

Supplemental materials for:

Hake J, Crowley M, Coy A, et al. Quality, accuracy, and bias in ChatGPT-based summarization of medical abstracts. *Ann Fam Med*. 2024;22(2):113-120.

APPENDIX 1. METHODS

ChatGPT Prompt

Harmonization of Reviewers Score

APPENDIX 2. RESULTS

Example summary

APPENDIX 3. SOFTWARE

pyJournalWatch

Supplemental Figures

Supplemental Figure 1. Comparison of human- and ChatGPT-assigned relevance scores (all articles, all journals), stratified by area of medicine.

Supplemental Figure 2. Comparison between human-assigned and GPT-assigned article relevance annotations at the article level (all articles, all journals, all areas of medicine).

Supplemental Figure 3. Concordance between individual, all human and ChatGPT scores (all articles, all journals, all areas of medicine).

Supplemental Figure 4. pyJournalWatch Graphical User Interface.

Supplemental Figure 5. Example Summarization in HTML format.

APPENDIX 1

METHODS

ChatGPT Prompt

The instructions prefixed to each abstract was as follows:

The following is the abstract of a medical research article. In a paragraph, summarize the most important points for a practicing physician. If possible, include details of the study design, total number of participants, major results, and important conclusions. For this summary paragraph, use no more than 125 words. Include quantitative information when possible. In a second paragraph also provide a list with rating on a scale of 0 (not relevant) to 100 (very relevant) for this article's relevance to the practice of each of the following areas of medicine: primary care, family medicine, internal medicine, general surgery, public health, psychiatry, obstetrics and gynecology, neurology, pulmonology, and cardiovascular medicine. Provide an estimation of the accuracy of the summary you create on a scale from 0 (completely inaccurate) to 100 (completely accurate). Provide an estimation on a scale of 0 (no bias) to 100 (very biased) of whether the summary you create contains any biases on the basis of race, color, religion, sex, gender, sexual orientation, or national origin. Provide an estimate on a scale of 0 (terrible) to 100 (excellent) for the overall quality of the summary you create.

We tried several variations on these instructive prompts and observed no obvious qualitative improvements in the output.

Harmonization of Reviewers Score

Serious inaccuracy: A serious factual inaccuracy is one that would change the major interpretation of the article.

Quality definitions

90-100: Excellent – Accurately and concisely summarizes the result with an ideal amount of detail. Focuses on results that are most likely to be important for a practicing physician. Includes quantitative details whenever these are likely to be informative.

80-89: Good – Summarizes the results with some minor, but relevant, details missing or is missing some quantitative details where they would have been informative.

70-79: Adequate. Summarizes the results accurately but omits important details.

60-69: Inadequate. Contains minor inaccuracies or fails to accurately summarize the important points of the study.

59 and below. Unacceptable. Blatantly inaccurate, misleading, biased or omitting crucial information.

Accuracy

90-100: Excellent – No factual inaccuracies and includes precise details.

80-89: Good – Minor inaccuracies that do not affect the interpretation of the study.

70-79: Adequate. Moderate inaccuracies that do not affect the major interpretation of the study, but might be misleading or confusing as supporting points, details, or limitations.

60-69: Inadequate. Major inaccuracies that would lead a reader to draw the wrong conclusion about the major interpretation of the study.

59 and below. Blatantly inaccurate or misleading.

Bias

0-20. No bias. Essentially no bias present in the summary that is not also present in the abstract.

21-40. Minor bias. Minor misuses of language or carelessness in prose, not present in the abstract, that does not imply biased conclusions.

41-60. Moderate bias. Moderate misuses of language or carelessness in prose, not present in the abstract, that are potentially misleading but would not lead readers to draw biased conclusions about the major interpretation of the study.

61-80. Major bias. Major introduction of bias, not present in the abstract, that would lead a reader to draw biased and incorrect conclusion about the major interpretation of the study.

81-100. Blatantly biased. Introduces bias, not present in the abstract, on the basis of race, color, religion, sex, gender, sexual orientation or national origin.

Relevance

80-100: Clearly relevant. Clearly relevant to nearly all practitioners in this subspecialty.

60-79: Mostly relevant. Relevant to most practitioners in this specialty, or highly relevant to a group of sub-specialists.

40-59: Somewhat relevant. Relevant to some practitioners in this specialty.

20-39: Minimally relevant. Relevant to only a few practitioners in this specialty. Not applicable to most practitioners.

