Paper PP12

Shots to Stats: AI's Take on Alcohol Units

Liam Clothier, Southampton Clinical Trials Unit, University of Southampton, Southampton, UK
Mike Radford, Southampton Clinical Trials Unit, University of Southampton, Southampton, UK

ABSTRACT

The Timeline Follow Back (TLFB) method is an approach to quantifying alcohol intake. The data collected with the tool can be difficult to convert into analysable data, using up trial resources. The development of AI offers an opportunity to speed up data cleaning in this context. However, the accuracy of AI tools to convert, for example "two glasses of wine", into number of units is unknown. This study uses data from the AlcoChange clinical trial to test three different open source AI software and look at their consistency and accuracy compared to traditional human review. We report an updated analysis that also demonstrates development of AI over the last 12 months.

INTRODUCTION

Artificial Intelligence (AI) is an increasingly discussed topic in clinical data management and presents opportunities including data cleaning, data interpretation, data analysis and reporting. Electronic Data Capture (EDC) vendors have begun to offer integrated AI in their platforms but clinical data management staff at academic trial units are more likely to be familiar with freely available generative AI services such as ChatGPT. Though the use of these free services in clinical data management comes with considerable concerns, one potential use is with the interpretation of free text data; translating qualitative data into quantitative data.

The AlcoChange study (ISRCTN reference: ISRCTN10911773) investigates whether a smartphone app and digital breathalyser delivering personalized behaviour change techniques can reduce alcohol consumption in adults with alcohol-related liver disease compared to the current standard of care within the UK. During the study, participants are requested to fill out a diary (Timeline Follow Back; TLFB) detailing their daily alcohol consumption for the 30 days leading up to one of three visits. Guidance is given on how to fill this out, but it is fundamentally a free text field, which the study team must then convert into the number of alcohol units consumed. In order to derive the units, the amount of liquid consumed, along with the alcohol by volume (ABV) is required.

This investigation aims to evaluate how and if AI technology could be used to support deconstructing of the TLFB free text into something quantitatively analysable and whether the AI output is accurate enough to utilise as a validation tool or not.

METHOD

Participant Diary data consisted of over 17,000 individual free text records, leading to 1,383 unique text fields. There was considerable variety in the format of how drinks and amounts were described, e.g., some were comma separated, some were carriage return separated, while others had no spaces between words. There was also a variety of spelling mistakes. Many records did not include the precise measures consumed or ABV. The trial team therefore drew up a list of assumptions to be used when the text was not otherwise precise enough to calculate the units consumed on that day.

A Clinical Data Coordinator (CDC) within Southampton Clinical Trials Unit (SCTU) manually reviewed every line of the data and deconstructed the text into drink quantity, ABV and/or units, depending on what was available. Any record without sufficient information would be flagged for querying. Three AI systems were used and compared: OpenAI ChatGPT v4 (2025), Google Gemini v2.5 (2025), Microsoft Copilot v4 (2025).

The CDC and AI programs were both given background on the study data, along with examples, and given the same set of rules to follow for the translation including:

- The list of Assumptions of common units and sizes of alcoholic drinks to use when these were not included in the free text provided.
- Any drink under <1% ABV was to be ignored
- Where a range (e.g. 2-3 glasses of wine) was supplied, the midpoint was to be used
- If the units couldn't be determined, then mark for review by the study team

- Units were to be calculated using the following formula: [strength (ABV) x volume (ml)] ÷ 1000

Each AI system performed the decoding three times using a new chat each time, whilst storage or use of the data by the systems for future review or AI improvement was disabled. The same free text was provided to both the AI programs and to the CDC, with no amendments made for spelling or grammatical errors. No further guidance was provided to the AI system beyond that given to the CDC.

Output from the AI programs was requested in table format for ease of transference into Excel for comparisons. Only obvious mistakes in the output, e.g. missing, incomplete or incorrectly formatted units were rerun. No patient identifiable information was uploaded and the AI systems were not given the human calculations to compare their own calculations against.

The most consistent AI was taken forward to compare against the CDC entry. Differences between the AI value and the CDC value were reviewed manually by the senior study team to determine which value, if either, was accurate.

INITIAL AI INVESTIGATION

An investigatory step was done with the AI systems to determine how best to communicate with them. This revealed that if more than 50/60 records were given, or the Excel file upload functionality was used, each AI would then create python code in the background to determine the units, rather than using its LLM model. Such scripts were simplistic and unable to unpick free text into usable data. Therefore, a prompt was added to make it clear to the AI systems that they were expected to use their LLM models rather than writing a script. It was also necessary to paste lines of text in groups of 50 per message to mitigate against this.

