]. Pharmaceutical companies are very often required to submit robust evidence of the cost-effectiveness of the new intervention compared with the currently available treatments (comparators) using data that have been acquired in a systematic, non-biased and transparent way. Systematic literature reviews (SLRs) and meta-analyses are considered the most rigorous methods for gathering and synthesising such evidence [3]. In addition to HTAs, SLRs and meta-analysis are also integral for other evidence-based practices including informing clinical decision making, clinical guidelines, medical education and policy [4].

Extended author information available on the last page of the article

The HTA process generally involves several key steps [5], and an overview is shown in Fig. 1. Firstly, SLRs are conducted to identify relevant clinical studies of the intervention and comparators, preferably head-to-head evidence from randomized controlled trials (RCTs). Secondly, data from the studies identified in the SLR are extracted and then synthesized using statistical techniques to determine the relative effectiveness. Typically, due to an absence of direct head-to-head evidence for all the different comparators, indirect or mixed treatment comparisons using methods such as network meta-analyses (NMA) are employed at this stage, which compare multiple treatments simultaneously using direct and indirect evidence across a "network" of RCTs [6]. These methods require the assessment of clinical and methodological heterogeneity and transitivity of the trials included [7]. Finally, economic models are developed to determine the cost-effectiveness of the intervention versus the comparators, populated using the clinical data obtained, and statistical data generated in the above steps.

The HTA process is very labour and resource intensive, requiring a large team of experts in the field of health economic outcomes research (HEOR), including SLR analysts, statisticians and economic modellers [5, 8]. For example, to reduce errors, there is a requirement for at least three analysts to be involved in the SLR and data extraction: two analysts to perform dual independent screening and duplicated data extraction or data checking and a third analyst to resolve any issues [9]. Despite these quality-control measures, the processes can still be prone to error [10]. In addition, the HTA process is time-consuming, generally taking over a year to complete, with the SLR and NMA part of this process taking several months each [8, 11]. A faster HTA process would ultimately mean faster access to new treatments and, therefore, better outcomes for patients.

Artificial intelligence (AI) tools, especially generative AI with large language models (LLMs), have the potential to optimise many steps of the HTA workflow, making the process quicker, less costly, more efficient and less error-prone [12, 13]. The latest LLMs, including GPT-4 [14], have been trained using vast quantities of publicly available text data to understand, generate and manipulate human language by recognising patterns and relationships in language [15]. They are capable of human-level performance on some programming and analytics tasks [16]; however, the practical use of these AI models remains untested in HEOR.

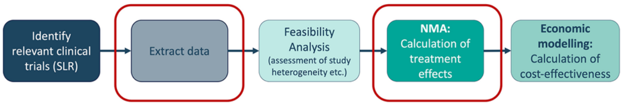

Identify relevant clinical trials (SLR)
Extract data
Feasibility Analysis (assessment of study heterogeneity etc.)
NMA: Calculation of treatment effects
Economic modelling: Calculation of cost-effectiveness

**Fig. 1** Overview of HTA process including NMA. Use of LLMs to automate the process have been applied within this study to those steps in the process highlighted with a red box. *HTA* health technology assessment, *LLM* large language model, *NMA* network meta-analysis

The aim of this study is to perform a feasibility assessment of LLM-based automated NMA, using previous manually conducted NMAs as case studies. The scope of the work was to develop LLM-based processes for automated data extraction, software programming to perform the NMA and interpretation of the results from the NMA (highlighted with red in Fig. 1). Manual assessments of study heterogeneity and suitability of studies for inclusion in an NMA were conducted prior to automation.

## 2 Methods

### 2.1 Case Studies

The ability of the LLM to replicate the results of manually conducted NMAs was tested using four case studies. For each of these case studies, a literature review had been conducted to identify relevant trials, followed by a feasibility analysis to determine which trials were appropriate to include in the NMA, i.e. involving a review of the study design, patient characteristics and outcomes to determine whether the trials were sufficiently similar to include in the NMA [17]. The four case studies spanned two disease areas (hidradenitis suppurativa [HS], which is a chronic, inflammatory skin disorder, and non-small cell lung cancer [NSCLC]) and two types of outcome (binary and time-to-event [survival]). These outcomes were chosen because if a prototype could be shown to work for binary or time-to-event outcomes, then it should be generalisable to other outcome types.

We have implicitly assumed that all studies included in the analyses were sufficiently homogeneous to be combined based on a previous publication [18] (NSCLC) and from a topline manual check of study design and characteristics (HS).

Case study 1 involved an indirect comparison of the efficacy of treatments for patients with moderate-to-severe hidradenitis suppurativa (unpublished literature review and analysis). The literature review had identified six relevant trials evaluating the clinical response to different treatments (adalimumab, secukinumab and bimekizumab) in this patient population, and the feasibility analysis determined that all six trials were suitable to include in the NMA. The network diagram for the analysis is shown in Fig. S1 and the trials and clinical response data are summarised in Table S1 (Online Resource).

Case studies 2, 3 and 4 concerned the efficacy of second-line treatments for patients with NSCLC. The SLR was originally conducted in 2018 and updated in 2021 [18]. Case study 2 involved treatments and outcome data (overall survival [OS]), which were used in the primary analysis (base case) of an economic model. The feasibility analysis identified five trials reporting on OS across relevant treatments (nivolumab, pembrolizumab and atezolizumab, with docetaxel as the common comparator treatment) that were appropriate for inclusion in the NMA (Fig. S2 and Table S2 [Online Resource]). Case study 3 involved an extra seven trials reporting OS for three additional treatments (nintedanib + docetaxel, pemetrexed and ramucirumab + docetaxel) that had been used in a sensitivity analysis of an economic model (Fig. S3 and Table S3 [Online Resource]). Case study 4 concerned the efficacy outcome of progression-free survival (PFS), and the feasibility analysis determined that the same five trials used in Case study 2 were appropriate for the NMA for this outcome (Fig. S4 and Table S4 [Online Resource]).

## 2.2 Overview of the LLM-Based Process for Automating the NMA

GPT-4 (Generative Pre-trained Transformer 4, developed by OpenAI [14]) was selected as the LLM engine for this study, as it was considered superior to other publicly available LLMs at the time of study. However, the method developed for interacting with GPT-4 in this study can, in theory, use different LLMs.

To allow the LLM to generate text to specify an NMA, it was decided to choose a programming language whereby the analysis and the data could be contained within one script. R was chosen as the software in which the AI-generated analysis would be built, as it is freeware and platform (operating system) independent. To implement an NMA in R, the 'multinma' package was used, which implements network meta-analysis, network meta-regression and multilevel network meta-regression models [19]. Models were estimated in a Bayesian framework using Stan [20].

LLMs require the user to provide 'prompts', i.e. instructions stating what the user wants the LLM to do and the output required. Interaction with the LLM was achieved through application programming interface (API) calls (a way for two or more computer programs to communicate with each other) written in a Python script. The outline of the process, as shown in Fig. 2a, is as follows:

- For data extraction from the publications, a prompt including text from the publication and requesting extraction of all relevant data from the supplied publication text was sent via an API call to the LLM for each publication needed for the NMA.
- To produce an R script with code to run the NMA, a prompt requesting generation of an R script was passed to the LLM via an API call, along with the data from all publications and an example R script (sourced from the Vignettes for the 'multinma' package) [19].
- To produce a small report containing a description of the disease, a description of the analysis conducted, the results of the analysis and an interpretation of the results, the LLM-generated R script was called from the Python script and the results of the NMA, along with a prompt requesting generation of a small report, were sent via an API call to the LLM.

Example prompts are provided in the Online Resource.