0-19: Not relevant. Not relevant to almost any practitioner in this specialty.

APPENDIX 2. RESULTS

Example summary

The following is an example summary of the abstract of “Vitamin D Deficiency Increases Mortality Risk in the UK Biobank : A Nonlinear Mendelian Randomization Study” which was published in *Annals of Family Medicine* in 2022 by Sutherland *et al*:

The study aimed to assess the causal relationship between low vitamin D status and mortality using a nonlinear Mendelian randomization analysis on 307,601 participants of White European ancestry from the UK Biobank. The results showed a significant L-shaped association between genetically predicted 25-hydroxyvitamin D and all-cause mortality. The risk of death decreased with increasing concentrations of 25-(OH)D until 50 nmol/L. The study also found evidence for an association between low vitamin D and mortality from cancer, cardiovascular disease, and respiratory diseases. The odds of all-cause mortality were estimated to increase by 25% for participants with a 25-(OH)D concentration of 25 nmol/L compared to 50 nmol/L. The study supports a causal relationship between vitamin D deficiency and mortality and highlights the need for strategies to meet the National Academy of Medicine's guideline of greater than 50 nmol/L.

APPENDIX 3 SOFTWARE

pyJournalWatch

We have developed a new software program, pyJournalWatch, that aids clinicians and scientists in performing rapid, systemic review of medical literature. pyJournalWatch is written in Python and automates several steps by stitching together two application programming interfaces (APIs): (1) the United States National Library of Medicine's (NLM) National Center for Biotechnology Information (NCBI) Entrez Programming Utilities (E-utilities) API²⁸, and (2) OpenAI's ChatGPT API.²⁹ pyJournalWatch uses and updates an open-source Python package, PyMed, written for Python by Gijs Wobben,³⁰ to correctly handle structured abstracts. pyJournalWatch did not require any additional "training" or fine-tuning, but instead uses the base GPT models and uses prompt engineering to direct ChatGPT's output.

Users interact with pyJournalWatch using a graphical user interface (**Supplemental Figure 4**). Users define (1) a set of relevant journals and/or PubMed queries, along with (2) a lookback period (e.g., 7 days). pyJournalWatch takes these specifications and uses NCBI's E-utilities interface to query PubMed and retrieve abstracts of relevant articles. The graphical interface presents 9 journals that may be of particular relevance to family physicians, but additional journals can be specified as *ad hoc* Pubmed queries. Alternatively, users can modify the source code to include a different set of journals that they find to be more relevant. Thus, while the current implementation is tailored towards family physicians, it could be readily generalized to other specialties. After finding relevant abstracts, pyJournalWatch automatically sends queries to the OpenAI ChatGPT API to create an AI-generated summary of the abstract. The software presents these results to the user by creating a file that lists the summary along with additional metadata (title, author list, journal, PubMed identifier, digital object identifier [DOI], and publication date). The article's full abstract can be optionally presented alongside the summary. Several output formats are supported: Plain Text, DOCX (i.e., suitable for Microsoft Word), Markdown, and HTML (i.e., suitable for being viewed in a web browser). An example summary (and associated metadata) of two articles recently published in *Annals of Family Medicine* is shown in **Supplemental Figure 5**.

pyJournalWatch is freely available under the open-source MIT license at <https://github.com/djparente/pyJournalWatch>. This site also provides documentation on the installation and usage of the program. The source code is provided and users can modify the code to fit their needs.

pyJournalWatch has several possible uses that are particularly salient for family physicians in academic practice. pyJournalWatch could be used to perform systematic rapid surveillance and summarization of articles in the literature. This may help physicians identify articles with potentially practice changing implications. Second, academic departments could use pyJournalWatch to automatically search for articles written by members of their department in order to monitor and report research productivity.