A rounding error of 0.1 was allowed as there were rounding inconsistencies both within and across AI systems. It was noted that results for individual records differed across the three runs within each AI. If at least two of the runs matched on a given record, to within 0.1 units, the mean of those values were calculated and selected as the "validation value" to take forward to compare against the CDC's value. If none of the values from the 3 runs matched to within 0.1 of each other no validation value was output for that record. See example table below. ").

Example text	Run one output	Run two output	Run three output	Validation value	Comments
1/2 BOTTLE 37.5% 35CL VODKA DAILY	13.13	13.13	13.125	13.128	All three matched within 0.1 unit. Mean value of all three values taken as validation value.
1 BOTTLE JAMAICAN WHITE RUM (200MLS) 63%1 PINT STELLA. 8X568MLS LAGER	38.16	38.16	27.26	38.16	Two matched within 0.1 unit. Mean value of these two was taken as validation value.
18 CANS OF STELLA	36.22	39.6	36.43	n/a	No two outputs match within 0.1 unit. No validation value taken.

RESULTS

CONSISTENCY

ChatGPT provided the least number of validation values, with only 1010 (73.0%) of the 1383 records having two or more runs matching. Of those, roughly half matched for all three runs. Gemini was able to provide a validation value for 1330 (96.2%) of the records, while Copilot had a validation value at 1361 (98.4%). It was also the most consistent between its three runs.

Figure 1: Consistency of values provided by each AI per run.

ACCURACY

Each run was compared to the CDC results to gain an initial insight into AI performance. Figure 2 demonstrates that each AI software had two similar runs, with a third outlier run. This suggests that three runs are the minimum requirement to produce a validation value.

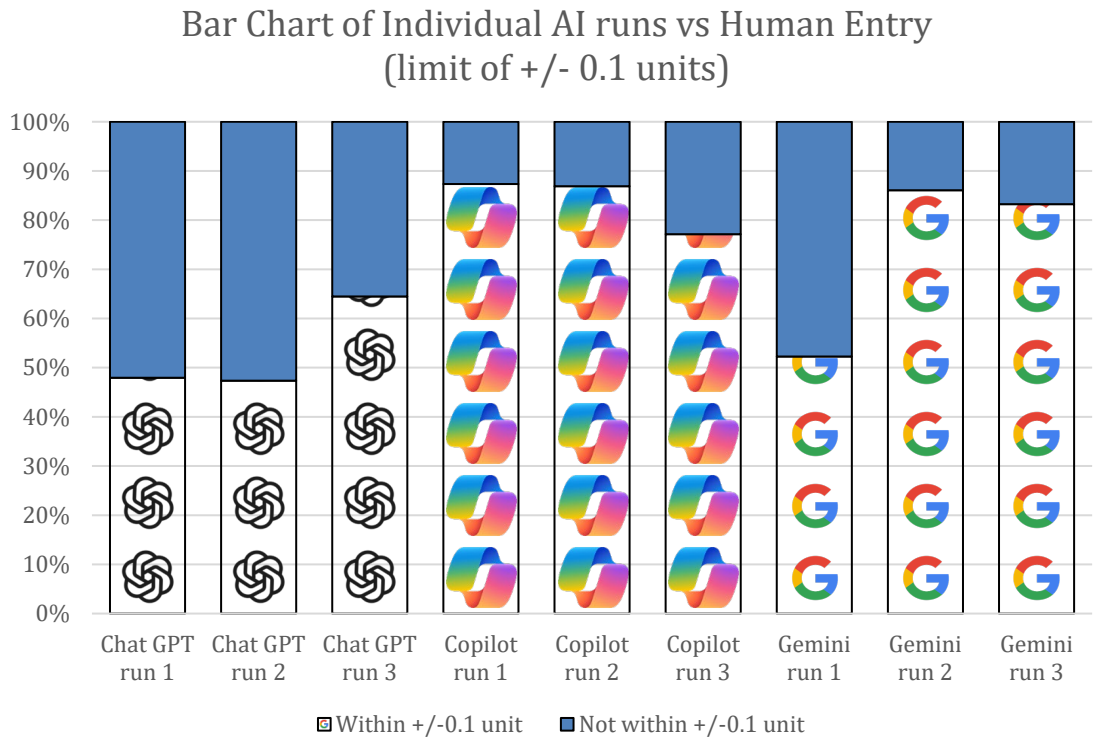
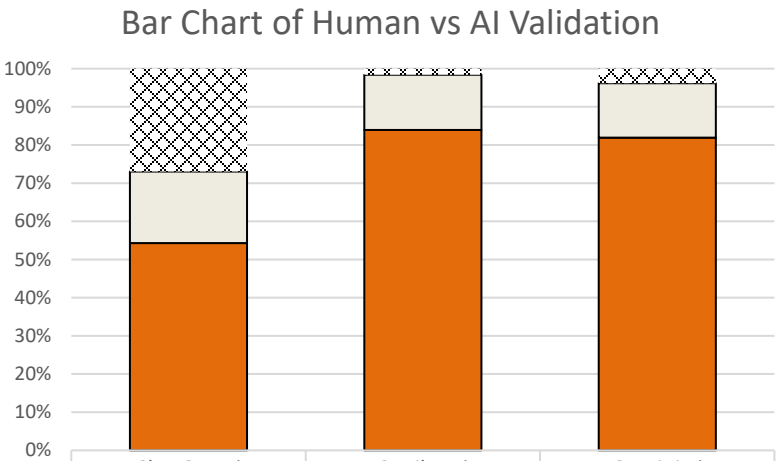


