Use of Generative AI in Health Economic Modeling:

Summary Report

May 8, 2025

NICE
National Institute for Health and Care Excellence

Value Analytics Labs

Value Analytics Labs Team

Jagpreet Chhatwal, PhD
Chief Scientific Officer, *Value Analytics Labs;*
Director, *Massachusetts General Hospital Inst. for Technology Assessment;*
Associate Professor, *Harvard Medical School*

I. Fatih Yildrim, MS
Director of Software Solutions, Value Analytics Labs

Jade Xiao, PhD
Data Scientist, Value Analytics Labs

Elif Bayraktar, BS
Software Engineer, Value Analytics Labs

Sumeyye Samur, PhD
VP, Head of Value and Access, Value Analytics Labs

Rachael Fleurence, PhD
Head of Evidence and AI Solutions, Value Analytics Labs

Turgay Ayer, PhD
Chief Technology Officer, *Value Analytics Labs;*
Professor & Director of Healthcare Analytics, CCHS, *Georgia Tech*

Akash Ramanarayanan
Intern**,** Value Analytics Labs

Jaykirit Palani, MS
Intern**,** Value Analytics Labs

Jagpreet Chhatwal served as the primary lead for this project. Sumeyye Samur co-led the initiative. Fatih Yildirim led the AI implementation and software development. Jade Xiao contributed to the country-specific adaptations. Elif Bayraktar supported software development efforts. Rachael Fleurence assisted with report editing. Turgay Ayer provided methodological guidance. Akash Ramanarayanan and Jaykirit Palani contributed to the model validations.

Table of Contents

1 Introduction

The integration of artificial intelligence (AI), particularly large language models (LLMs) and generative AI (GenAI), into health economic modeling has garnered increasing interest in recent years (1). With their unprecedented capacity to process large volumes of unstructured text and extract relevant information, these tools offer the potential to revolutionize how we construct, adapt, and validate health economic models. Traditional model development, especially for cost-effectiveness analyses (CEAs), is a time-intensive and expert-driven process that often involves significant manual effort to develop model structures, parameters, and assumptions from heterogeneous sources such as clinical trials, health technology assessment (HTA) reports, and peer-reviewed literature (1). As demands for localized economic evidence grow, so does the need for more efficient, scalable, and reproducible methods of adapting models across jurisdictions.

Despite the promise of GenAI to accelerate these processes, its actual performance in health economic modeling contexts remains largely untested (1, 2). While AI has demonstrated capabilities in related areas such as literature review automation and evidence synthesis, its application to tasks such as model development, adaptation, and validation has not been comprehensively evaluated (3). Notably, health economic modeling poses unique challenges for AI, including the need to interpret domain-specific terminology, contextual understanding of modeling, methodological assumptions, and reconcile heterogeneous data formats (1). As such, both the strengths and limitations of GenAI in these applications are currently poorly understood.

Recognizing this gap, National Institute for Health and Care Excellence (NICE) and the Value Analytics Labs (VAL) team initiated a collaborative project to systematically assess the utility of GenAI in health economic model adaptation and validation. This initiative aligns with NICE's broader efforts to responsibly integrate AI into HTA processes, as outlined in their position statement on the use of AI in evidence generation (4).

This report documents the initial set of experiments conducted under this collaboration, structured across two distinct projects: **1) Country Adaptation** and **2) Model Validation**. In

the first project, we adapted a previously published Alzheimer's disease model from the Institute for Clinical and Economic Review (ICER) as the foundation to evaluate GenAI's capabilities across three key stages: (i) replicating a cost-effectiveness model from a published report; (ii) adapting the model to a different country—specifically, the United Kingdom; and (iii) updating the model and its code with UK-specific inputs to automatically generate localized outputs. In the second project, we applied a validation protocol to an independently developed Excel-based economic model for asthma developed by NICE, using scenarios aligned with NICE's recommended validation practices.

Through this collaborative initiative, we aim to produce evidence-based insights into the opportunities and limitations of AI-assisted modeling, and to inform future methodological guidance on the use of AI in health technology assessment and economic modeling.

2 Project I: Country Adaptation

2.1 Introduction

The Country Adaptation Project evaluated the feasibility and utility of using GenAI to adapt health economic models for different national healthcare systems. Specifically, this project aimed to explore whether GenAI could effectively streamline the adaptation of a US-based Alzheimer's disease model to the United Kingdom (UK) healthcare context. Traditionally, adapting models across countries is a labor- and resource-intensive process that requires substantial manual effort to ensure contextual accuracy and analytical integrity. This project investigated whether GenAI could reduce these resource burdens while maintaining methodological rigor and alignment with country-specific guidelines.

2.2 Overall Approach

In collaboration with NICE, Alzheimer's disease was selected as the target condition for this case study. The adaptation process was structured into three key stages:

Stage I: Model Replication using ValueGen.AI
Since we did not have direct access to an existing U.S.-based Alzheimer's disease model, we first replicated a published model from the Institute for Clinical and Economic Review (ICER). The ValueGen.AI, a proprietary GenAI-based platform was used to automate extraction of the model's structural components, including health states, transition probabilities, costs, and utilities, from the ICER report. Due to the complexity of some data formats (e.g., non-standard parameter definitions, multipliers), a human-in-the-loop approach was implemented to resolve ambiguities and verify extracted data.

Stage II: AI-Integrated Country Adaptation Roadmap
Next, we developed a roadmap for country adaptation that could be used by GenAI. Specifically, we identified model parameters that would need to be adapted to local settings (5-7). Key parameters adapted included discount rates, background mortality, costs, utilities, and disease progression rates. AI-based literature reviews and database queries were conducted to source relevant UK data, with expert oversight for parameter selection.

We identified and summarized UK-specific model parameters from NICE technology appraisals, HTA reports, published literature, and international databases.

We categorized parameters into:

1. Automatically updated parameters
    - Discount rate
    - Background mortality
2. Parameters adapted based on data availability
    - Costs
    - Utilities
    - Disease progression

For simplicity, we assumed that the natural history of Alzheimer's disease did not change for UK setting.

Stage III: Automated Model Update

After finalizing the UK-specific parameters, ValueGen.AI integrated these inputs to automatically update the US-based model code and generate localized (UK-specific) outputs. This automated workflow demonstrated the feasibility of GenAI-assisted model adaptation with minimal manual coding intervention.

2.3 Alzheimer's Model Replication using ValueGen.AI

2.3.1 ValueGen.AI Technical Details

ValueGen.AI is a proprietary GenAI tool designed to automate the development of health economic models from structured or semi-structured documents. The platform processes uploaded reports to extract key model components, generate model (Markov or partition survival) code in R, execute simulations, and display results in a user-friendly interface. Users can download the R code for further analysis.

Below we summarize the technical features of ValueGen.AI:

1. **Multi-modal generative AI models:** ValueGen.AI utilizes multiple generative AI models, including OpenAI GPT-4o and Gemini 2.5 Pro, to extract information from documents. These models are capable of understanding and processing various types of data, including text, images, and structured data, ensuring comprehensive information extraction. Without these models, the extraction process would be less accurate and more time-consuming, as it would rely on manual data entry and interpretation.
2. **Unified interface for language models:** The tool interacts with various language models using the LangChain library, providing a seamless and unified interface. This allows ValueGen.AI to leverage the strengths of different language models, ensuring accurate and contextually relevant information extraction. Without this unified interface, users would need to interact with each language model separately, leading to inconsistencies and inefficiencies in data extraction.
3. **Custom agentic data extraction:** To enhance the quality of extracted information, ValueGen.AI employs a custom agentic data extraction agent developed using the LangGraph library. This agent is designed to intelligently navigate and extract data from complex documents, ensuring high precision and reliability. Without this custom agent, the extraction process would be less precise and more prone to errors, as it would rely on generic extraction methods.
4. **Agentic and non-agentic systems:** The tool incorporates both agentic and non-agentic systems for data retrieval. Agentic systems are designed to actively seek out and extract relevant information, while non-agentic systems passively process and analyze data. This combination ensures comprehensive and accurate information extraction. Without these systems, the retrieval process would be less thorough and more likely to miss critical information.
5. **Semantic and vector searches:** Retrieval processes utilize both custom and general embedding models for semantic search and vector searches. Semantic search allows the tool to understand the meaning and context of the information, while vector searches enable efficient and accurate retrieval of relevant data. Without these search capabilities, the retrieval process would be less efficient and more likely to return irrelevant or incomplete results.
6. **Micro-service architecture:** ValueGen.AI is an in-house developed full-stack application with a micro-service software architecture design. This architecture allows for modular development, scalability, and easy maintenance, ensuring the

tool can handle large volumes of data and complex processing tasks. Without this architecture, the application would be less scalable and more difficult to maintain, leading to potential performance issues.