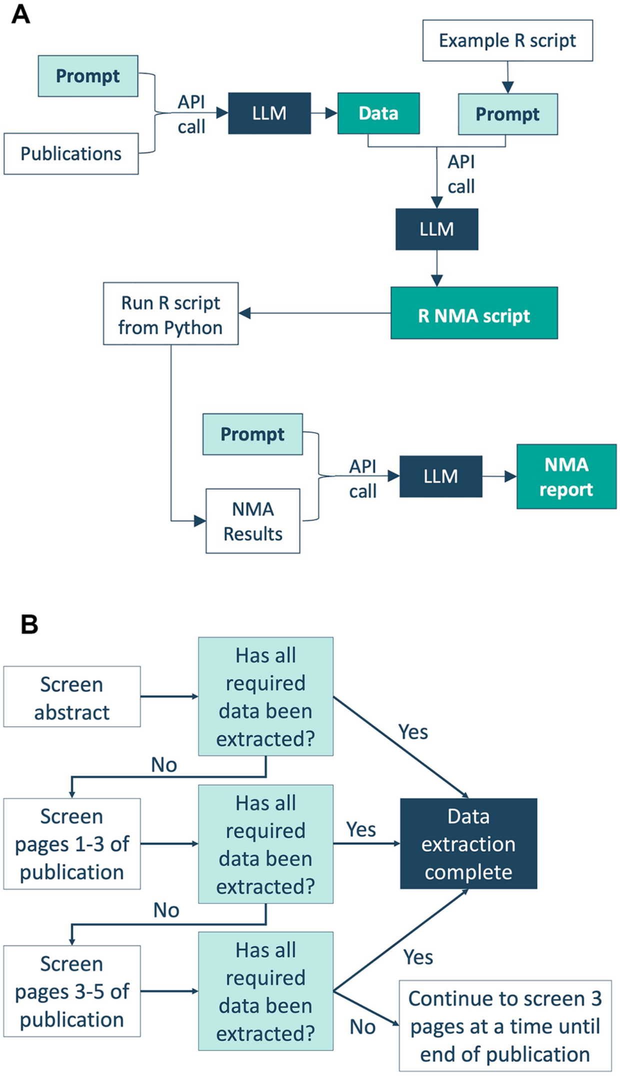

A
Example R script
Prompt
API call
LLM
Data
Prompt
Publications
API call
LLM
Run R script from Python
R NMA script
Prompt
API call
LLM
NMA report
NMA Results
B
Screen abstract
Has all required data been extracted?
Yes
No
Screen pages 1-3 of publication
Has all required data been extracted?
Yes
Data extraction complete
No
Screen pages 3-5 of publication
Has all required data been extracted?
Yes
No
Continue to screen 3 pages at a time until end of publication

**Fig. 2** **a** LLM-based process for automating the NMA. **b** Chunking approach to data extraction. *API* application programming interface, *LLM* large language model, *NMA* network meta-analysis

## 2.3 Prompts Used to Instruct the LLM and Hyperparameters

Prompts were developed that were used to instruct the LLM to:

- Extract the required data for the analysis from the abstracts of the publications.
- Determine if all data required was contained in the abstract and, if not, extract any missing data from the full publication.
- Infer missing data from other information, e.g. the number of patients affected from the proportion of patients affected and the number at risk as well as the number of patients at risk from total trial size and randomisation ratio.
- Transform extracted data to the correct format for inclusion into the model for analysis (number affected for binary outcomes and log scale for time-to-event outcomes).
- Generate an R script for NMA using generic script from the R ‘multinma’ package.
- Interpret the results of the analysis and write a small NMA report.

The Python script was used to pass the output of each prompt to the next, with the prompts loaded into Python as strings. Almost identical prompts were used for the four analyses conducted, with the following differences: use of relevant disease name (HS and NSCLC) and relevant outcome name (clinical response, OS and PFS); different R script examples were provided for binary and time-to-event outcomes [19], and additional contextual information was required for R script production for time-to-event outcomes (see Methods Sect. 2.4.3 below).

In addition to developing prompts, there was also a requirement to adjust some of the LLM’s hyperparameters, including role and temperature.

## 2.4 Prompt Development and Key Learnings

The prompts used have a significant impact on the output quality of the LLM. To evaluate the LLM’s capability to perform the required tasks, it was essential to create prompts of sufficient quality to obtain the required responses. Therefore, the following prompt creation process was followed: for each outcome type, initial prompts were generated and given to the LLM. The returned output was evaluated and, based on the contents, adjustments were made to the prompts. The adjusted prompts were then sent back to the LLM for further testing and evaluation. This process of output evaluation and prompt adjustment continued until no further improvements could be made and final prompts were reached. An example of the development of the OS data extraction prompt is given in Fig. S5 (Online Resource).

Several key learnings were uncovered through the prompt development process, which shaped the form of the final prompts. These were: using an iterative approach to data extraction, using multiple prompts and providing contextual information, as discussed in more detail below.

### 2.4.1 Chunking Approach to Data Extraction

A token is a chunk of text that an LLM reads or generates. At the time of the study, GPT-4 had a token limit of 8192 (approximately 6000 words), which restricted the amount of text that could be passed to, and be generated from, a single prompt. Since all the publications used for this study exceeded this limit, there was a need to cut publications into chunks before passing them to the LLM for data extraction. As shown in Fig. 2b, we asked the LLM to screen overlapping chunks of text from the main publication (e.g. pages 1–3, 3–5, 5–7, 7–9, etc.) to ensure that all text reviewed was in context and then asked the LLM to assess whether it had obtained all data required, before providing additional text for screening. It was possible for the LLM to get to the end of the publication without extracting all required data if it failed to identify that data.

### 2.4.2 Multiple Prompts

The first approach for creating an R script was to ask the LLM to write an initial R script using data from the first study, and then to ask it to add data from more trials. This approach worked well for the binary outcome, where the data required for the analysis in R is number at risk and number of patients affected in each arm. However, for the time-to-event outcomes (OS and PFS), the input is a hazard ratio and standard error for each treatment comparison, and the initial approach used did not produce the right format for this input, leading to incorrect results. Thus, for the time-to-event outcomes, we asked the LLM to gather the required data (hazard ratios, error measures, etc.) from all trials before writing the R script. For consistency, this approach was also used for the binary outcome. For the analysis input, different treatments were given numbers in the R script (Fig. S7 [Online Resource]) but the LLM did not always use the same numbering for the same treatment. Therefore, it was necessary to prompt the LLM to fix this in the initial script, to match the numbers with the names and doses of the treatments. Thus, multiple prompts were used to generate the required R script:

- Collate all data.
- Use this and the example R script to write an initial script.

- Tidy the initial script to ensure correct treatment numbering is used.

### 2.4.3 Contextual Information

For some tasks, the LLM was frequently observed to make general errors, such as not understanding statistical significance. These were not related to the content (disease and treatment) or language used in the included studies. Addressing these errors required the provision of contextual information in addition to the instructions. The contextual information was developed iteratively in the same manner as the prompts.

The LLM was initially not very successful at writing an executable R script or choosing the correct model to use for the analysis, for either outcome type. For example, the LLM sometimes invented R packages and functions that it included in the script. Including worked examples has previously been shown to improve the performance of LLMs in multi-step reasoning tasks [21]. Therefore, we provided an example script appropriate for the type of analysis needed, as contextual information for the LLM. The example scripts used were sourced from the online vignette of the 'multinma' package [19] (Fig. S6 [Online Resource]).

Similarly, when asked to write the R script for the time-to-event outcomes, the LLM did not always construct the input for the analysis in the correct way nor maintain the order of the treatment comparison. For instance, the LLM would try to construct a dataframe for the input data that had a row per treatment arm in the treatment names and number-at-risk columns but then would only include one row per study for the hazard ratios. It was therefore necessary to provide context to the LLM, which was achieved by including contextual statements within the code-writing prompt. For example, including the text "The order of the treatment comparison is important" ensured that the LLM maintained the treatment comparison order for each hazard ratio. Some of the trials included in the analyses treated patients with a combination of treatment plus placebo, e.g. treatment X plus placebo. Usually, when conducting an NMA, we would consider the treatment effects for these patients to be equivalent to patients treated only with treatment X. For the LLM to consistently make this assumption, and to therefore number the treatments correctly, we needed to provide contextual information, such as adding the statement, "We consider patients treated with 'treatment X plus placebo' to be treated with 'treatment X'", to the prompt asking the LLM to tidy the R script.

The LLM also required context for interpretation of the NMA results. The LLM reliably identified when a treatment outperformed the comparator, for both the binary and the time-to-event outcomes. However, we noticed that the LLM sometimes claimed that either all or none of the comparisons reached statistical significance, when in fact some did and some did not. Therefore, contextual statements, such as "A result is statistically significant if the lower and upper bound for the credible interval are either both greater than 1 or both less than 1", were included.

To summarise, the following contextual information was provided to the LLM: example R scripts for the analysis, the importance of the order of the treatment comparison when considering a hazard ratio, the assumptions generally made when considering equivalence of treatments and the definition of statistical significance.

## 2.5 Non-text-Based Publications

For all case studies considered, some of the publications needed for the case study were text-based, whilst some were photographs of presentations, or posters, or contained data within figures. Whilst it is now possible to ask an LLM to receive images as input (e.g. with GPT-4 Vision, Gemini), at the time of the study, GPT-4 was not able to receive images as input and thus was not able to extract any data for these publications. The trials that had image-based publications and the approach taken to obtain data are listed in Table 1.