Figure 2: Agreement (%) between CDC and AI per run of each AI software.

The validation values were then compared with the CDC data entry. Figure 3 shows the agreement between the CDC results and the validation values for each AI software. Copilot and Gemini were in agreement with the CDC in approximately 80% of cases, while ChatGPT was notably lower at 54%. The lack of agreement includes the 27% of cases where ChatGPT did not have a validation value; even excluding these, the rate of disagreement with the CDC was higher than the other software (26%).



	ChatGPT %	Copilot %	Gemini %
No Validation Value	27.0	1.6	3.8
Not Matching	18.7	14.5	14.2
Matching	54.3	83.9	81.9

Figure 3: Percentage agreement between CDC data entry and AI validation values.

Using Copilot as the exemplar, Figure 4 shows the magnitude of the differences between the CDC and Copilots values against the mean units of each record. The larger the alcohol intake, the more likely Copilot and the CDC were to disagree. The mean difference was 0.4, with 95% limits of agreement of -17.3, 18.2. This latter calculation assumes normality of differences; as we observed some extreme values, some beyond the scale of Figure 4, we also calculated the 95% empirical limits of agreement, which were -3.1 to 4.5 (i.e., 95% of our observed difference lie in this range).

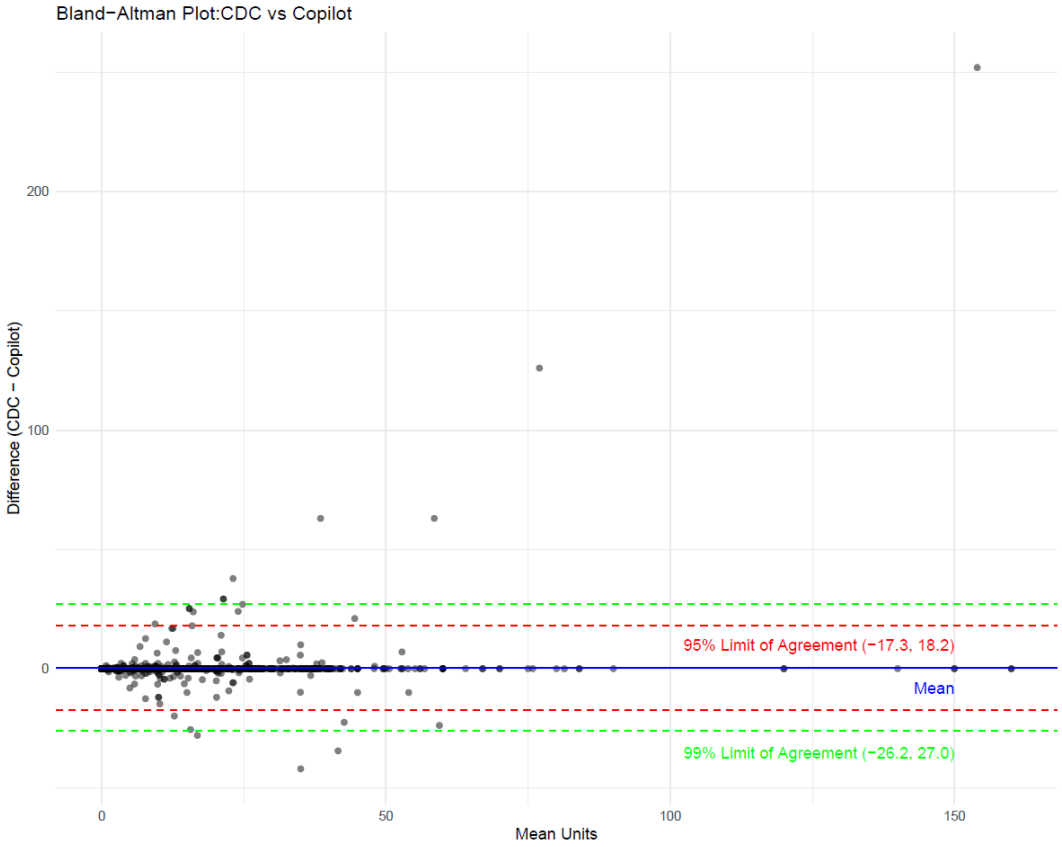


Figure 4: Bland-Altman Agreement Plot of CDC vs Copilot.

MANUAL ADJUDICATION

Copilot was found to be the most consistent AI system within itself, as well as aligning closet to the CDC's values, leaving only 222 differences. The senior study team went through each mismatch manually to determine which value was most accurate, Copilot or the CDC. Figure 5 describes the results from the CDC and CoPilot classified according to the final senior team review.

