

The Use of Artificial Intelligence in Systematic Reviews

James Thomas



EPPi Centre
Evidence for
Policy & Practice

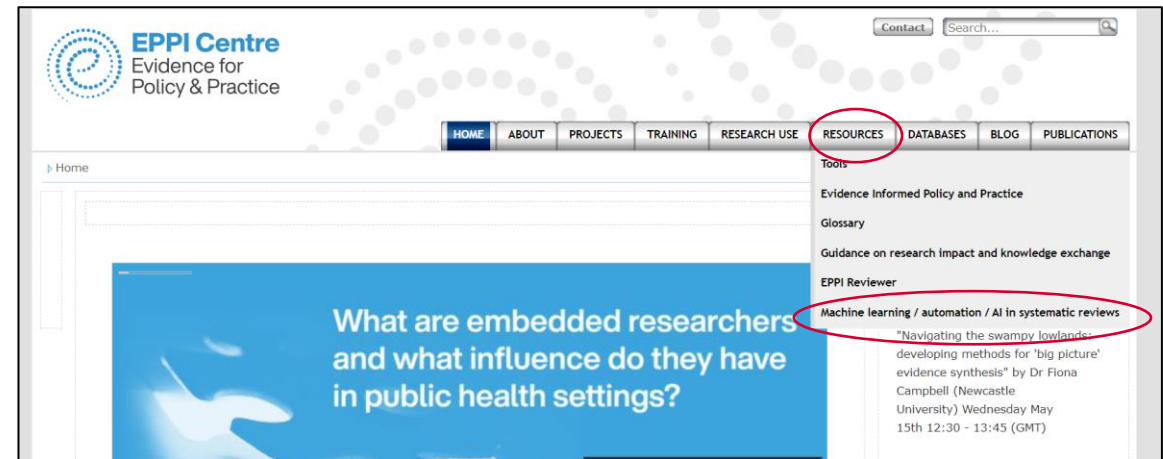
About me

- Worked in the EPPI-Centre, UCL for a long time
- Systematic reviews – mostly for Department of Health & Social Care / PHE
- Addressing questions beyond effectiveness
- Long-standing area of work in making the review process more efficient using new technologies



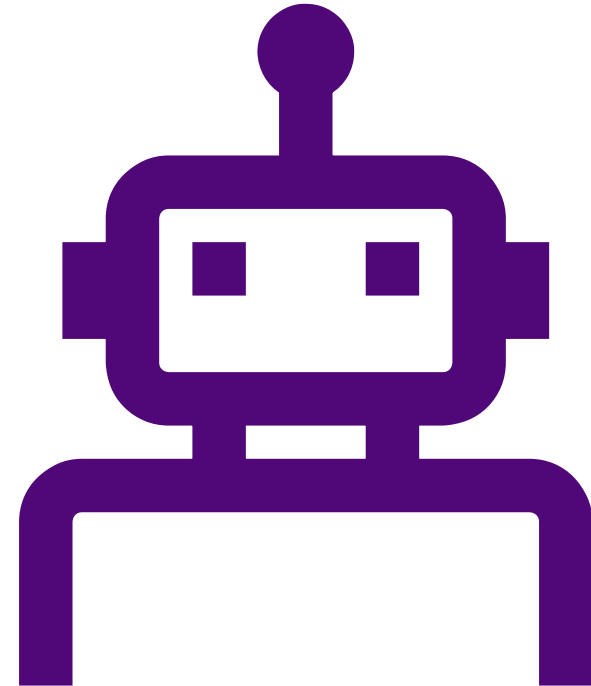
Outline

- This session
 - Introduction to automation / machine learning / AI in systematic reviews
- Next session
 - Tools and how to evaluate them
- Session three
 - Using and evaluating tools
- Session four
 - Feedback and discussion
- Slides and links to resources:
 - <https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3677>
 - Search for EPPI Centre website



In this session

- Introduction to AI / machine learning / automation tools for systematic reviews (and how they work)
- Please feel free to ask questions as we go
- Please also think about which tool you'd like to try out later this afternoon



Systematic review priorities

Systematic reviews are often used to inform decisions that affect people's lives

Systematic reviewers favour accuracy over efficiency

Highly sensitive searches are required to avoid selection bias

Highly accurate quality assurance processes are required to avoid human error

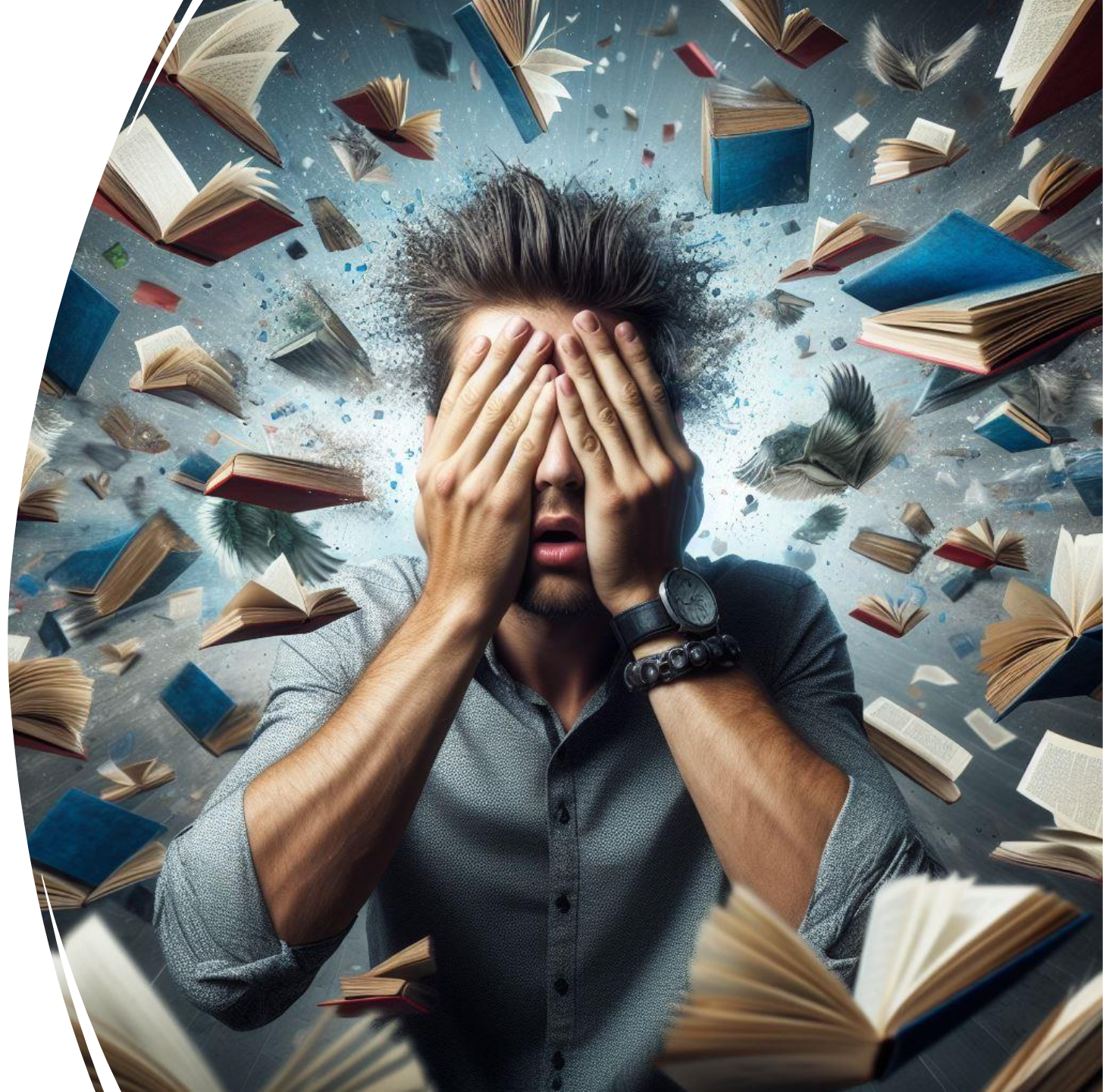


Impact of these priorities

- An inefficient, resource-intensive process has evolved that produces reliable, but expensive and time consuming, reviews
- We cannot keep pace with the deluge of new research being published
- E.g. in the Cochrane Reviews published March 2014, > 163k citations were screened; 6,599 full text reports were read; and 703 studies were included
- That's about 2 million records per year

This means

- Only a fraction of available studies are included in systematic reviews
- Systematic reviews do not cover all questions/ domains comprehensively
- We don't even know when systematic reviews *need* to be updated





Four machine learning / automation paradigms

- Rules-based approaches
 - (strictly speaking, not *machine learning*)
- Unsupervised approaches
- Supervised approaches
- Generative approaches ('Gen AI')

- Covering in terms of technology not purpose, so we can consider their strengths and weaknesses more easily

Rules-based approaches

As you might guess... a set of rules is constructed by humans and given to the machine



For example

Look up a simple set of words

Use of synonyms

If a given phrase is present, apply a given code

Many citation duplicate-checking algorithms



Rules can be accurate... but fragile



If you stick within the rules,
you get the anticipated
results



If you stray outside – even a
little bit – the rule can fail
altogether



No grey area – it works, or
completely fails

Rules are not fashionable!



Rules-based approaches

Designing and
running a
search strategy

Running an
automatic
deduplication
algorithm

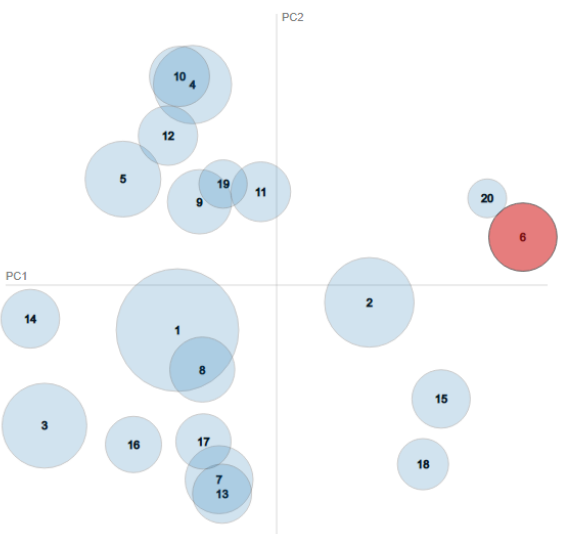
Polyglot search
translator



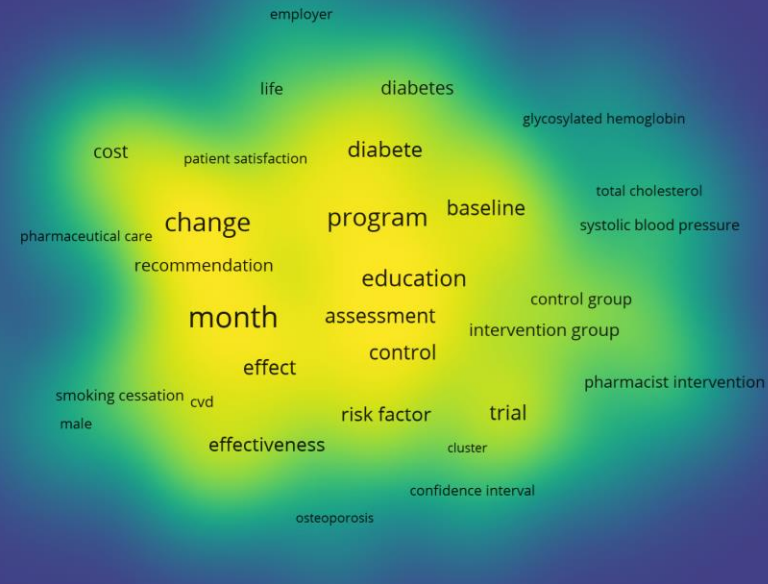
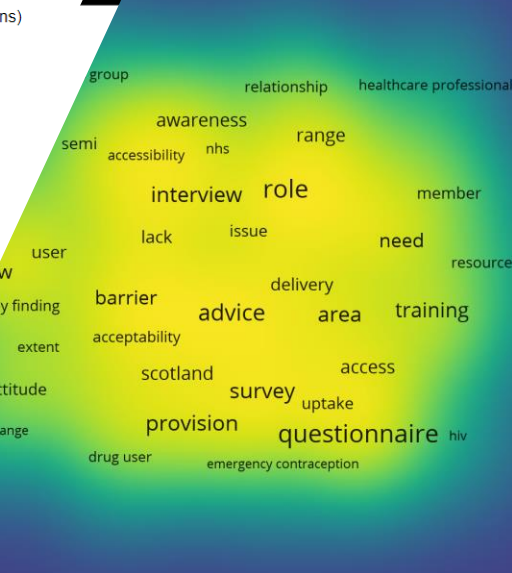
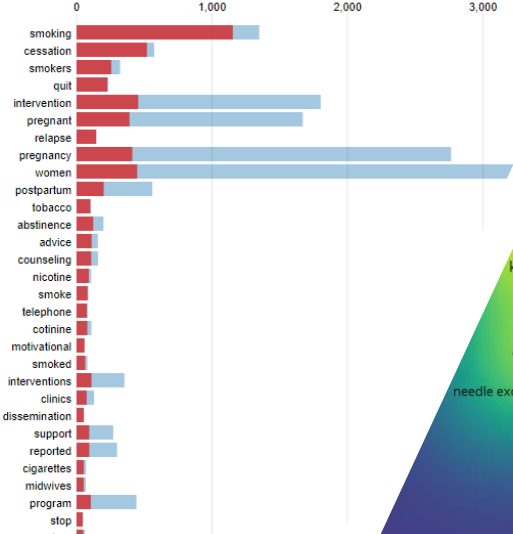
Unsupervised approaches