7. **Backend development:** The backend is developed using Python, featuring robust data privacy and security measures. Python's versatility and extensive libraries enable efficient data processing and integration with various systems, ensuring the tool's reliability and performance. Without these measures, the application would be more vulnerable to data breaches and less capable of handling complex data processing tasks.
8. **Frontend development**: The frontend of the application is built using React, offering a user-friendly and responsive interface. React's component-based architecture allows for dynamic and interactive user experiences, making it easy for users to interact with the tool and access its features. Without this frontend development, the user interface would be less intuitive and more difficult to navigate, leading to a poorer user experience.
9. **R packages and integration:** ValueGen.AI is integrated with R for building comprehensive health economic models. It supports development of cohort-level state-transition (Markov) models and partition survival models. Utilizing the heemod R package, w employed an R code template designed for creation of a Markov model our case study. The integration process is streamlined and efficient:
10. **Template utilization:** The integration begins with a pre-defined R code template for Health Economic Markov models, leveraging the capabilities of the heemod R package.
    a. **Parameter extraction:** ValueGen.AI extracts essential parameters—such as states, costs, utilities, and transition probabilities—from the uploaded document source.
    b. **Model generation:** Using the extracted parameters, ValueGen.AI generates a working R model code based on the template.
    c. **Simulation execution:** The generated R code is then executed to run the simulation model.
    d. **Result display:** Simulation results are displayed on screen, providing users with immediate insights.
    e. **Code download:** Users have the option to download the generated R code for further analysis or reuse.

This integration ensures that users can easily create, execute, and analyze Health Economic Markov models with minimal manual intervention, enhancing both efficiency and accuracy in health economics research and simulations.

Prompting Techniques and Approach

Our approach leveraged a sophisticated AI methodology, beginning with strategic prompt engineering. A range of prompting techniques was employed, including zero-shot for straightforward queries, few-shot learning where illustrative examples enhanced task performance (such as when we needed the output to follow a specific format), and Chain-of-Thought (CoT) reasoning for more complex, multi-step problem-solving. The selection of these techniques was tailored to the specific complexity and nature of each task.

Prompting followed an iterative process, wherein large or complex objectives were systematically decomposed into smaller, more achievable sub-tasks. This modular approach facilitated progressive refinement and the construction of increasingly sophisticated results. This iterative approach was underpinned by a robust technology stack, incorporating foundational libraries and tools such as the OpenAI API, orchestration frameworks like LangChain, and graph-based state management with LangGraph.

Furthermore, our AI pipeline integrated several key components to enhance performance and versatility. Retrieval Augmented Generation (RAG) was utilized to ground responses in relevant external knowledge, thereby improving accuracy and contextual relevance. Vector Databases, populated with embeddings, enabled efficient semantic search and retrieval of pertinent information. These core components were augmented by additional capabilities, including broad web search functionalities and targeted information retrieval from custom databases (e.g., NICE, Tufts CEA Registry, HTA repositories). A critical outcome of this integrated pipeline was the consistent generation of formatted outputs, ensuring data usability for subsequent analysis or integration.

2.3.2 Alzheimer's Disease Model Development Using ValueGen.AI

We replicated the ICER Alzheimer's disease model using ValueGen.AI, guided by human experts to interpret complex inputs like hazard ratios, cost multipliers, and disutilities.

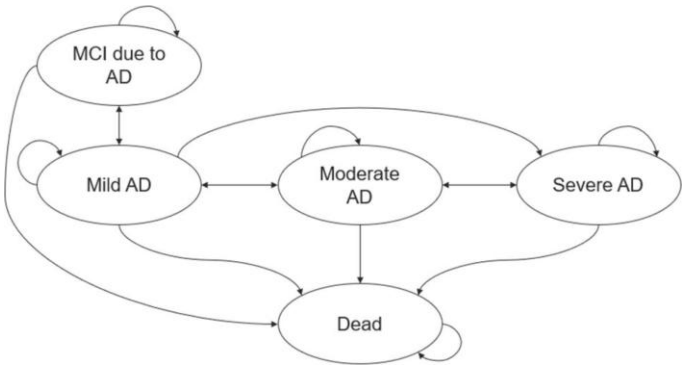

Figure 1. Model schematic for the ICER Alzheimer's disease model. Source: Lin et al. (2023).

Figure 1 shows the model schematic of Alzheimer's model available in ICER report (8). Several model parameters were reported indirectly (e.g., multipliers to a general population value, hazard ratios of disease progression) or in a non-standardized format (e.g., costs other than health state costs, health state disutilities rather than utilities). Therefore, human-in-the-loop was needed for ValueGen.AI to extract relevant parameter values. However, ValueGen.AI enables the user to easily modify model parameters.

In this experiment, human-in-the-loop was required to:

1. Define custom health states to incorporate the setting of Alzheimer's care;
2. Modify transition probabilities for the custom health states;
3. Apply hazard ratios of disease progression while taking Lecanemab;
4. Add transition probabilities for the transfer from Community to Long-Term Care;
5. Compute health state costs for the custom health states, including sourcing general population age-specific costs which were not stated nor referenced in the ICER report, and;
6. Compute health state utilities based on disutilities listed in the ICER report.

Figure 2 is a schematic of the final human-in-the-loop ValueGen.AI model.

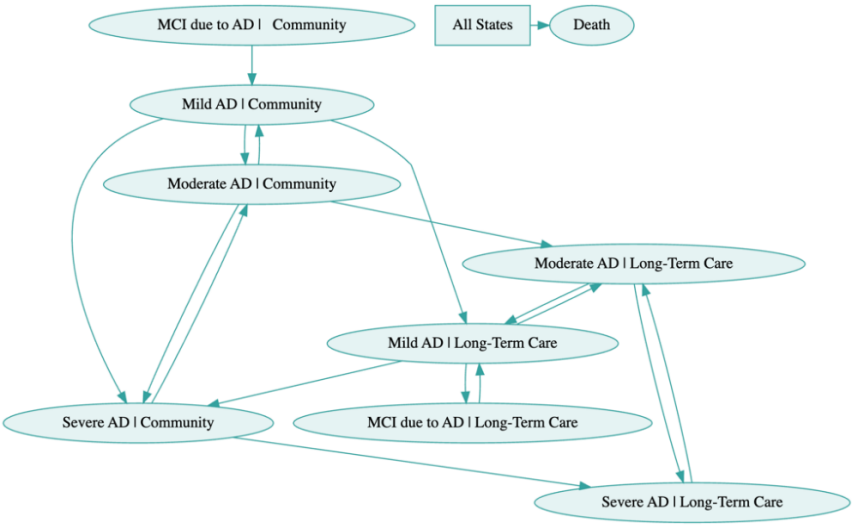

Figure 2. Model schematic for the ValueGen.AI Alzheimer's disease model.

Table 1 is a comparison of the structure of the two models.

Table 1. Comparison of model structures and parameters.

Model Aspect	ICER	ValueGen.AI
Health states	• MCI due to AD • Mild AD • Moderate AD • Severe AD • Dead	• MCI due to AD | Community • Mild AD | Community • Moderate AD | Community • Severe AD | Community • MCI due to AD | Community • Mild AD | Long-Term Care • Moderate AD | Long-Term Care • Severe AD | Long-Term Care • Dead
Attributes	**Setting** • Community • Long-term care	-
Cycle length	1 year	1 year
Time horizon	Lifetime	Lifetime

Initial population	• **Age:** 71 years
	• **Proportion female:** 52%
	• **Health state distribution:**
	  • MCI due to AD: 55%
	  • Mild AD: 45%
	• **Care setting distribution:**
	  • Community: 92%
	  • Long-term care: 8%
		• **Age:** 71 years
		• **Proportion female:** -
		• **Health state distribution:**
		  • MCI due to AD | Community: 55% x 92% = 50.6%
		  • Mild AD | Community: 45% x 92% = 41.4%
		  • MCI due to AD | Long-Term Care: 55% x 8% = 4.4%

		• Mild AD | Long-Term Care: 45% x 8% = 3.6%
Strategies	• Lecanemab • Supportive care	• Lecanemab • Supportive care

Transition probabilities	**Natural history**
	• MCI due to AD → MCI due to AD: 0.77
	• MCI due to AD → Mild AD: 0.23
	• Mild AD → MCI due to AD: 0.03
	• Mild AD → Mild AD: 0.58
	• Mild AD → Moderate AD: 0.35
	• Mild AD → Severe AD: 0.04
	• Moderate AD → Mild AD: 0.03
	• Moderate AD → Moderate AD: 0.55
	• Moderate AD → Severe AD: 0.42
	• Severe AD → Moderate AD: 0.02
	• Severe AD → Severe AD: 0.98
	
	Mortality
	• MCI due to AD: 1.82 x ACM
	• Mild AD: 2.92 x ACM
	• Moderate AD: 3.85 x ACM
	• Severe AD: 9.52 x ACM
	
	Transfer to long-term care
	• MCI due to AD: 0.024
	• Mild AD: 0.038
	• Moderate AD: 0.11
	• Severe AD: 0.259
	
	Treatment effectiveness
	HR of disease progression
	• MCI due to AD: 0.69
	• Mild AD: 0.69
	