## 2.6 LLM Hyperparameters

'Role' and 'temperature' are some of the hyperparameters that can be used to control the behaviour of GPT-4. Assigning a role to the LLM is a simple way to add context to a prompt, for example if you assign it a role of 'a poet', the style, and possibly the content, of the response will be different from that obtained if the role assigned is 'a surly teenager' [22]. Thus, there was a need to assign an appropriate role to GPT-4. We found that by telling GPT that it is a statistician and a medical researcher, we obtained the type and quality of responses that we needed.

The temperature parameter of GPT-4 is a number between 0 and 2 that determines the randomness of the generated output. A lower value for the temperature parameter will lead to a less random response, whilst a higher value will produce a more creative and/or surprising output. We wanted the responses to be as deterministic as possible, so we set the temperature to be 0.

Default values were used for all other hyperparameters.

## 2.7 Output Generation and Assessment

For each case study, a single Python script included the final set of prompts for interaction with the LLM and commands for the generated R script to run and to obtain the results of the analysis (Fig. 2a). Each Python script was run end to end, without human intervention, and produced

**Table 1** Non-text-based publications and approach to data extraction

| Case study | Trials | Approach taken |
|---|---|---|
| Case study 1 (HS) | BE HEARD I and II [31]<br>Photographs of a presentation | Data provided to GPT-4 within Python script (hard-coded into script), before R script generation |
| Case studies 2, 3 and 4 (NSCLC OS, OS sensitivity, PFS) | KEYNOTE-010 [32]<br>Photograph of poster | GPT-4 was asked to extract all data from an older publication [33] and then was provided with the updated hazard ratio (hard-coded into the Python script), before R script generation |
| Case study 3 (NSCLC OS sensitivity) | GFPC 05-06 [34]<br>Hazard ratio provided in a figure | GPT-4 was asked to extract all data from the text in the publication and then was provided with the hazard ratio (hard-coded into the Python script), before R script generation |

*HS* hidradenitis suppurativa, *NSCLC* non-small cell lung cancer, *OS* overall survival, *PFS* progression-free survival

an R script and a short report describing the disease area and method of analysis, presenting the NMA results and providing an interpretation of the results.

Reproducing results with LLMs can be difficult because of random elements at play that vary outputs over time [23], i.e. LLMs do not produce deterministic results. As previously mentioned (Sect. 2.6), temperature is one of several hyperparameters that control the behaviour of GPT-4. Despite setting the temperature of GPT-4 to 0 (least random), outputs were observed to vary when the same prompt set was used on multiple occasions. Therefore, we ran the Python script end to end 20 times for each analysis (80 runs in total) to capture variation in performance.

The performance of the LLM was assessed in three stages:

- Assessment of data extraction: for each run, did the LLM correctly extract all required data from each trial? This was evaluated by comparing outputs from the LLM with data extracted/checked by two of the investigators (SLR and NMA experts).
- Assessment of R script (evaluated by one of the investigators, an NMA expert familiar with R and who wrote the R script for the manually conducted NMAs):
    - o Did the LLM produce an R script that contained all relevant extracted data and the correct functions to conduct an NMA?
    - o Could the script be run without human intervention? If not, was minor (less than 2 minutes of work) or major (more than 2 minutes of work) editing required to enable this?
    - o Did the script produce results that matched the same NMA conducted by a human?
- Assessment of the NMA report (qualitatively assessed by one of the investigators, familiar with the disease area):
    - o Was a reasonable description of the disease area provided?
    - o Was the methodological description of the analysis correct?
    - o Were correct results presented?
    - o Was the interpretation of the results correct and informative?

## 3 Results

A summary of the LLM's success in data extraction across each case study is shown in Fig. 3a, and the quality of the generated R script produced is shown in Fig. 3b.

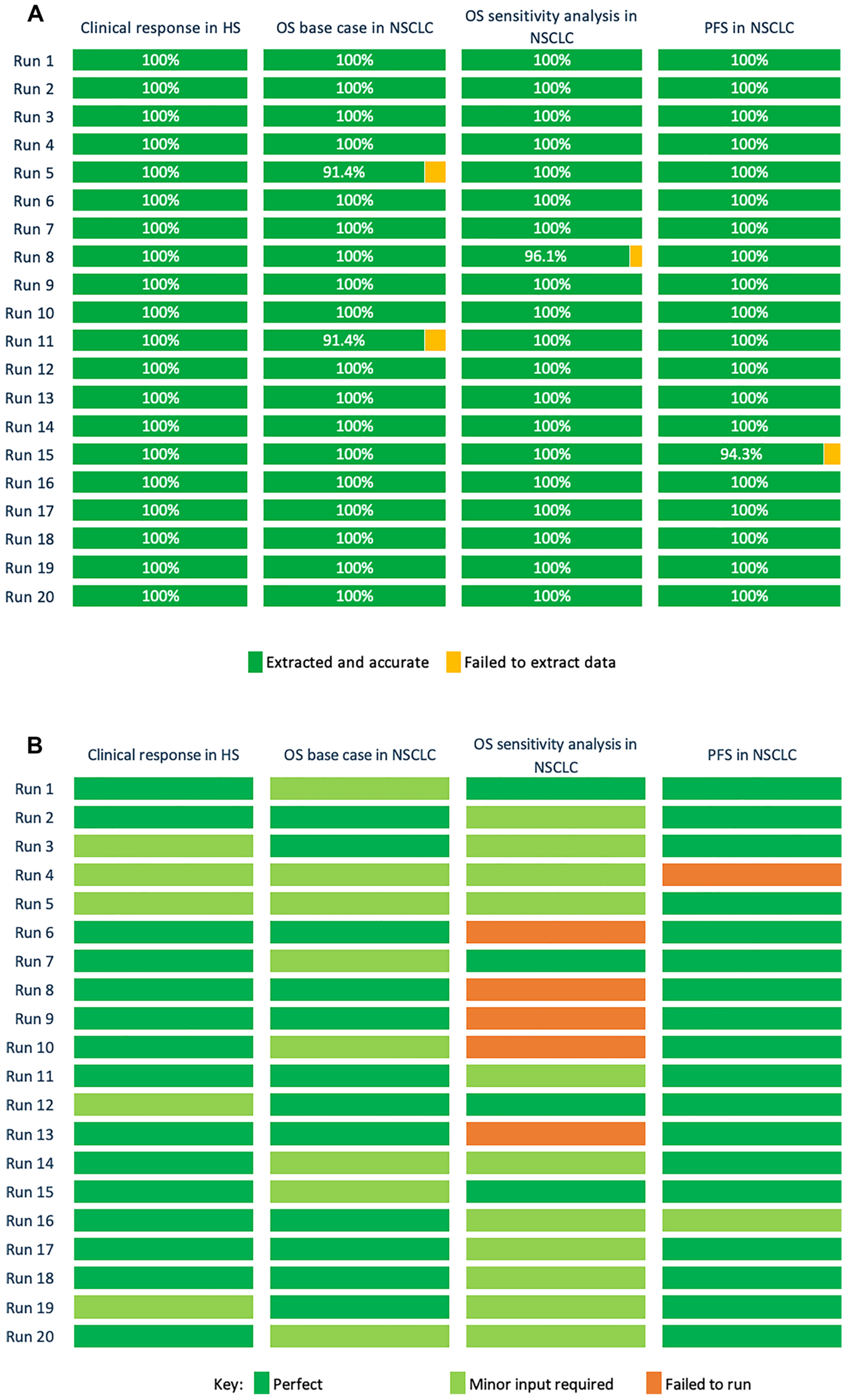
Four case studies were considered, and Fig. S1–S4 (Online Resource) display the network for each case study.

1. Case study 1 (indirect comparison of the efficacy of treatments for patients with moderate-to-severe HS): three publications reporting clinical response data for six trials (two trials per publication).
2. Case study 2 (indirect comparison of the efficacy of second-line treatment for patients with NSCLC: base case analysis of OS): four publications reported OS across five trials.
3. Case study 3 (indirect comparison of the efficacy of second-line treatment for patients with NSCLC: sensitivity analysis of OS): 11 publications reported OS data from 12 trials.
4. Case study 4 (indirect comparison of the efficacy of second-line treatment in patients with NSCLC: PFS): four publications reporting PFS across five trials.

### 3.1 Data Extraction

A summary of the LLM's overall performance for data extraction is presented in Fig. 3a.