- The machine is given no rules...
- And simply identifies patterns in the data
- E.g.
 - Relationships between words
 - Clustering documents

Intertopic Distance Map (via multidimensional scaling)

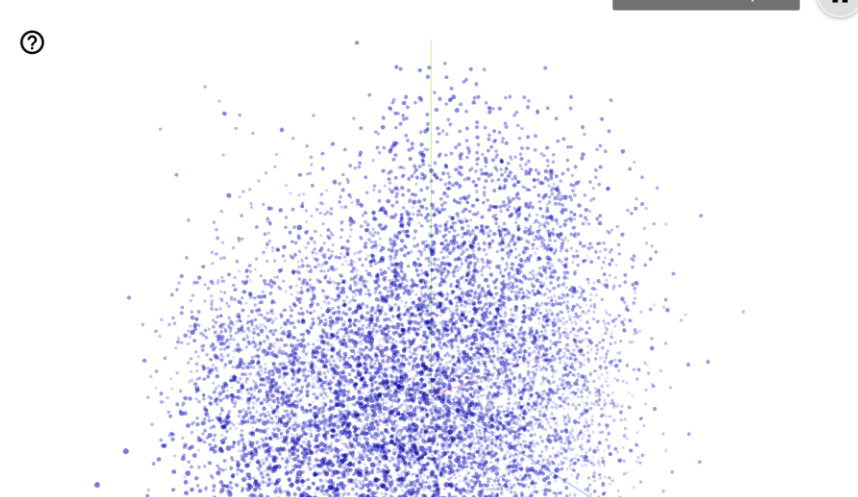


Top-30 Most Relevant Terms for Topic 6 (5% of tokens)



• Unsupervised approaches can help you explore patterns in your data
 • Attractive visualisations are possible

5 tensors found
 Word2Vec 10K
 Label by word Color by No color map
 Edit by word Tag selection as
 Load Publish Download Label
 Sphेरize data
 Checkpoint: Demo datasets
 Metadata: oss_data/word2vec_10000_200d_labels.tsv



Reset zoom to fit all points

Show All Data Isolate 101 points

Search

SMOKING CESSATION

ADOLESCENTS LUNG RELATIONSHIP CHRONIC CAUSE

TOBACCO SWINBERG'S FUNCTIONAL LIMITATIONS HEALTH BEHAVIORS LUNG CANCER PTOCHORDIAL IMPACTION MALE COPD/CLAP/USERS BIRTHDAY DISORDERS IDENTICAL TO STUDENTS EDUCATION ATTENTION KOREAN MENTALITY ABOVE PATIENTS OPAL MICROBIOME HONGKONG AIRPOLLUTANTS NA TUBERCULOSIS WOMEN'S EXPERIENCES OTHER TOPICS

Top 100 results of about 268158 for smoking

- 1 [Prospective, multi-centric benchmark study assessing delirium: prevalence, incidence and its correlates in hospitalized elderly Lebanese patients.](#)
 With the increase in the proportion of elderly Lebanese patients, little is known about delirium's prevalence, incidence and correlated factors. ... To identify the prevalence, incidence and factors associated with overall and incident delirium in hospitalized elderly Lebanese patients.
<http://www.ncbi.nlm.nih.gov/pubmed/31203529>
- 2 [Familial cancer of unknown primary.](#)
 Cancer of unknown primary site (CUP) is a deadly disease diagnosed through metastases at various organs without primary tumor identification. Despite the major molecular and technological advances, the carcinogenesis of CUP remains enigmatic which hampers adequate study design of treatments leading to survival improvement. To date, the pathogenesis of CUP is still debatable with one hypothesis considering CUP is simply a group of metastatic tumors with unidentified primaries and another considering it a distinct entity with specific genetic and phenotypic aberrations. Familial CUP seems to favor the first hypothesis due to common genetic predisposition factors between known primaries and CUP. Two clinical implications may be withdrawn from the pathogenesis of familial clustering of CUP. The detailed family history and environmental risk factors may orient towards the primary tumor identification. In cases of familial, **smoking** avoidance and adherence to general population guidelines for cancer screening would be strongly encouraged.
<http://www.ncbi.nlm.nih.gov/pubmed/31203526>
- 3 [Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: descriptive characteristics of the epidemiological studies within the IMI DIRECT Consortium.](#)
 Here, we describe the characteristics of the Innovative Medicines Initiative (IMI) Diabetes Research on Patient Stratification (DIRECT) epidemiological cohorts at baseline and follow-up examinations (18, 36 and 48 months of follow-up).
<http://www.ncbi.nlm.nih.gov/pubmed/31203377>
- 4 [Health behaviours and mental and physical health status in older adults with a history of homelessness: a cross-sectional population-based study in England.](#)
 This study compared (1) levels of engagement in lifestyle risk behaviours and (2) mental and physical health status in individuals who have previously been homeless to those of individuals who have not.
<http://www.ncbi.nlm.nih.gov/pubmed/31203244>
- 5 [Combined effects of lung function, blood gases and kidney function on the exacerbation risk in stable COPD: Results from the COSYCONET cohort.](#)
 Alterations of acid-base metabolism are an important outcome predictor in acute exacerbations of COPD, whereas sufficient metabolic compensation and adequate renal function are associated with decreased mortality. In stable COPD there is, however, only limited information on the combined role of acid-base balance, blood gases, renal and respiratory function on exacerbation risk grading.
<http://www.ncbi.nlm.nih.gov/pubmed/31203096>
- 6 ["Don't smoke in public, you look like trash": An exploratory study about women's experiences of smoking-related](#)

Query: smoking - Source: PubMed (100 results, 0 ms) - Cluster: Lingo
 v3.16.0-SNAPSHOT | build 277 | 2018-05-17 11:55 © 2002-2019 Stanislaw Ojnicki, David Hinton

Unsupervised approaches



'Mapping' characteristics of research automatically

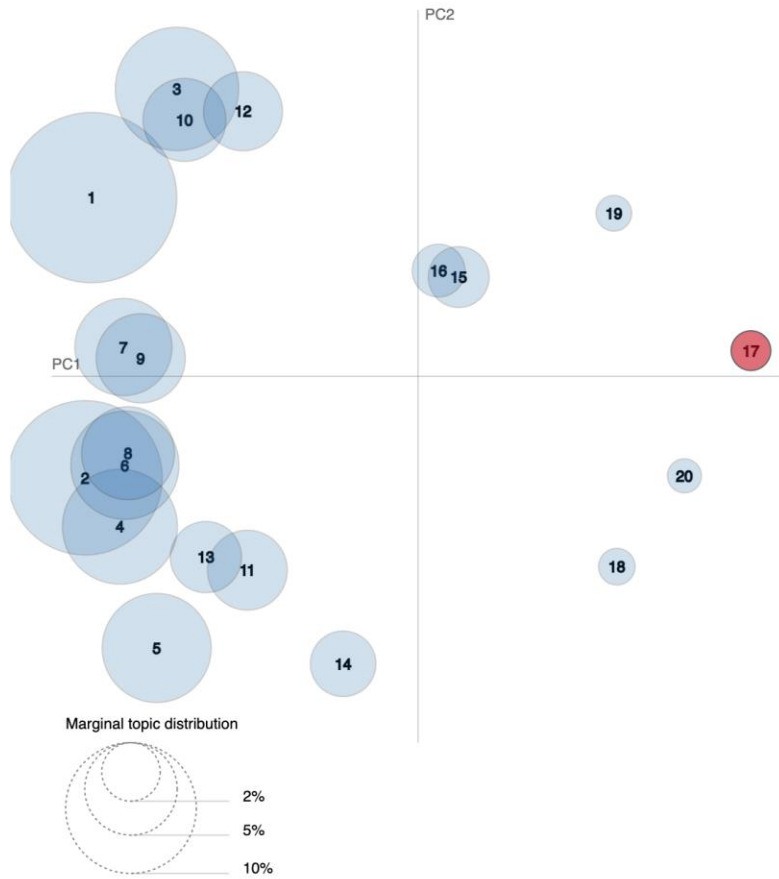


Identifying key terms from text data

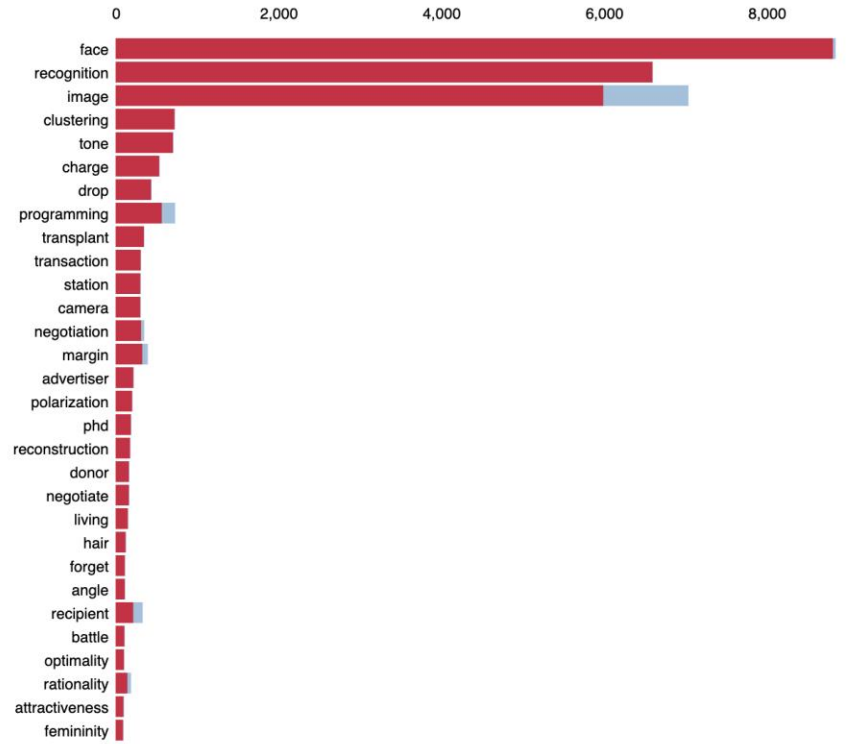
Selected Topic:

Slide to adjust relevance metric:(2) $\lambda = 0.36$

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 17 (0.9% of tokens)



Overall term frequency (blue bar)
 Estimated term frequency within the selected topic (red bar)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
 2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

Unsupervised approaches lack control



Very powerful – can reveal relationships in the data which are not necessarily obvious



Very efficient – data often need no preparation



But... you don't get to tell the machine which classifications to make

Supervised approaches



Humans prepare 'training' data – containing data + labels which describe the desired classification



For example

Image
recognition
Text
classification

-
- Image classification
 - Machines can be very good at this





**But can also
be easily
confused...**

- <https://www.freecodecamp.org/news/chihuahua-or-muffin-my-search-for-the-best-computer-vision-api-cbda4d6b425d/>

Despite this used e.g. to research stu

can be

The Nicotine Metabolite Ratio in Pregnancy Measured by trans-3'-Hydroxycotinine to Cotinine Ratio: Characteristics and Relationship With Smoking Cessation.

Vaz LR¹, Coleman T², Cooper S², Aveyard P³, Leonardi-Bee J⁴; SNAP trial team.

Author information

Abstract

INTRODUCTION: Nicotine replacement therapy (NRT) helps effective in pregnancy. As nicotine metabolism increases in hydroxycotinine to cotinine, the nicotine metabolite ratio (NM maternal characteristics and smoking cessation in pregnant

METHODS:

carbon mon regression in from smokin

RESULTS:

cigarette cor

Effect of nicotine patches in from the randomised, doub

Cooper S¹, Taggar J², Lewis S³, Marlow N⁴.

