# Ghost in the machine or monkey with a typewriter—generating titles for Christmas research articles in *The BMJ* using artificial intelligence: observational study

Robin Marlow,[1,2] Dora Wood[1]

[1]Bristol Royal Hospital for Children, Bristol, BS2 8BJ, UK
[2]Centre for Academic Child Health, University of Bristol, Bristol, UK

**Correspondence to:** R Marlow robin.marlow@bristol.ac.uk (or @robindmarlow on Twitter; ORCID 0000-0002-3192-3102)

Additional material is published online only. To view please visit the journal online.

**Cite this as:** *BMJ* 2021;375:e067732
http://dx.doi.org/10.1136/bmj-2021-067732

**Accepted:** 21 October 2021

## ABSTRACT

### OBJECTIVE

To determine whether artificial intelligence (AI) can generate plausible and engaging titles for potential Christmas research articles in *The BMJ*.

### DESIGN

Observational study.

### SETTING

Europe, Australia, and Africa.

### PARTICIPANTS

1 AI technology (Generative Pre-trained Transformer 3, GPT-3) and 25 humans.

### MAIN OUTCOME MEASURES

Plausibility, attractiveness, enjoyability, and educational value of titles for potential Christmas research articles in *The BMJ* generated by GPT-3 compared with historical controls.

### RESULTS

AI generated titles were rated at least as enjoyable (159/250 responses (64%) *v* 346/500 responses (69%); odds ratio 0.9, 95% confidence interval 0.7 to 1.2) and attractive (176/250 (70%) *v* 342/500 (68%); 1.1, 0.8 to 1.4) as real control titles, although the real titles were rated as more plausible (182/250 (73%) *v* 238/500 (48%); 3.1, 2.3 to 4.1). The AI generated titles overall were rated as having less scientific or educational merit than the real controls (146/250 (58%) *v* 193/500 (39%); 2.0, 1.5 to 2.6); this difference, however, became non-significant when humans curated the AI output (146/250 (58%) *v* 123/250 (49%); 1.3, 1.0 to 1.8). Of the AI generated titles, the most plausible was "The association between belief in conspiracy theories and the willingness to receive vaccinations," and the highest rated was "The effects of free gourmet coffee on emergency department waiting times: an observational study."

### CONCLUSIONS

AI can generate plausible, entertaining, and scientifically interesting titles for potential Christmas research articles in *The BMJ*; as in other areas of medicine, performance was enhanced by human intervention.

## WHAT IS ALREADY KNOWN ON THIS TOPIC

Recent parallel advances in technology and digitisation have led to a rapid development of artificial intelligence (AI) and machine learning

In medicine, early applications of AI have been based around image recognition and diagnostics but with great potential for broader use

The most recent AI systems are capable of advanced language recognition, interpretation, and generation

## WHAT THIS STUDY ADDS

Titles of potential Christmas research articles in *The BMJ* generated by AI were as attractive and entertaining to readers as real titles published in the Christmas issue of *The BMJ*

With an additional stage of human intervention, the titles also performed similarly in terms of potential scientific and educational value

AI could have a role in generating hypotheses or directions for future research

## Introduction

Recent developments in machine learning and artificial intelligence (AI) are likely to revolutionise aspects of medical practice over the next decade. Although simple human applied rule based algorithms have been used in medical settings for decades, more recent developments in computer processing power and exponential increases in available data have enabled the development of systems that can optimise their own performance without human intervention. These are already in routine use in non-medical settings—for example, to target advertisements or articles of interest on social media, and to generate art and music.

Increasing evidence shows that when given access to large imaging databases these algorithms can already be used effectively to diagnose breast and lung cancer, retinal disease, and intracranial haemorrhage, with similar accuracy to that of human experts.[1] Such tools are likely to be able to offer decision support in other areas of medical practice soon, and frameworks for reporting AI and machine learning research are being developed to match this need.[2]

A detailed description of how AI works is beyond the scope of this article, but essentially AI comprises multilayered neural networks, which themselves are a group of linked algorithms, with outputs tuned to collectively respond to a stimulus from a particular input. Most traditional AIs are task specific (ie, trained on one form of labelled data), so that they become experts at, for example, categorising images or playing chess. More recent methods allow unsupervised learning by identifying patterns within massive datasets. Once developed, however, AIs are metaphorical black boxes, with an input and output but an inability to explain or interrogate the workings; if trained on a dataset with an unknown inherent bias, the AI might inherit this in a way that is difficult to detect.[3]

