131

Automatically Extracting Numerical Results from Randomized Controlled Trials with LLMs

Hye Sun Yun, David Pogrebitskiy, Iain J Marshall, Byron C Wallace

{yun.hy, pogrebitskiy.d, b.wallace}@northeastern.edu, {iain.marshall}@kcl.ac.uk

MOTIVATION

Meta-analyses of randomized controlled trials (RCTs) provide robust estimates of treatment efficacy and require extraction of data elements from individual articles for synthesis

- **Can we fully automate "on-demand" meta-analysis of evidence relevant to a given clinical question?**
- **Are modern LLMs sufficiently capable of *numerical* data extraction to permit accurate, fully automated meta-analysis?**

DATA ANNOTATION

- **Intervention, Comparator, & Outcome (ICOs)** from PubMed RCT reports
- Annotations based on **Abstract + Results sections** of RCT
- Schema:
  - ***Type of outcome***: binary or continuous
  - ***Binary outcome***: events, group sizes for I & C
  - ***Continuous outcome***: means, standard deviations, group sizes for I & C

Example Annotation for Given ICOs

Intervention	Comparator	Outcome
Hypnotherapy	Relaxation	Smoking

Outcome Type	Intervention Events	Intervention Group Size	Comparator Events	Comparator Group Size
Binary	19	116	17	117

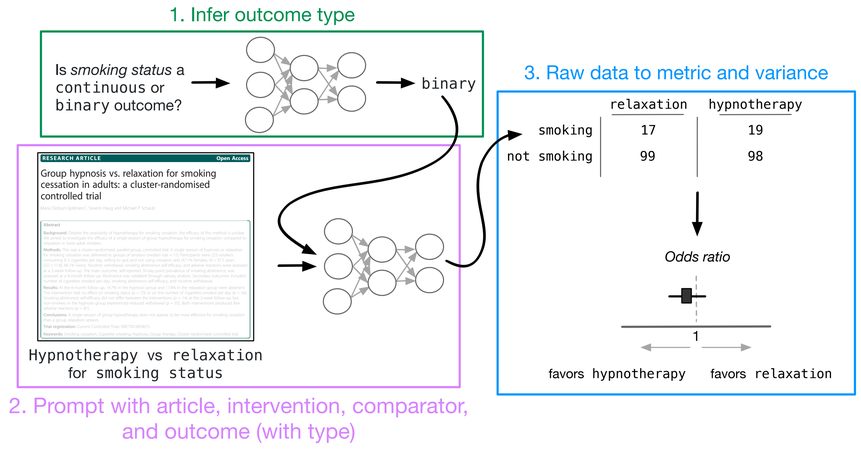
Metric	Dev	Test	Total
# PMC Articles	10	110	120
# Prompts (ICOs)	43	656	699
# Binary Outcomes	11	172	183
# Continuous Outcomes	32	484	516
% With Enough Data for Point Estimates	62.79	58.84	59.08
Mean Articles Tokens	3331	3603	3581

SUMMARY

- **Annotated dataset** for the task of extracting numerical clinical findings for conducting meta-analysis
- **Evaluation** of 8 modern LLMs using the annotated dataset
- **End-to-end case study** of a fully automated meta-analysis
- Binary outcomes extraction: LLMs with large input context windows (e.g. GPT-4) outperform smaller, open-source models
- Continuous outcomes extraction: LLMs perform poorly (<50% exact match)

This research was partially supported by National Science Foundation (NSF) grants RI-2211954 and IIS-1750978, and by the National Institutes of Health (NIH) under the National Library of Medicine (NLM) grant 2R01LM012086.

APPROACH



- Evaluated 8 LLMs on predicting outcome type and extracting binary and continuous outcomes in YAML format using **zero-shot approach**
- Python's `statsmodels` package for deriving point estimates and standard errors for meta-analysis

Data + Code

RESULTS

Part 1: Outcome Type

	GPT-4	GPT-3.5	Alpaca	Mistral	Gemma	OLMo	PMC LLaMA	BioMistral
Accuracy	0.713	0.607	**0.739**	0.201	0.665	0.290	0.732	0.133
F1 - Binary	**0.735**	0.680	0.000	0.576	0.590	0.424	0.124	0.275
F1 - Continuous	0.836	0.690	**0.851**	0.183	0.716	0.079	0.848	0.135
# Unknowns	155	152	1	489	**0**	5	15	409