Collaborators (69)

Author information

Erratum in

Lancet Respir Med. 2014 Nov;2(11):e22.

Abstract

BACKGROUND: The SNAP (Smoking

Cost-Effectiveness of Nicotine Patches f Randomized Controlled Trial (SNAP).

Essex HN¹, Parrott S², Wu Q², Li J², Cooper S³, Coleman T³.

Author information

Abstract

INTRODUCTION: Smoking during pregnancy is the most in miscarriage, premature birth, and low birth weight with hug published economic evaluations of smoking cessation inter not present incremental cost-effectiveness ratios (ICER). A therapy (NRT) in the general population, but this has yet to

METHODS: A cost-effectiveness analysis was undertaken alongside the smoking, nicotine, and preg behavioral support to behavioral support alone, for pregnant women who smoked.

RESULTS: At delivery, biochemically verified quit rates were slightly higher at 9.4% in the NRT group (odds ratio = 1.26, 95% CI = 0.82-1.96), at an increased cost of around £90 per participant. Higher attributable to the cost of NRT patches (mean = £46.07). The incremental cost-effectiveness ratio as quitter and a sensitivity analysis including only singleton births yielded an ICER of £4,156 per quitter. indicated a high level of uncertainty.

CONCLUSIONS: Without a specific willingness to pay threshold, and due to high levels of statistical uncer cost-effectiveness of NRT in this population. Furthermore, future research should address compliance issues, as potential effects of NRT, thus reducing the cost-effectiveness.

Protocol for study of financial incentives for smoking cessation in pregnancy (FISCP): randomised, multicentre study.

Berlin N¹, Goldzahl L², Jusot F², Berlin J³.

Author information

Abstract

INTRODUCTION: Maternal smoking during pregnancy is associated with adverse perinatal and postnatal health outcomes. The efficacy of ated; therefore new interventions should be is one of the promising options.

Birth weight differences between those offered financial voucher incentives for verified smoking cessation and control participants enrolled employing an intuitive approach an

McConnachie A¹, Haig C¹, Sinclair L², Bauld L², Ter

Author information

Abstract

BACKGROUND: The Cessation in Pregna pregnancy showed a clinically and statisti This study re-examines birth weight using information missed by intention-to-treat an

METHODS: CPIT offered financial incentiv non-smokers at primary outcome, compar randomised groups were split into three th quit even with incentives and potential qui weight gain with incentives is attributable

RESULTS: Mean birth weight of potential potential quitters in the control group (who 617, +803). The mean difference in birth v who managed to quit was 14.3%. Since the all women in the intervention group. Howe identical result, causal birth weight increas

CONCLUSIONS: Policy makers have grea clinically insignificant improvement in aver pregnant smokers who want to stop but ca

TRIAL REGISTRATION: ISRCTN Registry

This approach has many advantages over rules-based approaches:

- Data can be generated much more efficiently – we don't need to create detailed rules
- Data generated for other purposes can be reused
- The machine learning makes use of ALL the information in the abstract
 - This helps the model to generalise better than rule-based approaches
 - But can be a drawback...

Birth weight differences between those offered financial voucher incentives for verified smoking cessation and control participants enrolled in the **Cessation in Pregnancy Incentives Trial (CPIT), employing an intuitive approach and a Complier Average **Causal Effects** (CACE) analysis.**

McConnachie A¹, Haig C¹, Sinclair L², Bauld L², Tappin DM³.

⊕ Author information

Abstract

BACKGROUND: The Cessation in **Pregnancy** Incentives Trial (CPIT) which offered financial incentives for smoking cessation during pregnancy showed a clinically and statistically significant increase in birth weight. This study re-examines birth weight using an intuitive approach and a Complier Average Causal Effects (CACE) analysis to account for information missed by intention-to-treat analysis.

METHODS: CPIT offered financial incentives up to £400 to non-smokers at primary outcome, compared to 25 (8.7%) **randomised** groups were split into three theoretical sub-groups: potential quitters - required to quit even with incentives and potential quitters - required to quit even with incentives and potential quitters - required to quit even with incentives. Weight gain with incentives is attributable only to potential quitters.

RESULTS: **Mean** birth weight of potential quitters in the intervention group was 3617 g (SD 403) compared to 3546 g (SD 403) in potential quitters in the control group (who did not quit) (SD 403). The mean difference in birth weight between potential quitters who managed to quit was 14.3%. Since the intervention was offered to all women in the intervention group. However, "complier average causal effects" analysis showed an identical result, causal birth weight increase 21 g ÷ 0.14%.

CONCLUSIONS: Policy makers have great difficulty giving incentives to pregnant smokers who want to stop but cannot achieve so. This study shows a clinically insignificant improvement in average birth weight for potential quitters in the intervention group compared to potential quitters in the control group.

TRIAL REGISTRATION: ISRCTN Registry, ISRCTN87508788.

This means that:

- ALL of the text in the document can be used to 'learn' the classifications
- This increases the model's resilience to minor variations in wording that would break a rules-based system
- The disadvantage is that if you wanted to classify e.g. smoking cessation among young people, you'd need to ensure that the training data also covered young people – or performance would drop



**Good
supervision
is
required...**



Very dependent on quality
and coverage of training
data



Performance very
dependent on context



For example...

Study classification is a powerful tool



But very dependent on quality and coverage of training data



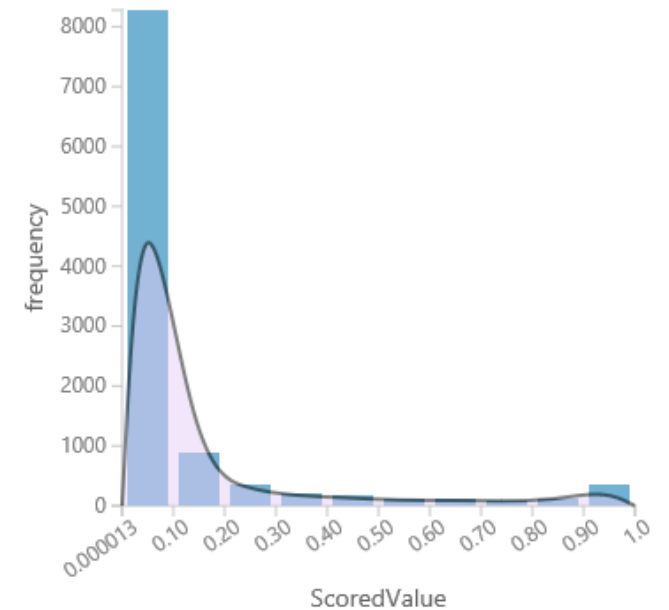
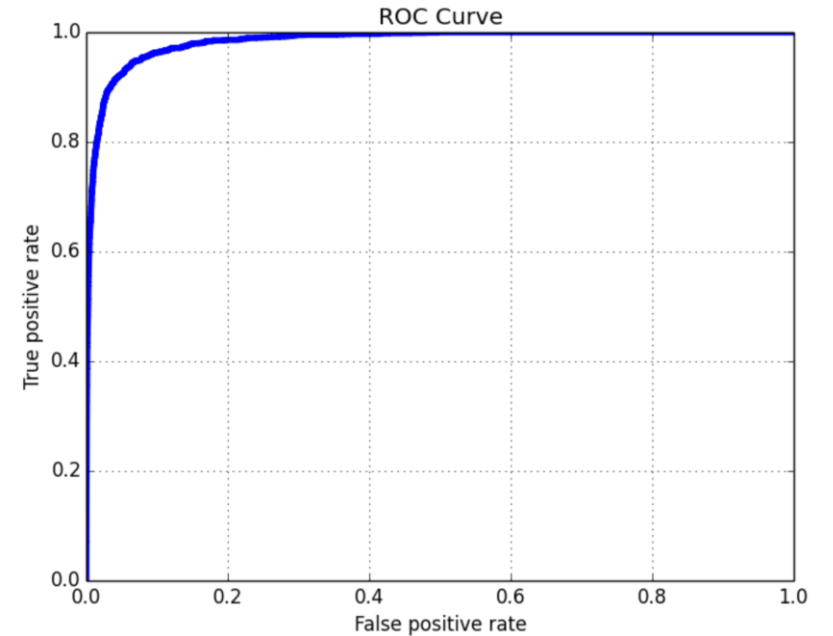
Performance dependent on context (e.g. Cochrane RCT classifier no good for education RCTs)



Creating high-quality training data can be expensive

Example of study classification: RCT Classifier

- A classifier was built using more than 280,000 records from Cochrane Crowd
- It is 'simply' applying single classification (RCT / not RCT)
- It has been calibrated to achieve a recall = 99% on the McMaster 'Hedges' dataset
 - Calibration = ranking the 'test' dataset by score
 - BUT precision is low
- It is very accurate!
 - But not all supervised learning can be so accurate, as lots of high-quality training data are needed



Priority screening

- Has received most R&D attention
- Diverse evidence base; difficult to compare evaluations
- ‘semi-automated’ approaches are the most common
- Possible reductions in workload in excess of 30% (and up to 97%)

Summary of conclusions

- Screening prioritisation
= Safe to use
- Machine as a ‘second screener’
= Use with care
- Automatic study exclusion
= Highly promising in many areas, but performance varies depending on the domain of literature being screened

O'Mara-Eves et al. *Systematic Reviews* 2015, 4:5
<http://www.systematicreviewsjournal.com/content/4/1/5>

