PP12

Shots to Stats: AI's Take on Alcohol Units

Liam Clothier, Mike Radford, Steve Nason on behalf of the Alcochange study group
Southampton Clinical Trials Unit, University of Southampton, Southampton, UK

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.

AIM: The main aim was to see **whether AI could provide a reliable alternative to double human data entry for validation** of free text fields.

Timeline Follow Back Questionnaire
(1383 unique records)

EXAMPLE FREE TEXT

2L CIDER 5%, 1.5L LAMBRINI 6%, 660ML TSINGTAO BEER 4.7%

EXAMPLE AI OUTPUT

Cider: 2000ml × 5 ÷ 1000 +
Lambrini: 1500ml × 6 ÷ 1000 +
Tsingtao Beer: 660ml × 4.7 ÷ 1000 =
10.00 + 9.00 + 3.10=
Total: 22.1 units

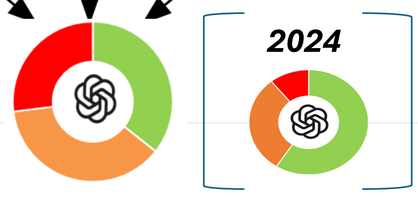
2024 Analysis: *Mike Radford et al, Unique to Units: Artificial Intelligence versus human translation of qualitative to quantitative data, poster at International Clinical Trials Methodology Conference, Edinburgh September 2024*

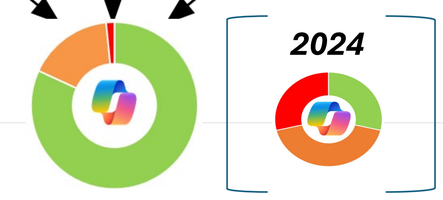
ChatGPT v5 | CoPilot v4 | Gemini v2.5 flash

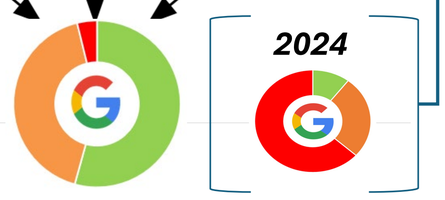
3 Unique runs per AI. Identical information provided for each run.

Run 1 | Run 2 | Run 3

A VALIDATION VALUE was considered if 2 or more run matched (within 0.1 unit)







Storage or use of the data by the systems for future review or AI improvement was disabled.

KEY:
All 3 runs match
2 runs match
Different units calculated each run

Despite being the most consistent in the 2024 run, **ChatGPT** was only able to match itself three times on **35.6%** of the records. **1010/1383 records with Validation Value.**

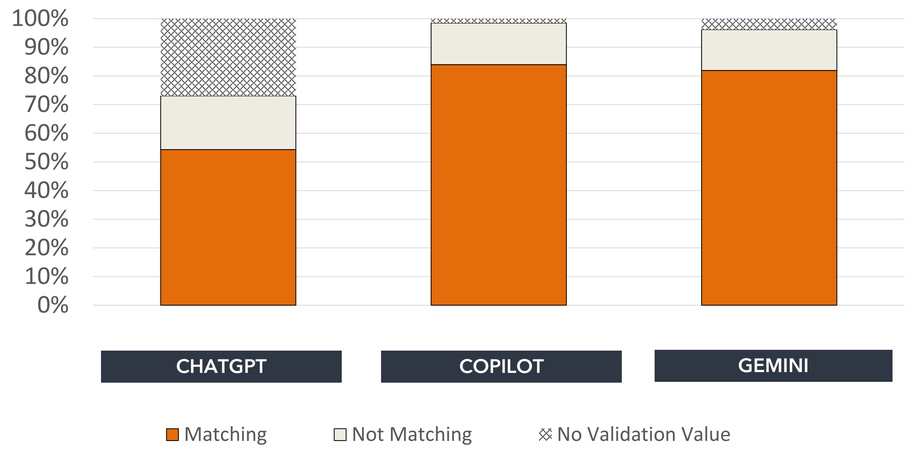
CoPilot was able to produce the same value for **81.9%** of the records given to it, with a further **16.5%** matching in two of the runs. **1361/1383 records with a Validation Value**

Gemini matched just over 50% of the time on all three runs. It was also able to produce a validation value (2+ matches) for **96.2%** of the records **(1330/1383)**.

AI Validation Values vs Human Data Entry (STEVE)

- All three AI's reported back around 15% (Range 14%-18%) of the 1383 records were incorrect from Steve.
- **ChatGPT** validation values matched Steve for only **54.3%**, mainly due to the 26.9% of records where no validation value was generated.
- **CoPilot** and **Gemini** were much higher at **83.9%** & **81.9%** respectively.

At this point, it was decided to **move forward with just CoPilot** due to it being more consistent, along with having a higher correlation to Steve.