The current most up to date general purpose language AI is the Generative Pre-trained Transformer 3 (GPT-3) developed by OpenAI (San Francisco, CA). GPT-3 was trained using 175 billion varied items of text,

including the entirety of Wikipedia and a collection of books and websites.[4] From a starting prompt GPT-3 is capable of translation, answering questions, and even writing newspaper articles.[5] GPT-3 is a commercial product, and, because of concerns about the potential for misuse, it can only be accessed by submitting a proposal and being accepted onto a Beta program.

Although traditionally computers have been thought incapable of innovative or independent thought, given the developments in technology it seemed timely to evaluate the capability of AI to generate worthwhile hypotheses for medical research. Since 1982 *The BMJ* has published a special Christmas edition, featuring articles in which evidence based science is combined with more light hearted or quirky themes.[6] In this study we determined whether AI generated titles for potential Christmas research articles in *The BMJ* would meet the brief of combining scientific merit with engaging and entertaining subject matter.

## Methods

We took the titles of the 13 most read Christmas research articles of the past 10 years in *The BMJ* and used these to construct a prompt instructing GPT-3 to generate similar titles (supplementary file). Both authors independently scored the 57 titles GPT-3 generated on a scale of 1 to 6 for scientific merit, entertainment, and plausibility. We used the mean composite scores from this process to rank the titles and select the 10 highest rated and 10 lowest rated newly generated titles.

Despite an extensive review of the literature on the use of AI to generate titles for Christmas research articles in *The BMJ*, we were unable to identify any articles that could provide the required sample size. For this small study to disprove our null hypothesis that AI would be incapable of generating plausible titles, we used a convenience sample of 25 medical doctors from a range of specialties and settings: paediatricians, physicians in adult medicine, general practitioners, and anaesthetists from Africa, Australia, and Europe.

The participants were required to self-declare that they were familiar with the usual content and format of the Christmas issue of *The BMJ*. They were then asked to complete an online survey containing 10 randomly selected titles of Christmas research articles obtained from the archive of *The BMJ* and the 10 highest rated and 10 lowest rated AI generated article titles (fig 1). The titles were presented to each participant in a random order, blinded to which of the three categories (real articles, AI generated 10 highest rated and 10 lowest rated titles) the articles belonged. The participants were told that the list contained a mixture of real and AI generated titles but not the proportion of each.

Using a seven level Likert scale (absolutely not, probably not, maybe not, unsure, maybe, probably, absolutely), the participants rated each paper according to four statements: This a real *BMJ* paper; I want to read this; This would be funny/enjoyable to read; and This would be scientifically/educationally useful. They were also asked to select which of the 30 titles was the most plausible overall and which the funniest.

We assessed the ability of GPT-3 to generate titles unaided by comparing the proportion of real titles with positive Likert scores (5 to 7) with the proportion of the 10 highest and 10 lowest rated titles combined with positive scores. To determine if human curation was beneficial to AI, we performed the same comparison between the real titles and the 10 highest rated titles. Ordinal regression was used to test statistical significance between groups. Data were analysed using R version 4.0.5,[7] the Tidyverse,[8] and Likert packages.

### Patient and public involvement

Although the topic of this paper does not directly apply to specific patient groups, we did speak to patients about the study. We also asked a member of the public to comment on our manuscript after submission.

## Results

### AI generated highest and lowest rated titles combined

When the titles of real Christmas research articles in *The BMJ* were compared with the combined list of highest and lowest rated AI generated titles (fig 2), the real titles were rated as more likely to be an actual article (182/250 responses (73%) *v* 238/500 responses (48%); odds ratio 3.1, 95% confidence interval 2.3 to 4.1; P<0.001) and more likely to be scientifically or educationally useful (146/250 (58%) *v* 193/500 (39%); 2.0, 1.5 to 2.6; P<0.001). AI generated titles were equally as attractive to read as the real article titles (176/250 (70%) *v* 342/500 (68%); 1.1, 0.8 to 1.4; P=0.49) and rated as equally enjoyable (159/250 (64%) *v* 346/500 (69%); 0.9, 0.7 to 1.2; P=0.55).