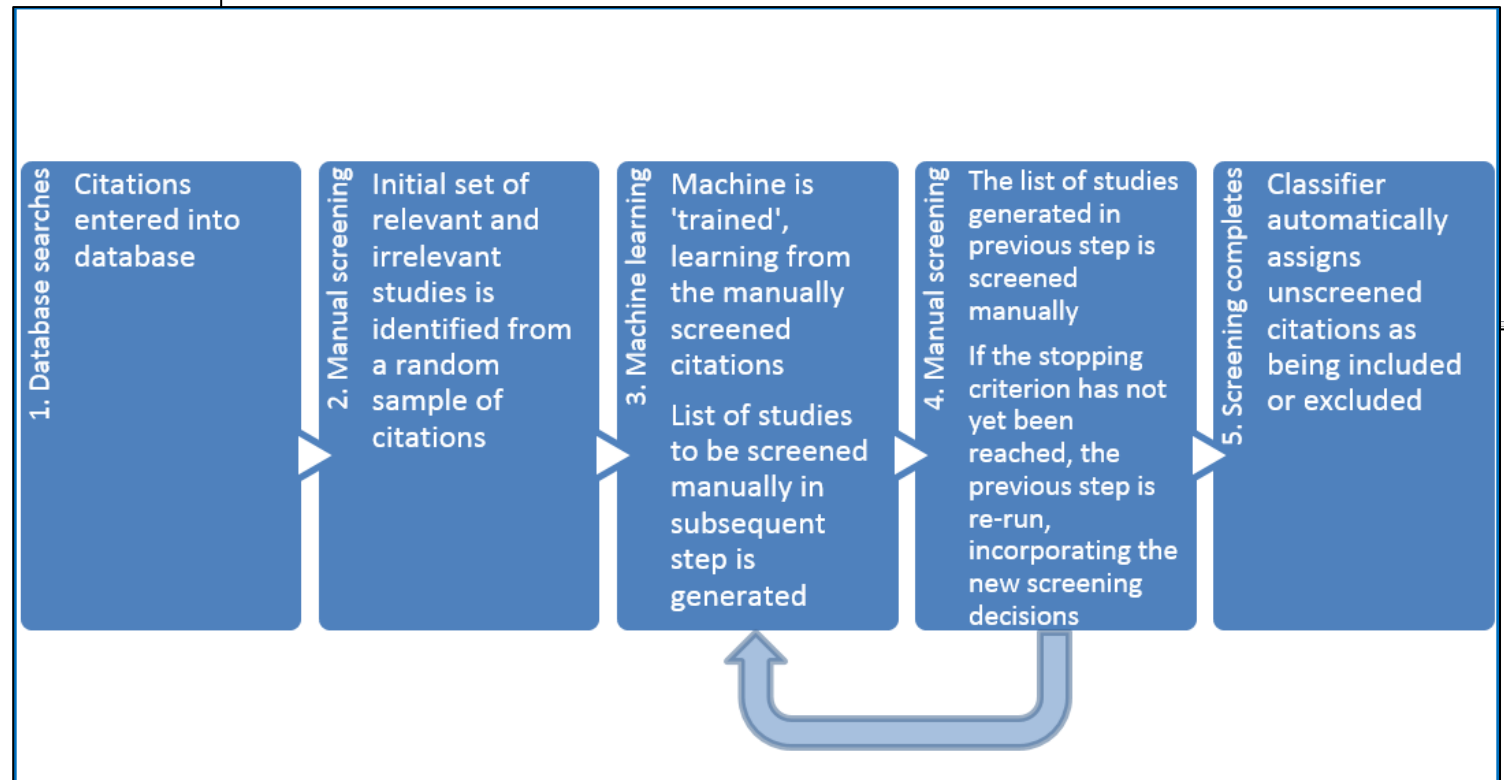


RESEARCH

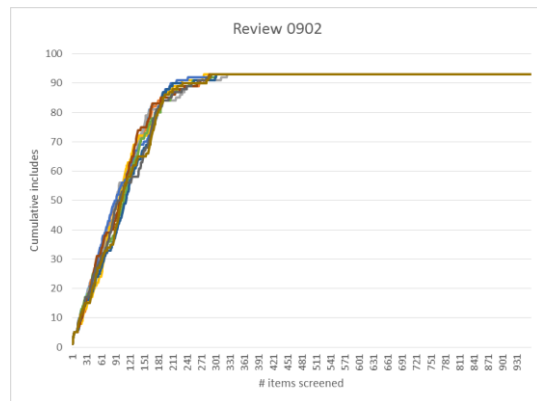
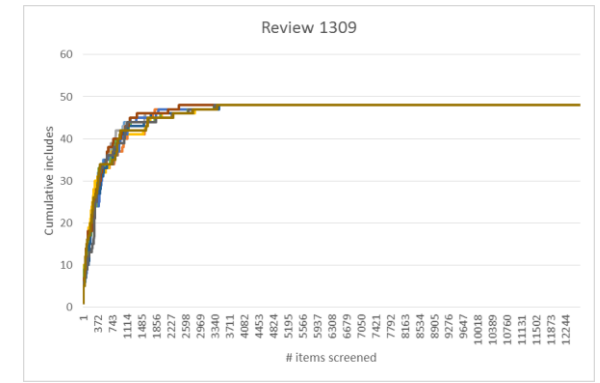
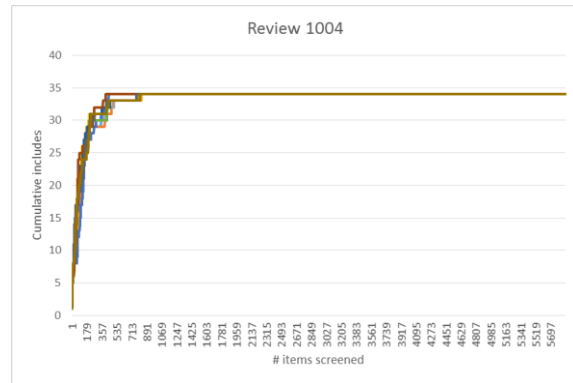
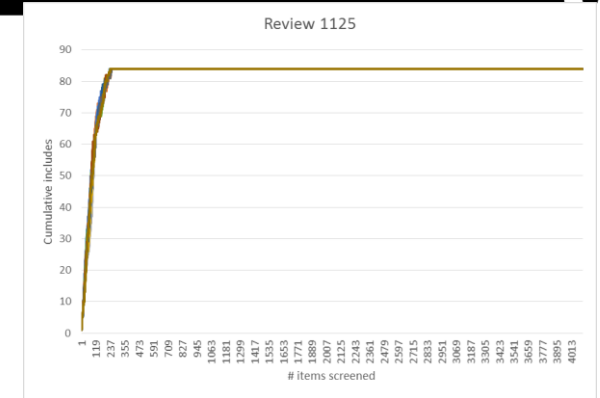
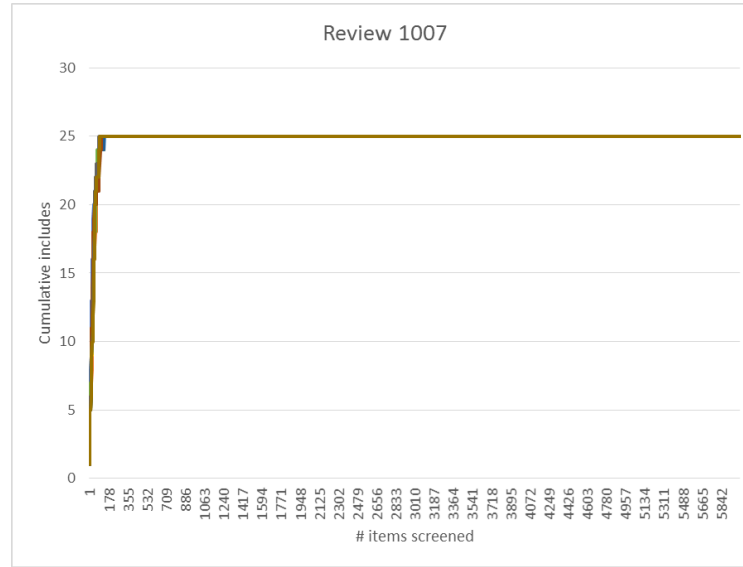
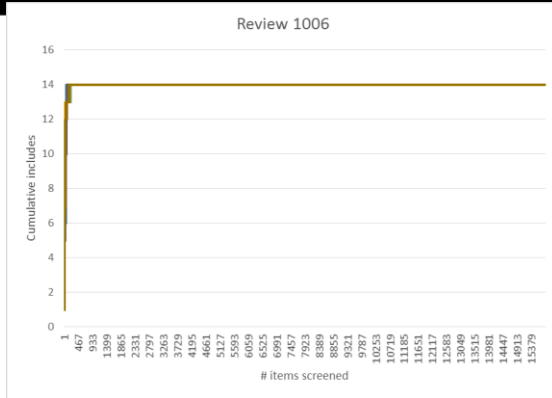
Open Access

Using text mining for study identification in systematic reviews: a systematic review of current approaches

Alison O'Mara-Eves¹, James Thomas^{1*}, John McNaught², Makoto Miwa³ and Sophia Ananiadou²



A validated stopping algorithm is needed to make best use of this technology

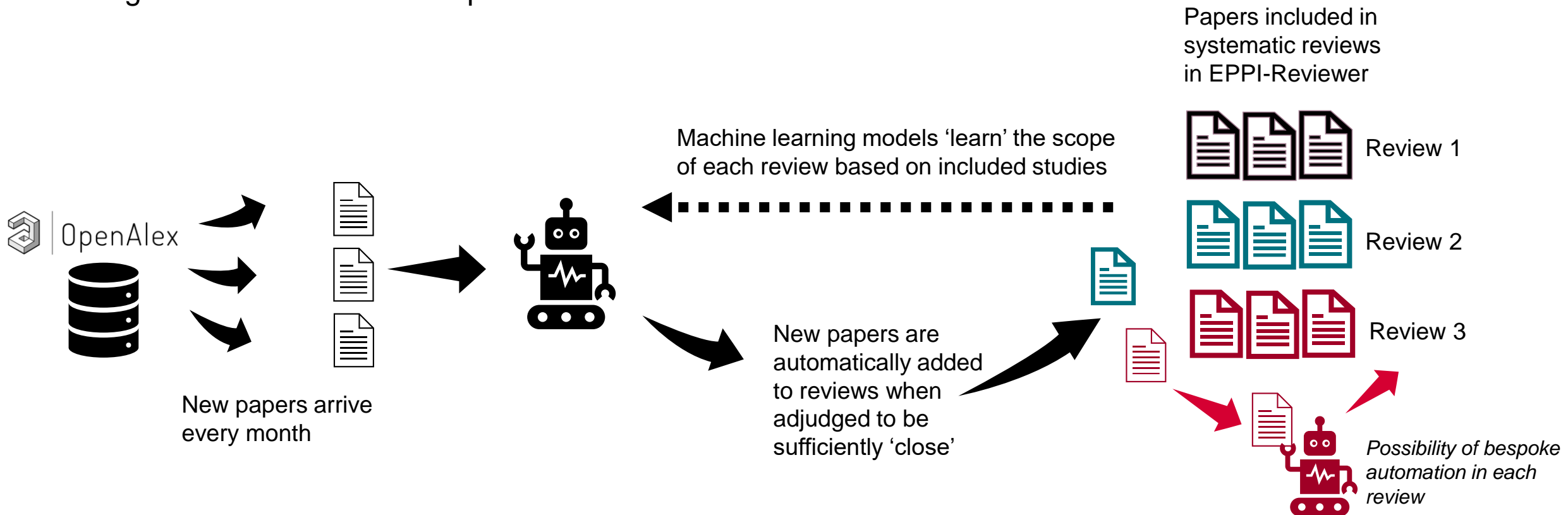


Does it work? e.g. reviews from Cochrane Heart Group

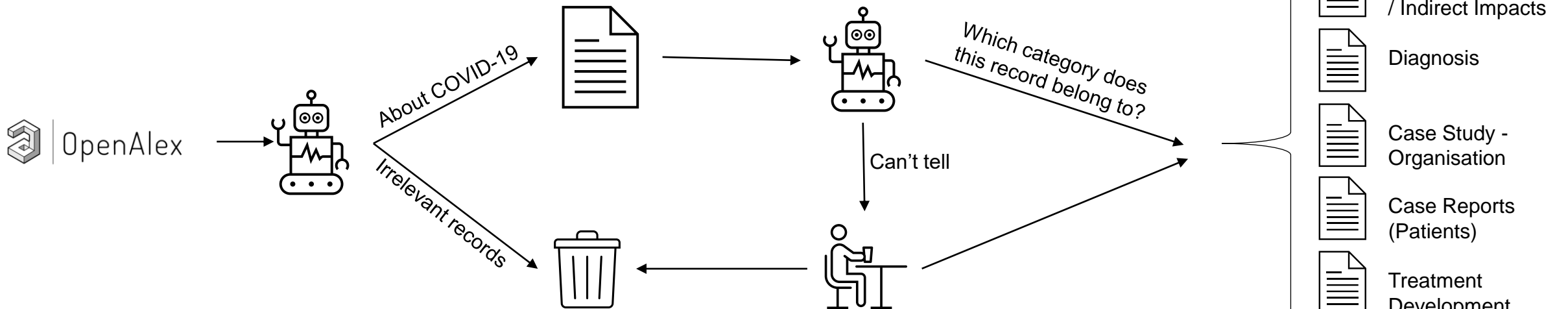
BUT when is it safe to stop..?

Continuous update of reviews in EPPI-Reviewer

Maintains a 'surveillance' of the literature as it emerges to maintain reviews up to date



For example... full workflow in our map of COVID-19 research



- Treatment Evaluation
- Genetics / Biology
- Transmission / Risk / Prevalence
- Social / Economic / Indirect Impacts
- Diagnosis
- Case Study - Organisation
- Case Reports (Patients)
- Treatment Development
- Mental Health Impacts
- Vaccine Development
- Long COVID

Human judgement required when machine is 'unsure'

From our initial purely manual workflow, we have now moved to a position where almost all of the work is carried out by automation tools and technologies

Wellcome Open Research | Wellcome Open Research 2021, 6:210 Last updated: 27 JUN 2022

RESEARCH ARTICLE

[Check for updates](#)

Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research [version 1; peer review: 2 approved with reservations]

Ian Shemilt^{1*}, Anneliese Arno^{1*}, James Thomas^{1*}, Theo Lorenc², Claire Khouja², Gary Raine², Katy Sutcliffe¹, D'Souza Preethy¹, Irene Kwan¹, Kath Wright², Amanda Sowden²

¹EPPI-Centre, UCL Social Research Institute, University College London, London, London, WC1H 0NR, UK
²Centre for Reviews and Dissemination, University of York, UK, York, Yorkshire, UK

* Equal contributors

V1 First published: 19 Aug 2021, 6:210

Open Peer Review

Supervised approaches



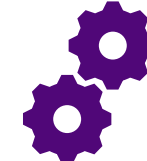
Using pre-built machine learning classifiers (e.g. RCT Classifier)