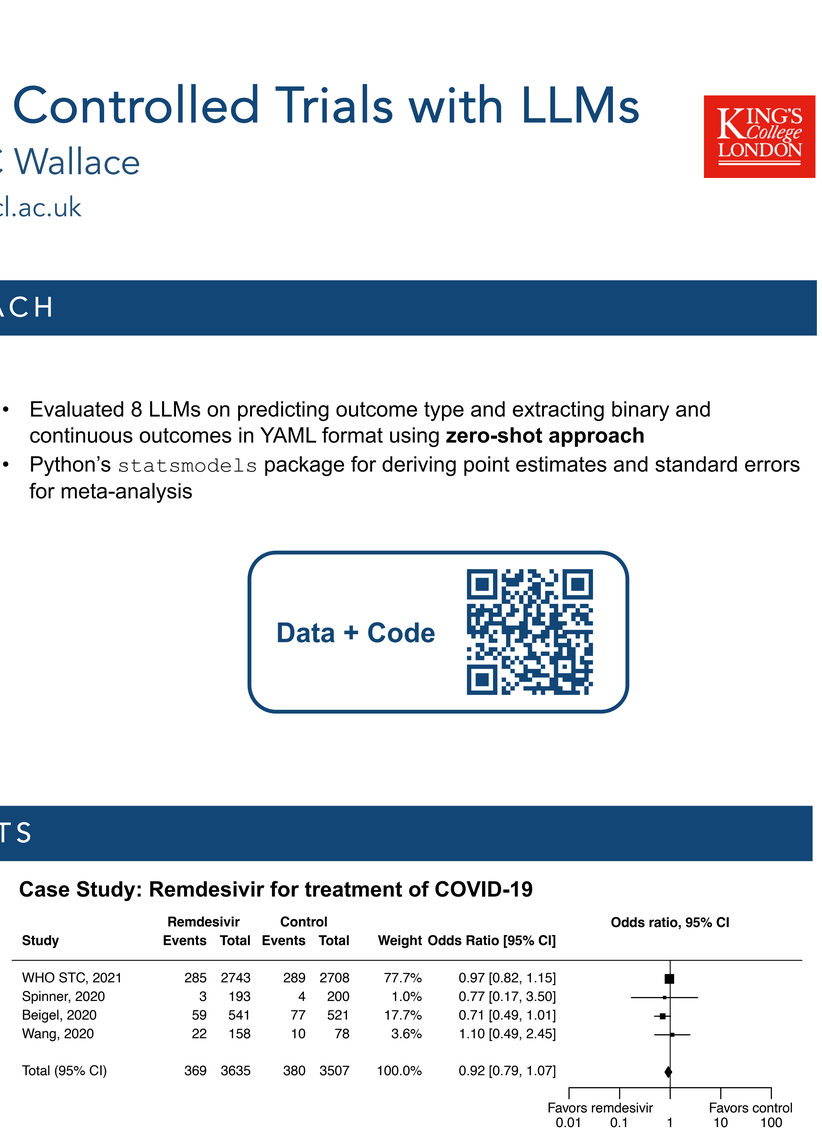
Part 2a: Binary Outcome Numerical Results Extraction

		GPT-4	GPT-3.5	Alpaca	Mistral	Gemma	OLMo	PMC LLaMA	BioMistral
Exact Match	*Total*	**0.655**	0.298	0.035	0.164	0.135	0.012	0.035	0.035
	IE	**0.749**	0.462	0.129	0.345	0.275	0.076	0.146	0.158
	IGS	**0.842**	0.655	0.094	0.515	0.509	0.170	0.088	0.053
	CE	**0.737**	0.392	0.129	0.333	0.275	0.123	0.158	0.158
	CGS	**0.830**	0.649	0.094	0.567	0.556	0.140	0.058	0.053
MSE		**0.101**	0.441	0.485	0.657	0.913	1.253	1.523	-
# Unknowns		41	145	490	**28**	90	319	524	612
% Complete		**87.94**	61.70	9.22	87.23	58.87	24.11	7.09	0.00

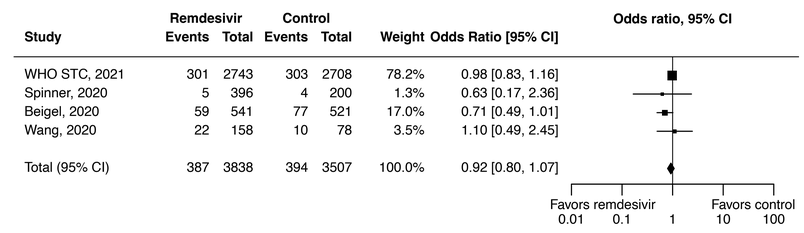
Part 2b: Continuous Outcome Numerical Results Extraction

		GPT-4	GPT-3.5	Alpaca	Mistral	Gemma	OLMo	PMC LLaMA	BioMistral
Exact Match	*Total*	**0.487**	0.280	0.039	0.095	0.087	0.035	0.039	0.041
	IM	**0.720**	0.538	0.309	0.348	0.328	0.221	0.369	0.390
	ISD	**0.751**	0.606	0.334	0.375	0.412	0.311	0.447	0.470
	IGS	**0.734**	0.641	0.216	0.507	0.534	0.190	0.107	0.087
	CM	**0.720**	0.526	0.330	0.361	0.324	0.227	0.390	0.402
	CSD	**0.738**	0.584	0.338	0.390	0.404	0.282	0.456	0.472
	CGS	**0.691**	0.608	0.181	0.427	0.447	0.184	0.109	0.087
MSE		**0.290**	0.951	6.257	1.138	3.466	1.738	-	-
# Unknowns		**422**	437	1169	483	775	1213	1778	1985
% Complete		**63.64**	62.40	31.82	62.81	40.08	11.98	4.96	0.00

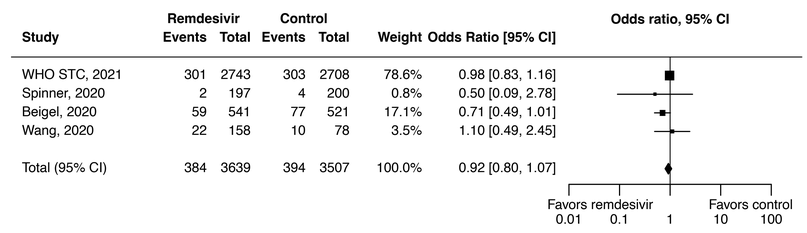
Case Study: Remdesivir for treatment of COVID-19



(A) Cochrane meta-analysis (reference)



(B) meta-analysis from GPT-4 outputs



(C) meta-analysis from Mistral Instruct 7B outputs