

# Machine learning and automation in reviews is now a reality; but do we yet know how to use these technologies?

**James Thomas, Sergio Graziosi, Jeff Brunton**  
Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre)  
Social science research unit, Department of Social Science  
UCL Institute of Education  
University College London

# Declaration of interests and funding

- James Thomas is co-lead of the Cochrane ‘Transform’ project, which is implementing some of the technologies discussed here. He also directs development & management of EPPI-Reviewer, the EPPI-Centre’s software for systematic reviews.
- Parts of this work funded by: Cochrane, JISC, Medical Research Council (UK), National Health & Medical Research Council (Australia), Wellcome Trust, Bill & Melinda Gates Foundation. All views expressed are my own, and not necessarily those of these funders.

# Objectives

- To introduce technologies for automating parts of the review process including: study selection; risk of bias assessment; and synthesis;
- For participants to try these technologies for themselves; and
- To discuss methodological issues concerning their use.

## Structure:

- Presentation (many slides...) + questions / discussion
- Experimentation with online tools and discussion (small groups)
- Whole group feedback (if there's time / need)
- (<http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3677>)

# Automation in systematic reviews – what can be done?

- Study identification:
  - Citation screening
  - Updating reviews
  - RCT classifier
- Mapping research activity
- Search strategy development
- Data extraction
  - Risk of Bias assessment
  - Other study characteristics
  - Extraction of statistical data
- Synthesis and conclusions





# Reducing workload during citation screening





The screenshot shows the EPPI-Reviewer4 software interface. The browser address bar indicates the URL is <http://eppi.ioe.ac.uk/eppireviewer4/>. The interface includes a menu bar with options like Documents, Search, Diagrams, Frequencies, Crosstabs, Reports, Meta-analysis, Collaborate, Mv info, and Screenings. Below the menu is a 'Document details' section with tabs for Citation details, Text document, Reference Search, Coding record, Linked records, and PDF. The 'Citation details' tab is active, showing the following information:

- Item 14 in current list
- Next (arrow)
- Cancel (X)
- Find on web (Globe icon)
- Title: **Smoking habits in secondary school students.**
- Author(s): Damas C; Monteiro S; Marinho A; Fernandes G
- Item IDs: Internal: 631675 Imported:
- Month: January
- Pub type: Journal, Article
- Year: 2009
- Included?:
- Abstract: BACKGROUND: Smoking is an important health risk in general, and responsible for diseases with high mortality and morbidity. Smoking habits start early and adolescence is a notorious time for starting smoking. AIM AND METHODS: To assess knowledge on smoking and smoking habits in a population of 8th grade students in Porto schools, using a confidential self administered questionnaire. Collected data were evaluated using SPSS 12 statistics program (2004 version). RESULTS: A total of 1,770 students aged 11-12 years old, mainly males (58%), answered. Most students (n=952, 54.6%) were unaware of signs or symptoms of smoking in their schools. The great majority (n=1639, 92.7%) considered themselves well informed about the harmful effects of smoking, but only 31.1% could list three or more tobacco-associated health risks. However, parents and friends were seen as privileged sources of information. Among these students, 11.1% were smokers and the average started to smoke at the age of 11.5 years. The majority of the smokers (57.2%) had parents who smoked and 96.4% had friends who smoked, versus 83.1% of non-smokers. There was a statistically significant difference ( $p < 0.001$ ). Pocket money was the means of acquiring cigarettes. Most (60.8%) considered themselves able to stop smoking at any time, while 11.4% of the smokers smoked more than one pack a day and 9.8% smoked the first cigarette within 5 minutes of waking, however, no dependence were found. Knowledge of the risks of smoking was poor and information on smoking in schools had an apparently low and variable impact. Parents' and friends' behaviour may have had an impact on the decision to start smoking.

On the left side, there is a 'Codes' panel with a tree view under 'Screening Criteria'. Several criteria are listed with checkboxes and 'Info' links. One criterion, 'T&A: Include Quantitative 98+ prevalence of sources', is checked. Below the codes panel is a 'Files' section with an 'Upload' button and a table with columns for Title, Document, Extension, Delete, View text, and Download.