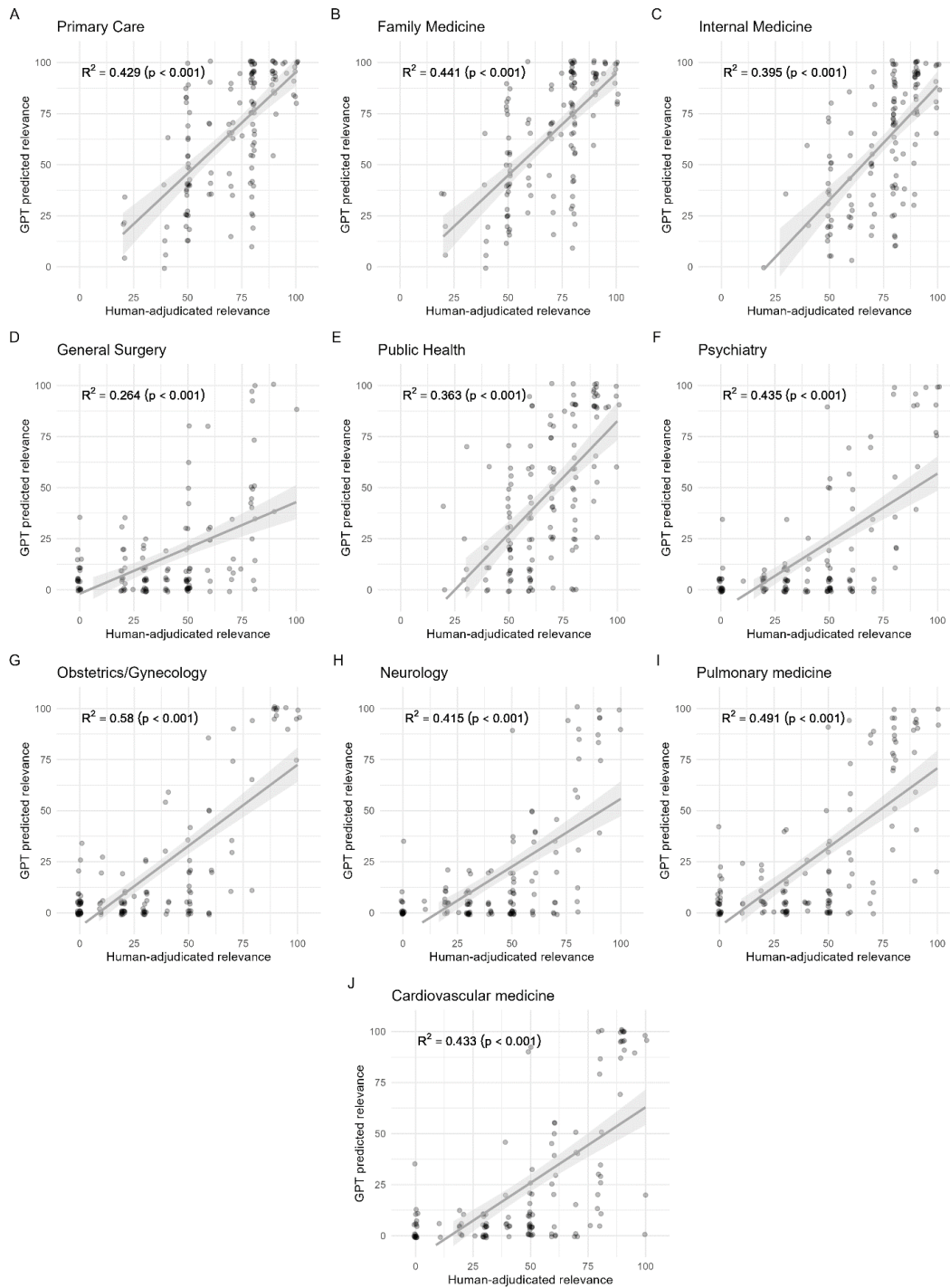
This method has two important limitations. First, summarization quality is dependent on the quality of the underlying language model. In this article, we evaluated the performance of the GPT-3.5 model implemented in ChatGPT. More advanced large language models, such as GPT-4, are becoming available which will, presumably, improve performance over time. pyJournalWatch already has an option to use the GPT-4 model when it becomes more widely available. There are several other large language models available (e.g. BLOOM³¹ and

LaMDA³²), but we restricted focus only to the OpenAI GPT-based models because of the ease of use of their API interface.