	Adverse events
	ARIA in the first year of treatment
	• Probability of any ARIA: 0.215
	• Probability of symptomatic ARIA: 0.035
	• Probability of treatment discontinuation halfway through the first year of treatment: 0.069
		Natural history
		Lecanemab
		• MCI due to AD | Community →
		  • Mild AD | Community: 0.15
		  • MCI due to AD | Long-Term Care: 0.03
		• Mild AD | Community →
		  • MCI due to AD | Community: 0.03
		  • Moderate AD | Community: 0.23
		  • Severe AD | Community: 0.03
		  • Mild AD | Long-Term Care: 0.03
		• Moderate AD | Community →
		  • Mild AD | Community: 0.03
		  • Severe AD | Community: 0.38
		  • Moderate AD | Long-Term Care: 0.09
		• Severe AD | Community →
		  • Moderate AD | Community: 0.02
		  • Severe AD | Long-Term Care: 0.2
		• MCI due to AD | Long-Term Care →
		  • Mild AD | Long-Term Care: 0.15
		• Mild AD | Community →
		  • MCI due to AD | Community: 0.03
		  • Moderate AD | Community: 0.23
		  • Severe AD | Community: 0.03
		• Moderate AD | Community →
		  • Mild AD | Community: 0.03
		  • Severe AD | Community: 0.38
		• Severe AD | Community →
		  • Moderate AD | Community: 0.02
		Supportive care
		• MCI due to AD | Community →
		  • Mild AD | Community: 0.22
		  • MCI due to AD | Long-Term Care: 0.03
		• Mild AD | Community →
		  • MCI due to AD | Community: 0.03
		  • Moderate AD | Community: 0.34
		  • Severe AD | Community: 0.04
		  • Mild AD | Long-Term Care: 0.03
		• Moderate AD | Community →
		  • Mild AD | Community: 0.03
		  • Severe AD | Community: 0.38
		  • Moderate AD | Long-Term Care: 0.09
		• Severe AD | Community →
		  • Moderate AD | Community: 0.02
		  • Severe AD | Long-Term Care: 0.2
		• MCI due to AD | Long-Term Care →

		• Mild AD | Long-Term Care: 0.22
		• Mild AD | Community →
		  • MCI due to AD | Community: 0.03
		  • Moderate AD | Community: 0.34
		  • Severe AD | Community: 0.04
		• Moderate AD | Community →
		  • Mild AD | Community: 0.03
		  • Severe AD | Community: 0.38
		• Severe AD | Community →
		  • Moderate AD | Community: 0.02
		
		Mortality
		• MCI due to AD | Community: 1.82 x ACM
		• Mild AD | Community: 2.92 x ACM
		• Moderate AD | Community: 3.85 x ACM
		• Severe AD | Community: 9.52 x ACM
		• MCI due to AD | Long-Term Care: 1.82 x ACM
		• Mild AD | Long-Term Care: 2.92 x ACM
		• Moderate AD | Long-Term Care: 3.85 x ACM
		• Severe AD | Long-Term Care: 9.52 x ACM

Costs	**Health states** Cost multipliers relative to an individual from the general US population (base costs are unstated and unreferenced) • MCI due to AD: 1.12 • Mild AD: 1.56 • Moderate AD: 1.93 • Severe AD: 1.93 **Interventions** • Drug acquisition: $26,500 annual WAC • Drug administration: $78.35 per administration **Long-term care costs** $7,394 per month	**Health states** Lecanemab • MCI due to AD | Community: $36,366 • Mild AD | Community: $39,038 • Moderate AD | Community: $11,758 • Severe AD | Community: $11,678 • MCI due to AD | Long-Term Care: $125,094 • Mild AD | Long-Term Care: $127,766 • Moderate AD | Long-Term Care: $100,486 • Severe AD | Long-Term Care: $100,406 • Death: $0 Supportive care • MCI due to AD | Community: $6,784 • Mild AD | Community: $9,457 • Moderate AD | Community: $11,758 • Severe AD | Community: $11,678 • MCI due to AD | Long-Term Care: $95,512 • Mild AD | Long-Term Care: $98,185 • Moderate AD | Long-Term Care: $100,486 • Severe AD | Long-Term Care: $100,406 • Death: $0
Utilities	**Health states | Community** Disutilities • MCI due to AD: -0.17 • Mild AD: -0.22 • Moderate AD: -0.36 • Severe AD: -0.53 **Health states | Long-Term Care** Disutilities	**Health states** Utilities • MCI due to AD | Community: Age-related utility x 0.83 • Mild AD | Community: Age-related utility x 0.76 • Moderate AD | Community: Age-related utility x 0.60

	• MCI due to AD: -0.17 • Mild AD: -0.19 • Moderate AD: -0.42 • Severe AD: -0.59 **Adverse events** Disutility for the average duration of ARIA (12 weeks) -0.14	• Severe AD | Community: Age-related utility x 0.42 • MCI due to AD | Long-Term Care: Age-related utility x 0.83 • Mild AD | Long-Term Care: Age-related utility x 0.76 • Moderate AD | Long-Term Care: Age-related utility x 0.60 • Severe AD | Long-Term Care: Age-related utility x 0.42 • Death: 0

Abbreviations: ACM = all-cause mortality; AD = Alzheimer's disease; ARIA = amyloid-related imaging abnormalities; MCI = mild cognitive impairment.

Tables 2 and **3** are replications of the model outcomes and cost-effectiveness outcomes from the ICER report. In the base case, the total cost and QALYs in the **Lecanemab** arm were $489,000 and 3.84, respectively; the total cost and QALYS in the **Supportive Care** arm were $363,000 and 5.77, respectively. This equates to an ICER of $236,000 per QALY gained for Lecanemab.

Table 2. Model cost and health outcomes for the ICER model. Source: Lin et al. (2023).

Treatment	Intervention Cost*	Total Cost	Life Years	QALYs	evLYs	Years in the Community
Lecanemab	$109,000	$489,000	6.23	3.84	3.96	4.20
Supportive Care	$0	$363,000	5.77	3.34	3.34	3.69

Abbreviations: eVLY = equal-value life year; QALY = quality-adjusted life year.

Table 3. Cost-effectiveness outcomes for the ICER model. Source: Lin et al. (2023).

Lecanemab vs. Supportive Care			
Perspective	**Cost per Life Year Gained**	**Cost per QALY Gained**	**Cost per evLY Gained**
Health Care Sector	$277,000	$254,000	$204,000
Societal	$265,000	$236,000	$183,000

Abbreviations: eVLY = equal-value life year; QALY = quality-adjusted life year.

Table 4 presents cost-effectiveness outcomes for the ValueGen.AI model. In the base case, the total cost and QALYs in the **Lecanemab** arm were $268,356 and 3.33, respectively; the total cost and QALYS in the **Supportive Care** arm were $154,409 and 3.01, respectively. This equates to an ICER of $361,265 per QALY gained for Lecanemab. We believe the difference in QALYs between the ICER model and AI-generated model are attributed to different age-

specific utilities used by the two models. The discrepancy in total costs is likely driven by the models using different age-specific base costs prior to application of the health state multipliers.

Table 4. Cost-effectiveness outcomes for the ValueGen.AI model.

Treatment	Total Cost	Life Years	QALYs	Cost per QALY Gained
Without age-based utilities				
Lecanemab	$268,356	6.63	4.24	$281,339
Supportive Care	$154,409	6.22	3.84	-
With age-based utilities				
Lecanemab	$268,356	6.63	3.33	$361,265
Supportive Care	$154,409	6.22	3.01	-

Abbreviations: QALY = quality-adjusted life year.

Because our primary objective was to conduct AI-assisted model adaptation—and not model replication—the AI-developed Alzheimer's model was deemed satisfactory for our purpose. Therefore, we proceeded with adapting our, AI-developed, version of US-based Alzheimer's model for adaptation to UK.

2.3 AI-Integrated Country Adaptation Roadmap

Successful country adaptation of health economic models requires careful adjustment of key model parameters to reflect local epidemiology, clinical practice, and economic conditions. We developed a structured roadmap to guide GenAI in identifying, sourcing, and integrating UK-specific inputs, combining automated extraction with expert review.

In consultation with NICE and review of published sources on model adaptation, we identified the following data sources that GenAI would utilize.

2.3.1 Data Sources to Inform Costs

1. **HTA agencies:** These agencies often publish cost-effectiveness analyses and reports that include jurisdiction-specific cost data.

2. **NICE clinical guidelines:** These guidelines often contain economic models and cost information that have been reviewed and accepted by NICE, making them a reliable resource for health economic modeling. For instance, the general dementia guideline can provide relevant cost data for dementia-related analyses.
3. **Literature review:** Conduct AI-based literature searches to find published studies that report on healthcare costs relevant to the model.
4. **Specific sources like tariffs and reference costs:** The national tariff payment system sets prices for healthcare services in England, and NHS reference costs provides average unit costs for defined services.
5. **WHO-CHOICE approach:** Use OneHealth tool for country-specific cost estimates (https://www.who.int/tools/onehealth)
6. **Purchasing power parity (PPP) adjustments:** Use data from international organizations like the OECD or a national organization like HMRC to adjust costs based on purchasing power parity, which can help in comparing costs across different countries.
7. **Price index comparisons**: In the absence of current cost data, adjust costs using price index data, as recommended by NICE Guidelines, to ensure alignment with current economic conditions

2.3.2 Data Sources to Inform Utilities

1. **Literature reviews:** Conduct AI-based reviews to identify and synthesize utility values from various studies, including cost-effectiveness papers
2. **Health surveys:** National health surveys or specific disease registries may include utility data collected from patients, which can be useful for the model.
3. **Utility databases:** Some organizations maintain databases of utility values, such as the Tufts CEVR CEA Registry, which compiles cost-effectiveness analyses and associated utility data.
4. **Clinical trials:** Review data from clinical trials that may report health-related quality of life (HRQoL) outcomes, which can be converted into utility values.
5. **Local HTA reports:** Review data from past HTA reports, as they often include utility values in their evaluations of new health technologies.

