The Best of Both Worlds? Hybrid Programmatic-AI Workflows for HTA-grade Evidence Generation for Time-to-event Analysis Interpretation and Reporting

MSR232

Simone Rivolo[1], Tereza Lanitis[2], George Bungey[3], Michalis Galanakis[4], Ivan Houisse[5], Jack Ishak[6]

[1]Thermo Fisher Scientific, San Felice Segrate, Italy; [2]Thermo Fisher Scientific, Limassol, Cyprus; [3]Thermo Fisher Scientific, London, UK; [4]Thermo Fisher Scientific, Chania, Greece; [5]Thermo Fisher Scientific, Budapest, Hungary; [6]Thermo Fisher Scientific, Waltham, MA, USA

Background

- Generative artificial intelligence (GenAI) can streamline research tasks required for health technology assessment (HTA) evidence generation, including interpretation and report drafting of time-to-event (TTE) statistical analyses.[1]
- Given the concerns around consistency and accuracy of GenAI outputs, TTE report drafting requires a repeatable workflow that:
  - Standardises the presentation of all HTA required TTE analyses, using GenAI where it adds most value;
  - Facilitates subject matter expert review[2] (human in the loop) for report finalisation.

Objectives

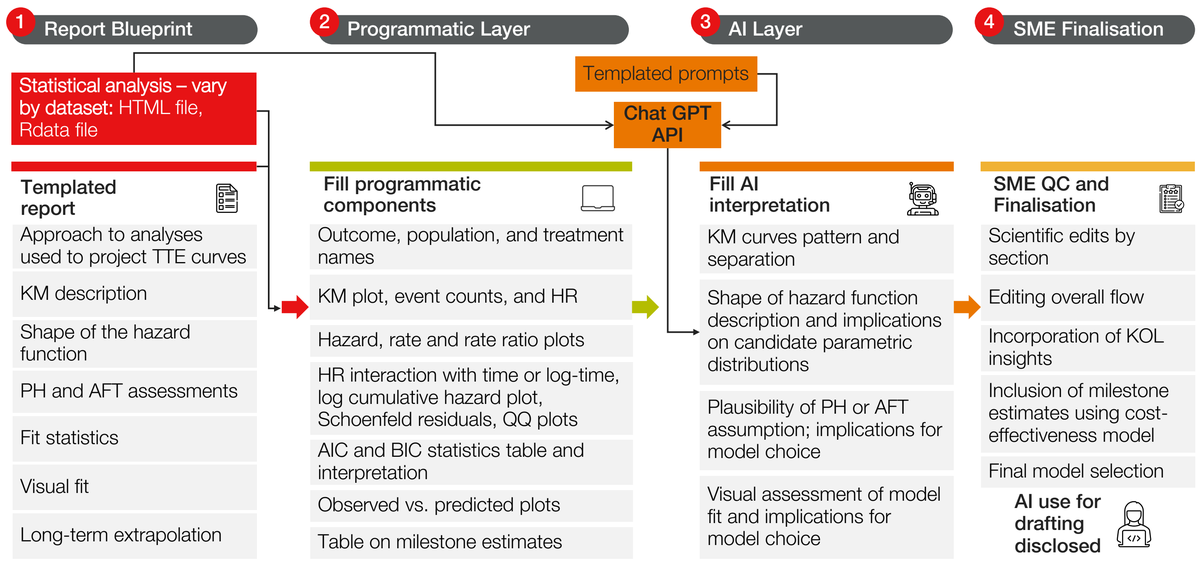
- To demonstrate how HTA-grade reporting of statistical analyses can be achieved through a hybrid programmatic-AI workflow using a TTE survival-analysis case study.

Methods

Workflow overview and development

- An overview of the overall TTE report drafting workflow (TTERep-AI) that was developed in R is shown in **Figure 1**.
- The report drafting workflow uses outputs from parametric survival analyses applied to TTE data for a single analysis (referred as a scenario). A scenario can include one or multiple outcomes (e.g., progression-free survival [PFS], overall survival [OS]) for a single dataset, either single arm or with two interventions.
- A report blueprint was developed in Microsoft Word® for each of the six sections shown in **Figure 1** (Box 1). The blueprint included description of the methodology for selecting the most plausible parametric models, based on NICE DSU TSD guidance.[3,4]
- Bookmarks indicate where the programmatic layer should insert scenario-specific results across each section. Contents populated by the programmatic layer are summarised in **Table 1**.