### Curated AI generated titles

When the real titles were compared with the top ranked AI generated ones curated by humans (fig 3), the real titles were still believed to be more likely to represent an actual article (182/250 (73%) *v* 147/250 (59%); 2.2, 1.6 to 3.0; P<0.001) and were considered as educationally useful (146/250 (58%) *v* 123/250 (49%); 1.3, 1.0 to 1.8; P=0.08). The selected group of top ranked AI titles were still rated as equally attractive to read as the real titles (176/250 (70%) *v* 185/250 (74%); 0.9, 0.6 to 1.2; P=0.45) and as enjoyable (159/250 (64%) *v* 180/250 (72%); 0.8, 0.6 to 1.1; P=0.25).

When the participants were asked to choose the single most plausible title, 10 (40%) chose one that had been AI generated—the most popular being "The association between belief in conspiracy theories and the willingness to receive vaccinations." For the single funniest title only six (24%) participants chose a real article (fig 4).

## Discussion

In this small study, AI generated titles for potential Christmas research articles in *The BMJ* were at least as entertaining and attractive to readers in our sample


doi: 10.1136/bmj-2021-067732 | *BMJ* 2021;375:e067732 | the bmj

| Real Christmas titles | AI generated titles: Highest rated | AI generated titles: Lowest rated |
|---|---|---|
| Efficacy of educational video game versus traditional educational apps at improving physician decision making in trauma triage: randomized controlled trial[10] | The association between belief in conspiracy theories and the willingness to receive vaccinations | Superglue your nipples together and see if it helps you to stop agonising about erectile dysfunction at work |
| Are "armchair socialists" still sitting? Cross sectional study of political affiliation and physical activity[11] | Chicken soup prevents the development of pneumonia in children: randomized, double-blind, placebo controlled trial | What would happen if we stopped wiping our bottoms? |
| Stormy weather: a retrospective analysis of demand for emergency medical services during epidemic thunderstorm asthma[12] | The effects of free gourmet coffee on emergency department waiting times: an observational study | (Un)controlled explosions: assessing the risks of teaching pharmacy trainees about explosives by live firing |
| Working 9 to 5, not the way to make an academic living: observational analysis of manuscript and peer review submissions over time[13] | A double blind randomized placebo controlled trial of sleep deprivation by general physicians on intensive care unit mortality | Top ten reasons for repeated failed carjacking: a retrospective observational study |
| Effect of therapeutic suggestions during general anaesthesia on postoperative pain and opioid use: multicentre randomised controlled trial[14] | The multinational study of free-form dancing on hospital wards: a multicentre, randomized, controlled, observational trial | Where is Harold Shipman's coffin? |
| The survival time of chocolates on hospital wards: covert observational study[15] | Is Jack Frost nipping at your nose? Observational study of the times of day and night people make emergency dental appointments | The fire-hose carrying capacity of a Yorkshire farmer: observational study |
| Televised medical talk shows—what they recommend and the evidence to support their recommendations: a prospective observational study[16] | Are teddy bears bored by oral presentations? A cross sectional study of teddy bear gaze and attention seeking behaviour in paediatricians' offices | The evolution of homeopathy and other interesting stuff from The Lancet |
| Morphology and size of stem cells from mouse and whale: observational study[17] | The clinical effectiveness of lollipops as a treatment for sore throats: randomized controlled trial | Laughing gas, Santa Claus, and tooth fairy: a medical myth? |
| Intellectual engagement and cognitive ability in later life (the "use it or lose it" conjecture): longitudinal, prospective study[18] | An epidemiological and economic evaluation of Santa Claus for prophylaxis of festive vomiting in children: time series analysis | Using the stethoscope as a lie detector |
| Following celebrities' medical advice: meta-narrative analysis[19] | Does chocolate affect how long you live? A historical cohort study | Playing "spot the consultant": an observational study of the use of reflective pin badges in hospital consultants |

**Fig 1 | Ten randomly selected actual titles of Christmas research articles in *The BMJ* and 10 highest rated and 10 lowest rated artificial intelligence (AI) generated articles**

as titles of actual articles that were published in the Christmas issue of *The BMJ*. Real titles performed significantly better than AI generated ones (both curated and non-curated by the human participants) in terms of plausibility, although it was not possible to differentiate inherent plausibility from the participants' familiarity with previous published Christmas research articles in *The BMJ*. A small number of well known articles included by chance in our sample could have substantially skewed the results.