2.3.3 Data Sources to Inform Disease Progression

1. **Epidemiological studies:** Search for epidemiological studies that focus on the incidence, prevalence, and natural history of the disease. These studies can provide insights into how the disease progresses in different populations.
2. **Cohort studies:** Look for cohort studies that track patients over time. These studies can provide valuable longitudinal data on disease progression and associated factors.
3. **HTA reports:** HTA reports often include analyses of disease progression and treatment effectiveness. These reports can provide a comprehensive overview of the disease and its management.
4. **Country-specific sensitivity analysis:** If sensitivity analyses indicate that the model's outcomes are highly sensitive to certain assumptions or parameters, it may be beneficial to revisit the model structure to address these uncertainties and improve robustness.

2.4 Development of UK-based Model using Roadmap

We assumed that the natural history of Alzheimer's disease remains unchanged between the US and the UK. For other parameters, we utilized GenAI to provide UK specific values. Specifically for cost data, GenAI successfully extracted relevant information from available NICE reports, guidelines, and UK-specific publications within the Tufts CEA Registry. It also estimated the purchasing power parity (PPP) and applied it to cost data from the ICER report to derive UK-specific cost estimates. Similarly, GenAI extracted utility values from UK-specific publications available in the Tufts CEA Registry. For UK-specific background mortalities by age and sex, we utilized the **Heemod** package (9) and the **RGHO** package (10) to access data from the World Health Organization's Global Health Observatory (GHO). In cases where age- and sex-specific mortality data were unavailable for the UK, we substituted global average mortality rates to maintain completeness.

For NICE health technology assessment reports and disease guidelines, ValueGen.AI maintains an internal repository that includes these documents. This repository is automatically updated every six months for all covered disease areas, ensuring that the data

and recommendations remain up-to-date, an essential requirement for developing accurate and reliable health economic models.

For cost conversion using PPP, ValueGen.AI applied OECD-provided PPP multipliers, which are integrated within the platform. This approach adjusts cost estimates to reflect country-specific price levels, providing more accurate and comparable cost estimates across countries. As for the Tufts CEA Registry, ValueGen.AI performed a targeted online search based on the disease and country, retrieves open-access publications,and extracts relevant information. This process ensures the integration of the most up-to-date and context-relevant parameter values, supporting robust evaluations of the cost-effectiveness of health interventions.

We utilized healthcare perspective for our use case.

2.4.1 UK-specific model parameters

Table 5 summarizes the model structure and features of UK-based Alzheimer's model developed by ValueGena.AI. The platform successfully extracted data from publicly available HTA reports and open-access publications within the Tufts CEA Registry to inform cost and utility inputs from relevant studies. These details are summarized in the Supplement. Selecting appropriate parameter values for the Alzheimer's disease model requires contextual knowledge and clinical understanding of the disease. As such, determining which parameter values are most suitable for model incorporation necessitates a human-in-the-loop approach. Given the proof-of-concept nature of this experiment, we used a simpler approach for converting US-based input parameters to UK setting.

First, we updated the model to account for UK-specific background mortality rates and the discount rate of 3.5% for both cost and QALY outcomes.

Additionally, cost inputs were adapted using the PPP approach, as previously described. **Table 6** summarizes UK-based cost inputs used in the model. Note that the cost of Lecanemab is not available in UK. For our analysis, we assumed that the annual cost of Lecanemab was £17,264 using PPP ratio between UK and US.

Table 5. Model Summary generated by ValueGen.AI

Parameter	**Value**
Disease	Alzheimer's Disease
Country	United Kingdom
Intervention	Lecanemab
Health States	MCI due to AD | Community, Mild AD | Community, Moderate AD | Community, Severe AD | Community, MCI due to AD | Long-Term Care, Mild AD | Long-Term Care, Moderate AD | Long-Term Care, Severe AD | Long-Term Care, Death
Time Horizon	Lifetime
Cycle Length	1 Year
Modeling Approach	Markov
Baseline Population Age	71.0
Annual Discount Rate (%)	3.5

Table 6. Updated costs using the PPP approach, i.e. ICER-reported costs adapted to the UK setting

State	**Lecanemab Cost** (£)	**Supportive Care Cost** (£)
MCI due to AD | Community	23683.18	4418.107
Mild AD | Community	25423.67	6158.599
Moderate AD | Community	7657.588	7657.588
Severe AD | Community	7605.344	7605.344
MCI due to AD | Long-Term Care	81467.64	62202.57
Mild AD | Long-Term Care	83208.13	63943.06
Moderate AD | Long-Term Care	65442.05	65442.05

State	Lecanemab Cost (£)	Supportive Care Cost (£)
Severe AD | Long-Term Care	65389.81	65389.81
Death	0.0	0.0

2.4.2 UK-specific model outcomes

Table 7 presents the total costs, LYs, and QALYs for both treatment strategies, modeled with and without age-based utility adjustments. Without incorporating age-based utility decrements, lecanemab resulted in an incremental cost of £75,504. Lecanemab provided 0.38 additional life years (7.30 vs. 6.92) and 0.40 additional QALYs (4.50 vs. 4.10). The resulting incremental cost-effectiveness ratio (ICER) was £189,419 per QALY gained. When age-based utilities were applied, the QALY gain associated with lecanemab was attenuated to 0.31 (3.53 vs. 3.22), while life years remained unchanged. Under this scenario, the ICER increased to £246,470 per QALY gained.

Table 7. Cost-effectiveness outcomes from the ValueGen.AI model.

Treatment	Total Cost	Life Years	QALYs	Cost per QALY gained
Without age-based utilities				
Lecanemab	£187,120	7.30	4.50	£189,419
Supportive Care	£111,616	6.92	4.10	-
With age-based utilities				
Lecanemab	£187,120	7.30	3.53	£246,470
Supportive Care	£111,616	6.92	3.22	-

Abbreviations: QALY = quality-adjusted life year.

The UK-adapted economic model indicates that lecanemab increases both costs and health outcomes relative to supportive care in patients with early Alzheimer's disease. The cost per QALY gained for lecanemab exceeded commonly cited willingness-to-pay thresholds in the UK, particularly when accounting for age-related declines in utility. These findings suggest that, under current annual price of £17,264 and assumptions, lecanemab is unlikely to be cost-effective in the UK setting.

2.5 Key Insights

Structured AI-Assisted Adaptation

One of the primary insights from this project is that GenAI can be effectively leveraged to structure and streamline the process of country adaptation for health economic models. Traditionally, adapting a model to a new country's context requires gathering data from disparate sources such as published literature, HTA reports, clinical guidelines, and national databases. This process is often fragmented and time-consuming, with risks of omission or inconsistency. By contrast, GenAI platforms like ValueGen.AI can aggregate and summarize data from multiple sources into a centralized format, enabling faster identification of country-specific inputs. However, this efficiency is contingent on the system being guided by a well-defined framework specifying which parameters need localization. In our project, this structured AI-assisted approach allowed us to synthesize relevant cost, utility, and epidemiological data across sources, significantly reducing the manual effort typically required for literature reviews.

Semi-Automated Workflow with Human Oversight

A critical finding was that full automation of country adaptation remains elusive without expert human oversight. While GenAI successfully extracted a substantial proportion of parameters, it often encountered ambiguities—such as interpreting non-standardized parameter definitions, missing baseline values, or identifying the correct source among conflicting data points. In these cases, human intervention was necessary to verify extracted data, apply appropriate assumptions, and fill data gaps. This highlights that AI is most effective when integrated into a semi-automated workflow, where AI accelerates data aggregation and preliminary extraction, and domain experts retain control over final data selection and validation. This hybrid approach ensures that the adapted model maintains analytical integrity while still gaining efficiency benefits from automation.

Prompt Engineering as a Critical Factor

Our work demonstrated that the quality and comprehensiveness of AI-extracted data are highly sensitive to the prompting strategy used. Open-ended prompts tended to yield broader, more exploratory outputs but required significant manual filtering to eliminate irrelevant or tangential results. Conversely, highly specific prompts narrowed the scope of

retrieved data but risked missing nuanced or context-dependent inputs. Striking the right balance between specificity and openness in prompt design was key to maximizing the usefulness of AI outputs.

AI's Current Limitations in HEOR Domain Knowledge

Despite its capabilities, GenAI exhibited notable limitations in domain-specific understanding of HEOR concepts and conventions. For example, the AI occasionally confused input cost parameters with model's cost outcomes and utility inputs with QALYs, and transition probabilities with event rates. These errors were not due to technical failure but rather reflected the AI's lack of embedded HEOR-specific schema and contextual awareness. Therefore, explicit guidance, clear framing of the modeling task, and domain-specific prompt customization were essential to mitigate these issues. This insight suggests that while GenAI is a powerful tool for accelerating technical tasks, specialized human expertise remains indispensable to ensure conceptual correctness and alignment with health economic modeling standards.