1. Read title & abstract
2. Click include / exclude
3. Click 'next' and move on to the next reference
4. Repeat...

# Citation 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%)

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

Alison O'Mara-Eves<sup>1</sup>, James Thomas<sup>1\*</sup>, John McNaught<sup>2</sup>, Makoto Miwa<sup>3</sup> and Sophia Ananiadou<sup>2</sup>

### Abstract

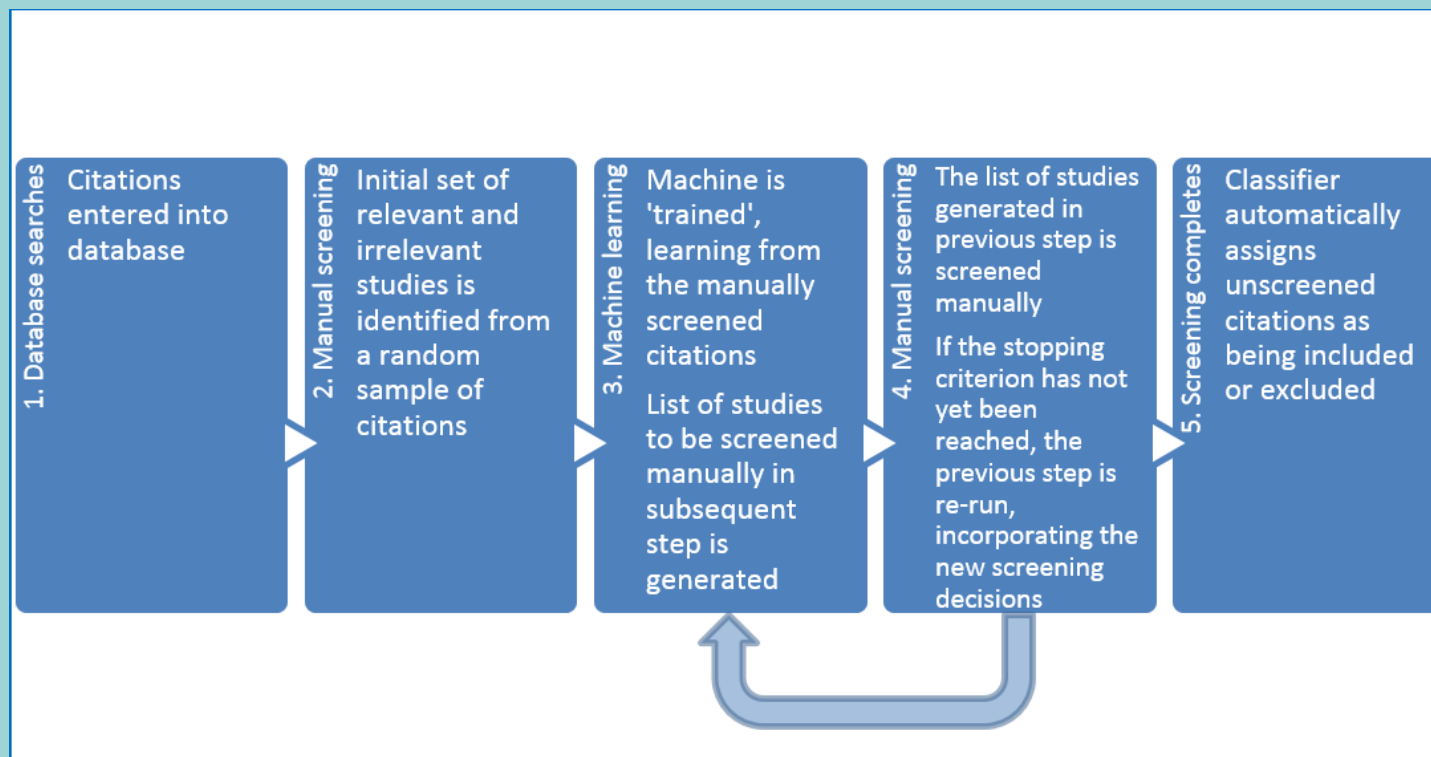
**Background:** The large and growing number of published studies, and their increasing rate of publication, makes the task of identifying relevant studies in an unbiased way for inclusion in systematic reviews both complex and time consuming. Text mining has been offered as a potential solution: through automating some of the screening process, reviewer time can be saved. The evidence base around the use of text mining for screening has not yet been pulled together systematically; this systematic review fills that research gap. Focusing mainly on non-technical issues, the review aims to increase awareness of the potential of these technologies and promote further collaborative research between the computer science and systematic review communities.