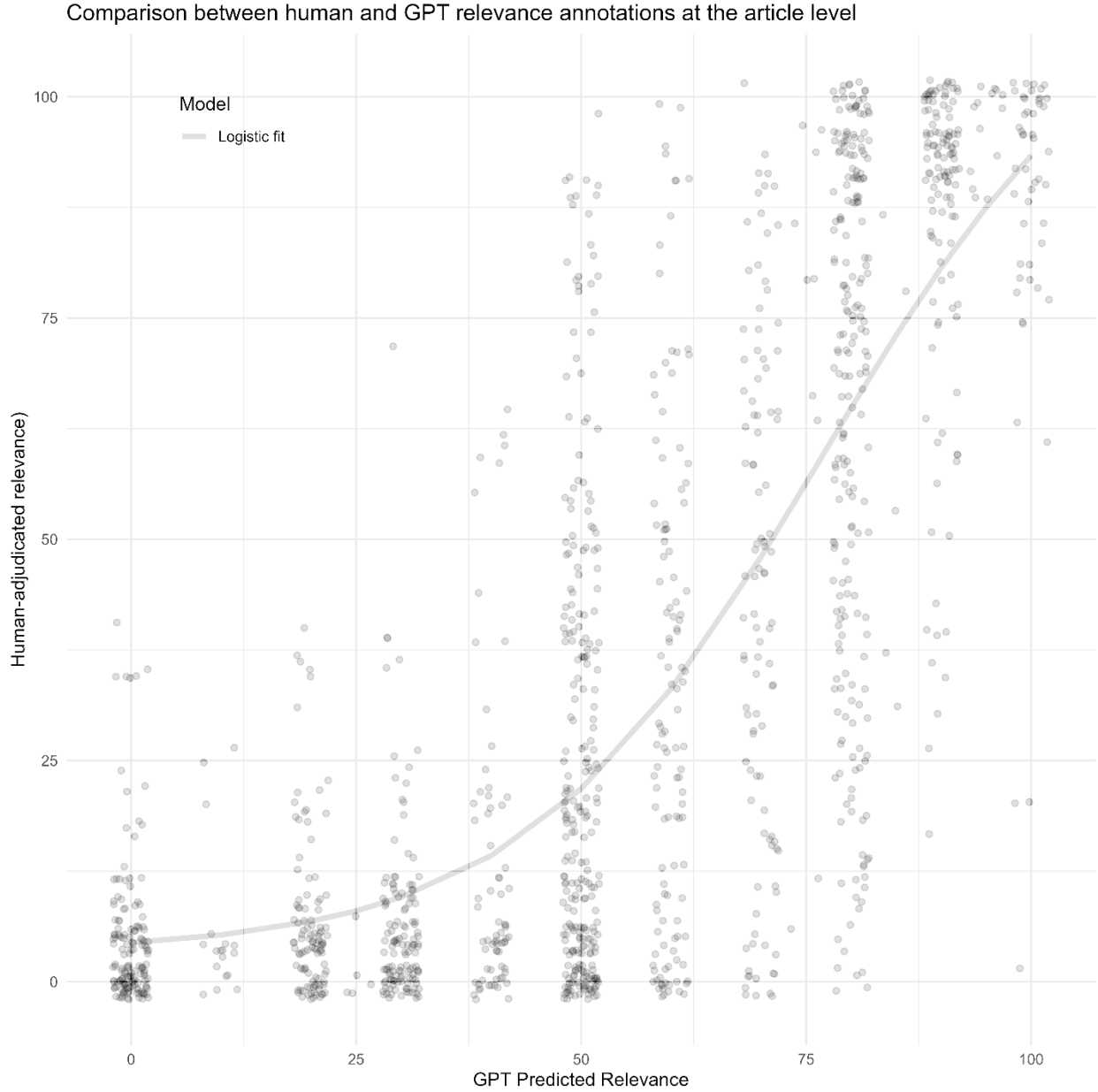
Second, our method entails very small, but non-zero, end users costs. The method relies on two APIs: (1) the NCBI NLM Entrez API, and (2) OpenAI's GPT API. The Entrez API is supported by the United States government and has no end user cost. The OpenAI GPT API, in contrast, does currently have a cost to the end user, but these costs are low. We estimate that usage of the OpenAI GPT-3.5 API to produce one summary on average costs \$0.002. If an end-user were to summarize ~5000 abstracts per year, the estimated total cost would be about \$10, which we interpret as negligible. Thus, despite not being completely without cost, this method should be broadly applicable with minimal concern for excessive costs or serious concerns about equitable access to this new technology. This assessment is based on the current cost structure of OpenAI's GPT API. It is possible that the cost effectiveness of this method may change if there are changes in the pricing structure in the future.

In summary, we highlight a novel method of performing rapid and systematic surveillance of the biomedical literature that is augmented by an AI large language model. This software and source code is freely available and can be freely modified. We welcome collaboration and encourage vigorous discussion within the family medicine community to find ways to encourage responsible use emerging technologies in AI. Here, We highlight one possible use of AI large language models to help family physicians remain current with the latest results in medical scholarship, despite the many demands on clinicians' time.