A Replicable and Generalizable Adaptation Framework

An encouraging outcome of the project was that the adaptation framework we developed proved replicable and generalizable beyond Alzheimer's disease and the UK context. The structured roadmap specifying parameter categories, recommended data sources, and validation steps can be applied to other disease areas or countries with minimal modification. This framework serves as a blueprint for scaling AI-assisted country adaptations across therapeutic areas and geographies, which is particularly valuable for HTA in multi-country submissions or global access planning. The ability to replicate this process offers a pathway toward standardized, transparent, and more efficient adaptation workflows.

Efficiency Gains Without Compromising Quality

Importantly, the project demonstrated that GenAI integration can deliver meaningful efficiency gains without compromising the analytical rigor of the adapted model. Compared to a traditional manual adaptation process, the AI-assisted workflow reduced the time required for data gathering, parameter extraction, and code generation while maintaining fidelity to core model structures and assumptions. Although human oversight was still required for validation, the overall process was accelerated, and the burden on expert time

was reduced. This finding suggests that GenAI can play a transformative role in scaling health economic modeling efforts, particularly in resource-constrained environments or projects requiring rapid turnaround.

3. Project II: Model Verification and Validation

3.1 Introduction

Health economic models are critical tools utilized in HTAs for informing reimbursement and policy decisions. Ensuring the credibility and correctness of these models is paramount, yet technical errors frequently undermine their reliability. Verification—the process of ensuring a model accurately represents its conceptual design—is essential but often overlooked due to its resource-intensive nature.

This report outlines the utilization of OpenAI's O4-Mini reasoning mode to validate health economic models under different scenarios. Our approach leverages the reasoning capabilities of the AI to compare inputs and outputs, ensuring the robustness of the economic models used in healthcare decision-making.

The TECHnical VERification (TECH-VER) checklist, a structured approach previously outlined by Nasuh et al. (2019), categorizes verification efforts into five domains: input calculations, event-state calculations, result calculations, uncertainty analyses, and overall checks (11). These involve hierarchical verification procedures including black-box tests, white-box testing, and replication-based testing to systematically identify model errors and their root causes. Despite its comprehensiveness, manual application of TECH-VER can be burdensome, motivating exploration of automation using generative AI (GenAI). It is important to clarify that validation and verification represent distinct but complementary processes in the assessment of health economic models (12). Verification addresses whether the model has been implemented correctly and is free from technical errors, essentially answering the question, "*is the model built right?*". In contrast, validation evaluates whether the model adequately represents the real-world system it aims to simulate, addressing the question, "*is the right model built?*". Both processes are necessary to establish model credibility, but confusion between these terms often leads to inadequate attention to verification processes.

This report documents experiments conducted using GenAI to automate TECH-VER-based verification, undertaken in collaboration with NICE.

3.2 Methods

3.2.1 Health Economic Model for Experiments

We used an Excel-based health economic model that evaluates the cost-effectiveness of two asthma management strategies—Maintenance and Reliever Therapy (MART) and Inhaled Corticosteroids combined with Long-Acting Beta Agonists (ICS/LABA)—specifically for adults with uncontrolled asthma. This model was developed and provided by NICE. Utilizing a life table (Markov-type) modeling approach, the analysis assesses the health outcomes and associated costs over a 5-year time horizon, with annual cycle lengths, from the perspective of the UK's National Health Service (NHS) and Personal Social Services (PSS).

The model tracks patient progression through three distinct health states—"Alive, symptomatic," "Alive, non-symptomatic," and "Dead"—capturing changes in health status, quality of life, and mortality. Key outcomes measured include total costs, quality-adjusted life years (QALYs), incremental cost-effectiveness ratios (ICERs), and net monetary benefits (NMB). These metrics allow for an economic comparison of MART and ICS/LABA strategies, providing critical insights into their relative value and effectiveness.

3.2.2 Overall Approach

Our approach involved integrating GenAI into the TECH-VER verification workflow through a structured and automated pipeline:

1. **Model parsing:** An Excel-based cost-effectiveness model was loaded using backend Python scripts designed to programmatically access and interact with spreadsheet data.
2. **Checklist implementation:** Selected TECH-VER verification items were systematically encoded within Python, directly translating each TECH-VER criterion

into executable verification rules. The checklist implementation followed established verification principles outlined in NICE guidelines, which emphasize systematic checking of inputs, logical and mathematical processes, and outputs (13).

3. **Parameter identification and updates:** Model parameters, including their exact Excel cell references or named ranges, were clearly specified within Python scripts. Parameters subjected to verification tests—such as discount rates, utilities, efficacy values, and cost inputs—were automatically adjusted within the model to predefined test conditions.
4. **Model execution:** Post parameter adjustments, the Excel model was rerun automatically using integrated Python-based scripts that triggered recalculation within the spreadsheet environment.
5. **Output extraction:** Results from the model recalculations were programmatically extracted from specific sheets and ranges. This enabled efficient capture of verification outcomes.
6. **LLM-based verification:** Extracted outputs were then fed into Large Language Models (LLMs), which were prompted specifically to interpret the extracted model results against predetermined verification criteria. The prompts were structured to clearly elicit correct or incorrect responses based on the numerical and logical expectations of each test scenario.
7. **Response integration:** Responses from the LLM were systematically captured via integrated APIs, automatically recording the outcomes of each verification step and flagging discrepancies or validation failures for further investigation.

3.2.3 Validation Checklist

We implemented the following checklist as a proof-of-concept:

1. Set discount rate to 0% for costs and outcomes
2. Increase discount rate
3. Set treatment efficacy equal in both arms
4. Set all cost inputs to 0
5. Set all utility values to 1 and adverse event (AE) disutilities to 0
6. Set all utility values to 0
7. Set mortality to 0

8. Increase mortality rate
9. Increase starting age
10. Decrease starting age
11. Reduce time horizon
12. Increase time horizon
13. Set AE rates to 0
14. Increase the efficacy of the intervention

The approach captures key NICE verification methods, including analyses with null and extreme values, ensuring that model results can be explained logically, and confirming that predictions of intermediate and final endpoints are plausible (13).

3.2.4 Selection of the AI Model

We selected the O4-Mini reasoning mode from OpenAI due to its exceptional performance in complex reasoning tasks. Reasoning LLMs outperform simpler models by handling logical dependencies and contextual inference, making them valuable for interpreting validation tests and flagging issues beyond surface-level discrepancies. This advanced generative AI model was ideally suited for validating economic models, as it proficiently generates outputs and analyses input-output relationships. Additionally, we integrated the OpenAI API with our custom development to create a solution precisely tailored to our specific requirements, instead of utilizing a chatbot interface like ChatGPT.

3.2.5 System Prompt and Execution

The validation process begins with the Python program altering the input parameters of the health economic model stored in an Excel file. The modified inputs are then used to execute the model, generating output tables that reflect the new scenario. These output tables, alongside the baseline results, are sent to the O4-Mini AI model. The AI model's task was to reason through the provided data, comparing the validation output against the expected results. It determines whether each validation test has passed or failed based on the alignment of outputs with predefined criteria.

The system prompt provided to the AI was intentionally generic. It included the validation test description, expected output, and necessary output tables. This approach allowed the model to focus on the reasoning task without being influenced by overly specific instructions.

The system prompt used for validation

You are a helpful health economics and outcomes research assistant. Your primary goal is to analyze tabular result data from different modeling simulations and help answer questions by comparing the results from experiment 1 and experiment 2. The Inputs will be in this format:

Test Action: The Validation Test Action Expected Outcome: Validation test Expected Outcome Simulation Results: Simulation run results that needs to be validated LLM Evaluation: Your evaluation of the result whether the results satisfy the outcome. First say "PASSED" or "FAILED" then Give a detailed explanation to explain your reasoning

Example user prompt for The First Validation Test

Action: Set discount rate to 0% for costs and outcomes
Outcome: Discounted and undiscounted costs and outcomes are the same
Simulation Results:

Within time horizon?	Cycle	Age	MART						
Tx cost		Monitoring cost		Exacerbation cost		Cost	Disc. Cost		
1	0	40.65	-	-	-	-	-		
1	1	41.65	£262.22	£27.24	£44.46	£333.93	£333.93	...	
1	2	42.65	£261.85	£27.20	£44.40	£333.45	£333.45	...	
1	3	43.65	£261.44	£27.16	£44.33	£332.93	£332.93	...	
1	4	44.65	£260.99	£27.11	£44.25	£332.36	£332.36	...	
1	5	45.65	£260.50	£27.06	£44.17	£331.73	£331.73	...	

0	6	46.65	£259.95		£27.01	£44.07	£331.03	£331.03	...
0	7	47.65	£259.36		£26.94	£43.97	£330.28	£330.28	...
...	...	...	...	...	...	...	...		
0	72	112.65	£0.50	£0.05	£0.08	£0.63	£0.63	...	