Mismatches where both the CDC and Copilot were correct (n=20) were identified as records where it was not possible to derive units. Here Copilot had specified this in a derivation column, whilst entering a result of 0 units in the result column, while the CDC simply stated that the units could not be derived. This may have been resolved with more detailed instructions when using AI. Most disagreements were a result of human error (130/222=59%) rather than CoPilot. Only 19 records were identified where both the CDC and AI were incorrect, meaning an overall error rate for the CDC stood at 10.77% (149/1383). Copilots error rate stood at 5.21% (72/1383).

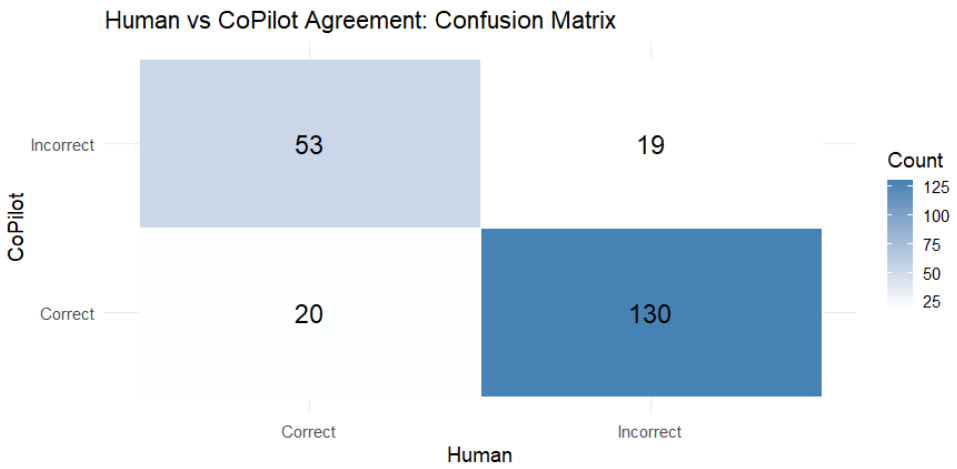


Figure 5: Classification of disagreements between CDC data entry and CoPilot validation values.

OTHER FINDINGS

AI performed best when asked to review the data in batches of around 50 lines at a time. However, it did need prompting to use their LLM functionality. Without this, it would write itself a python script that gave greatly reduced accuracy.

All AI systems skipped rows where the records were similar as well as skipping multiple records it deemed to have no alcohol, making re-combining less efficient. Both Copilot and Gemini would occasionally guess at unit consumption whilst ChatGPT would correctly state “More information needed” in these scenarios. However, ChatGPT also often stated “more information needed” for records where there was actually enough information to make an informed decision.

