Session Title: 155 - RWE Methods and Applications

Innovations in Automated Survival Curve Selection and Reporting of Survival Analyses through Generative AI

Authors: Yunchou Wu (Estima Scientific), Sven Klijn (BMS), Siguroli Teitsson (BMS), Bill Malcolm (BMS), Cheryl Jones (Estima Scientific), Will Rawlinson (Estima Scientific)

Affiliations:

ISPOR EU, Barcelona, Spain
17th–20th November 2024
Session: P24, Monday 18th November 5pm

What is the problem we want to solve?

- **Survival analysis** is a crucial step involved in cost-effectiveness modelling (CEM), particularly in disease areas such as oncology
- Experts often spend **significant amounts of time** considering both survival analysis outputs and external data to ensure an appropriate curve is selected
- Experts must provide **clear rationale and justifications** for the curve selected
- The selection of a curve is a highly influential, and often sensitive, driver of CEM results

Abbreviations; *CEM, cost-effectiveness modelling; LLM; Large Language Models*

What we require of an automated system?

01

Automation to significantly reduce time spent on curve selection

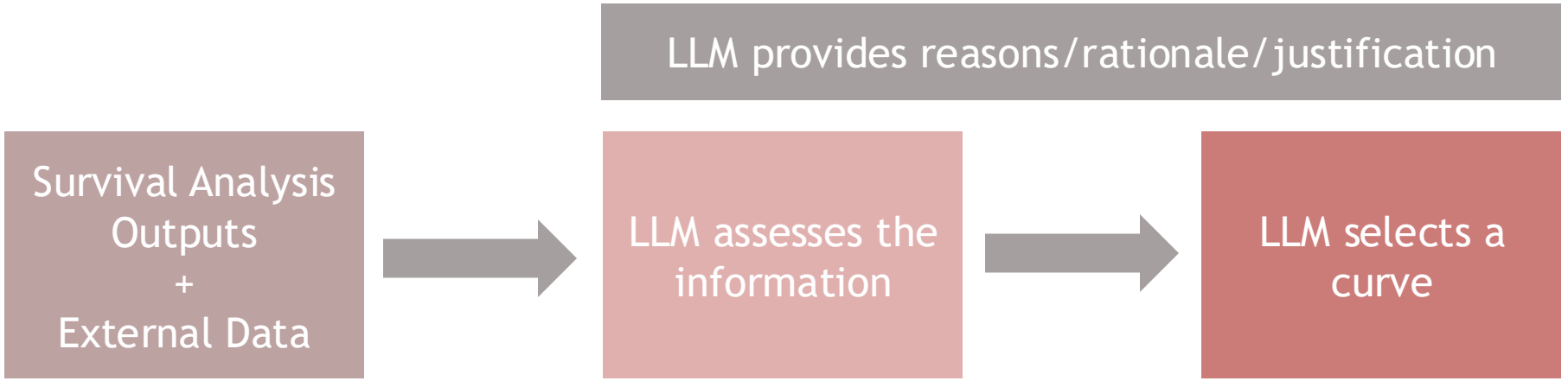
02

Require reasons and justifications behind the curves that have been selected

Abbreviations; *LLM; Large Language Models; HTA: Health Technology Assessment; HTD: Health Technology Developer*

Why are LLMs suitable for automating survival curve selection?

- Vision capabilities of LLMs have significantly improved → LLMs can interpret graphs
- Reasoning capabilities of LLMs → LLM can provide reasons for the assessments and decisions made



Abbreviations; *LLM; Large Language Models*

Study Objective

To leverage LLMs and explore automation of survival curve selection following published best practice guidelines[5]

LLMs were used to:

Assess data generated by a survival analysis and relevant external data

Recommend an appropriate extrapolation curve

Provide rationale and justifications for all decisions made

5. Latimer, N., 2011 (updated 2013). NICE DSU Technical Support Document 14: Survival analysis for Economic Evaluation alongside Clinical Trials – Extrapolation with Patient-Level Data. Decision Support Unit.