**Fig. 3** **a** Summary of the data extraction performance. **b** Summary of R script quality. *HS* hidradenitis suppurativa, *NSCLC* non-small cell lung cancer, *OS* overall survival, *PFS* progression-free survival



For case study 1, the LLM accurately and consistently extracted the correct data from the text-based publications for each of the 20 runs. For the text-based publications available for case study 2 (Sect. 2.5), the LLM accurately and consistently extracted the correct data most times but failed to extract number at risk data for the KEYNOTE-010 trial (three items of data) in two of the 20 runs. Since the LLM needed to extract 35 separate pieces of data for each run, this equates to a very high overall extraction success rate of 99.1%. Similarly, for the text-based documents available for case study 3, the LLM accurately and consistently extracted the correct data for all but one run, where it did not manage to extract the number of patients at risk (three items of data) for the KEYNOTE-010 trial. Since the LLM needed to extract 77 separate pieces of data for each run, this equates to a very high overall extraction success rate of 99.8%. For the text-based publications available for case study 4, the LLM accurately and consistently extracted the correct data, except for one run, where it failed to extract the number at risk data for the CheckMate017 trial (two data items; Fig. 3a). Since the LLM needed to extract 35 separate pieces of data for each run, this equates to an overall extraction success rate of 99.7%.

It is not clear why the LLM failed to extract number of patients at risk from the KEYNOTE-010 publication in two runs for case study 2 and one run for case study 3 and why it failed to extract the number of patients at risk from the CheckMate017 trial for one run of case study 4, and there did not appear to be a systematic pattern for this failure. However, it may be due to the language used to describe patient assignment in these publications, and it may be possible to enhance performance further using improved prompting, or by running the data extraction multiple times, and using the responses where the LLM has found the data required.

Thus, the LLM achieved a data extraction success rate of over 99% for each case study. The LLM did not report incorrect data on any occasion, it only intermittently failed to extract data from two trials. This level of performance exceeds the performance seen for human data extraction, where between 8 and 42% of data extraction errors have been observed [10].

## 3.2 R Script Generation

All R scripts generated with the LLM were well commented and the script was easy to read and interpret (Fig. S7 and Fig. S8 [Online Resource]). The LLM-generated R scripts ran with no or very minor amendments (Tables 2, 3, 4, 5) in each of the 20 runs of case study 1 and case study 2, 15 runs of case study 3 and 19 runs of case study 4 (Fig. 3b). In these cases, once any required amendments had been implemented, the scripts ran and produced correct results (Sect. 3.3).

**Table 2** Summary of intervention required for HS analysis R script

| Number of runs | Description of intervention required |
|---|---|
| 4 (20%) | Conversion of output of R *nma* function to a dataframe produced a dataframe with slightly different column headings than were expected. So, the formula to convert mean log odds ratio and lower and upper credible limits to natural scale failed. This was easily fixed by either changing the way that the dataframe was produced or by changing the column headings within the conversion formula |
| 1 (5%) | The SUNSHINE and SUNRISE publication reported non-whole numbers for the patients achieving clinical response. GPT-4 had tried to include these non-integers in the analysis. This error was easily fixed by converting these values to whole numbers |

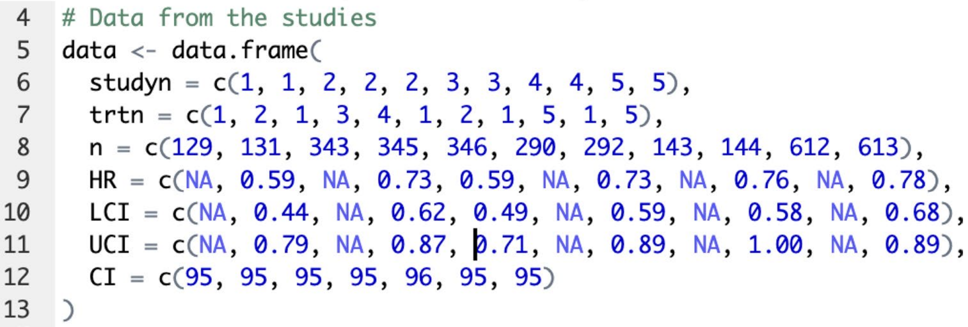
*HS* hidradenitis suppurativa

For case study 1, the R script failed to run (producing an error message) on five occasions. Four scripts contained very minor errors that took less than 2 min to fix (Table 2). For the fifth script, the LLM had included data directly from a publication that reported non-integer values for number of events [24]. On all other occasions, it followed instructions in the prompt to only use integer values and thus calculated the number of events from the percentage given. Again, this was a minor error, requiring a quick and very simple fix (Table 2).

For case study 2, the R script failed to run on eight occasions and required human input. For each of these occasions, the script contained very minor errors that took less than 2 min to fix (Table 3).

Case study 3 was the most complicated of all analyses conducted: the number of trials and treatments included in the network for the sensitivity analysis of OS were greater than for the base case analysis of OS (12 trials and eight treatments versus 5 trials and five treatments [Fig. S3, Online Resource]), and, whilst all trials included an arm where patients were treated with docetaxel, not all hazard ratios were reported with docetaxel as the reference treatment. Four of the generated R scripts ran without requiring human input, whilst 16 failed to run (producing an error message). Of these, 11 scripts contained very minor errors that took less than 2 min to fix (Table 4). For the remaining five scripts, the LLM had incorrectly constructed the dataframe, used as input to the R *set_agd_contrast* function: the column containing (log) hazard ratios should have one entry per treatment in each trial, with ‘NA’ provided for the reference treatment, but the LLM only included one ‘NA’ in the whole column, so, without prior knowledge of the network and the data, it was not possible to update the

**Table 3** Summary of intervention required for base case OS analysis R script

| Number of runs | Description of intervention required |
|---|---|
| 8 (40%) | Some of the publications reported confidence intervals for the hazard ratio that were not 95%. GPT-4 had constructed a dataframe that included all treatments, hazard ratios and confidence interval limits and had included a column in this to record the confidence levels. Whilst the hazard ratio and confidence interval limit columns included a row per treatment in each trial, the confidence level column only contained one value per hazard ratio:<br><br># Data from the studies<br>data <- data.frame(<br>  studyn = c(1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 5),<br>  trtn = c(1, 2, 1, 3, 4, 1, 2, 1, 5, 1, 5),<br>  n = c(129, 131, 343, 345, 346, 290, 292, 143, 144, 612, 613),<br>  HR = c(NA, 0.59, NA, 0.73, 0.59, NA, 0.73, NA, 0.76, NA, 0.78),<br>  LCI = c(NA, 0.44, NA, 0.62, 0.49, NA, 0.59, NA, 0.58, NA, 0.68),<br>  UCI = c(NA, 0.79, NA, 0.87, 0.71, NA, 0.89, NA, 1.00, NA, 0.89),<br>  CI = c(95, 95, 95, 95, 96, 95, 95)<br>)<br>This was easily fixed by editing the values in this column to include the correct number of rows |

*OS* overall survival

**Table 4** Summary of intervention required for sensitivity analysis of OS R script

| Number of runs | Description of minor intervention required |
|---|---|
| 4 (20%) | GPT-4 had constructed a dataframe that included all treatments, hazard ratios and confidence interval limits and a column for the trials. Whilst the other columns included a row per treatment in each trial, GPT-4 had only included one row per trial in the studies column. This was easily fixed by editing the values in this column to include the correct number of rows |
| 7 (35%) | When tidying the script, GPT-4 did not update the treatment numbering correctly, leading to a disconnected network in some cases. This was easily fixed by editing the treatment numbers in the script |

*OS* overall survival

script to enable it to run and produce correct results. It may be possible to avoid this error by feeding the LLM with more domain knowledge, along with the prompt.

For case study 4, two of the generated R scripts failed to run, with one requiring only a minor correction to the format of treatment numbers for one trial. For the remaining run, the dataframe used within R's *set_agd_contrast* function did not have the correct format for an analysis of trial-level data (e.g. hazard ratios; Table 5).

Thus, in many cases, the LLM generated an R script that ran without human input, and in the majority of cases, a script was generated that ran following minor human input (< 2 min of effort). All errors caused the R script to fail and produce an error message, no script ran and generated erroneous results.

## 3.3 NMA Results

Where the R scripts ran, the NMA results calculated very closely matched those of the human-conducted NMA.