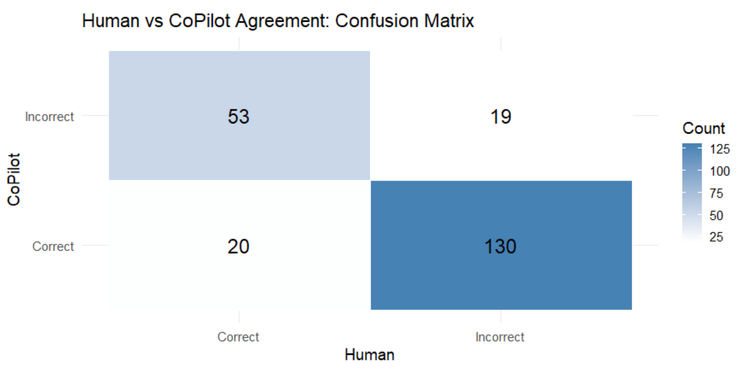
- **Steve** was given the same background on the study data and an identical set of rules to follow as the three AI systems.
- The same free text was provided to both the AI programs and to Steve, with no amendments made for spelling or grammatical errors.

- Only obvious mistakes in the output, e.g. missing, incomplete or incorrectly formatted units were rerun.
- No patient identifiable information was uploaded.

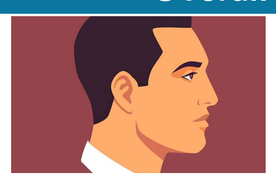
STEVE vs COPILOT

Mismatch Adjudication

(Remaining 222 discrepancies)



Overall Error Rates

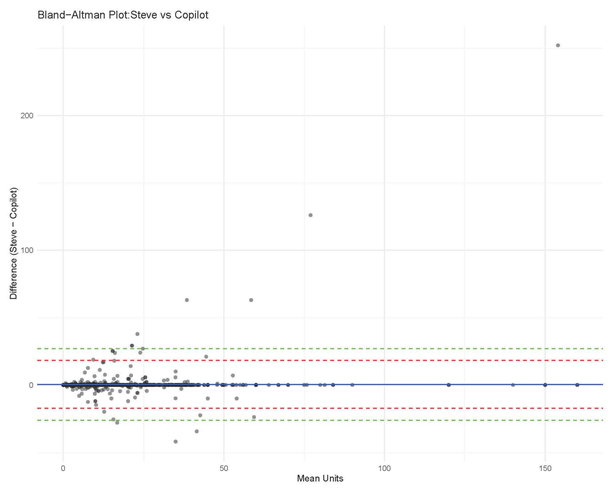


vs

10.77% **5.21%**

- **Steve's** errors were mainly due to typo's or where the assumptions were not followed.
- **CoPilot** was also better at determining ABV from brand names.
- There didn't seem to be much consistency to the CoPilot errors.
- In a few instances CoPilot worked out the correct formula but was unable to compute the correct number of units.
- There were several occasions where one of the CoPilot runs was correct but didn't have a second correct value to put forward a validation value.

Agreement Plot



95% Limit of agreement (-17.3, 18.2)

99% Limit of agreement (-26.2, 27.0)

CONCLUSION

Data entry is a very laborious task that is extremely prone to errors. Here, we have shown that, although not perfect, **AI has provided a solid response** to assist in the validation of this task. The overall error rates are comparable between **CoPilot** (**5.21%**) and the **human** (**10.77%**) entry, suggesting we would get a similar outcome had we used a second human for the data entry validation.

The **time taken** for three runs of CoPilot were superior to the time taken for the human entry. It was determined the time taken to do the comparison would've been similar to that taken to validate via double data entry.

Within the past year since the last run of this analysis, **AI has improved dramatically**. The consistency between the three runs has got significantly closer and the number of records matching the human input 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, with its consistency and accuracy it certainly seems to have proved its usefulness for this interpretative function.

OTHER FINDINGS

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

CoPilot was the most consistent in its output, leading to **each run** taking approximately **3 hours**, however it was not able to produce a precise audit trail.
- **Steve** logged around **37 hours** for his data entry.

All AI's skipped rows where the records were similar.
- They also would skip multiple records it deemed to have no alcohol, making re-combining less efficient.
- At times, both CoPilot and Gemini would **guess unit consumption**. ChatGPT would correctly state "More information needed", although it ended up doing this for records where there was enough information to make an informed decision.

Message limits proved to be an obstacle.
- **ChatGPT** would run out of permitted messages per day, so runs were completed over multiple days.
- **CoPilot** ran out of messages on all three chats, so we would be unable to re-use these chats if more records are received in the future.
- **Gemini** would often 'forget' where it got to and had to restart from earlier in the run.

FUNDED BY NIHR National Institute for Health and Care Research

University of Southampton
UKCRC Registered Clinical Trials Units
For more information, please email:
l.p.clothier@soton.ac.uk or
m.j.radford@soton.ac.uk