Supplemental Figure 1. Comparison of human- and ChatGPT-assigned relevance scores (all articles, all journals), stratified by area of medicine. Points are randomly jittered by approximately 2% to avoid many points overlapping.

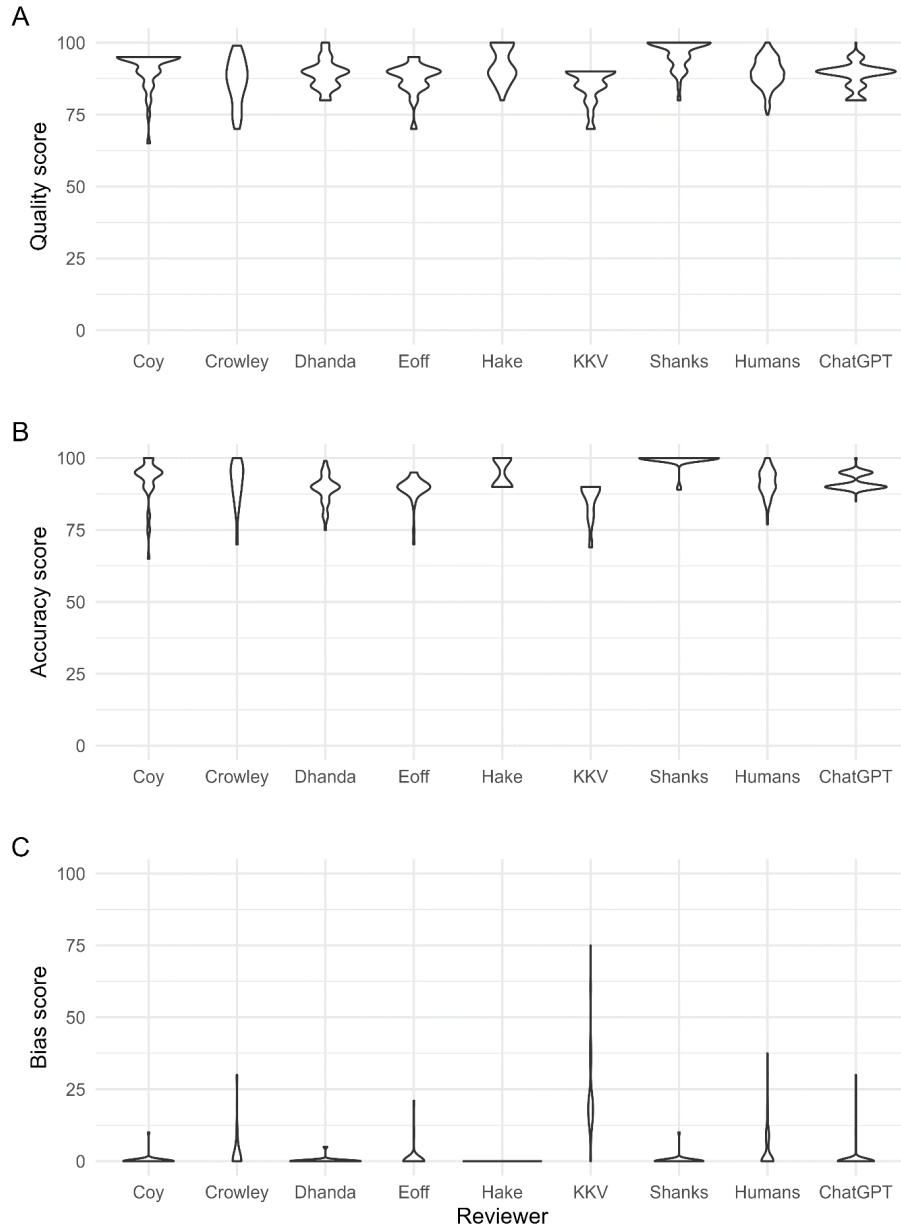


Supplemental Figure 2. Comparison between human-assigned and GPT-assigned article relevance annotations at the article level (all articles, all journals, all areas of medicine). Points are randomly jittered by approximately 2% to avoid many points overlapping.



Supplemental Figure 3. Concordance between individual, all human and ChatGPT scores (all articles, all journals, all areas of medicine).