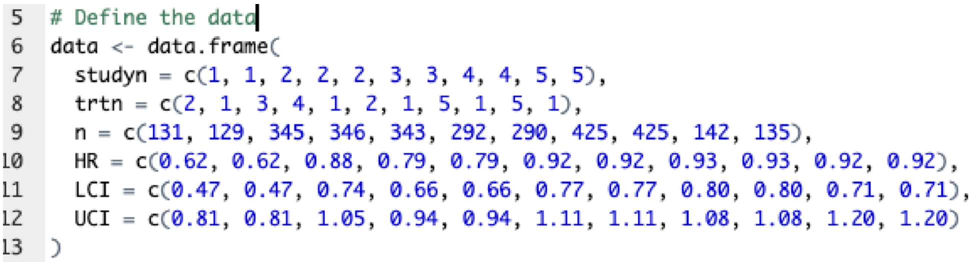
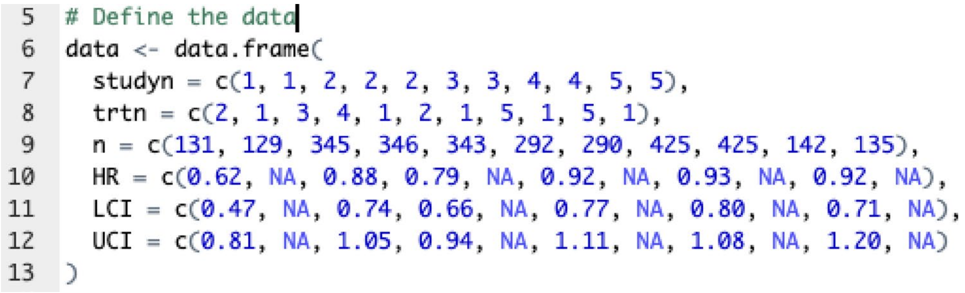
The mean odds ratios for treatments versus placebo, calculated using the LLM-generated R scripts for case study 1, were very close to those calculated by the manual (human) NMA for all 20 runs (Table 6). All differences observed were within the range of expected variability (< 1%), since results can vary slightly when running an NMA multiple times using the same R code, due to Monte Carlo error occurring when the random seed is not set within the analysis (these differences are also observed when running a human-written script several times) [25].

For case studies 2, 3 and 4, the mean hazard ratios for treatments versus docetaxel, produced by the LLM-generated R scripts, were identical to those calculated by the manual NMA, whilst the limits of the credible interval varied slightly but only within the realms of the variability obtained if a human ran the NMA, using the same R code multiple times (due to Monte Carlo error; Tables 7, 8 and 9, respectively).

## 3.4 Report Writing and Interpretation of Results

Using the results from the R scripts, the NMA reports generated with the LLM were clearly written and included a good summary of the disease area (hidradenitis suppurativa for case study 1 [Fig. S9 (Online Resource)] and NSCLC for case studies 2–4 [Fig. S10 (Online Resource)]). The methods of analysis were summarized clearly and at an appropriate level of detail, as we asked the LLM to create a concise report (i.e. we did not request the methods to be as elaborate as they might

**Table 5** Summary of intervention required for PFS R script

| Number of runs | Description of minor intervention required |
|---|---|
| 1 (5%) | When tidying the script, GPT-4 did not update the treatment numbering correctly. This was easily fixed by editing the treatment numbers in the script |
| 1 (5%) | GPT-4 did not construct the dataframe used within R's set_agd_contrast function to have the correct format for an analysis of trial-level data:<br><br>`# Define the data`<br>`data <- data.frame(`<br>`  studyn = c(1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 5),`<br>`  trtn = c(2, 1, 3, 4, 1, 2, 1, 5, 1, 5, 1),`<br>`  n = c(131, 129, 345, 346, 343, 292, 290, 425, 425, 142, 135),`<br>`  HR = c(0.62, 0.62, 0.88, 0.79, 0.79, 0.92, 0.92, 0.93, 0.93, 0.92, 0.92),`<br>`  LCI = c(0.47, 0.47, 0.74, 0.66, 0.66, 0.77, 0.77, 0.80, 0.80, 0.71, 0.71),`<br>`  UCI = c(0.81, 0.81, 1.05, 0.94, 0.94, 1.11, 1.11, 1.08, 1.08, 1.20, 1.20)`<br>`)`<br>How the dataframe should have looked:<br><br>`# Define the data`<br>`data <- data.frame(`<br>`  studyn = c(1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 5),`<br>`  trtn = c(2, 1, 3, 4, 1, 2, 1, 5, 1, 5, 1),`<br>`  n = c(131, 129, 345, 346, 343, 292, 290, 425, 425, 142, 135),`<br>`  HR = c(0.62, NA, 0.88, 0.79, NA, 0.92, NA, 0.93, NA, 0.92, NA),`<br>`  LCI = c(0.47, NA, 0.74, 0.66, NA, 0.77, NA, 0.80, NA, 0.71, NA),`<br>`  UCI = c(0.81, NA, 1.05, 0.94, NA, 1.11, NA, 1.08, NA, 1.20, NA)`<br>`)` |

*PFS* progression-free survival

**Table 6** Results of the manual NMA and example results of LLM generated NMA for clinical response in patients with moderate-to-severe hidradenitis suppurativa

| Treatment versus placebo | Odds ratio [95% CrI] obtained from manual NMA | Odds ratio [95% CrI] obtained from GPT-4 NMA |
|---|---|---|
| Adalimumab | 2.84 [2.06, 3.97] | 2.84 [2.06, 3.90] |
| Bimekizumab every 2 weeks | 2.26 [1.52, 3.36] | 2.24 [1.52, 3.30] |
| Bimekizumab every 4 weeks | 2.21 [1.44, 3.37] | 2.20 [1.44, 3.38] |
| Secukinumab every 2 weeks | 1.61 [1.18, 2.20] | 1.61 [1.20, 2.16] |
| Secukinumab every 4 weeks | 1.65 [1.22, 2.27] | 1.65 [1.22, 2.19] |

For each case study, the set of results produced by the first R script generated by GPT-4 were used for the examples given here. These results were then compared to the remaining 19 sets of results and assessed by one of the investigators (NMA expert)

*CrI* credible interval, *LLM* large language model, *NMA* network meta-analysis

need to be for an HTA submission). The interpretation of the results was correct in all 20 runs for each case study, with the correct treatment identified as the best treatment in the network. Whether treatment effects were statistically significant was also correctly stated. Variation was seen in the amount of detail generated by the LLM in different runs. Examples of the interpretation summary are shown in Fig. 4.

## 4 Discussion

We present a novel LLM-based process for automating the data extraction, software script construction and results interpretation for an NMA, which required only trial publications as the input. Using four previously conducted NMAs as case studies, we demonstrated that an LLM (GPT-4) was capable of extracting data to a high standard and could produce quality R script, which included all required data, and could be run end to end with little or no human intervention. We also demonstrated that the LLM could successfully

**Table 7** Results of the manual NMA and example results of LLM-generated NMA for base case analysis of overall survival in patients receiving second-line treatment for NSCLC

| Treatment versus docetaxel | Hazard ratio [95% CrI] obtained from manual NMA | Hazard ratio [95% CrI] obtained from GPT-4 NMA |
|---|---|---|
| Atezolizumab | 0.78 [0.69, 0.88] | 0.78 [0.69, 0.88] |
| Nivolumab | 0.68 [0.58, 0.80] | 0.68 [0.58, 0.80] |
| Pembrolizumab 10 mg/kg | 0.59 [0.49, 0.70] | 0.59 [0.49, 0.71] |
| Pembrolizumab 2 mg/kg | 0.73 [0.62, 0.86] | 0.73 [0.62, 0.86] |

For each case study, the set of results produced by the first R script generated by GPT-4 were used for the examples given here. These results were then compared to the remaining 19 sets of results and assessed by one of the investigators (NMA expert)

*CrI* credible interval, *LLM* large language model, *NMA* network meta-analysis, *NSCLC* non-small cell lung cancer

**Table 8** Results of the manual NMA and example results of LLM-generated NMA for sensitivity analysis of overall survival in patients receiving second-line treatment for NSCLC

| Treatment versus docetaxel | Hazard ratio [95% CrI] obtained from manual NMA | Hazard ratio [95% CrI] obtained from GPT-4 NMA |
|---|---|---|
| Atezolizumab | 0.78 [0.68, 0.87] | 0.78 [0.69, 0.88] |
| Nintedanib + docetaxel | 0.94 [0.83, 1.06] | 0.94 [0.83, 1.06] |
| Nivolumab | 0.68 [0.59, 0.79] | 0.68 [0.59, 0.78] |
| Pembrolizumab 10 mg/kg | 0.59 [0.49, 0.71] | 0.59 [0.49, 0.71] |
| Pembrolizumab 2 mg/kg | 0.73 [0.62, 0.86] | 0.73 [0.62, 0.86] |
| Pemetrexed | 0.97 [0.87, 1.10] | 0.97 [0.86, 1.09] |
| Ramucirumab + docetaxel | 0.86 [0.76, 0.97] | 0.86 [0.75, 0.98] |