**Methods:** Five research questions led our review: what is the state of the evidence base; how has workload reduction been evaluated; what are the purposes of semi-automation and how effective are they; how have key contextual problems of applying text mining to the systematic review field been addressed; and what challenges to

## 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 significantly depending on the domain of literature being screened

# How the machine learns...

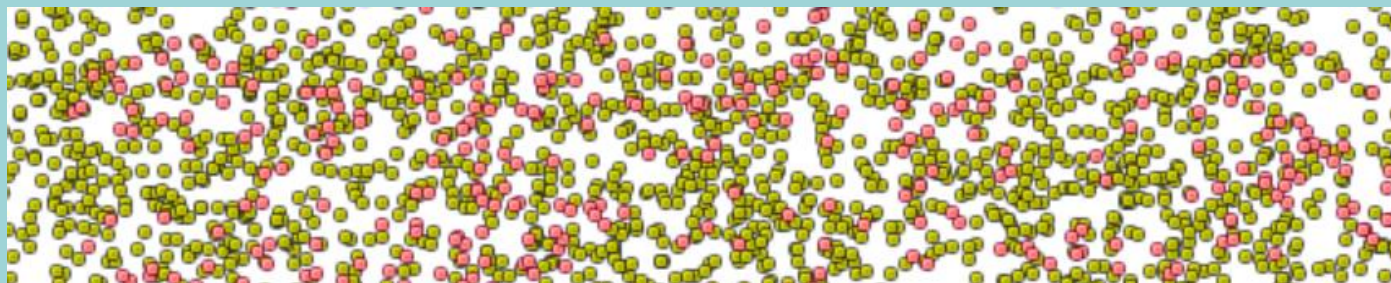


And it can work quite well...