Message limits proved to be an obstacle, with ChatGPT running out of permitted messages per day, meaning runs were completed over multiple days. Copilot also ran out of messages on all three chats, so any further records needed to be done from a new chat. Additionally, Gemini would often ‘forget’ where it got to and have to restart from earlier in the run.

When considering time taken, the CDC logged around 37 hours for their entire data entry. If we take Copilot as the validation AI system, each run taking approximately 3 hours, however it was not able to produce a precise audit trail. Additional time was also required to extract the output into a usable format, e.g. Excel. Thus time taken for AI to complete the task totalled about 11 hours.

The results on consistency presented in Figure 1 were in stark contrast to a similar investigation from 2024 (Radford et al, 2024) where ChatGPT looked the most promising, with Gemini struggling to provide consistent responses (Figure 6). In the 2024 investigation, none of the three AIs were considered acceptable to provide sufficient validation of the human data entry.

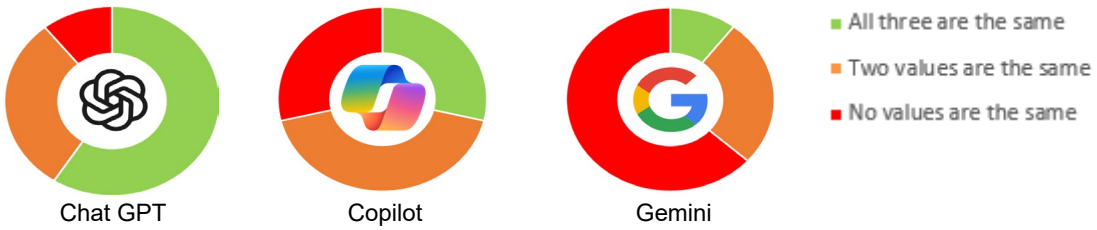


Figure 6: Consistency of values provided by each AI across 3 runs – 2024 analysis.

DISCUSSION

Data entry is a very laborious task that is extremely prone to errors. . In clinical trials where free text fields are used for critical data, the gold standard is double data entry or 100% independent quality control review. Here we have shown that, although not perfect, some AI software has an error rate lower (5.2%) than a single human review (10.8%) and takes less time than human data entry, even when considering the resource required to combine the

three outputs. This suggests that using AI as an approach to validating human review (or vice versa) presents a viable alternative.

The AI review demonstrated results across runs where it was unable to produce a validation value or provide the correct number of units for reasons that were not clear. At times it would produce the correct derivation but was unable to do the arithmetic in order to get the correct result, suggesting it can be as fallible as a human. However very little prompt engineering was performed for this study and with more experimentation to improve prompts, it is possible the error rates for AI would decrease

Furthermore, we showed that over the past year since the initial 2024 analysis, AI has improved dramatically. The consistency between the three runs has become significantly greater and the number of records matching the CDC data entry has also improved. In the 2024 analysis, it was suggested that AI still had a long way to go to be used for tasks such as this. However, its improved consistency and accuracy in this investigation demonstrates increased proficiency for handling such data. One potential caveat is given how rapidly performance has changed, it is unclear whether this performance level should be expected increase, be maintained or indeed decrease again with natural fluctuations as AI models are updated. The fact that all three AI systems evaluated demonstrated a 3rd outlier run also plants seeds of doubt in AI's capabilities. Risks of using AI should therefore be mitigated against with significant human oversight if AI is to be used, especially in a clinical trial setting.

CONCLUSION

AI presents a viable, cost-saving solution for data entry from free text as well as an efficient validation tool to replace human quality control checks but still requires human oversight and adequate controls in place to ensure accuracy. Future work should look to improve prompts, improve workflow efficiencies and investigate other use cases.

REFERENCES

Sobell, L.C., Sobell, M.B. (1992). Timeline Follow-Back. In: Litten, R.Z., Allen, J.P. (eds) Measuring Alcohol Consumption. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-4612-0357-5_3

Mike Radford et al, Unique to Units: Artificial Intelligence versus human translation of qualitative to quantitative data, poster at International Clinical Trials Methodology Conference, Edinburgh September 2024

ACKNOWLEDGMENTS

Steve Nason, for his manual adjudication.

Special thanks to the participants on the AlcoChange trial.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Contact the author at:
Liam Clothier
Southampton Clinical Trials Unit, University of Southampton
CCI Building, MP131, Southampton General Hospital, Tremona Road, Southampton, SO16 6YD
L.P.Clothier@soton.ac.uk
www.southampton.ac.uk/ctu