The only two titles to be rated both as the most plausible and the funniest were "The survival time of chocolates on hospital wards: covert observational study" (which was the third most accessed Christmas research article in the month of its publication, with 298 841 readers) and "The effects of free gourmet coffee on emergency department waiting times: an observational study," now our potential submission for the 2022 Christmas issue of *The BMJ*.

When we considered the perceived scientific value of the articles in our sample, AI generated titles not selected by humans performed noticeably more poorly than real titles. When a subsequent step of human curation was applied, the performance of the AI generated titles came within the range of the real titles.

This finding fits with previous work on AI, suggesting that the best results come from combining machine learning with human oversight.[9] Both human and machine decision making are limited by the quality and quantity of inputs. Humans are psychologically

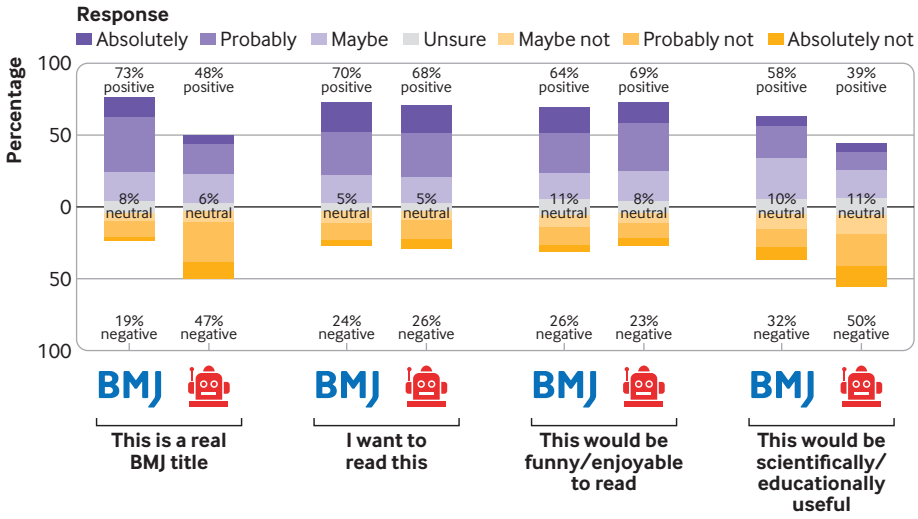


**Fig 2 | Real titles of Christmas research articles in *The BMJ* compared with top 10 and bottom 10 ranked AI generated titles using seven point Likert scales**

limited by how much data they can review, retain, and process, whereas machines are more likely to be constrained by the method of input. In our study, GPT-3 "knew" about the subject matter, wording, and associations of previously successful article titles but did not have the experience of clinical practice shared by the authors and study participants. Although humans might see the real world application of a study about clinician sleep deprivation on mortality in the intensive care unit, AI, with its inputs, sees this as no more or less useful than understanding the effects of applying superglue to nipples as a distraction from erectile dysfunction at work, nor can it understand if the titles are offensive. One limitation of our study is that we compared articles that had been accepted by, rather than submitted to, *The BMJ* for publication in its Christmas issue with the outputs of GPT-3. The performance of GPT-3 might have been better if this broader sample had been used.

Although our study might be the first to consider the use of AI to generate titles of research articles and to determine the attractiveness of those articles to potential readers, interest in the use of AI to generate research hypotheses is growing. For example, it has been proposed that the Euretos platform, mainly used by preclinical researchers to identify potential targets and biomarkers, could be used to generate hypotheses based on published papers, with subsequent expert review determining which of these are appropriate research directions to pursue.[10]

The findings of our study reinforce the essential role humans have in directing AI and curating its output. It is overwhelmingly likely, however, that recent developments in AI and machine learning will change the way work is done in healthcare, whether this is through improving diagnostic speed and accuracy, decision support, or reducing medical error. AI has the potential to change the way we