# Screening prioritisation: Changing the distribution of studies

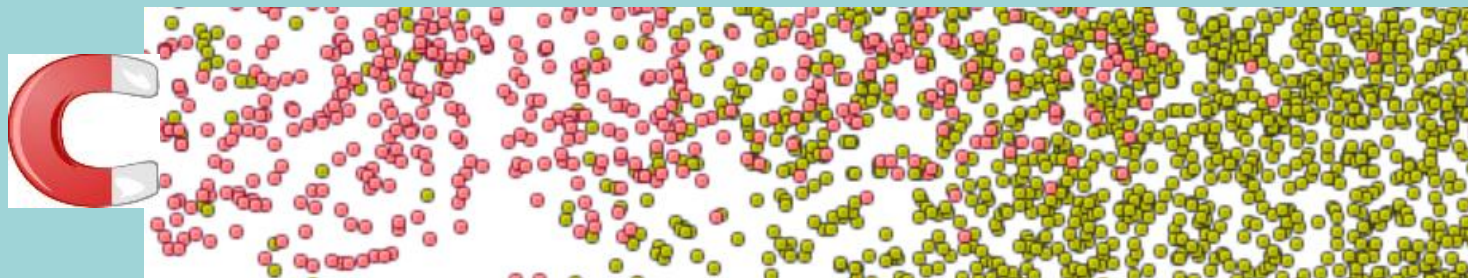
Traditional screening



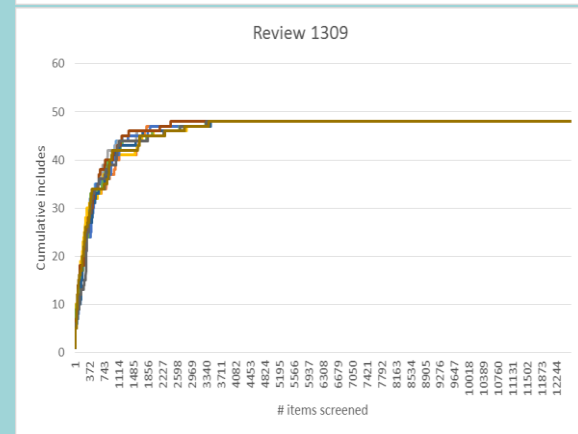
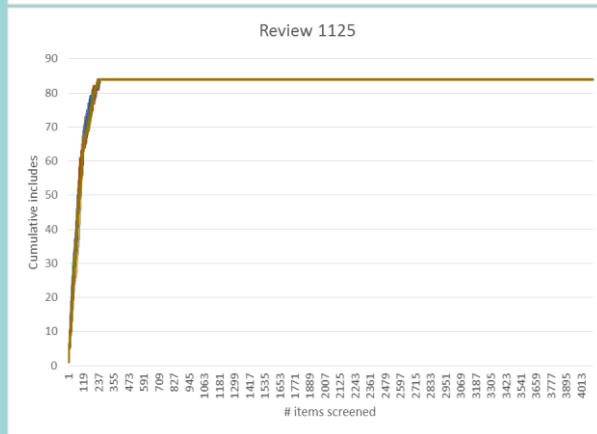
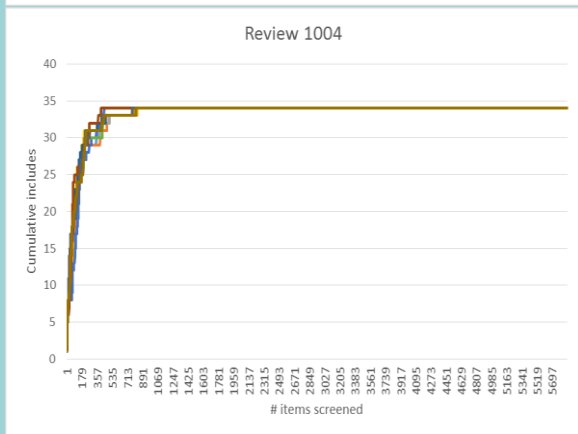
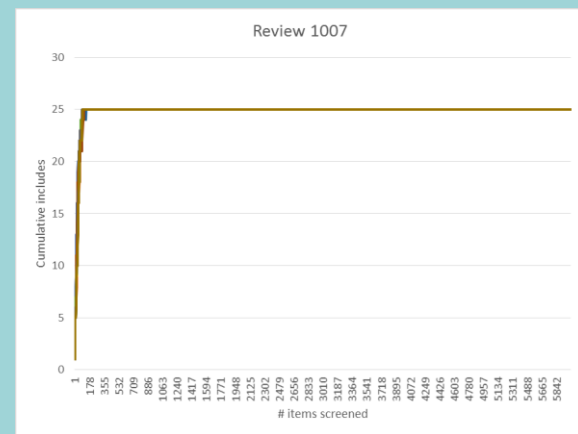
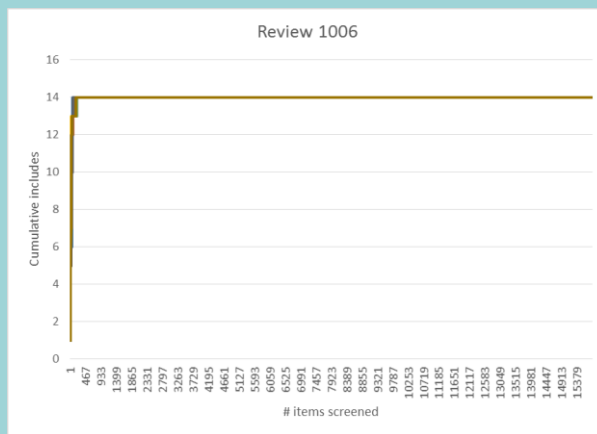
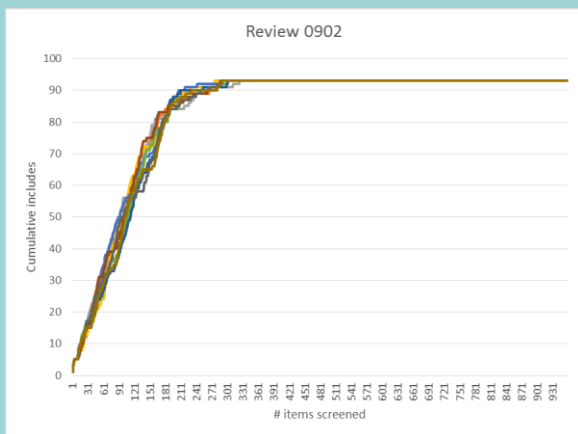
Screening process (red = eligible study)