For each case study, the set of results produced by the first R script generated by GPT-4 were used for the examples given here. These results were then compared to the remaining 19 sets of results and assessed by one of the investigators (NMA expert)

*CrI* credible interval, *LLM* large language model, *NMA* network meta-analysis, *NSCLC* non-small cell lung cancer

**Table 9** Results of the manual NMA and example results of LLM-generated NMA for progression-free survival in patients receiving second-line treatment for NSCLC

| Treatment versus docetaxel | Hazard ratio [95% CrI] obtained from manual NMA | Hazard ratio [95% CrI] obtained from GPT-4 NMA |
|---|---|---|
| Atezolizumab | 0.93 [0.82, 1.05] | 0.93 [0.81, 1.06] |
| Nivolumab | 0.81 [0.70, 0.95] | 0.81 [0.69, 0.95] |
| Pembrolizumab 10 mg/kg | 0.79 [0.67, 0.94] | 0.79 [0.66, 0.95] |
| Pembrolizumab 2 mg/kg | 0.88 [0.74, 1.04] | 0.88 [0.74, 1.05] |

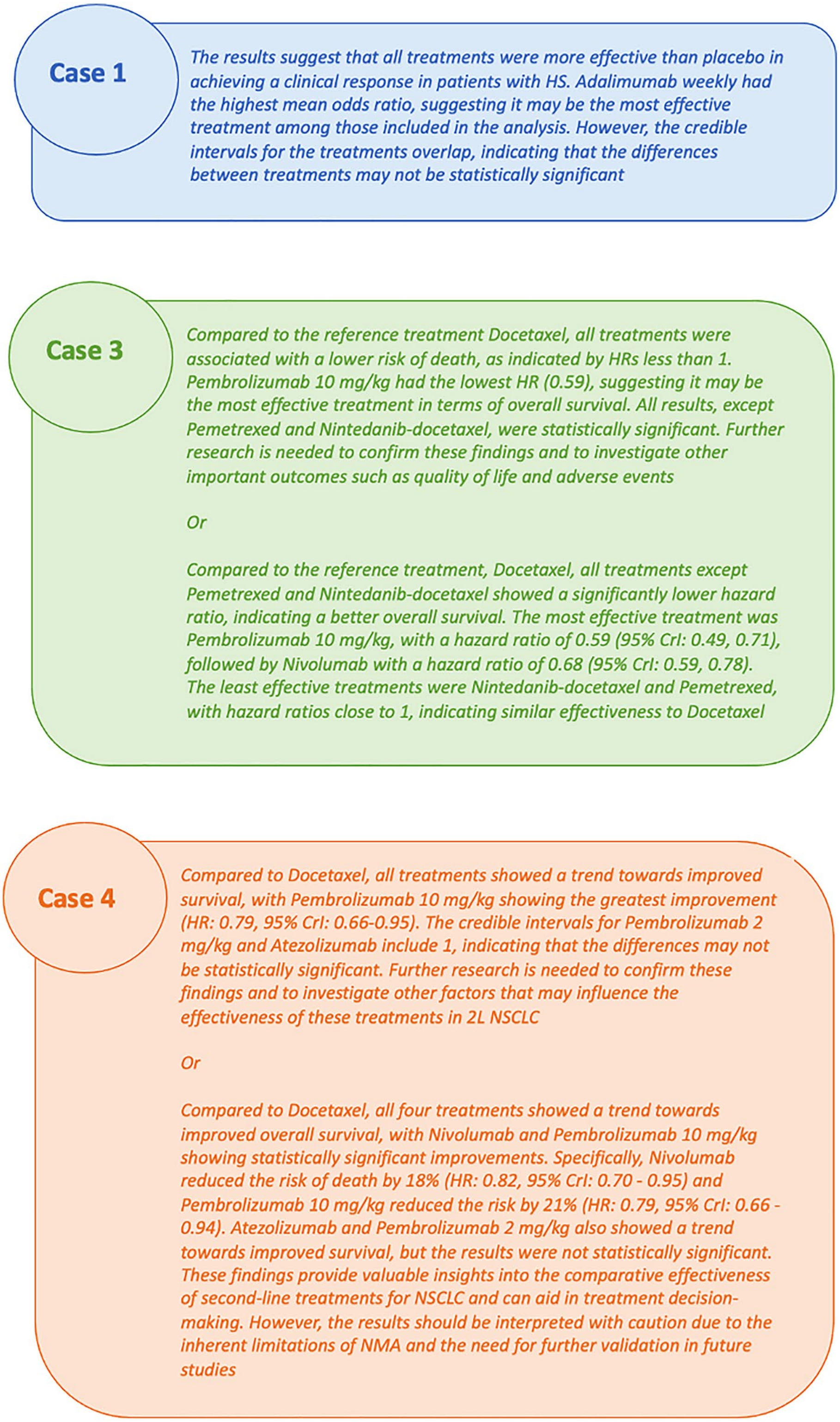
For each case study, the set of results produced by the first R script generated by GPT-4 were used for the examples given here. These results were then compared to the remaining 19 sets of results and assessed by one of the investigators (NMA expert)

*CrI* credible interval, *LLM* large language model, *NMA* network meta-analysis, *NSCLC* non-small cell lung cancer

produce good quality text that summarised the disease area, summarised the analysis method used, presented the results in an informative manner and interpreted the results correctly.

Whilst some studies have considered use of automation or AI within HEOR [12, 13, 16], we believe that this is the first study to consider the use of an LLM to automate data extraction and NMA end to end. As an early feasibility

**Fig. 4** Report writing: examples of the LLM's (GPT-4) interpretations of the results. *CrI* credible interval, *HR* hazard ratio, *HS* hidradenitis suppurativa, *NSCLC* non-small cell cancer



**Case 1**

*The results suggest that all treatments were more effective than placebo in achieving a clinical response in patients with HS. Adalimumab weekly had the highest mean odds ratio, suggesting it may be the most effective treatment among those included in the analysis. However, the credible intervals for the treatments overlap, indicating that the differences between treatments may not be statistically significant*

**Case 3**

*Compared to the reference treatment Docetaxel, all treatments were associated with a lower risk of death, as indicated by HRs less than 1. Pembrolizumab 10 mg/kg had the lowest HR (0.59), suggesting it may be the most effective treatment in terms of overall survival. All results, except Pemetrexed and Nintedanib-docetaxel, were statistically significant. Further research is needed to confirm these findings and to investigate other important outcomes such as quality of life and adverse events*

*Or*

*Compared to the reference treatment, Docetaxel, all treatments except Pemetrexed and Nintedanib-docetaxel showed a significantly lower hazard ratio, indicating a better overall survival. The most effective treatment was Pembrolizumab 10 mg/kg, with a hazard ratio of 0.59 (95% CrI: 0.49, 0.71), followed by Nivolumab with a hazard ratio of 0.68 (95% CrI: 0.59, 0.78). The least effective treatments were Nintedanib-docetaxel and Pemetrexed, with hazard ratios close to 1, indicating similar effectiveness to Docetaxel*

**Case 4**

*Compared to Docetaxel, all treatments showed a trend towards improved survival, with Pembrolizumab 10 mg/kg showing the greatest improvement (HR: 0.79, 95% CrI: 0.66-0.95). The credible intervals for Pembrolizumab 2 mg/kg and Atezolizumab include 1, indicating that the differences may not be statistically significant. Further research is needed to confirm these findings and to investigate other factors that may influence the effectiveness of these treatments in 2L NSCLC*

*Or*

*Compared to Docetaxel, all four treatments showed a trend towards improved overall survival, with Nivolumab and Pembrolizumab 10 mg/kg showing statistically significant improvements. Specifically, Nivolumab reduced the risk of death by 18% (HR: 0.82, 95% CrI: 0.70 - 0.95) and Pembrolizumab 10 mg/kg reduced the risk by 21% (HR: 0.79, 95% CrI: 0.66 - 0.94). Atezolizumab and Pembrolizumab 2 mg/kg also showed a trend towards improved survival, but the results were not statistically significant. These findings provide valuable insights into the comparative effectiveness of second-line treatments for NSCLC and can aid in treatment decision-making. However, the results should be interpreted with caution due to the inherent limitations of NMA and the need for further validation in future studies*

assessment, this study points to several potential benefits of automated data extraction and NMA. The primary potential benefit is in time savings in the data extraction process, which could enable quicker and less costly decision making in healthcare, which may ultimately speed up patient access to medicines. Specifying the user prompts and data required for the process should take no longer than 1–2 h. This time includes the time required to update the prompts and run the Python script but does not include checking of data extraction, assessment of study or population heterogeneity and feasibility analysis. The time taken will depend on the complexity of the treatment network and the run time of the process (which runs without human attention); the run time for the four case studies considered was approximately 10 min for the three smaller NMAs (including 5 or 6 treatments and trials) and approximately 15 min for the larger NMA (8 treatments, 12 trials). The time taken by the LLM is substantially lower than the time it took to manually extract data, write the R script and run it (approximately half a day per outcome).