Building bespoke machine learning classifiers



Using 'priority screening' to rank, and re-rank, records for screening



Using pre-built machine learning algorithms to assess risk of bias

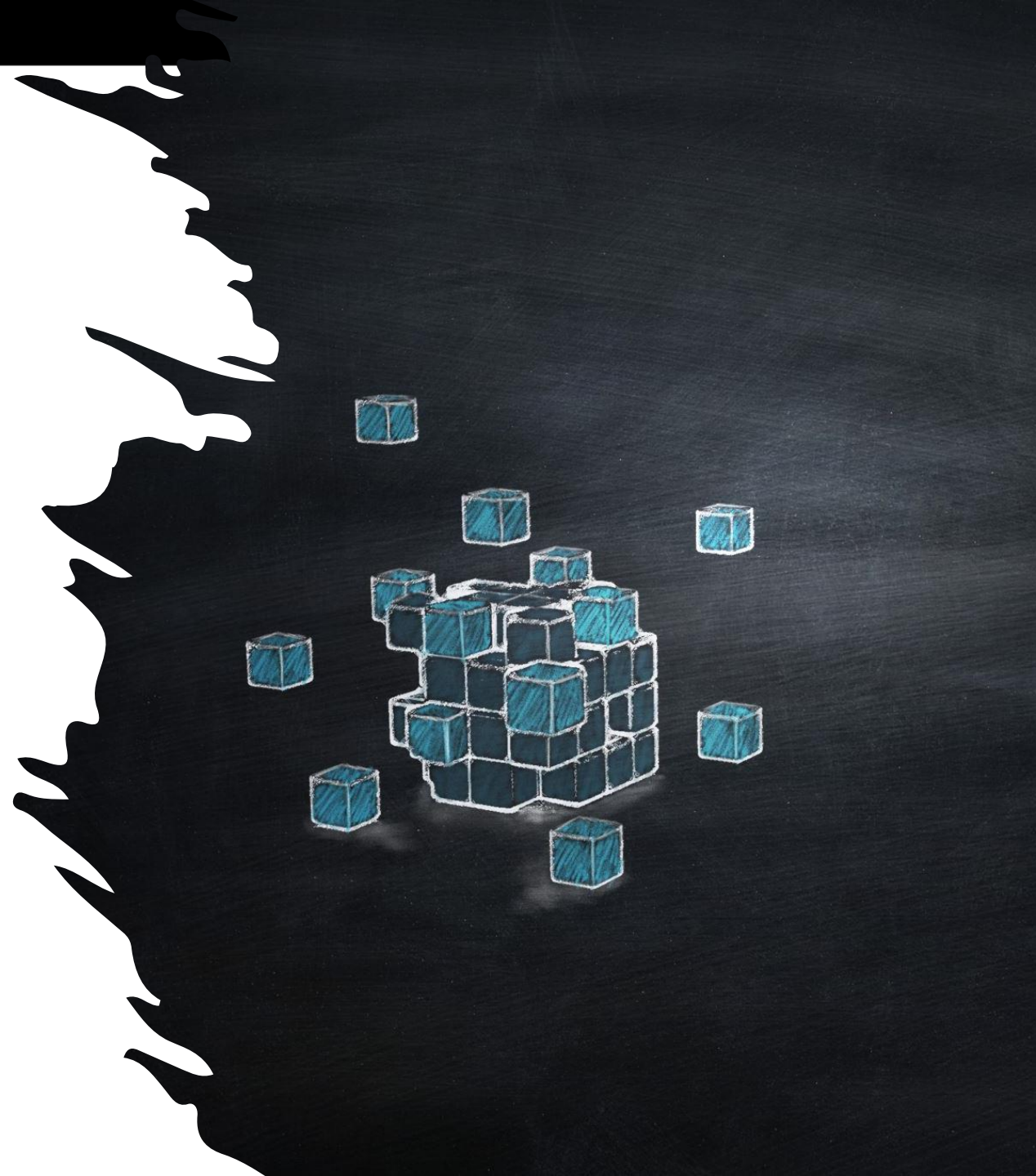


old hat?

Slang = Old-fashioned or out-of-date

New approaches: more contextually 'aware' classification

- The theory:
 - When a human reads, they read in the light of their pre-existing knowledge
 - The previous examples do not do that
 - Is it possible to address this using machine learning?
- Word embeddings
 - E.g. Word2Vec
- Transformer models
 - E.g. BERT (Bidirectional Encoder Representations from Transformers)
 - LARGE 'generative' language models
- Key to bear in mind: these are all (sophisticated) statistical representations of words / phrases that tend to 'go together'



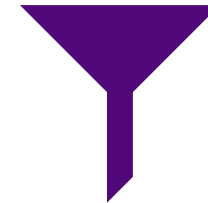
Generative approaches



ChatGPT (or other LLM chatbot)



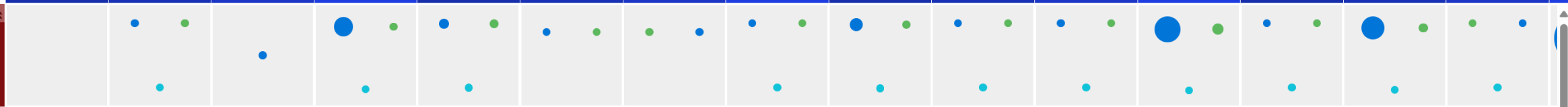
LLM-based database querying and summarisation



LLM-based information extraction

Explosion of work on generative AI

- There are numerous tools being developed using Generative LLMs ('Gen AI')
- Articles have started to be published that report testing out Gen AI in systematic reviews (& expect exponential growth)
- So far there is lots of potential, but no validated tools
- There is an urgent need for robust evaluation to inform deployment and future development of these tools



- As an experiment, we mapped the literature on AI and equity – using AI
- 26,000 records identified through conventional searches
- 6,228 records included in the map
- GPT4 used for screening and mapping
- Took a few days to screen & code
- Cost £800 in OpenAI API fees
- Evaluation found:
 - Sensitivity 95%; specificity 100% (screening)
 - Classification of records: 86% no errors; 12% minor errors; 2% major errors

In summary

Rule-based	Unsupervised	Supervised	Generative
<ul style="list-style-type: none">• Not fashionable• Potentially powerful• Very demanding in time• Rules can be fragile	<ul style="list-style-type: none">• Very little time effort required to create rules or training data• No control over classifications	<ul style="list-style-type: none">• Can utilise lots of training data which can be generated efficiently• Makes use of data created for other purposes• Does not break as easily as rule-based approaches• Can predict specific classification terms (unlike unsupervised)	<ul style="list-style-type: none">• Considered current 'state of the art'• Huge research focus• Sometimes beats simpler models (though sometimes only marginally)• Concerns about bias and other negative outcomes



Time for a break...

A futuristic blue robot arm is shown in a dynamic, hovering position. The arm is composed of various mechanical segments, including joints and fingers, all rendered in a metallic blue color. It is positioned above a glowing digital data stream that flows across the bottom of the frame. The data stream consists of numerous small, bright blue and white points of light, creating a sense of movement and information processing. The background is dark, with some blurred red and blue light trails, suggesting a high-tech or data center environment. The overall lighting is cool and futuristic, with a strong emphasis on blue tones.

AI Tools and how to evaluate them



Example presentation

- Google presentation
- https://www.youtube.com/watch?v=sPiOP_CB54A

**So...
what
did you
think?**



Can we all just go and use Gemini now and forget about how we 'used' to do systematic reviews?



What else might we want to know?

Important questions to ask of any machine learning system

- Where did the data come from?
- Are the data biased in some way?
- For supervised approaches:
 - Were there sufficient training data to build robust models?
 - How similar are the training data to my use scenario?
 - Was the evaluation internally valid?
- For all approaches:
 - How can I tell if the tool is fit for my purpose?

Starting points



Decisions that affect people's lives should be informed by reliable research



Individual research studies can be atypical; we need to draw on the sum of current knowledge

Therefore we use evidence synthesis



Evidence syntheses can be unreliable for two reasons:

They have been conducted badly
The research they contain is unreliable

Critical questions to ask when considering using a new tool for evidence synthesis



Does it enable me to draw on the sum of current knowledge?

Or does it present an incomplete or biased picture?



Does it enable me to distinguish between reliable and unreliable research?

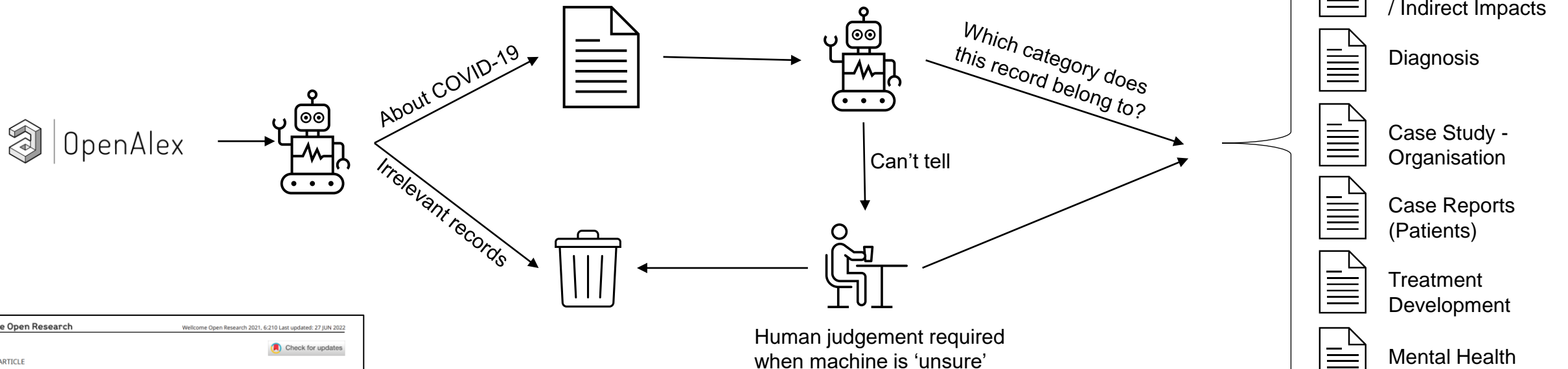
Or does it treat all research as equally reliable?



Is the tool itself reliable?

Do we know how it works?
Is it validated for my use case?

First example... workflow in our map of COVID-19 research



Wellcome Open Research

Wellcome Open Research 2021, 6:210 Last updated: 27 JUN 2022

Check for updates

RESEARCH ARTICLE

Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research [version 1; peer review: 2 approved with reservations]

Ian Shemilt^{1*}, Anneliese Arno^{1*}, James Thomas^{1*}, Theo Lorenc², Claire Khouja², Gary Raine², Katy Sutcliffe¹, D'Souza Preethy¹, Irene Kwan¹, Kath Wright², Amanda Sowden²

¹EPPI-Centre, UCL Social Research Institute, University College London, London, London, WC1H 0NR, UK
²Centre for Reviews and Dissemination, University of York, UK, York, Yorkshire, UK

* Equal contributors

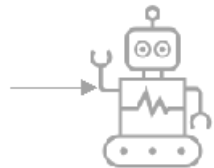
1 First published: 19 Aug 2021, 6:210

Open Peer Review

From our initial purely manual workflow, we have now moved to a position where almost all of the work is carried out by automation tools and technologies

For example... full workflow map of COVID-19 research

OpenAlex



Does it enable me to draw on the sum of current knowledge? ✓

Does it enable me to distinguish between reliable and unreliable research? ✓



Human judgement required when machine is 'unsure'

Wellcome Open Research

RESEARCH ARTICLE

Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research [version 1; peer review: 2 approved with reservations]

Ian Shemilt^{1*}, Anneliese Arno^{1*}, James Thomas^{1*}, Theo Lorenc², Claire Khouja², Gary Raine², Katy Sutcliffe¹, D'Souza Preethy¹, Irene Kwan¹, Kath Wright², Amanda Sowden²

¹EPPI-Centre, UCL Social Research Institute, University College London, London, London, WC1H 0NR, UK
²Centre for Reviews and Dissemination, University of York, UK, York, Yorkshire, UK

* Equal contributors

V1 First published: 19 Aug 2021, 6:210

Open Peer Review

From our initial purely manual workflow, we have now moved to a position where almost all of the work is carried out by automation tools and technologies

- Treatment Evaluation
- Case Reports (Patients)
- Treatment Development
- Mental Health Impacts
- Vaccine Development
- Long COVID

Why is this trustworthy?



Not too far from 'traditional' methods



Its dataset has been validated as being sufficiently comprehensive for this task



It uses machine learning, but in 'standard' ways: training data are used to build a model and a transformer language model is used, but not in a 'generative' way

But...

- Training data was needed
- The digital evidence synthesis tools were partly developed *for* the project
- The evidence synthesis team had technical development team working with them
- What about more generic and less tailored tools?





ChatGPT

JA

Are mindfulness-based interventions effective for smoking cessation among people who smoke?



Mindfulness-based interventions (MBIs) have shown some promise as a complementary approach to smoking cessation for people who smoke. While more research is needed to establish their long-term effectiveness and determine how they compare to other smoking cessation methods, there is evidence to suggest that MBIs can be helpful for certain individuals. Here are some key points to consider:



JA

Are mindfulness-based interventions effective for helping people to stop smoking?