Implementation Using Python

The entire validation process was managed by a Python program specifically developed for this project. The steps involved are as follows:

- Modifying the input parameters of the Excel-based health economic model using the `xlwings` and `openpyxl` libraries to edit the model Excel file.
    - `xlwings`: This library allows easy interaction with Excel from Python, providing the capability to call Excel macros, read/write data, and automate Excel functions directly from Python.
    - `openpyxl`: This library is used for reading and writing Excel 2010 xlsx/xlsm/xltx/xltm files. It allows for modification of spreadsheets without the need for Excel itself.
- Executing the model to produce new output tables by running an Excel macro that runs the simulation model.
- Extracting the output tables from the validation run using `xlwings`.
- Sending the validation output along with baseline results to the language model. This was done using LangChain to interface with the language model.
    - `LangChain`: This framework facilitates the integration between various language models and applications. It streamlines the process of interacting with language models like OpenAI's GPT-3 by providing a flexible API and tools for managing prompts and responses.
- Requesting the language model to evaluate whether the validation test passed or failed by sending system prompts and output tables to the OpenAI API and retrieving the response from the language model.
    - `OpenAI API`: This API provides access to advanced language models by OpenAI, such as GPT-3, enabling applications to leverage natural language

understanding and generation capabilities for various tasks including evaluation and analysis.

3.3 Results

The GenAI-driven TECH-VER verification process produced detailed outcomes across various verification checks:

1. **Discount rate verification (passed test)**
    - **Action:** Discount rate set to 0%.
    - **Outcome:** The LLM correctly identified that discounted and undiscounted values matched exactly, confirming correct discounting implementation.
2. **Cost input verification (passed test)**
    - **Action:** All cost inputs set to 0.
    - **Outcome:** The LLM successfully confirmed that total costs were effectively zero across all model cycles, demonstrating accurate cost input resetting.
3. **Utility value verification (passed test)**
    - **Action:** All utility values set to 0.
    - **Outcome:** The LLM verified correctly that total QALYs calculated by the model were zero, as logically expected.
4. **Utility-disutility verification (failed test)**
    - **Action:** Utility values set to 1 and AE disutilities set to 0.
    - **Outcome:** The LLM highlighted a failure in this verification step, noting persistent utility values below 1. This indicated that utility resetting was incomplete or incorrectly implemented, leading to a clear identification of a potential implementation error.
5. **Discount rate sensitivity verification (passed test)**
    - **Action:** Discount rate increased.
    - **Outcome:** The LLM correctly identified that total discounted results decreased, confirming appropriate discounting sensitivity.
6. **Mortality deactivation verification (passed test)**
    - **Action:** Mortality rate set to 0.

- **Outcome:** The LLM confirmed that no deaths occurred in the model, verifying proper mortality implementation.

7. **Mortality sensitivity verification (passed test)**
   - **Action:** Mortality rate increased.
   - **Outcome:** The LLM observed an increase in deaths at each time point, validating sensitivity to changes in mortality inputs.
8. **Starting age sensitivity – increase (passed test)**
   - **Action:** Starting age increased.
   - **Outcome:** The LLM confirmed a decrease in total life-years, reflecting appropriate model response to older starting age.
9. **Starting age sensitivity – decrease (passed test)**
   - **Action:** Starting age decreased.
   - **Outcome:** The LLM confirmed an increase in total life-years, indicating correct age-based calculations.
10. **Time horizon sensitivity – reduction (passed test)**
    - **Action:** Time horizon reduced.
    - **Outcome:** The LLM verified that total costs, QALYs, and life-years all decreased accordingly, demonstrating correct time horizon behavior.
11. **Time horizon sensitivity – extension (passed test)**
    - **Action:** Time horizon increased.
    - **Outcome:** The LLM confirmed that total costs, QALYs, and life-years increased, reflecting appropriate model behavior over extended time.

Table 8 summarizes key verification tests performed on the model, outlining the expected outcomes for each test and the actual results. These tests were designed to verify the logical integrity of model components, with results confirming correct implementation in most cases.

Table 8. Summary of the model verification tests

Verification Test	Expected Outcome	Test Result
Set discount rate to 0% for costs and outcomes	Discounted and undiscounted costs and outcomes are the same	PASS
Increase discount rate	Total discounted results decrease	PASS
Set all cost inputs to 0	Total costs outcomes are 0	PASS

Set all utility values to 1 and AE disutilities to 0	Total QALYs are equal to total life-years	FAIL
Set all utility values to 0	Total QALYs are 0	PASS
Set mortality to 0	No deaths observed	PASS
Increase mortality rate	The number of deaths at each time point increases	PASS
Increase starting age	Total LYs decrease	PASS
Decrease starting age	Total LYs increase	PASS
Reduce time horizon	Total costs, QALYs and LYs decrease	PASS
Increase time horizon	Total costs, QALYs and LYs increase	PASS

Abbreviations: AE = adverse events; LYs = life years; QALY = quality-adjusted life years.

Insights from the Failed Test

The LLM not only flagged a failure in the verification checklist but also provided a diagnostic explanation for the issue. Specifically, it identified that the user had failed to instruct the model to reset age- and sex-based utility values to 1. As a result, the model's calculated QALYs did not align with expected life expectancy benchmarks under full-health assumptions. The LLM clarified that the discrepancy arose from an incomplete or incorrect implementation of utility resetting, highlighting an implementation error rather than a structural or conceptual flaw in the model itself. This capability to pinpoint the source of verification failure demonstrates the LLM's utility not just in pass/fail assessment but also in guiding users toward precise corrective actions in model verification.

3.3.1 Comparison with Manual Verification

The GenAI-driven verification process demonstrated several advantages over traditional manual verification approaches:

1. **Efficiency:** While we did not conduct an experiment where one group conducts verification manually and another group uses AI-driven process to measure time efficiency, we nevertheless expect efficiency gains from the following capabilities. First, the automated process completed verification checks in a fraction of the time required for manual verification after the initial setup. This efficiency gain is consistent with findings from recent studies on the application of GenAI in health economic modeling (14). Another efficiency gain of GenAI-driven verification process will likely occur from repeated iterations of the verification checklist at the model development

stage. Finally, efficiency gains using the GenAI approach will likely occur from automated detection of errors in the verification process as noted in the above Table.

2. **Consistency:** The LLM-based approach ensured systematic and consistent application of verification criteria across all checks, reducing the risk of oversight or human error. This consistency is particularly valuable in the context of complex health economic models, where manual verification can be challenging and time-consuming (15).
3. **Error detection:** The automated verification process successfully identified implementation errors, demonstrating its effectiveness as a quality assurance tool. This aligns with findings from studies where GenAI could detect discrepancies (3). This capability could lead to substantial time gains and increases in accuracy.
4. **Documentation:** The automated approach generated comprehensive documentation of verification outcomes, facilitating transparency and reproducibility. This addresses a key challenge in health economic modeling, where lack of transparency and reproducibility has been identified as a significant limitation (15).

3.3.2 Current Challenges and Limitations

Despite the promising results, several challenges and limitations were identified in the application of GenAI to health economic model validation:

1. **Technical limitations:** The current implementation was limited to Excel-based models and specific verification tests. Extending the approach to other modeling platforms and more comprehensive validation frameworks will require additional technical development.
2. **GenAI capabilities:** While GenAI demonstrated effectiveness in routine verification tasks, its capabilities in complex interpretative validations were limited. This aligns with broader challenges in the application of AI in healthcare, where human oversight remains essential for complex clinical and methodological assessments.
3. **Data privacy and security:** The application of GenAI in health economic model validation raises concerns about data privacy and security, particularly when handling sensitive data, such as patient level data or confidential company data. These concerns require careful attention to data management practices and compliance with regulations such as HIPAA and GDPR.

4. **Validation framework integration:** The current implementation focused on automating the TECH-VER verification framework and did not fully integrate with other validation frameworks, such as those defined by the ISPOR-SMDM Modeling Good Research Practices Task Force. Such integration would be valuable for comprehensive model validation.
5. **Stakeholder acceptance:** The acceptance of GenAI-driven validation approaches by stakeholders, including HTA authorities, healthcare providers, and patients, remains uncertain. Engaging stakeholders in the development and implementation of GenAI-driven validation frameworks would be essential for their successful adoption.

3.4 Key Insights

This study provides a proof-of-concept of integration of GenAI within the model verification framework illustrating substantial potential to automate and enhance verification processes, particularly for straightforward numeric and logical checks. Future enhancements should focus on improving the interpretative capacities of GenAI, particularly around complex methodological validations, the explanation of identified discrepancies, and pinpointing the exact sources of model errors. Future work will also include reporting our findings using the ELEVATE-AI-LLMs reporting standards, developed by the ISPOR Working Group on AI, to promote standardized, high-quality reporting in AI-driven research (16).

Expanding GenAI applications to other forms of validation, as recommended by ISPOR-SMDM guidelines, offers another promising avenue for enhancing model validation processes. These include face validity (expert assessments of model structure and assumptions), cross-validation (comparisons with other models analyzing the same problems), external validation (comparison with observed real-world outcomes), and predictive validation (forecasting future events and subsequently comparing with actual outcomes). Such advancements would further minimize manual intervention, substantially enhancing the credibility, reliability, and overall efficiency of health economic model verification.

Research into fully automated model construction already demonstrates promising results, with studies showing GPT-4 can replicate published models with 93-100% accuracy (14,

17). This suggests potential for "double programming" validation approaches, where GenAI independently constructs a model based on specifications to verify human-built models. Such approaches could transform validation practices by providing independent verification without additional human resources.