### 4.1 Limitations

The novel approach developed has been tested on a single LLM (GPT-4) and on a limited number of case studies. The treatment networks were relatively simple; all trials included a common treatment and the LLM was not asked to check whether the proportional hazards assumption held for the three analyses of time-to-event (survival) data.

Further research is needed to determine the level of additional work required to use our approach with LLMs other than GPT-4 and whether improved accuracy and/or processing speed can be achieved with other LLMs or alternative prompting strategies and context.

The responses given by LLMs are not always consistent when asking the same question multiple times, and the responses may change over time, as the LLM learns from the questions it is being asked. This means that it could be difficult to reproduce results. However, if using general access LLMs, this uncertainty could be reduced by ensuring that LLM parameters are set to reduce the level of randomness in responses by asking the LLM to repeat all tasks and then identifying (and discarding) outliers in the responses and by using a specific version of an LLM, e.g. GPT-4 at 1 January 2024. Alternatively, using open-source LLMs, for example through the ‘Ollama’ package [26, 27], would allow tighter LLM version control and allow better reproducibility of responses. Many open-source LLMs can be downloaded and used locally [26], which would enable the same LLM model version to be used.

We believe that the prompts developed within this study are generalisable to NMAs in different disease areas, and similar prompts can be used for continuous outcomes; however, there is a need to demonstrate that this is the case. We would also want to demonstrate the methodology and developed prompts with NMAs with larger and more complicated networks. Additionally, there is a need to investigate using LLMs to support further NMA tasks, such as choosing statistical models; choosing fixed effects or random effects models; determining whether the models have converged (Gelman–Rubin diagnostic); testing the proportional hazards assumption (very important when considering time-to-event outcomes); and deciding the approach to analysis, use of fractional polynomials, conducting feasibility analyses, etc.

Whilst the LLM used in this study (GPT-4) showed great promise for extracting data, it was not 100% perfect on every single run. Future studies should investigate the practicality of and effect on performance when collating data from repeated extractions on each publication and taking the mode of results. Due to the token limit of GPT-4, it was necessary to pass the publication text in chunks to the LLM, which may have affected data extraction performance. Thus, consideration should also be given to improvements in performance and effect on speed when using LLMs capable of processing whole documents at a time (e.g. GPT-4 32k context model [32,768 tokens] or GPT-4 Turbo [128,000 tokens]).

Since the study was conducted, image capabilities have been rolled out for GPT-4 [28], and multi-modal functionality has been developed for other LLMs (e.g. Google’s Gemini) [29]. Thus, there is now a need to determine whether this allows for data extraction from the image-based publications that we encountered.

In the materials used for this study, the extracted data were reported within the main text of the documents, not within tables. This meant that we could use a Python package (PyPDF2) to convert the PDF documents to text and then pass this text to the LLM for data extraction. Investigating data extraction from tables was out of scope for the current study, but it may be possible to use the same Python package to parse table text from the PDF and for this to be provided in the correct order, to use other Python packages, such as OpenCV, or to pass the table as an image to a multi-modal LLM, which may then be able to extract the relevant data.

Overall, the LLM produced R scripts of high quality; however, there were a few occasions when a useable script was not produced. Whilst the errors in these scripts were very easy to spot and fix and would be identified when conducting quality assurance checks of all input data and software (as is currently applied to human-generated software), there is a need to investigate the effect of enhanced prompt engineering and/or fine tuning on the quality of the R script generated. Fine tuning allows users to train the model, making it follow instructions better, and consistently format responses, a crucial aspect for applications demanding a specific response format, such as code completion [30], a preview version of which was made available at the end of 2023 for GPT-4 [30].

The summary report generated with the LLM was always informative and accurate, but, whilst the overall sentiment of the writing was preserved, the level of detail provided and the exact text used varied each time this was requested. The report produced was not sufficiently detailed to be used for an HTA appraisal but did provide an easy-to-read high-level review of the analysis and could be useful in settings such as early HTA planning and scientific advice. Whether fine tuning can be used to improve the consistency of the report should be investigated, along with the LLM's ability to generate sections of text for a full technical report.

The approach developed herein is not yet of sufficient robustness to meet the methodological rigour required by HTA bodies, and it does not include some important steps of NMA, e.g. heterogeneity assessment. However, there is likely to be rapid advancement in the use of LLMs to a point where they could be used for the majority of the process. Heterogeneity assessment and more detailed reporting would need to be further explored to implement fully automated NMAs for HTA purposes. Furthermore, once complete automation is achieved, there will still be a need for human involvement due to the requirement for human accountability in the process.

## 4.2 Recommendations and Future Research

Whilst we found that the LLM is currently not 100% consistent, it is one of the first types of general purpose LLMs, and thus, there is scope for LLMs to improve with time and additional training. The ability to use fine tuning and the introduction of better models and larger word limits are likely to improve the success rate seen in this study. The novel approach developed in this study has potential to be developed further but could already be applied to data extraction, software script generation and to aid result interpretation within the NMA process. This would reduce the time taken to conduct an NMA and reduce the level of human error.

We would recommend that the same level of (human) checking be applied to LLM extracted data and LLM generated software, as would be applied during a manual NMA process, i.e. that data extraction is checked, input data and software used are subjected to quality assurance and any text and outputs produced by the LLM is sense-checked and, if necessary, corrected and improved by a human.

Given the level of accuracy that we have observed, we believe that elements of AI should start to be incorporated into the SLR and NMA workflow now. We firmly believe that the accuracy issues we have encountered in this study are very likely to improve, and even disappear completely, with time.

Further research could also investigate the use of LLMs to extract data from published tables, to add generation of graphical outputs normally used to interpret NMA to the R script (e.g. forest plots and network diagrams) and to provide additional standard outputs (e.g. surface under the cumulative ranking curve SUCRA). Additionally, further research could investigate the use of LLMs to support further upstream tasks, such as choosing an appropriate statistical model, using deviance information criteria (DIC) to determine fixed or random effects, node splitting to determine heterogeneity, determining effect modification due to covariates, deciding the approach to analysis and conducting feasibility assessments, as well as scaling the methods developed in this study to conduct larger NMAs. Whether an LLM is capable of producing a more detailed report or generating sections of text for a full technical report could also be investigated.

# 5 Conclusions

This study offers evidence for the use of LLMs like GPT-4 in automating data extraction and NMA. The use of generative AI to automate NMAs offers great potential to enable quicker and less costly decision making in healthcare, and this potential should be further developed, so that it might be harnessed and used to deliver faster patient access to medicines.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s41669-024-00476-9.

**Acknowledgements** The SLRs and feasibility analyses used to identify publications and data for the non-small cell lung cancer analyses (case studies 2–4) were conducted by Precision-HEOR.

## Declarations

**Funding** This study was supported by Bristol Myers Squibb.

**Conflict of interest** Tim Reason, Emma Benbow, Julia Langham, and Andy Gimblett are consultants at Estima Scientific and have worked on behalf of Bill Malcolm and Sven L. Klijn, who are employees and shareholders of Bristol Myers Squibb.

**Availability of data and material** All data generated or analysed during this study are included in this published article (and its supplementary information files).

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Code availability** Examples for using R's 'multinma' package were provided to GPT-4 for context for the binary outcome (Beta-blockers example) and time-to-event outcomes (Parkinson's disease example). Vignettes, including these examples, are provided here: https://cran.r-project.org/web/packages/multinma/index.html.