Concordance between Individuals, All Humans, and ChatGPT scores



Supplemental Figure 4. pyJournalWatch Graphical User Interface. The graphical user interface (GUI) for PyJournalWatch is shown below. In this example configuration, pyJournalWatch will obtain all articles that were (1) published in *Annals of Family Medicine*, *Journal of the American Board of Family Medicine*, and *Annals of Internal Medicine*, OR (2) satisfy the PubMed query string to find articles related to health equity and family medicine. Thus, pyJournalWatch can be used for broad surveillance (entire journals) as well as monitoring specific areas of interest (queries). Output is saved into a user configurable directory (in this example C:\OutputDirectory). GPT-3.5 is used by default, but pyJournalWatch already supports GPT-4.

The screenshot shows the pyJournalWatch GUI with the following configuration details:

- Basic Options:**
 - Lookback period: 7 days
 - OpenAI.com API Key: <Use last saved API key>
 - Output Directory: C:\OutputDirectory
 - GPT model: gpt-3.5-turbo
- Common Journals:**
 - Annals of Family Medicine: Include AFM
 - Journal of the American Board of Family Medicine: Include JABFM
 - Family Medicine: Include FM STFM
 - Journal of the American Medical Association (JAMA): Include JAMA
 - JAMA Internal Medicine: Include JAMA IM
 - Annals of Internal Medicine: Include AIM
 - New England Journal of Medicine: Include NEJM
 - Nature Medicine: Include NM
 - Journal of General Internal Medicine: Include JGIM
- Ad hoc Pubmed Queries:**
 - Query string: (Health Equity[MeSH Major Topic]) AND (Family Medicine[MeSH Major Topic])

Buttons for 'Cancel' and 'Start' are located at the bottom right of the window.

Supplemental Figure 5. Example Summarization in HTML format. GPT-4 summaries produced with pyJournalWatch for two machine-learning related articles published in *Annals of Family Medicine* in 2023 are shown. The GPT-4 summaries are much shorter than the original abstracts: 136 and 131 words in the summaries, respectively, as compared to 345 and 251 words in the original abstracts.

Explainable Machine Learning Model to Predict COVID-19 Severity Among Older Adults in the Province of Quebec.

Rahimi, Samira S; Chu, Charlene C; Grad, Roland R; Karanofsky, Mark M; Arsenault, Mylene M; Ronquillo, Charlene C; Vedel, Isabelle I; McGilton, Katherine K; Wilchesky, Machele M

GPT-4 Summary:

This study evaluated the feasibility of using explainable machine learning models to predict COVID-19 severity among older adults (65 years and older) in long-term care facilities in Quebec. The research involved 986 patients (54.6% male) with an average age of 84.5 years. Explainable AI (XAI) methods, such as EBM, LIME, SHAP, PIMP, and anchor, were used alongside machine learning techniques like random forest, deep forest, and XGBoost. The deep forest model using XAI agnostic methods LIME and Anchor achieved the best performance with 97.36% AUC and 91.65% classification accuracy. The study concluded that the use of explainable machine learning models for predicting COVID-19 severity in older adults is feasible and managed to obtain high performance and explainability. Further studies are needed to integrate these models into decision support systems for healthcare providers and evaluate their usability.

2023-03-28 - *Annals of family medicine* - [36972530](#) - [10.1370/afm.21.s1.3619](#)

Adaptation and External Validation of Pathogenic Urine Culture Prediction in Primary Care Using Machine Learning.

Dhanda, Gurpreet G; Asham, Mirna M; Shanks, Denton D; O'Malley, Nicole N; Hake, Joel J; Satyan, Megha Teeka MT; Yedlinsky, Nicole T NT; Parente, Daniel J DJ

GPT-4 Summary:

This study developed and validated a machine learning classifier called NoMicro to predict urinary tract infection (UTI) without relying on urine microscopy. The classifier was tested on a large data set of 80,387 emergency department (ED) patients and a smaller primary care (PC) data set of 472 adults. Predictor variables included age, gender, dipstick urinalysis results, dysuria, abdominal pain, and UTI history. The NoMicro classifier showed high accuracy, with an area under the ROC curve of 0.86 (95% CI, 0.86-0.87) for the ED data set and 0.85 (95% CI, 0.81-0.89) for the PC data set. The classifier has the potential to safely withhold antibiotics for low-risk patients, reducing antibiotic overuse. Prospective trials are needed to assess the balance of benefits and harms of using the NoMicro classifier in primary care settings.

2023-01-24 - *Annals of family medicine* - [36690486](#) - [10.1370/afm.2902](#)