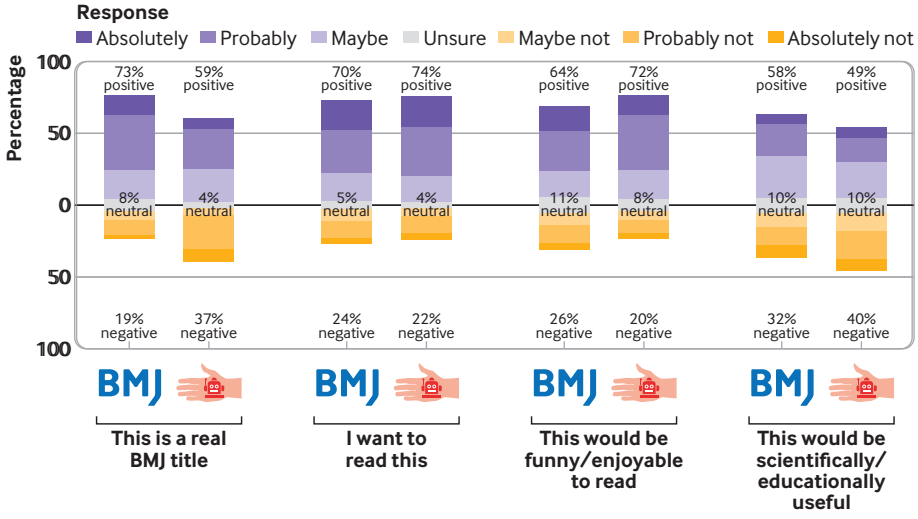


**Fig 3 | Real titles of Christmas research articles in *The BMJ* compared with curated top 10 ranked AI generated titles using seven point Likert scales**


doi: 10.1136/bmj-2021-067732 | *BMJ* 2021;375:e067732 | the bmj

| Most plausible title | | | Funniest title | |
|---|---|---|---|---|
| Rank | | No of humans (%) | | No of humans (%) |
| 1 | The survival time of chocolates on hospital wards: covert observational study | 8 (32) | Superglue your nipples together and see if it helps you to stop agonising about erectile dysfunction at work | 9 (36) |
| 2 | The association between belief in conspiracy theories and the willingness to receive vaccinations | 4 (16) | Are "armchair socialists" still sitting? Cross sectional study of political affiliation and physical activity | 3 (12) |
| 3 | Effect of therapeutic suggestions during general anaesthesia on postoperative pain and opioid use: multicentre randomised controlled trial | 3 (12) | Are teddy bears bored by oral presentations? A cross sectional study of teddy bear gaze and attention seeking behaviour in paediatricians' offices | 3 (12) |
| 4 | Efficacy of educational video game versus traditional educational apps at improving physician decision making in trauma triage: randomized controlled trial | 3 (12) | The survival time of chocolates on hospital wards: covert observational study | 3 (12) |
| 5 | The clinical effectiveness of lollipops as a treatment for sore throats: randomized controlled trial | 3 (12) | The effects of free gourmet coffee on emergency department waiting times: an observational study | 2 (8) |
| 6 | Chicken soup prevents the development of pneumonia in children: randomized, double-blind, placebo controlled trial | 1 (4) | What would happen if we stopped wiping our bottoms? | 2 (8) |
| 7 | Stormy weather: a retrospective analysis of demand for emergency medical services during epidemic thunderstorm asthma | 1 (4) | (Un)controlled explosions: assessing the risks of teaching pharmacy trainees about explosives by live firing | 1 (4) |
| 8 | The effects of free gourmet coffee on emergency department waiting times: an observational study | 1 (4) | The multinational study of free-form dancing on hospital wards: a multicentre, randomized, controlled, observational trial | 1 (4) |
| 9 | Using the stethoscope as a lie detector | 1 (4) | Top ten reasons for repeated failed carjacking: a retrospective observational study | 1 (4) |

**Fig 4 | Most plausible and funniest titles chosen by participants. Logo with Santa's hat indicates titles of real Christmas research articles i*n The BMJ***

select and interact with the medical literature; our study is an early demonstration of the way these technologies might also change the way we produce that literature.

## Conclusion

Even in the context of quirky titles such as those that appear in the Christmas issues of *The BMJ*, AI has the potential to generate plausible outputs that are engaging and could attract potential readers. Attracting interest can only be done with expert guidance, however, as some of the article titles in our study were irrelevant or offensive. This finding mirrors the potential use of AI in clinical medicine, as decision support rather than as outright replacement of clinicians.