Two Case Studies in Oncology

Two survival reports previously accepted by NICE:

01 **NICE TA817:** Survival analysis for patients treated with Nivolumab vs. Placebo for resectable urothelial cancer (PD-L1 ≥1%) [CheckMate 274]

- **External data:** Published Kaplan Meier plot for placebo in similar patient population

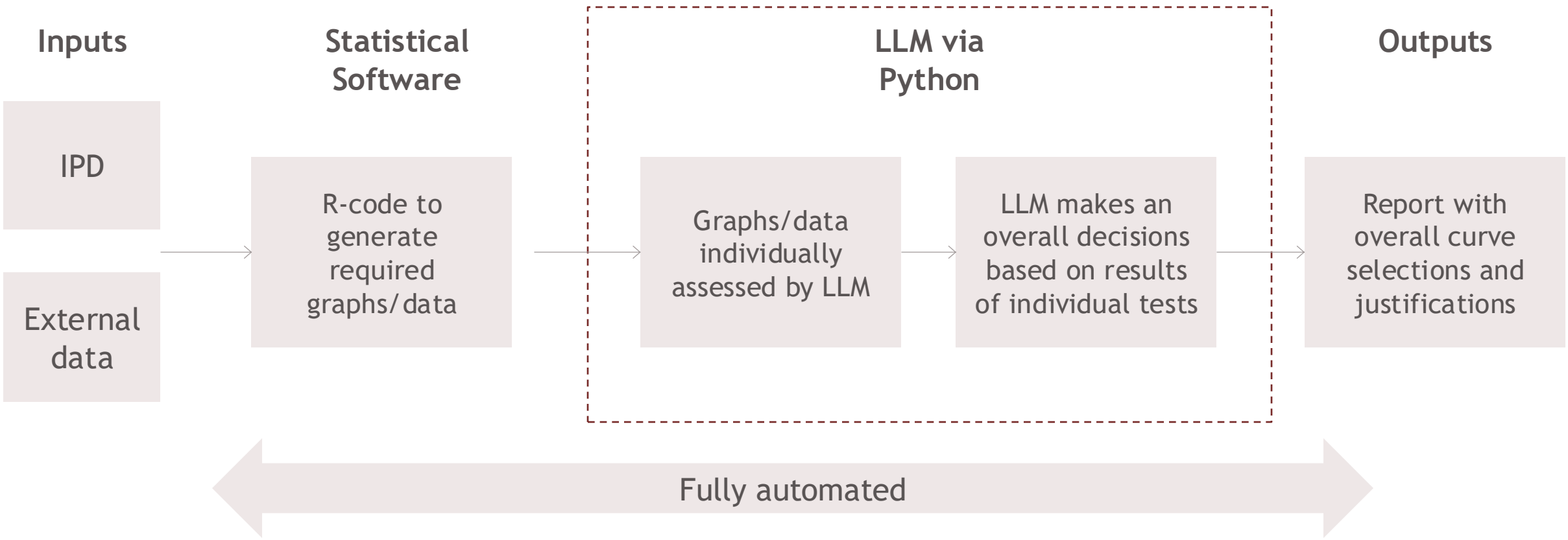
02 **NICE TA417:** Nivolumab vs. Everolimus for previously treated advanced Renal Cell Carcinoma (RCC) [CheckMate 025]

- **External data:** Clinical expert opinion – landmark survival estimates at year 5

Abbreviations: NICE: National Institute for Health and Care Excellence; PD-L1: Programmed Death-Ligand 1; TA: Technology Appraisal

Methods

Automated process for curve selection and reporting



Abbreviations; LLM: Large Language Models; IPD: Individual Patient Data

Structure of prompts to guide the LLM

Proportional Hazards

Log Cumulative Hazard Plot

Schoenfeld Residual's Plot

Grambsch-Therneau Test

Overall Decision

PH Holds Dependent Models	PH does not hold Independent Models	PH Unclear Both

Goodness-of-fit

AIC/BIC

Landmark Survival Probabilities

Visual Fit of Extrapolations (graph)

External Data

Overall Decision - Select Curve

Abbreviations; AIC: Akaike Information Criterion; BIC: Bayes Information Criterion; PH: Proportional Hazards | 5. Latimer, N., 2011 (updated 2013). NICE DSU Technical Support Document 14: Survival analysis for Economic Evaluation alongside Clinical Trials – Extrapolation with Patient-Level Data. Decision Support Unit.