Finally, collaborative validation efforts that engage multiple stakeholders, including HTA authorities, healthcare providers, patients, and technology developers, will be essential for the successful adoption of GenAI-driven validation approaches. Such collaboration would align with the trend toward increased international cooperation in health economic model validation and support the development of standardized validation approaches that enhance the credibility and transparency of health economic models used in healthcare decision-making.

4. Supplement

The findings below summarize the results of AI-assisted data extraction from HTA reports and open-access publications within the Tufts CEA Registry.

Reports from the NICE Repository

1) Donepezil, galantamine, rivastigmine and memantine for the treatment of Alzheimer's disease (2011), NICE technology appraisal guidance TA217 (18)

Cost Item	Cost (£)	Source Reference
Monthly pre-FTC/institutionalized care costs	£724	"Monthly pre-FTC/institutinalized care costs £724 Dependent on severity, >£724."
Monthly FTC/institutional care costs	£3201	"Monthly FTC/institutional care costs £3201 £2117"
Monthly memantine drug costs	£64.80	"Monthly memantine drug costs £64.80 £71.28"
Monthly drug costs for Donepezil	£97	"Monthly drug costs Donepezil £97 £97"
Monthly drug costs for Rivastigmine	£74	"Monthly drug costs Rivastigmine £74 £98"
Monthly drug costs for Galantamine	£83	"Monthly drug costs Galantamine £91 £83"
6-monthly monitoring visit cost	£158	"6-monthly monitoring visit cost £108 £158"

Monthly pre-FTC/inst cost	£2051	"Monthly pre-FTC/inst cost £328 £2051"
Monthly FTC/inst cost	£2117	"Monthly FTC/inst cost £937 £2117 (£2941 * 72%)"

Utility Item	Utility Value	Source Reference
Pre-FTC (Full-Time Care)	0.60	"Utilities from the original SHTAC model for AChEIs: 0.34 for FTC, 0.60 for pre-FTC."
FTC (Full-Time Care)	0.34	"Utilities from the original SHTAC model for AChEIs: 0.34 for FTC, 0.60 for pre-FTC."
FTC (Full-Time Care)	0.336	"Utility for the FTC state was calculated using the FTC sub-group of moderate to severe patients from the LASER-AD and was 0.336."
Death	0.00	"Utility for death was assumed equal to 0."

2) Dementia Guidance: assessment, management and support for people living with dementia and their carers (2018), NICE technology appraisal guidance NG97 (19)

Treatment/Strategy	Absolute Cost	Incremental Cost	Source Reference

No Olanzapine	$34,215	-	"The study considered both ‘direct’ and ‘indirect’ costs, though did not specify what each comprised. The results presented here are from analyses of ‘direct’ costs only, as these were judged less likely to include items that are beyond the NICE reference case."
Olanzapine	$39,781	$5,566	"The base-case analysis (Table 84) suggests that treatment with olanzapine incurs additional costs (primarily due to cost of the drug itself) but also provides QALY gains, with an ICER of $37,104 per QALY."
Placebo	$4,923	-	"The base-case analysis suggests that olanzapine is dominated by placebo as placebo is cheaper and produces more health benefits."
Olanzapine	$6,480	$1,557	"Olanzapine $6,480 0.12 QALYs $1,557 -0.02 QALYs Dominated"
Quetiapine	$7,839	$2,916	"Quetiapine $7,839 0.15 QALYs $2,916 0.01 QALYs ext. dom."

Risperidone	$10,215	$5,292	"Risperidone $10,215 0.16 QALYs $5,292 0.02 QALYs $264,600 /QALY"

Utility Value	Source Text
0.14 QALYs	"Placebo $4,923 0.14 QALYs"
0.12 QALYs	"Olanzapine $6,480 0.12 QALYs"
0.15 QALYs	"Quetiapine $7,839 0.15 QALYs"
0.16 QALYs	"Risperidone $10,215 0.16 QALYs"
0.60	"non-institutionalised = mean for mild–moderate AD (0.60)"
0.34	"institutionalised = mean for severe, profound or terminal AD (0.34)"
0.0982 + 0.0298 × MMSE	"Utility = 0.0982 + 0.0298 × MMSE"
0.202 + 0.008 × ADCS-ADL	"Utility = 0.202 + 0.008 × ADCS-ADL"
0.359 + 0.00745 × MMSE + 0.00394 × DAD − 0.0054 × NPI	"HUI-3 Total Score = 0.359 + 0.00745 × MMSE + 0.00394 × DAD − 0.0054 × NPI"

Publications from Tufts CEA Registry

1) The Impact of Including Costs and Outcomes of Dementia in a Health Economic Model to Evaluate Lifestyle Interventions to Prevent Diabetes and Cardiovascular Disease (20)

Summary of the Paper:

Background:

- Lifestyle interventions aim to prevent diabetes and cardiovascular disease (CVD) but have not traditionally included dementia.
- These interventions can reduce dementia risk and extend life expectancy, affecting healthcare costs.

Objective:

- Demonstrate the feasibility of including dementia in public health cost-effectiveness analysis and quantify the overall impacts considering competing effects.

Methods:

- Adapted the School for Public Health Research (SPHR) diabetes prevention model to include dementia.
- Used data from primary care databases, health surveys, and dementia trials.
- Evaluated the NHS diabetes prevention program (NHS DPP) from an NHS/personal social services perspective with three scenarios: no dementia, dementia only, and reduced dementia risk.
- Conducted subgroup, parameter, and probabilistic sensitivity analyses.

Results:

- Lifetime cost savings per patient: £145 (no dementia), £121 (dementia only), £167 (reduced dementia risk).
- QALY gains: 0.0006 (dementia only), 0.0134 (reduced dementia risk).
- Dementia inclusion did not alter the cost-effectiveness recommendation of the NHS/DPP.

Conclusion:

- Including dementia in lifestyle intervention models is feasible but does not change policy recommendations or significantly modify health economic outcomes.
- The largest impact on health economic outcomes occurs when a direct impact on dementia incidence is assumed, especially in elderly populations.

Cost Item	Cost (in £)	Source Reference
Health care cost (mild dementia)	£3,103	"Health care cost mild dementia (MMSE 21–26) £3103 Assumed 10% (23)"
Health care cost (moderate dementia)	£8,293	"Health care cost moderate dementia (MMSE 10–20) £8293 Assumed 10% (23)"
Health care cost (severe dementia)	£9,841	"Health care cost severe dementia (MMSE 0–9) £9841 Assumed 10% (23)"
Social care cost (mild dementia)	£5,674	"Social care cost mild dementia (MMSE 21–26) £5674 Assumed 10% (23)"
Social care cost (moderate dementia)	£22,703	"Social care cost moderate dementia (MMSE 10–20) £22703 Assumed 10% (23)"
Social care cost (severe dementia)	£23,466	"Social care cost severe dementia (MMSE 0–9) £23466 Assumed 10% (23)"
Cost of dementia diagnosis	£687.82	"The most recent cost study of Dementia for the UK estimated the cost of Dementia diagnosis

		at £650 in 2012/13 prices (42) inflated to £687.82."

MMSE Range	Utility Value	Source Reference
21–25	0.93	Utility decrement MMSE 21–25 0.93 Assumed 10%
15–20	0.725	Utility decrement MMSE 15–20 0.725 Assumed 10%
10–14	0.710	Utility decrement MMSE 10–14 0.710 Assumed 10%
0–9	0.478	Utility decrement MMSE 0–9 0.478 Assumed 10%

2) Does Structured Exercise Improve Cognitive Impairment in People with Mild to Moderate Dementia? A Cost-Effectiveness Analysis from a Confirmatory Randomised Controlled Trial: The Dementia and Physical Activity (DAPA) (21)

Summary of the Paper:

Background:

- Previous studies suggest physical exercise could slow dementia progression.
- Evidence for cost-effectiveness of structured exercise is conflicting and based on small trials.

Objective:

- Compare the cost-effectiveness of a tailored, structured, moderate- to high-intensity exercise programme versus usual care in people with mild to moderate dementia.

Methods:

- Economic evaluation from the UK NHS and personal social services perspective.
- Data from a large randomised controlled trial.
- Primary clinical outcome: ADAS-Cog at 12 months.
- Costs collected over a 12-month follow-up.
- Bivariate regression of costs and QALYs with multiple imputation of missing data.
- Sensitivity and subgroup analyses conducted.

Results:

- Participants (n = 494) randomised to exercise plus usual care or usual care only.
- Mean ADAS-Cog score worsened slightly in both groups.
- Mean costs: £5945 for exercise vs. £4597 for control.
- Mean QALY estimates: 0.787 for exercise vs. 0.826 for control.
- Probability of cost-effectiveness < 1% across thresholds.
- INMBs ranged between -£2601 and £2158 at thresholds between £15,000 and £30,000 per QALY.
- Results robust to sensitivity and subgroup analyses.

Conclusion:

- The structured exercise programme does not slow cognitive impairment in mild to moderate dementia and is not cost-effective.