**Contributors:** RM and DW designed the study and drafted the paper. They are both guarantors. RM collected and analysed the data. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

**Funding:** None.

**Competing interests:** All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.docx and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

**Ethical approval:** Not required.

**Data sharing:** Dataset and full reproducible code are available at https://doi.org/10.5281/zenodo.5681251.

The corresponding author (RM) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; no important aspects of the study have been omitted; that any discrepancies from the study as planned have been explained.

**Dissemination to participants and related patient and public communities:** Results were sent to the participants after taking part. We will share our results with the wider community through social media channels, educational meetings, and press release.

**Provenance and peer review:** Not commissioned; externally peer reviewed.



1 Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med* 2021;4:65. doi:10.1038/s41746-021-00438-z.
2 Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;368:l6927. doi:10.1136/bmj.l6927
3 Quinn TP, Jacobs S, Senadeera M, Le V, Coghlan S. The three ghosts of medical AI: Can the black-box present deliver?*Artif Intell Med* 2021;102158. doi:10.1016/j.artmed.2021.102158
4 Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. *arXiv:200514165 [cs]*.https://arxiv.org/abs/2005.14165 (accessed 26 Jul 2021).
5 GPT-3. A robot wrote this entire article. Are you scared yet, human? www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3 (accessed 27 Jul 2021).
6 Ladher N. Christmas crackers: highlights from past years of The BMJ's seasonal issue. *BMJ* 2016;355:i6679. doi:10.1136/bmj.i6679
7 R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2014. www.R-project.org/
8 Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *JOSS* 2019;4:1686. doi:10.21105/joss.01686
9 Alvin Powell. AI revolution in medicine. The Harvard Gazette. 2020. https://news.harvard.edu/gazette/story/2020/11/risks-and-benefits-of-an-ai-revolution-in-medicine/ (accessed 27 Jul 2021).
10 Sasselli V, Koers H. How big data and AI can help you generate your scientific hypothesis. www.elsevier.com/connect/how-big-data-and-ai-can-generate-your-scientific-hypothesis (accessed 27 Jul 2021).
11 Mohan D, Farris C, Fischhoff B, et al. Efficacy of educational video game versus traditional educational apps at improving physician decision making in trauma triage: randomized controlled trial. *BMJ* 2017;359:j5416. doi:10.1136/bmj.j5416
12 Bauman A, Gale J, Milton K. Are "armchair socialists" still sitting? Cross sectional study of political affiliation and physical activity. *BMJ* 2014;349:g7073. doi:10.1136/bmj.g7073
13 Andrew E, Nehme Z, Bernard S, et al. Stormy weather: a retrospective analysis of demand for emergency medical services during epidemic thunderstorm asthma. *BMJ* 2017;359:j5636. doi:10.1136/bmj.j5636
14 Barnett A, Mewburn I, Schroter S. Working 9 to 5, not the way to make an academic living: observational analysis of manuscript and peer review submissions over time. *BMJ* 2019;367:l6460. doi:10.1136/bmj.l6460
15 Nowak H, Zech N, Asmussen S, et al. Effect of therapeutic suggestions during general anaesthesia on postoperative pain and opioid use: multicentre randomised controlled trial. *BMJ* 2020;371:m4284. doi:10.1136/bmj.m4284
16 Gajendragadkar PR, Moualed DJ, Nicolson PLR, et al. The survival time of chocolates on hospital wards: covert observational study. *BMJ* 2013;347:f7198. doi:10.1136/bmj.f7198
17 Korownyk C, Kolber MR, McCormack J, et al. Televised medical talk shows--what they recommend and the evidence to support their recommendations: a prospective observational study. *BMJ* 2014;349:g7346. doi:10.1136/bmj.g7346
18 Hoogduijn MJ, van den Beukel JC, Wiersma LCM, Ijzer J. Morphology and size of stem cells from mouse and whale: observational study. *BMJ* 2013;347:f6833. doi:10.1136/bmj.f6833
19 Staff RT, Hogan MJ, Williams DS, Whalley LJ. Intellectual engagement and cognitive ability in later life (the "use it or lose it" conjecture): longitudinal, prospective study. *BMJ* 2018;363:k4925. doi:10.1136/bmj.k4925
20 Hoffman SJ, Tan C. Following celebrities' medical advice: meta-narrative analysis. *BMJ* 2013;347:f7151. doi:10.1136/bmj.f7151

**Supplementary file:** Writing prompt and full list of 57 generated titles