**Author contributions** T.R.: idea conceptualisation, development of approach and editorial input. E.B.: development of approach, critical appraisal of SLR and NMA and drafting versions of the paper. J.L.: critical appraisal of SLR and NMA and editorial input. A.G.: Python and API technical input. S.L.K.: technical input and editorial input. W.M.: technical input and editorial input.

## References

1. Angelis A, Lange A, Kanavos P. Using health technology assessment to assess the value of new medicines: results of a systematic review and expert consultation across eight European countries. Eur J Health Econ. 2018;19(1):123–52. https://doi.org/10.1007/s10198-017-0871-0.
2. Jenei K, Raymakers AJN, Bayle A, Berger-Thürmel K, Cherla A, Honda K, et al. Health technology assessment for cancer medicines across the G7 countries and Oceania: an international, cross-sectional study. Lancet Oncol. 2023;24(6):624–35. https://doi.org/10.1016/S1470-2045(23)00175-4.
3. Barratt A, Irwig L, Glasziou P, Cumming RG, Raffle A, Hicks N, et al. Users' guides to the medical literature: XVII. How to use guidelines and recommendations about screening. JAMA. 1999;281(21):2029. https://doi.org/10.1001/jama.281.21.2029.
4. Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. BMC Med Res Methodol. 2018;18(1):5. https://doi.org/10.1186/s12874-017-0468-4.
5. Goodman C. HTA 101: essential information for newcomers. Health Technology Assessment International. https://htai.org/wp-content/uploads/2023/06/The-Newcomers-Guide-to-HTA-A-collection-of-resources-for-early-career-professionals-2.pdf. Accessed 05 Jan 2024.
6. Rouse B, Chaimani A, Li T. Network meta-analysis: an introduction for clinicians. Intern Emerg Med. 2017;12(1):103–11. https://doi.org/10.1007/s11739-016-1583-7.
7. McKenzie J, Brennan S, Ryan R, Thomson H, Johnston R. Chapter 9: summarizing study characteristics and preparing for synthesis in Cochrane handbook for systematic reviews of interventions version 64. www.training.cochrane.org/handbook. Accessed 05 Jan 2024.
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. https://doi.org/10.1136/bmjopen-2016-012545.
9. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Cochrane handbook for systematic reviews of interventions version 6.4. Cochrane; 2023. http://www.training.cochrane.org/handbook. Accessed 05 Jan 2024.
10. Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol. 2017;17(1):152. https://doi.org/10.1186/s12874-017-0431-4.
11. Khangura S, Konnyu K, Cushman R, Grimshaw J, Moher D. Evidence summaries: the evolution of a rapid review approach. Syst Rev. 2012;1(1):10. https://doi.org/10.1186/2046-4053-1-10.
12. Rueda JD, Cristancho RA, Slejko JF. Is artificial intelligence the next big thing in health economics and outcomes research? Value Outcomes Spotlight Magazine. 2019:22–4. https://www.ispor.org/docs/default-source/publications/value-outcomes-spotlight/march-april-2019/vos-heor-articles---rueda.pdf?sfvrsn=18cb16f5_0. Accessed 05 Jan 2024.
13. van Dinter R, Tekinerdogan B, Catal C. Automation of systematic literature reviews: a systematic literature review. Inf Softw Technol. 2021;136: 106589. https://doi.org/10.1016/j.infsof.2021.106589.
14. OpenAI. GPT-4. https://openai.com/chatgpt. Accessed 05 Jan 2024.
15. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys Syst. 2023;3:121–54. https://doi.org/10.1016/j.iotcps.2023.04.003.
16. Poldrack RA, Lu T, Beguš G. AI-assisted coding: experiments with GPT-4. 2023. https://arxiv.org/abs/2304.13187. Accessed 05 Jan 2024.
17. Cope S, Zhang J, Saletan S, Smiechowski B, Jansen JP, Schmid P. A process for assessing the feasibility of a network meta-analysis: a case study of everolimus in combination with hormonal therapy versus chemotherapy for advanced breast cancer. BMC Med. 2014;12(1):93. https://doi.org/10.1186/1741-7015-12-93.
18. Cope S, Mojebi A, Popoff E, Hertel N, Korytowsky B, McKenna M, et al. Comparative efficacy and safety of nivolumab versus relevant treatments in pretreated advanced non-small cell lung cancer: a systematic literature review and indirect treatment comparison of randomized controlled trials. Copenhagen: ISPOR; 2019. https://www.ispor.org/docs/default-source/euro2019/cope-et-al---comparative-efficacy-and-safety-of-nivolumab-pdf.pdf?sfvrsn=5634556b_0. Accessed 05 Jan 2024.
19. Phillippo DM. multinma: Bayesian network meta-analysis of individual and aggregate data https://cran.r-project.org/web/packages/multinma/index.html. Accessed 05 Jan 2024.
20. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. J Stat Softw. 2017;76(1). http://www.jstatsoft.org/v76/i01/. Accessed 05 Jan 2024.
21. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. 2022. https://arxiv.org/abs/2201.11903. Accessed 05 Jan 2024.
22. OpenAI. API reference. https://platform.openai.com/docs/api-reference/chat/create. Accessed 25 Jan 2024.
23. Edwards B. As ChatGPT gets "lazy," people test "winter break hypothesis" as the cause. ARS Technica. 2023. https://arstechnica.com/information-technology/2023/12/is-chatgpt-becoming-lazier-because-its-december-people-run-tests-to-find-out/. Accessed 05 Jan 2024.
24. Kimball AB, Jemec GBE, Alavi A, Reguiai Z, Gottlieb AB, Bechara FG, et al. Secukinumab in moderate-to-severe hidradenitis suppurativa (SUNSHINE and SUNRISE): week 16 and week 52 results of two identical, multicentre, randomised, placebo-controlled, double-blind phase 3 trials. Lancet. 2023;401(10378):747–61. https://doi.org/10.1016/S0140-6736(23)00022-3.
25. Cohen D. Applied Bayesian statistics using Stan and R: reproducibility. Methods Bites Tutorial. https://www.mzes.uni-mannheim.de/socialsciencedatalab/article/applied-bayesian-statistics/#reproducibility. Accessed 30 Jan 2024.
26. Ollama. https://ollama.ai. Accessed 30 Jan 2024.
27. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. ArXiv Comput Sci. 2023. https://doi.org/10.48550/arXiv.2307.09288.
28. OpenAI. ChatGPT can now see, hear, and speak. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak. Accessed 05 Jan 2024.

29. Google. Gemini AI. https://deepmind.google/technologies/gemini/#introduction. Accessed 05 Jan 2024.
30. OpenAI. GPT-3.5 Turbo fine-tuning and API updates. https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates. Accessed 05 Jan 2024.
31. Kimball A, Zouboulis C, Sayed C. Bimekizumab in patients with moderate-to-severe HS: 48-week efficacy and safety from BE HEARD I & II, two phase 3, randomized, double-blind, placebo controlled, multicenter studies. In: American Academy of Dermatology Annual Meeting, 2023, New Orleans, LA.
32. Herbst RS, Baas P, Kim DW, Felip E, Perez-Gracia JL, Han JY, et al. Factors associated with better overall survival (OS) in patients with previously treated, PD-L1-expressing, advanced NSCLC: Multivariate analysis of KEYNOTE-010. J Clin Oncol. 2017;35(15_suppl):9090–9090. https://doi.org/10.1200/JCO.2017.35.15_suppl.9090.
33. Herbst RS, Baas P, Kim DW, Felip E, Pérez-Gracia JL, Han JY, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. Lancet. 2016;387(10027):1540–50. https://doi.org/10.1016/S0140-6736(15)01281-7.
34. Vergnenegre A, Corre R, Berard H, Paillotin D, Dujon C, Robinet G, et al. Cost-effectiveness of second-line chemotherapy for non-small cell lung cancer: an economic, randomized, prospective, multicenter phase iii trial comparing docetaxel and pemetrexed: the GFPC 05–06 study. J Thorac Oncol. 2011;6(1):161–8. https://doi.org/10.1097/JTO.0b013e318200f4c1.

## Authors and Affiliations

**Tim Reason[1] · Emma Benbow[1] · Julia Langham[1] · Andy Gimblett[1] · Sven L. Klijn[2] · Bill Malcolm[3]**

✉ Tim Reason
tim.reason@estima-sci.com

[1] Estima Scientific, Mediaworks, 191 Wood Lane, London W12 7FP, UK

[2] Bristol Myers Squibb, Princeton, NJ, USA

[3] Bristol Myers Squibb, Uxbridge, UK