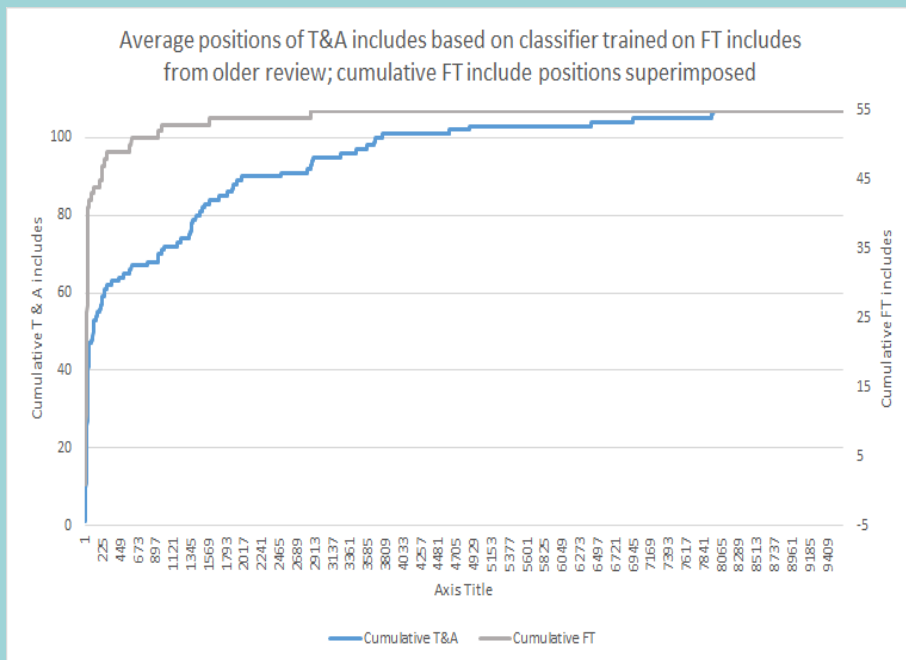
Screening aided by text mining



# e.g. reviews from Cochrane Heart Group



# Updating existing reviews



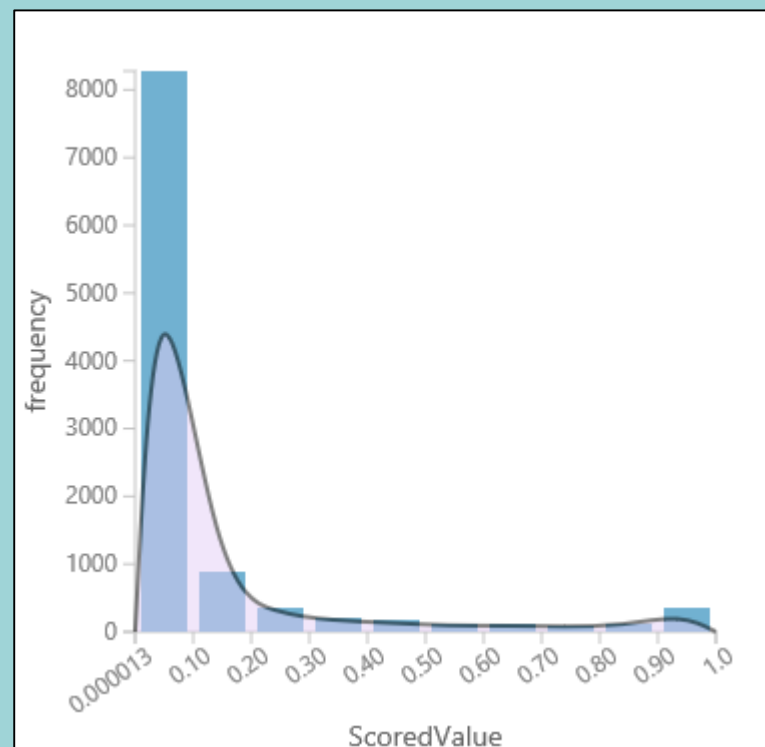
Weightman A, Thomas J, Baker P, Lovie-Toon Y, Francis D, O'Mara-Eves A (2014) Text mining for screening efficiency? Testing within a Cochrane public health review. Poster presented at Cochrane Colloquium 2014, Hyderabad