Prompt Development

Prompting techniques used were:

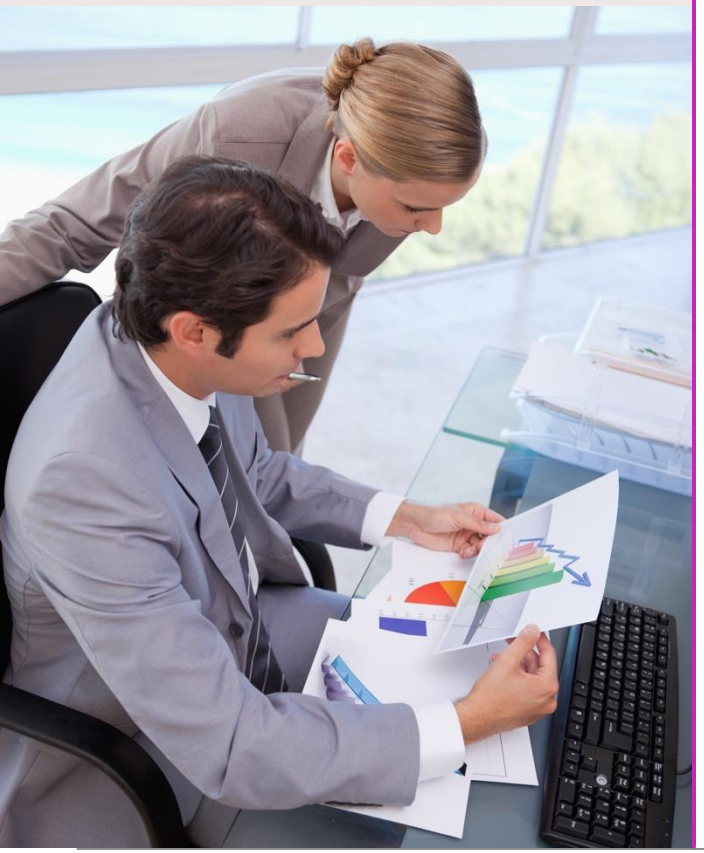
- Few-shots – Provide the LLM with a few relevant examples to guide its generation
- Chain-of-thought - Break down a complex problem into a sequence of logical steps for the LLM to follow

Prompts were designed to be generalizable (i.e., apply the same prompts to both case studies)

Claude 3.5 Sonnet [version 1] was used to generate the results

Abbreviations; LLM: Large Language Models

Assessing Accuracy of Responses



Assess the **rationales and/or justifications** provided by the LLM for decisions made against the expert opinion of 3 health economists

01 Interpretation of survival analysis outputs

02 Overall decisions made, taking account of external data

Abbreviations; LLM: Large Language Models; NICE: National Institute for Health Care Excellence

Results

Appropriate justifications provided by LLM

Assess the justifications for decisions made by the LLM against the expert opinion of 3 health economists

