Decoding Participant Voices With Artificial Intelligence: A Pilot Analysis of Free-Text Participant-Reported Outcome Data From the PURPOSE 1 Study of Lenacapavir for HIV Pre-Exposure Prophylaxis

MSR4

PURPOSE 1

Saeid Shahraz[1], **Ryan Thaliffdeen**[1], **JeanPierre Coaquira**[1], **Aaditya Rawal**[2], **Veda Donthireddy**[2,*], **Shubhi Pathak**[2], **Dylan Mezzio**[1]

[1]Gilead Sciences, Inc., Foster City, CA, USA; [2]Costello Medical, Boston, MA, USA
*Affiliation at the time the analysis was conducted

Don't miss our Panel Discussion!
Wednesday, May 20
10:00 AM - 11:00 AM
Please scan the QR code on the right to access the abstract

Conclusions

- This pilot analysis suggests that artificial intelligence (AI) could offer a viable approach to analyzing free-text data, despite mismatches between human benchmark and Copilot categorization into the most common concepts
- Classification tasks surpass the capabilities of general-purpose, foundational, large language models (LLMs), potentially resulting in biased outputs
  - LLMs integrated into closed-loop systems that iteratively learn from their own outputs, as implemented in this study, may not consistently produce reliable insights for classification tasks
- Implementation of appropriate AI workflows to contextually customize LLMs may substantially improve the efficiency and reliability of participant-reported outcome (PRO) analyses, and could unlock further opportunities to leverage free-text data
- LLMs should be carefully selected and specifically modified to suit classification needs

Plain Language Summary

- Participant-reported outcome (PRO) questionnaires are special 'tools' that are sometimes used in clinical trials. They ask about how study participants feel while on the study medication
- PROs are usually measured using questionnaires that include multiple-choice and free-text answers
- Free-text responses in PRO questionnaires are important to capture participants' perspectives because they provide context and reveal insights that fixed-choice responses might not identify
- Free-text responses provide valuable information, but analyzing them is time consuming
- Artificial intelligence (AI) can help analyze large amounts of free-text datasets
- A PRO questionnaire in the PURPOSE 1 trial of injectable versus oral HIV pre-exposure prophylaxis (PrEP) asked participants about which form of PrEP administration they favored (twice-yearly injections versus daily pills) and included a free-text question to say why they liked that option better
- This test compared human versus AI-based sorting of free-text reasons for how they want PrEP to be given, among PURPOSE 1 participants at Week 52
- Sorting of reasons for how they prefer to take PrEP was different between humans and AI: almost 3 out of 10 responses were sorted by AI as "other" while humans sorted 3% of the responses as "other"
  - AI was able to spot "perceived efficacy" as a reason for preference much less often than humans
- This study suggests that AI may be a useful tool to evaluate free-text PRO data with human input, but choosing the right AI tool is important to optimize output

Introduction

- Free-text responses in PRO datasets can reveal deeper, more detailed perspectives than predefined response choices[1-4]
- However, distilling consistent meaningful concepts from large numbers of free-text responses is very time consuming and labor intensive,[5] and AI could facilitate automated analysis of large free-text datasets
- PURPOSE 1 (NCT04994509) was a Phase 3, double-blind, randomized controlled trial conducted in South Africa and Uganda
  - Cisgender women aged 16-25 years who were HIV negative were randomized 2:2:1 to receive subcutaneous (SC) lenacapavir (927 mg as two 1.5-mL injections) every 26 weeks, oral emtricitabine/tenofovir alafenamide (F/TAF; 200 mg/25 mg) daily, or oral emtricitabine/tenofovir disoproxil fumarate (F/TDF; 200 mg/300 mg) daily, along with the alternate SC or oral placebo[6]
- A PRO questionnaire in PURPOSE 1 asked participants about their pre-exposure prophylaxis (PrEP) administration preferences (daily pills vs twice-yearly injections) at baseline, Week 26, and Week 52, and included a free-text question to explain their reasons for these preferences (**Figure 1**)

Figure 1. Three Key Items From the PrEP Impacts and Administration Preference Questionnaire (23 Items)

If I could take just one kind of PrEP medication, knowing they both worked equally well, I would prefer to take PrEP medication:

By injection every 6 months	I have no preference one way or the other	By a daily pill

I would rate my preference for the PrEP medication I prefer as[a]:

Slight preference	Moderate preference	Strong preference

In a few words, explain the reason for your preference (or why you have no preference):

_____ (short-text-length, open-ended response)

Focus of manual versus AI analysis

[a]Participants who reported no preference for PrEP administration type did not answer this question.
AI, artificial intelligence; PrEP, pre-exposure prophylaxis.