Cost Item	Exercise Group (Mean)	Usual Care Group (Mean)	Mean Difference
Total NHS and Personal Social Services Costs	£5945 (US$8501)	£4597 (US$6534)	£1347 (US$1926)
Total Societal Costs	£6063 (US$8670)	£4761 (US$6808)	£1301 (US$1860)
Intervention Costs	£1269 (US$1815)	-	-
Patient Accommodation	£187.6	£54.4	£133.2
Hospital Services	£2019.3	£1827	£192.3
Day-care Services	£33.9	£49.2	-£15.3
General Community Health Services	£366.3	£347.6	£18.7
Community Mental Health Services	£163.6	£150.2	£13.4
Social Care Services	£647	£759.9	-£112.9
Equipment, Adaptations/Repairs	£1.9	£14.5	-£12.6
Participant Travel	£5.7	£7	-£1.3
Concomitant/Prescription Medications	£1046.2	£1067.4	-£21.2

Other Costs	£204.5	£321	-£116.5

ity Item	ity Value	rce Text
5D-3L utility score for trial pants	220	e EQ-5D-3L utility score for each rticipant... Resulting utility scores rom −0.59 to 1.0, with 0 enting death and 1.0 representing alth; values below 0 are indicative th states worse than death."
5D-3L utility score for carer-ed	0665	outed attributable costs and , covariate- and baseline-adjusted eported EQ-5D utility score"
5D-3L utility score including ioner travel costs	220	outed attributable costs and , covariate- and baseline-adjusted utility score including practitioner costs"

5D-3L utility score excluding hire costs	220	buted attributable costs and , covariate- and baseline-adjusted utility score, excluding venue hire

3) Pragmatic Multi-Centre Randomised Trial of Reminiscence Groups for People with Dementia and their Family Carers: Effectiveness and Economic (22)

Summary of the Paper:

Background:

- Joint reminiscence groups for people with dementia and their family carers are popular, but evidence of their effectiveness is limited.

Objective:

- Assess the effectiveness and cost-effectiveness of joint reminiscence groups compared to usual care.

Methods:

- Design: Multi-centre, pragmatic randomised controlled trial with two parallel arms.
- Participants: 488 people with mild to moderate dementia and their carers, recruited from NHS Memory Clinics and community mental health teams.
- Intervention: Weekly joint reminiscence groups for 12 weeks, followed by monthly sessions for 7 months.
- Primary Outcomes: Quality of life for people with dementia (QoL-AD) and psychological distress for carers (GHQ-28).
- Secondary Outcomes: Autobiographical memory, activities of daily living, carer stress, mood, relationship quality, service use, and costs.

Results:

- No significant differences in primary or secondary outcomes between intervention and control groups.

- Carers in the intervention group reported increased anxiety at the ten-month endpoint.
- Compliance analyses suggested benefits for people with dementia attending more sessions, but increased stress for carers.
- Economic analysis indicated joint reminiscence groups are unlikely to be cost-effective.

Conclusion:

- The trial does not support the clinical or cost-effectiveness of joint reminiscence groups. Benefits for people with dementia are offset by increased anxiety and stress in carers. Further exploration of these outcomes is needed.

Cost Item	Reminiscence Mean Total Costs (£)	Control Mean Total Costs (£)	Difference in Mean Total Costs (£)
Community Care (Participants)	1,072	1,170	-98
Day Care (Participants)	1,098	610	488
Hospital Use (Participants)	2,719	2,529	190
Total (Participants)	4,889	4,309	580
Community Care (Carers)	258	283	-25
Day Care (Carers)	7	34	-27

Hospital Use (Carers)	1,266	1,043	223
Total (Carers)	1,531	1,360	171
Grand Total	6,419	5,667	751

Utility Item	Utility Value	Source Reference
EQ-5D (proxy) Complete case	0.588	"EQ-5D (proxy) Complete case 17 0.588 0.626 -0.038 (-0.10, 0.03) 0.37 0.06"
EQ-5D (proxy) Imputed data	0.575	"EQ-5D (proxy) Imputed data 0 0.575 0.596 -0.021 (-0.07, 0.03) 0.44 0.04"
EQ-5D VAS (proxy) Complete case	62.1	"EQ-5D VAS (proxy) Complete case 12 62.1 63.2 -1.11 (-5.44, 3.22) 0.59 0.03"
EQ-5D VAS (proxy) Imputed data	60.55	"EQ-5D VAS (proxy) Imputed data 0 60.55 62.43 -1.88 (-5.66, 1.9) 0.33 0.05"

EQ-5D Complete case	0.804	"EQ-5D Complete case 39 0.804 0.806 -0.001 (-0.06, 0.06) 0.72 0.00"
EQ-5D Imputed data	0.772	"EQ-5D Imputed data 0 0.772 0.769 0.004 (-0.05, 0.06) 0.88 0.01"

5. References

1. Fleurence RL, Bian J, Wang X, Xu H, Dawoud D, Higashi M, et al. Generative AI for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations - an ISPOR Working Group Report. Value Health. 2024.
2. Reason T, Klijn S, Rawlinson W, Benbow E, Langham J, Teitsson S, et al. Using Generative Artificial Intelligence in Health Economics and Outcomes Research: A Primer on Techniques and Breakthroughs. PharmacoEconomics - Open. 2025.
3. Sharp S, Lokuge K, Elvidge J, Hudson T, Dawoud D. Systematic literature review of the use of generative AI in health economic evaluation. medRxiv. 2025:2025.04.25.25326412.
4. Use of AI in evidence generation: NICE position statement, https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation--nice-position-statement, Accessed by April 2025.
5. NICE health technology evaluations: the manual, https://www.nice.org.uk/process/pmg36/chapter/economic-evaluation-2, Accessed by April 2025.
6. Daniel Mullins C, Onwudiwe NC, Branco de Araújo GT, Chen W, Xuan J, Tichopád A, et al. Guidance Document: Global Pharmacoeconomic Model Adaption Strategies. Value Health Reg Issues. 2014;5:7-13.
7. Drummond M, Barbieri M, Cook J, Glick HA, Lis J, Malik F, et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. Value Health. 2009;12(4):409-18.
8. Lecanemab for Early Alzheimer's Disease, https://icer.org/wp-content/uploads/2023/04/ICER_Alzheimers-Disease_Final-Report_For-Publication_04172023.pdf, Accessed by May 2025.
9. Filipović-Pierucci A, Zarca, K., & Durand-Zaleski, I. Markov models for health economic evaluations: the R package heemod. arXiv preprint arXiv:170203252. 2017.
10. RGHO: Access WHO Global Health Observatory Data from R, https://cran.r-project.org/web/packages/rgho/index.html, Accessed by May 2025.
11. Büyükkaramikli NC, Rutten-van Mölken MP, Severens JL, Al M. TECH-VER: a verification checklist to reduce errors in models and improve their credibility. Pharmacoeconomics. 2019;37:1391-408.
12. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model Transparency and Validation A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force–7. Medical Decision Making. 2012;32(5):733-43.
13. Developing NICE guidelines: the manual, https://www.nice.org.uk/process/pmg20/chapter/incorporating-economic-evaluation, Accessed by April 2025.
14. 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.

15. Feenstra T, Corro-Ramos I, Hamerlijnck D, van Voorn G, Ghabri S. Four Aspects Affecting Health Economic Decision Models and Their Validation. Pharmacoeconomics. 2022;40(3):241-8.
16. Fleurence RL, Dawoud D, Bian J, Higashi MK, Wang X, Xu H, et al. The ELEVATE-AI LLMs Framework: An Evaluation Framework for Use of Large Language Models in HEOR: an ISPOR Working Group Report. arXiv:250112394. 2024.
17. Chhatwal J, Samur S, Yildirim I, Bayraktar E, Ermis T, Ayer T. EE247 Fully Replicating Published Markov Health Economic Models Using Generative AI. Value in Health. 2024;27(12):S102.
18. Donepezil, galantamine, rivastigmine and memantine for the treatment of Alzheimer's disease , https://www.nice.org.uk/Guidance/TA217, Accessed by May 2025.
19. Dementia: assessment, management and support for people living with dementia and their carers, https://www.nice.org.uk/guidance/ng97, Accessed by May 2025.
20. Breeze P, Thomas C, Thokala P, Lafortune L, Brayne C, Brennan A. The Impact of Including Costs and Outcomes of Dementia in a Health Economic Model to Evaluate Lifestyle Interventions to Prevent Diabetes and Cardiovascular Disease. Medical Decision Making. 2020;40(7):912-23.
21. Khan I, Petrou S, Khan K, Mistry D, Lall R, Sheehan B, et al. Does Structured Exercise Improve Cognitive Impairment in People with Mild to Moderate Dementia? A Cost-Effectiveness Analysis from a Confirmatory Randomised Controlled Trial: The Dementia and Physical Activity (DAPA) Trial. PharmacoEconomics - Open. 2019;3(2):215-27.
22. Woods RT, Orrell M, Bruce E, Edwards RT, Hoare Z, Hounsome B, et al. REMCARE: Pragmatic Multi-Centre Randomised Trial of Reminiscence Groups for People with Dementia and their Family Carers: Effectiveness and Economic Analysis. PLOS ONE. 2016;11(4):e0152843.