Figure 1. Overview of How TTERep-AI Builds the Draft Report



Abbreviations: AFT = accelerated failure time; AI = artificial intelligence; AIC = Akaike information criterion; API = application programming interface; BIC = Bayesian information criterion; HR = hazard ratio; KM = Kaplan-Meier; KOL = key opinion leader; PH = proportional hazards; QC = quality control; QQ = quantile-quantile; SME = subject matter expert; TTE = time-to-event

Table 1. Summary of Content Filled by the Programmatic Layer

Section	Content
KM plots with descriptions	KM plots with event counts by arm and hazard ratio summarising relative effectiveness across arms (excluded in case of single-arm reporting)
Shape of the Hazard Function	KM hazard-shape interpretations based on: • Smoothed hazard plots over the trial period • Average hazard rates and rate ratio over time across survival quintiles (i.e., 20th, 40th, 60th, 80th percentiles) of times where events were observed across all groups
PH and AFT assessments	PH and AFT assessments with: • Interaction between effect of treatment (log-HR) over time and log-time • Log cumulative hazard plots • Schoenfeld residuals • QQ plots
AIC and BIC fit statistics	AIC and BIC fit statistics tabulation and automated summary, using the following rules of thumb[5–7] to interpret relative fit compared to the models with the lowest AIC and BIC values, adapted from the literature: • Models with a difference of <4 point in AIC or BIC from the lowest value are classified as having a statistical fit "equivalent" to that of the model with the lowest AIC/BIC model • Models with differences of 4–7 points are considered to have a "comparable" statistical fit relative to the model with the lowest AIC/BIC model • Models with differences of >7 points in AIC or BIC are considered to have an "inferior" statistical fit relative to model with the lowest AIC/BIC model
Visual fit assessment of each parametric model compared to the observed data	Visual summary, over the trial period, of observed KM overlaid with the parametric fits included in the TTE analysis (i.e., independent and joint standard parametric fits)
Long-term extrapolation	Tabulated long-term predicted milestones by model type (joint vs. independent), treatment, and standard parametric distributions fitted

Abbreviations: AFT = accelerated failure time; AIC = Akaike information criterion; BIC = Bayesian information criterion; HR = hazard ratio; KM = Kaplan-Meier; OS = overall survival; PFS = progression-free survival; PH = proportional hazards; QQ = quantile-quantile; TTE = time-to-event

- The AI layer is then used to generate narrative descriptions and interpretations using section-specific prompts iteratively developed by subject matter experts (SMEs) using two TTE datasets.
  - For each section requiring interpretation, the following are provided to ChatGPT 5.2 through an application programming interface (API):
    - Behavioural instruction (persona, role, behaviour, tone, expected output)
    - Section-specific templated prompt with relevant figures or tables from the TTE analysis output
- To facilitate and focus SME review, the programmatic-layer results and the AI-drafted paragraphs are highlighted in different colours in the draft report.

Methods (cont.)

Workflow testing

- The programmatic layer was checked against the original TTE analysis outputs to ensure the results were transferred correctly to the draft reports.
- The AI-generated narrative descriptions and interpretation of results were evaluated across six separate TTE datasets by two independent SMEs who were not involved in developing the original TTERep-AI workflow.
  - Narrative quality and time savings attributable to the AI layer, compared with human-only report drafting, were rated using a 5-point Likert scale; results are summarised in **Table 2**.

Table 2. Score Descriptions for the 5-Point Likert Scale: Report Quality and Time Savings

Score	Report Quality: Score description	Time Savings: Score description
5	Practically no changes needed, except for minor tweaks (e.g., minor adjustments to language based on SME preference, deletion of at most one sentence since not needed)	Over 90% reduction in drafting time
4	Minor changes needed (e.g., more than simple language tweaks but in few specific sentences of no more than 1-2 paragraphs)	70–90% reduction in drafting time
3	Intermediate level of changes, with the SME needing to rewrite some of the sentences across paragraphs or most sentences in one of the paragraphs.	50–70% reduction in drafting time
2	Major changes required across most paragraphs, with only few sentences retained	30–50% reduction in drafting time
1	The paragraphs produced were not useful, with most/all requiring the SME to completely rewrite them	Essentially no reduction in drafting time (<30%)