Objective

- This pilot analysis compared human manual versus AI-based categorization of free-text reasons for PrEP administration preference among PURPOSE 1 participants who indicated a preference for injections at Week 52

Methods

- Using behavioral theories,[7] we developed an ontology reflecting human reasoning behind PrEP administration preference (**Figure 2**)[8]

Figure 2. Ontology of Concepts Underlying People's Preferences for Injectable or Oral PrEP

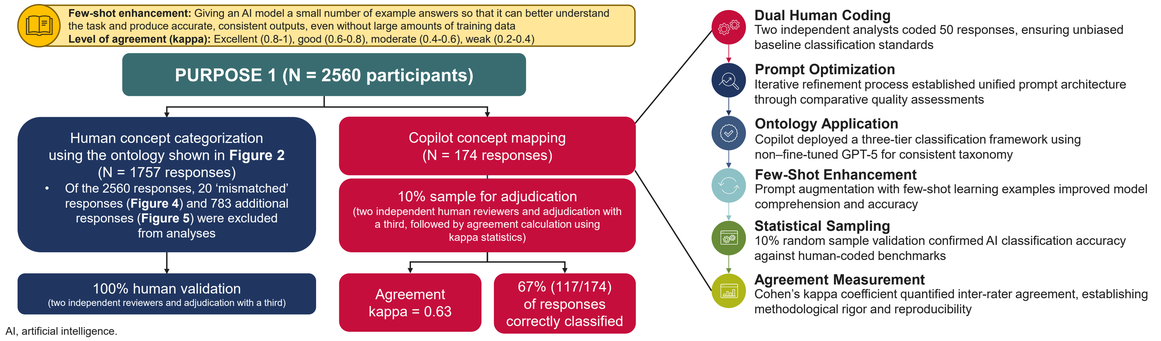
Ontology: Structured map of ideas that defines the key categories, concepts, and relationships within a topic, which helps people or AI to understand and organize information consistently

Tier 1 (Domain)	Tier 2 (Concept)	Tier 3 (Subconcept)[a]
Personal factors	Cognitive evaluations	Perceived risk and safety
		Perceived efficacy
	Behavioral alignment	Adherence feasibility and forgetfulness (or lack thereof)
		Routine integration
		Convenience and logistical effort
		Lifestyle compatibility or burden
	Psychological responses	Emotional response to mode
		Affective forecasting and anticipated regret
		Pain and discomfort
		Body autonomy and control
Contextual factors	Social influences	Social norms
		Partner attitudes
		Stigma and disclosure concerns
		Social support
	Systemic context	Trust in system and technology
	Cultural context	Cultural beliefs about medication
		Cultural acceptance and symbolism
		Other

[a]Results for Tier 3 subconcepts are not shown on the poster for simplicity.
AI, artificial intelligence; PrEP, pre-exposure prophylaxis.

- Two independent reviewers manually categorized 1757 free-text responses into 17 ontological concepts, with arbitration by a third reviewer; responses could be sorted into multiple categories (**Figure 3**)
- The same free-text dataset was then provided to Copilot (GPT-5; Microsoft Corporation, Redmond, WA, USA), along with a small set of human-categorized examples and the ontological categories included in the prompt, to generate AI-based concepts (**Figure 3**)
  - A random 10% sample (174/1757) was used to validate Copilot's classifications: two raters reviewed the Copilot-generated concepts for this sample, with a third rater adjudicating disagreements. Agreement was quantified using Cohen's kappa coefficient
- Finally, the most frequent concepts into which humans and Copilot categorized the free-text responses were compared quantitatively to assess alignment between human and AI categorization

Figure 3. Methodology for Comparing Human Versus Copilot-Based Concept Mapping

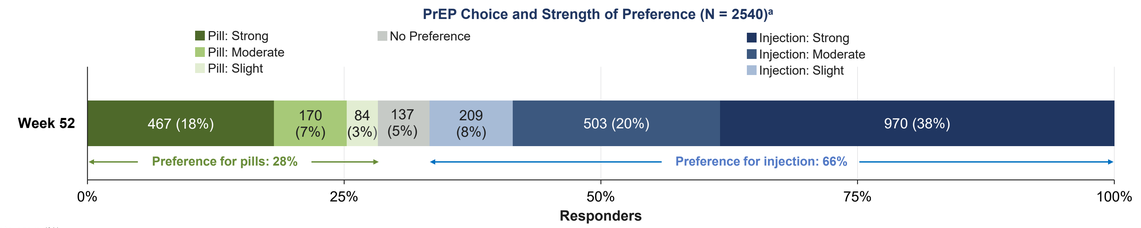


AI, artificial intelligence.