Mindfulness-based interventions (MBIs) have shown promise as a new approach to smoking cessation for people who smoke. To establish their long-term effectiveness and determine how to best integrate them into existing cessation methods, there is evidence to suggest that MBIs can be helpful for certain individuals. Here are some key points to consider:

Does it enable me to draw on the sum of current knowledge?

Does it enable me to distinguish between reliable and unreliable research?

Acceptance-Based Treatment for Smoking Cessation

Prior works

Search... Expand

Origin paper

Acceptance-Based Treatment for Smoking Cessation
E. Gifford, B. Kohlenberg, S. Hayes, D. Antonuccio, M.... 2004

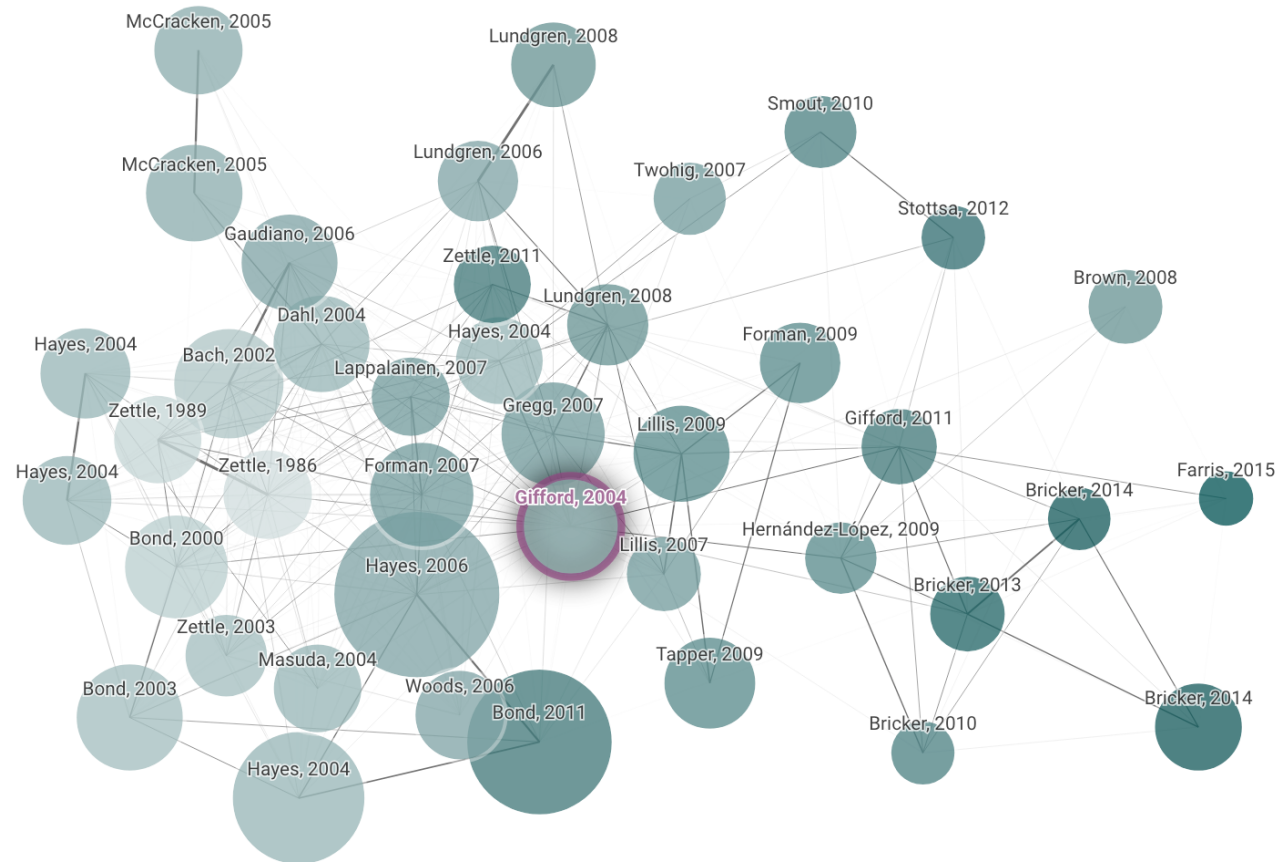
A Preliminary trial of twelve-step facilitation and acceptance and commitment therapy with...
S. Hayes, K. Wilson, E. Gifford, R. Bissett, M. Piasecki, ... 2004

Acceptance and commitment therapy for smoking cessation: a preliminary study of its effectiveness i...
Mónica Hernández-López, M. C. Luciano, J. Bricker, J.... 2009

Improving diabetes self-management through acceptance, mindfulness, and values: a randomize...
J. Gregg, G. M. Callaghan, S. Hayes, June L. Glenn-... 2007

Does acceptance and relationship focused behavior therapy contribute to bupropion outcomes? A...
E. Gifford, B. Kohlenberg, S. Hayes, Heather M Pierson... 2011

Acceptance and commitment therapy and the treatment of persons at risk for long-term disability...
J. Dahl, K. Wilson, A. Nilsson 2004



Acceptance-Based Treatment for Smoking Cessation

Search... Expand

Origin paper

Acceptance-Based Treatment for Smoking Cessation
E. Gifford, B. Kohlenberg, S. Hayes, D. Antonuccio, M.... 2004

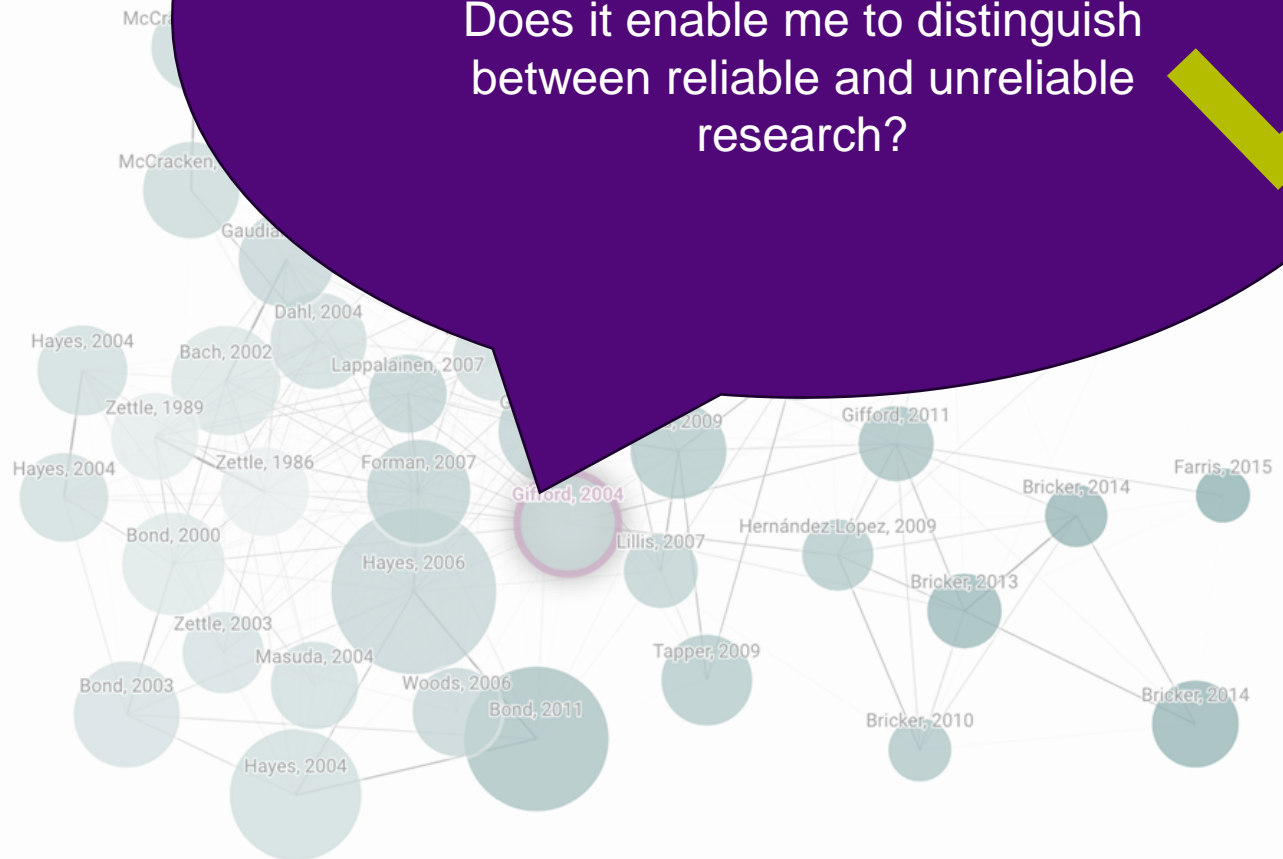
A Preliminary trial of twelve-step facilitation and acceptance and commitment therapy with...
S. Hayes, K. Wilson, E. Gifford, R. Bissett, M. Piasecki, ... 2004

Acceptance and commitment therapy for smoking cessation: a preliminary study of its effectiveness i...
Mónica Hernández-López, M. C. Luciano, J. Bricker, J.... 2009

Improving diabetes self-management through acceptance, mindfulness, and values: a randomize...
J. Gregg, G. M. Callaghan, S. Hayes, June L. Glenn-... 2007

Does acceptance and relationship focused behavior therapy contribute to bupropion outcomes? A...
E. Gifford, B. Kohlenberg, S. Hayes, Heather M Pierson... 2011

Acceptance and commitment therapy and the treatment of persons at risk for long-term disability...
J. Dahl, K. Wilson, A. Nilsson 2004



Does it enable me to draw on the sum of current knowledge?

Does it enable me to distinguish between reliable and unreliable research?



Welcome to Elicit, your AI research assistant

Are mindfulness-based interventions effective for smoking cessation among people who smoke?

- Input search query
- Searching for papers
- Summarizing 8 abstracts
- Final answer



Summary of top 8 papers

The research on mindfulness-based interventions for smoking cessation is promising but mixed. Several studies found that mindfulness training helps smokers quit or reduce smoking. [Vinci 2020](#) found that both cognitive behavioral therapy and mindfulness-based interventions are effective for smoking cessation, especially for certain populations. [Oikonomou 2017](#) conducted a meta-analysis of four RCTs and found that 25.2% of smokers who received mindfulness training quit smoking for over 4 months, compared to only 13.6% of those who received usual care. [Davis 2007](#) found that 56% of smokers who received mindfulness training quit smoking for 6 weeks. [Bowen 2009](#) found that smokers who received brief mindfulness training smoked fewer cigarettes over a week compared to controls. However, other studies found limited or no effect of mindfulness on smoking cessation. [Carim-Todd 2013](#) reviewed 14 studies and found that while the results were promising, the studies had too many limitations and methodological problems to draw a definitive conclusion. [Maglione 2017](#) conducted a meta-analysis of 5 RCTs and found no significant effect of mindfulness meditation on smoking abstinence or number of cigarettes smoked compared to controls. The studies were too heterogeneous and low quality to find an effect. [Garrison 2015](#) proposes an RCT to evaluate a smartphone-based mindfulness intervention for smoking cessation, indicating the research is still ongoing. In summary, while several initial studies found promising effects of mindfulness on smoking cessation and reduction, the research is limited by a small number of studies, methodological weaknesses, and heterogeneity across interventions and measures. Higher quality, larger RCTs that evaluate specific types of mindfulness interventions are still needed to determine if and how mindfulness effectively helps people quit smoking.

Welcome to Elicit, your AI research assistant

Are mindfulness-based interventions effective for smoking cessation?