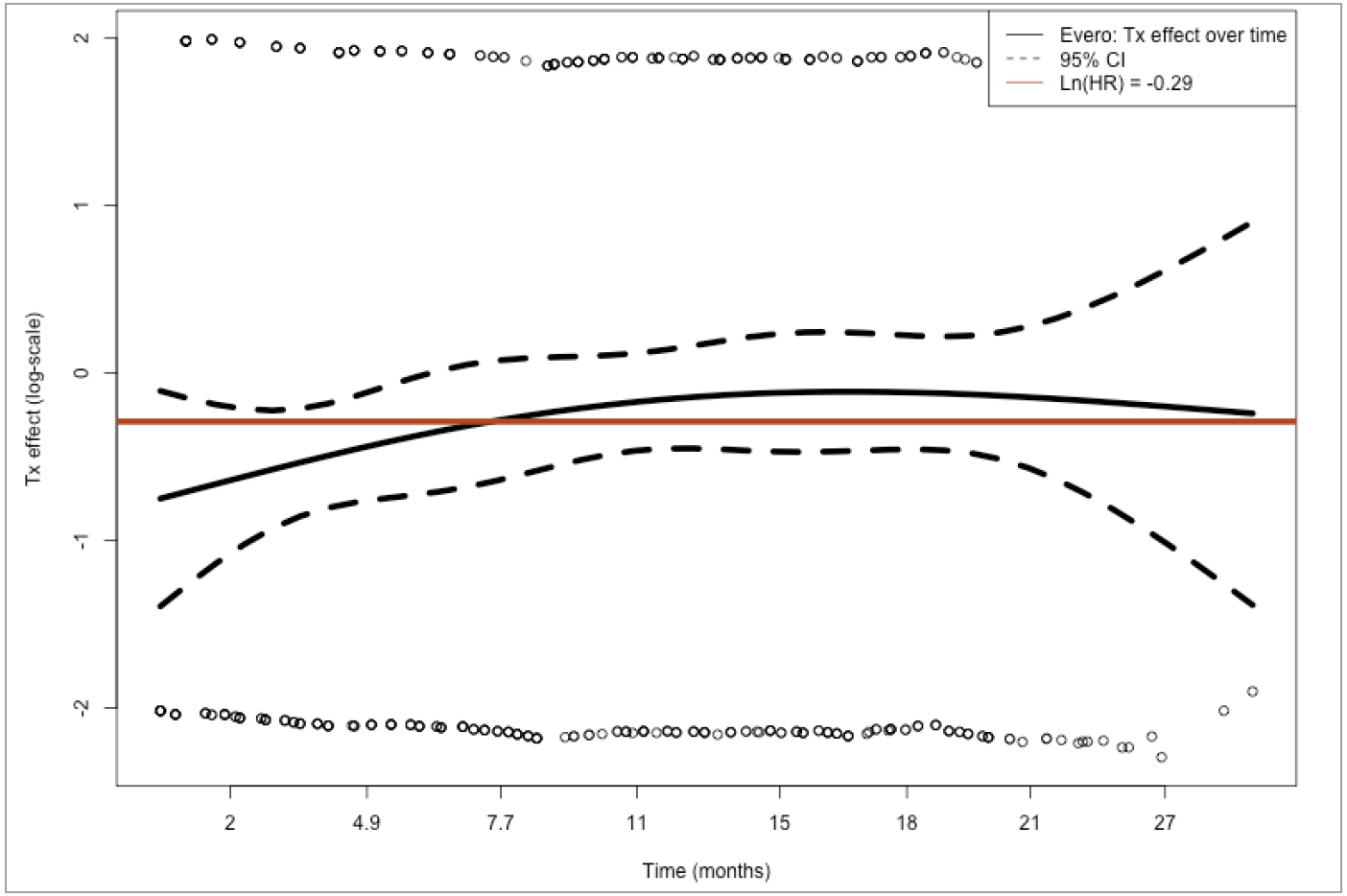
Trial	Statement Type	Number of Statements by LLM	Number of Appropriate Statements
NICE TA817	Interpretation of survival data	28	28 (100.0%)
	Justifications for overall decisions	14	14 (100.0%)
NICE TA417	Interpretation of survival data	28	27 (96.4%)
	Justifications for overall decisions	14	14 (100.0%)

Abbreviations; LLM: Large language models; NICE: National Institute for Health Care Excellence; TA: Technology Appraisal

Limitations of Vision Capabilities

- LLM stated the dashed 95% CI crossed the red-line
- Did not affect the overall interpretation and decision related to the plot
- Signals potential vision issues

Schoenfeld Residual Plot



Abbreviations: CI: Confidence Intervals; LLM: Large Language Models

Discussion

Automated Curve Selection is Possible

Our research suggests LLMs are able to:

- Interpret evidence generated by survival analysis
- Produce sensible rationales for its decisions
- Use survival analysis outputs and external evidence to inform appropriate curve selections

Abbreviations; LLM: Large Language Models

Further Research

01 Concerns about vision capabilities of Claude 3.5 Sonnet [version 1],[6] may impact reliability

- Claude 3.5 Sonnet [version 2] was released with enhanced vision capabilities
- Send data and images to the LLM

02 For the automated process to be most effective, it must also accommodate complex models, including Splines and Cure models

03 Incorporate a “human-in-the-loop” approach to add further context (e.g. clinical)

Abbreviations; HTA: Health Technology Assessment; LLM: Large language models | 6. Rahmanzadehgervi et al., 2024 “Vision language models are blind”. https://vlmsareblind.github.io

Conclusion

- Our research shows LLMs can be leveraged to automate the assessment of survival analysis outputs and select a curve to predict long-term outcomes for use in cost-effectiveness modelling
- LLMs can provide an additional view-point which may be valuable to experts selecting a curve
- Ultimately an automated survival analysis process can speed-up curve selection process and allowing patients faster access to life-saving treatments

Abbreviations; LLM: Large Language Models

Declaration of Interests

This study was supported by Bristol Myers Squibb

Wu Y, Jones C, Rawlinson W:

None

Klijn S, Teitsson S, Malcolm W:

Employees of Bristol Myers Squibb