Abbreviation: SME = subject matter expert

Results

- TTERep-AI was able to generate draft reports for SME review within minutes. The programmatic layer ensured correct handling of statistical analysis outputs, including automated description of AIC/BIC fit statistics.
  - The programmatic layer's description and interpretation of AIC/BIC fit statistics were satisfactory during testing. Using AI to expand or replace this interpretation was therefore considered unnecessary.
- SMEs reported the following quality scores for AI-generated interpretive text.
  - Only minor edits were required for the KM and hazard-shape description sections (scores ≥4). These edits included refining obvious statements (e.g., "The KM curves decline over time…").
  - PH/AFT assessments required revisions ranging from none to an intermediate level, depending on the dataset (scores ≥3).
  - AI drafting was least effective for visual-fit interpretation, which required major SME revision. This may reflect the subjective nature of the task and the difficulty of distinguishing acceptable visual fits from poor visual fits.
- SMEs estimated time savings attributable to AI drafting were greatest for PH/AFT assessments (50% to >90% time savings), followed by hazard-function shape descriptions (50% to 90%) and KM descriptions (30% to 90%). Visual-fit interpretation yielded limited (<30%) or no time savings since it required substantial SME rework.
- All draft reports required SME editing for finalisation. Despite the challenges noted in some of the sections, SMEs reported several benefits of TTERep-AI workflow, including:
  - Focusing expert time on analysis interpretation and model-selection implications, rather than copy-and-paste work and drafting the report from scratch;
  - Reducing the chance of copy-and-paste errors and supporting consistent organisation across outcomes;
  - Reducing effort required to produce clear technical language;
  - Completing reports faster, even after the SME needed to refine the AI interpretation.

Conclusions

- Our study indicates that hybrid programmatic-AI workflows can improve consistency, transparency, and efficiency in generating TTE statistical reports, while maintaining alignment with HTA reporting requirements.
- Future developments should improve AI performance in visual-fit assessment, extend the TTERep-AI workflow to more flexible parametric models (e.g., splines), and include AI as clinical assistant[8] to guide final model selection based on the plausibility of long-term extrapolations.
  - As the AI foundational models improve, the TTERep-AI workflow could delegate more drafting to AI, including running the programmatic workflow and completing long-term extrapolation. AI could also support final model selection using published literature and key opinion leader (KOL) insights.

References

1. Fleurence RL, Bian J, Wang X, et al. *Value Health*. 2025;28(2):175-83.
2. National Institute for Health and Care Excellence. Use of AI in evidence generation: NICE position statement. Published August 15, 2024. https://www.nice.org.uk/corporate/ecd11/chapter/our-position-on-the-use-of-ai-in-evidence-generation-and-reporting
3. Latimer N. NICE DSU Technical Support Document 14: Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data. 2011. http://nicedsu.org.uk/wp-content/uploads/2016/03/NICE-DSU-TSD-Survival-analysis.updated-March-2013.v2.pdf
4. Rutherford MJ, Lambert PC, Sweeting MJ, et al. NICE DSU Technical Support Document 21: Flexible methods for survival analysis. 2020. https://sheffield.ac.uk/nice-dsu/tsds/full-list
5. Burnham KP, Anderson DR. *Sociol Methods Res*. 2004;33:261-304.
6. Burnham KP, Anderson DR, Huyvaert KP. *Behav Ecol Sociobiol*. 2011;65:23-35.
7. Kass RE, Raftery AE. *J Am Stat Assoc*. 1995;90:773-95.
8. Ishak KJ, Caro JJ. MSR22: AI-assisted time-to-event projection: a case study and broader potential. *Value Health*. 2025;28(suppl 2):MSR22.

Disclosures

All authors are employees or contractors of PPD™ Evidera™ Health Economics & Market Access, Thermo Fisher Scientific, and may hold stocks or stock options in the company. Funding for this poster is provided by Thermo Fisher Scientific.

Acknowledgments

Editorial and graphic design support were provided by Carrie Collins-Cann and Shani Berger of Thermo Fisher Scientific.

Contact: simone.rivolo@thermofisher.com

© 2026 Thermo Fisher Scientific Inc. All rights reserved. All trademarks are the property of Thermo Fisher Scientific and its subsidiaries unless otherwise specified. This information is not intended to encourage use of these products in any manner that might infringe the intellectual property rights of others.