Results

- Most participants expressed a preference for the twice-yearly injection (66%) over daily oral pills (28%) at Week 52 (**Figure 4**)

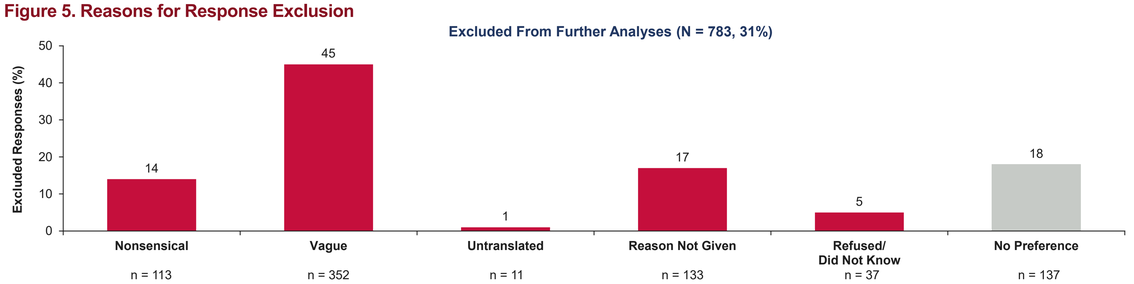
Figure 4. Participants' Preference for Injection Over Pill at Week 52



Data are n (%).
[a]n = 20 'mismatched' responses (responses in which the free text clearly states a preference that is different from what was selected in the dropdown preference question) were excluded from the analysis.
PrEP, pre-exposure prophylaxis.

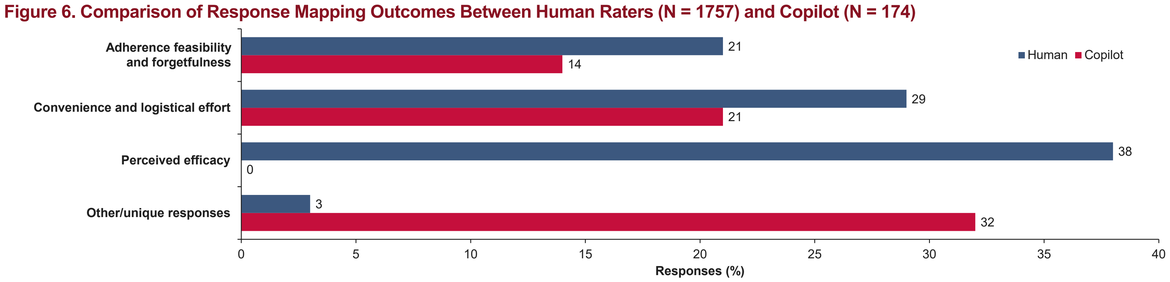
- A third of responses to the reasons for preference question (N = 783) were excluded from further analyses by human reviewers due to the reasons shown in **Figure 5**

Figure 5. Reasons for Response Exclusion



- Copilot classified 32% of responses as "other/unique responses" (vs 3% by humans), 0% as "perceived efficacy" (vs 38%), 21% as "convenience and logistical effort" (vs 29%), and 14% as "adherence feasibility and forgetfulness" (vs 21%) (**Figure 6**)
- In the validation sample, Copilot classified 67% (117/174) of responses accurately, with kappa = 0.63 indicating good agreement

Figure 6. Comparison of Response Mapping Outcomes Between Human Raters (N = 1757) and Copilot (N = 174)



References: **1.** Abraham TH, et al. *Eval Program Plann*. 2020:78;101733; **2.** Kilborn K, et al. *Pain Manag Nurs*. 2023;24:201-8; **3.** Riiskjaer E, et al. *Int J Qual Health Care*. 2012;24:509-16; **4.** Rich JL, et al. *PLOS One*. 2013;8:e68832; **5.** Young KJ, et al. *Chiropr Man Therap*. 2024;32:2; **6.** Bekker L-G, et al. *N Engl J Med*. 2024;391:1179-92; **7.** Michie S, et al. *ABC of Behaviour Change Theories: An Essential Resource for Researchers, Policy Makers and Practitioners*. 2nd ed. Silverback Publishing; 2022; **8.** Mansoor LE, et al. Poster TUPE057 presented at: IAS; July 13-17, 2025; Kigali, Rwanda.

Acknowledgments: We extend our thanks to the PURPOSE 1 participants, their families, and all participating investigators. This study was funded by Gilead Sciences, Inc. Medical writing support was provided by Thierry Deltheil, PhD (Aspire Scientific Ltd, UK), and was funded by Gilead Sciences, Inc.

Disclosures: **SS**, **RT**, **JPC**, and **DM** are employees and shareholders of Gilead Sciences, Inc. **AR** and **SP** are employees of Costello Medical. **VD** was an employee of Costello Medical at the time the analyses were conducted and is now a part-time employee at Varosync.

Correspondence: Saeid Shahraz, Saeid.Shahraz@gilead.com.