- Input search query
- Searching for papers
- Summarizing 8 abstracts
- Final answer

Summary of top 8 papers

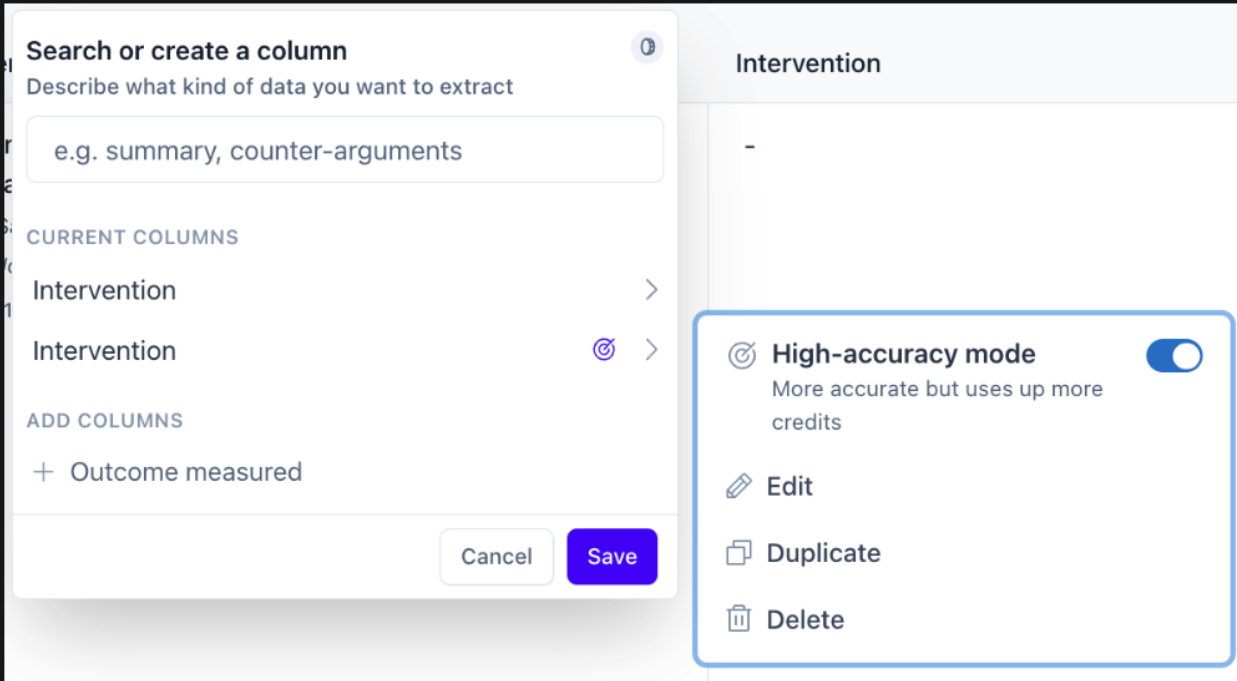
The research on mindfulness-based interventions for smoking cessation is mixed. Mindfulness training helps smokers quit or reduce smoking. Some studies show that mindfulness-based interventions are effective for smoking cessation. A meta-analysis of four RCTs and found that 25.2% of smokers who received mindfulness-based interventions quit smoking for 6 weeks compared to only 13.6% of those who received usual care. [Davis 2007](#) found that smokers who received brief mindfulness training over a week compared to controls. However, other studies found limited or no effect of mindfulness on smoking cessation. [Carim-Todd 2013](#) reviewed 14 studies and found that while the results were promising, the studies had many limitations and methodological problems to draw a definitive conclusion. [Maglione 2017](#) conducted a meta-analysis of 5 RCTs and found no significant effect of mindfulness meditation on smoking abstinence or number of cigarettes smoked compared to controls. The studies were too heterogeneous and low quality to find an effect. [Garrison 2015](#) proposes an RCT to evaluate a smartphone-based mindfulness intervention for smoking cessation, indicating the research is still ongoing. In summary, while several initial studies found promising effects of mindfulness on smoking cessation and reduction, the research is limited by a small number of studies, methodological weaknesses, and heterogeneity across interventions and measures. Higher quality, larger RCTs that evaluate specific types of mindfulness interventions are still needed to determine if and how mindfulness effectively helps people quit smoking.

Does it enable me to draw on the sum of current knowledge?

Does it enable me to distinguish between reliable and unreliable research?

What is high-accuracy mode?

High-accuracy mode gives better results when adding columns and extracting data. In our testing, high-accuracy mode had about 1/2 the error rate of standard columns. High-accuracy mode is particularly useful for conducting systematic reviews and meta-analyses.



High-accuracy mode is only available to Elicit Plus subscribers, and costs about 250 credits per answer.

Learn more about high-accuracy mode [here](#).

Improvements

As of today, we're using a new technique for high-accuracy mode. Our testing found that our new technique reduces the error rate by about 8% compared to our old technique.


- Apparently Elicit can be used in 'high accuracy mode' for systematic reviews and meta-analyses
- Apparently the error rate is reduced by 8% compared with... something else
- Elicit is not alone in providing LLM-based tools with no evaluations to support their use
- This is becoming the industry norm


Mindfulness for smoking cessation

✉ [Sarah Jackson](#), [Jamie Brown](#), [Emma Norris](#), [Jonathan Livingstone-Banks](#),
Authors' declarations of interest

Version published: 14 April 2022 [Version history](#)

<https://doi.org/10.1002/14651858.CD013696.pub2> 

Does it enable me to draw on the sum
of current knowledge? 

Does it enable me to distinguish
between reliable and unreliable
research? 

Can mindfulness help people to stop smoking?

Key messages

- There is currently no clear evidence that mindfulness-based treatments help people to stop smoking or improve their mental health and well-being.
- However, our confidence in the evidence is low or very low, and further evidence is likely to change our conclusions.

Data (information) extraction

- Earlier language models lacked precision & limited context 'window'
- Newer models have larger windows and offer impressive early results
- E.g. Claude2, published by Anthropic

The screenshot shows the Anthropic website homepage. At the top left is the 'ANTHROPIC' logo. To the right are navigation links for 'Product', 'Research', 'Company', 'News', and 'Careers'. The main headline reads 'AI research and products that put safety at the frontier'. Below this are two featured sections. The left section is titled 'NEW' and 'Claude in Beta, now available!', with a subtext 'Your friendly assistant. Fast, capable, and truly conversational.' and a 'Talk to Claude' button. The right section is titled 'ENTERPRISE' and 'Build with Claude', with a subtext 'Start using Claude and unlock business value with AI.' and a 'Submit business interest' button.

ANTHROPIC

Product Research Company News Careers

AI research and products that put safety at the frontier

NEW

Claude in Beta, now available!

Your friendly assistant. Fast, capable, and truly conversational.

Talk to Claude

ENTERPRISE

Build with Claude

Start using Claude and unlock business value with AI.

Submit business interest

Here is a report of a randomized trial. State: the population recruited; the intervention(s) evaluated; the main outcome(s) measured; and the statistical results of the primary outcome

JT

According to the paper, this randomized controlled trial had the following key details:

Population:

- 4613 participants aged 18+ years who smoked daily and were willing to make a serious quit attempt in the UK.
- Recruited online, mainly via the English Department of Health website.
- Stratified into low socioeconomic status (n=2142) and high socioeconomic status (n=2471) based on occupation.

Interventions:

- Interactive internet-based smoking cessation website called StopAdvisor.
- Control group received a 1-page static website with brief standard smoking cessation advice.

Main outcome:

- Primary outcome was biochemically verified 6-month sustained abstinence from smoking (self-report of smoking ≤ 5 cigarettes in the past 6 months plus saliva cotinine < 15 ng/mL).

Results:

- No overall difference between groups in 6-month abstinence (StopAdvisor 10% vs control 10%, RR 1.06, 95% CI 0.89-1.27, $p=0.49$).
- In the low SES subgroup, 6-month abstinence was higher in StopAdvisor vs control (8% vs 6%, RR 1.36, 95% CI 1.00-1.86, $p=0.0499$).
- No difference between groups in the high SES subgroup (StopAdvisor 12% vs control 13%, RR 0.95, 95% CI 0.77-1.17, $p=0.61$).

Internet-based intervention for smoking cessation (StopAdvisor) in people with low and high socioeconomic status: a randomised controlled trial

Jamie Brown, Susan Michie, Adam WA Geraghty, Lucy Yardley, Benjamin Gardner, Lion Shahab, John A Stapleton, Robert West

Summary

Background Internet-based interventions for smoking cessation could help millions of people stop smoking at very low unit costs; however, long-term biochemically verified evidence is scarce and such interventions might be less effective for smokers with low socioeconomic status than for those with high status because of lower online literacy to engage with websites. We aimed to assess a new interactive internet-based intervention (StopAdvisor) for smoking cessation that was designed with particular attention directed to people with low socioeconomic status.

Methods We did this online randomised controlled trial between Dec 6, 2011, and Oct 11, 2013, in the UK. Participants aged 18 years and older who smoked every day were randomly assigned (1:1) to receive treatment with StopAdvisor or an information-only website. Randomisation was automated with an unseen random number function embedded in the website to establish which treatment was revealed after the online baseline assessment. Recruitment continued until the required sample size had been achieved from both high and low socioeconomic status subpopulations. Participants, and researchers who obtained data and did laboratory analyses, were masked to treatment allocation. The primary outcome was 6 month sustained, biochemically verified abstinence. The main secondary outcome was 6 month, 7 day biochemically verified point prevalence. Analysis was by intention to treat. Homogeneity of intervention effect across the socioeconomic subsamples was first assessed to establish whether overall or separate subsample analyses were appropriate. The study is registered as an International Standard Randomised Controlled Trial, number ISRCTN99820519.

Findings We randomly assigned 4613 participants to the StopAdvisor group (n=2321) or the control group (n=2292); 2142 participants were of low socioeconomic status and 2471 participants were of high status. The overall rate of smoking cessation was similar between participants in the StopAdvisor and control groups for the primary (237 [10%] vs 220 [10%] participants; relative risk [RR] 1.06, 95% CI 0.89-1.27; $p=0.49$) and the secondary (358 [15%] vs 332 [15%] participants; 1.06, 0.93-1.22; $p=0.37$) outcomes; however, the intervention effect differed across socioeconomic status subsamples (1.44, 0.99-2.09; $p=0.0562$ and 1.37, 1.02-1.84; $p=0.0360$, respectively). StopAdvisor helped participants with low socioeconomic status stop smoking compared with the information-only website (primary outcome: 90 [8%] of 1088 vs 64 [6%] of 1054 participants; RR 1.36, 95% CI 1.00-1.86; $p=0.0499$; secondary outcome: 136 [13%] vs 100 [10%] participants; 1.32, 1.03-1.68, $p=0.0267$), but did not improve cessation rates in those with high socioeconomic status (147 [12%] of 1233 vs 156 [13%] of 1238 participants; 0.95, 0.77-1.17; $p=0.61$ and 222 [18%] vs 232 [19%] participants; 0.96, 0.81-1.13, $p=0.64$, respectively).



Lancet Respir Med 2014

Published Online
September 25, 2014
[http://dx.doi.org/10.1016/S2213-2600\(14\)70195-X](http://dx.doi.org/10.1016/S2213-2600(14)70195-X)

See Online/Comment
[http://dx.doi.org/10.1016/S2213-2600\(14\)70214-0](http://dx.doi.org/10.1016/S2213-2600(14)70214-0)

Cancer Research UK Health Behaviour Research Centre, Department of Epidemiology and Public Health (J Brown PhD, B Gardner DPhil, L Shahab PhD, Prof R West PhD) and Department of Clinical, Educational, and Health Psychology (Prof S Michie DPhil), University College London, London, UK; National Centre for Smoking Cessation and Training, London, UK (Prof S Michie, Prof R West); Primary Care and Population Sciences (A W A Geraghty PhD) and School of Psychology (Prof L Yardley PhD), University of Southampton, Southampton, UK; Addictions Department, Institute of Psychiatry, Kings College London, London, UK (J A Stapleton MSc)
Correspondence to: Dr Jamie Brown, Health

The overall rate of smoking cessation was similar between participants in the StopAdvisor and control groups for both the primary (237 [10%] vs 220 [10%] participants; relative risk [RR] 1.06, 95% CI 0.89-1.27; p=0.49) and the secondary (358 [15%] vs 332 [15%] participants; 1.06, 0.93-1.22; p=0.37) outcomes. However, analysis of the interaction between intervention and socioeconomic status showed clear evidence of non-ignorable heterogeneity of intervention effect by both primary (RR 1.44, 95% CI 0.99-2.09; p=0.0562) and secondary (1.37, 1.02-1.84; p=0.0360) cessation measures. This finding was evident before and after adjustment for all other baseline characteristics (adjusted data not shown). Consequently, the analysis of outcome was done separately within each of the two socioeconomic status subsamples.

In the subsample of participants with low socioeconomic status, a benefit of StopAdvisor was evident for both primary and secondary measures compared with the information-only website, whereas in those with high socioeconomic status, no evidence of a difference was shown (table 2). Adjustment for all baseline characteristics had a negligible effect on these comparisons (table 2). In a post-hoc sensitivity analysis, we re-examined the effect of StopAdvisor on biochemically verified smoking cessation in participants with low socioeconomic status, after exclusion of those in full-time education from the classification (n=237). The benefit of StopAdvisor compared with the information-only website remained in both unadjusted (primary outcome 87 [9%] of 967 vs 60 [6%] of 938 participants; RR 1.41, 95% CI 1.03-1.93; p=0.0346; secondary

who did not have post-16 education. In this smaller subsample (n=1687), the results were consistent with the primary analyses for StopAdvisor versus information only, but were non-significant in both the unadjusted (primary outcome 818 participants; RR 1.21, 0.93-1.58; p=1.27, 0.92-1.75; secondary outcome 1.21, 0.93-1.58; p=0.17) analysis, we re-examined self-reported rather than biochemically verified smoking cessation on the basis of similar verification criteria. New analyses showed no difference reported in table 2 (primary outcome 141 [13%] of 1088 compared with information only; unadjusted RR 1.20, RR 1.23, 0.97-1.56; prevalence 227 [21%] vs 200 [18%]; unadjusted RR 1.13, 0.95-1.34; p=0.07). That statistic failed to reach significance in participants with high socioeconomic status. decreased power between the two groups.

State, in tabular form, the outcomes of the study

Here is a table summarizing the key outcomes of the study:

	StopAdvisor	Control	Relative Risk (95% CI)	p-value
Overall				
6-month abstinence	10%	10%	1.06 (0.89-1.27)	0.49
Low SES				
6-month abstinence	8%	6%	1.36 (1.00-1.86)	0.0499
High SES				
6-month abstinence	12%	13%	0.95 (0.77-1.17)	0.61

Copy

	StopAdvisor	Control	Relative risk (95% CI)	Odds ratio (95% CI)*	Percentage-point difference (95% CI)	p value†
Primary outcome (abstinence for 6 months)						
High SES	147/1233 (12%)	156/1238 (13%)	0.95 (0.77 to 1.17)	0.94 (0.74 to 1.19)	-0.68 (-3.27 to 1.91)	0.61
Adjusted	0.97 (0.78 to 1.19)‡	0.95 (0.75 to 1.22)‡	..	0.75
Low SES	90/1088 (8%)	64/1054 (6%)	1.36 (1.00 to 1.86)	1.39 (1.00 to 1.94)	2.20 (0.02 to 4.38)	0.0499
Adjusted	1.43 (1.05 to 1.96)‡	1.46 (1.04 to 2.05)‡	..	0.0238
Secondary outcome (point prevalence of smoking cessation)						

The study authors did not report a risk of bias assessment using the Cochrane RoB tool. However, here are some observations about potential risks of bias based on the details provided in the paper:

- Random sequence generation: Participants were randomized using an automated unseen random number function, so this should have been adequate.
- Allocation concealment: The randomization was automated with no experimenter involvement, so allocation was likely adequately concealed.
- Blinding of participants/personnel: Not reported, but blinding of participants may not have been feasible given the difference in intervention. Personnel were likely not blinded.
- Blinding of outcome assessment: Outcome assessors were reported to be blinded to treatment allocation.
- Incomplete outcome data: Follow-up rates were 77% overall and similar between groups, with missing data treated as smoking, so the risk appears low.
- Selective reporting: The outcomes specified in the protocol appear to be reported, suggesting low risk of bias.
- Other biases: The study appears to have been adequately powered and analyzed as pre-specified, with low risks of other biases.

So in summary, the lack of blinding of participants and personnel, as well as the lack of formal risk of bias assessment by the authors, suggest some uncertainties but overall the study seems reasonably well conducted.

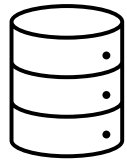
Technologies behind the tools



Is a language model,
not a database



Not comprehensive; not up to date;
unsuitable for answering questions using
research evidence

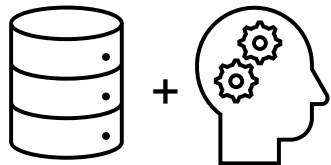


ConnectedPapers

A database building on *Open
Access* data



Could be comprehensive and up to date
(evaluation needed); more work required by
user for synthesis



Elicit, EPPI Reviewer,...

Database + language model +
machine learning ('RAG')



Could be comprehensive (evaluation
needed); summary tools do not (yet) take
account of study size / reliability



Using a large language model for
information (data) extraction

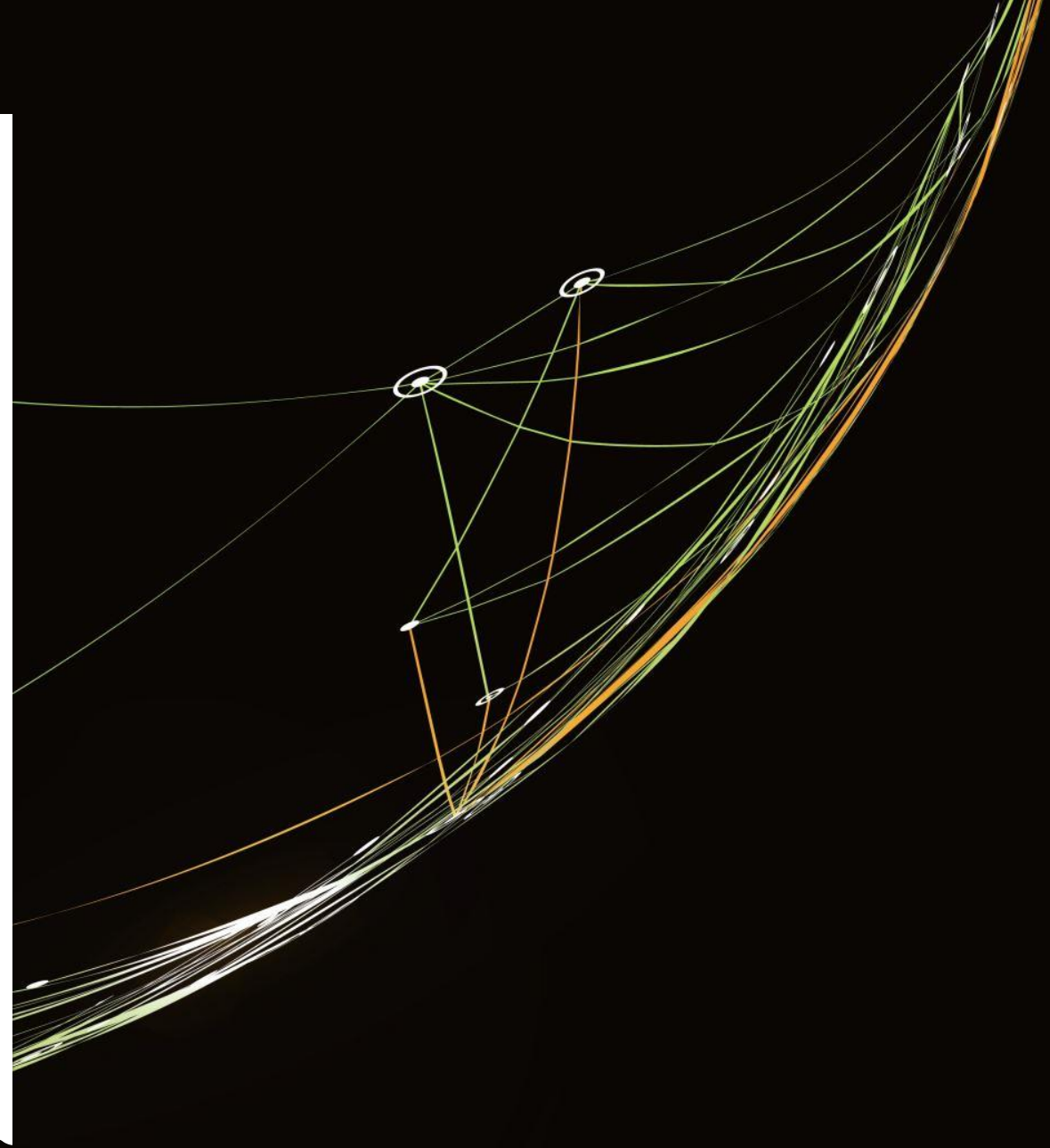


Constraining LLM to 'look' only at the
document looks promising. Key is to limit
possibility for 'hallucinations'. (More
research needed)

Claude 2 / ChatGPT

Critical points for internally valid evaluation

- Training and evaluation data *must* be as similar as possible to the data that the tool will be used on
- Evaluation data must *never* be used for training
 - This includes developing 'prompts' for LLMs
- Always check that a tool works in the specific review context that you intend to use it for





Now it's your turn!

To try a tool (and evaluate it)



Ideas...

1. Try out a prebuilt classifier:
 - a. RCT Classifier
 - b. Systematic reviews
 - c. Economic evaluations
2. RobotReviewer for assessing Risk of Bias of RCTs
3. Compare the performance of RobotReviewer with ChatGPT for extracting PICO and / or Risk of Bias information
4. Try using GPT-4 for classifying studies (on the web or via EPPI Reviewer)
5. Try using ChatGPT for data extraction
6. Try another tool of your choice...

See the resources on the website for links and further tools



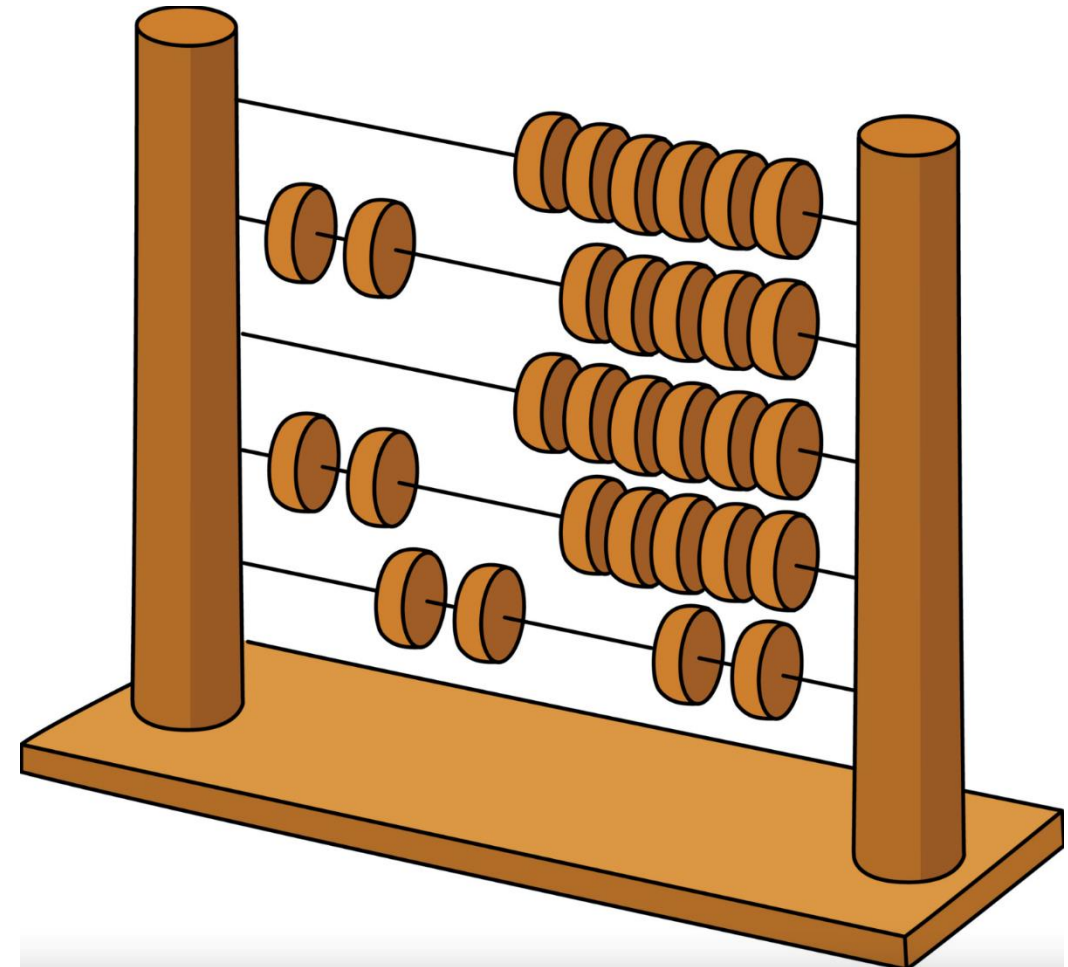
Time for a break...



Questions and discussion

Summing up

- Most evidence synthesis still uses almost entirely manual processes
- Machine learning is only used in some
- While many tools are promising there are barriers to implementation for some tools
- There are some great tools that are ready for use
- The promise of GenAI is currently only a promise
- We need lots of rigorous evaluation before we can see the promise realised



Thank you

James Thomas

EPPI-Centre website: <http://eppi.ioe.ac.uk>

Email

james.thomas@ucl.ac.uk

Twitter

James_M_Thomas

EPPI-Centre
Social Science Research Unit
Institute of Education
University of London
18 Woburn Square
London WC1H 0NR

Tel +44 (0)20 7612 6397
Fax +44 (0)20 7612 6400
Email eppi@ioe.ac.uk
Web eppi.ioe.ac.uk/