# RCT Classifier



# RCT Classifier

- ‘Trained’ on more than 280,000 human classifications from Cochrane Crowd
- Recall = 99.879% if all citations  $> 0.1$  are screened manually; 60% workload saving??





http://localhost/WcfHostPortal/EppiReviewer4.a EPPI-Reviewer4 (V.4.6.1.2)

Documents Search Diagrams Frequencies Crosstabs Reports Meta-analysis Collaborate My info Screening

New search Refresh search list Delete selected Combine AND OR NOT (included) NOT (excluded)

Title	Created by	Date	Hits	Select
176 Items classified according to model: RCT	James Thomas	10/18/2016	22	Select
175 Items classified according to model: RCT				
174 Coded with: EXC1-				
173 Coded with: EXC1-				
172 "low income" AND				
171 low income team				
170 food club				
166 moynihah				
165 roe				
164 kennedy				
163 friends with food				
162 hodgson				
161 McGlone				
160 white and carlin				
159 white				
158 carter				
157 lloyd				
156 Dutton				
155 cook for Life				
154 cook4Life				
153 gregg				
152 Barton	Rebecca Rees	4/21/2011	15	Select
151 Symon	Rebecca Rees	4/21/2011	5	Select
150 149 AND 57	James Thomas	4/19/2011	1400	Select
149 Coded with: To screen pages 1 and 3	James Thomas	4/19/2011	1400	Select

**Machine learning classifier**

The machine learning classifier built into EPPI-Reviewer enables you automatically to assign a code to new items based on those you have assigned to others. There are two stages to the process: first the classifier 'learns' to apply the code by building a model (which is saved for future use); and then the model is applied to new items. One 'pre-built' classifier is available: the 'RCT model', which will automatically identify RCTs for you; it has been trained on more than 280,000 records screened by the Cochrane Crowd.  
(This function enables you to build a linear classifier from a bag of words representation of your studies, using the scikit-learn python library.)

**Stage 1: build the model**

Learn to apply this code

Distinguish from this code

Name for your model

(Go straight to stage 2 if you are applying the RCT model.)

**Stage 2: apply the model**

Title	Applies	Compared with	Precision	Recall
Check screening	Include	Exclude	0.17	1.00

Apply above selected model      Apply to all items in review  
 Apply RCT model      Apply to items with this code  
 Apply to items from this source

Codes Sources Review statistics

Status: Normal. The Online Shop is now active in the Account Manager pages. It... [Show More] | User: James Thomas | Review: Cooking Skills Review (Expired)

[Documents](#) | [Search](#) | [Diagrams](#) | [Frequencies](#) | [Crosstabs](#) | [Reports](#) | [Meta-analysis](#) | [Collaborate](#) | [My info](#) | [Screening](#)

29 documents loaded.

Showing: Items classified according to model: RCT

Filter:

	<input type="checkbox"/>	Author	Title	Year	Score
Go	<input type="checkbox"/>	I Horner S	Enhancing Asthma Self-Management in Rural School-Aged Children: A Randomized Controlled Trial	2015	99
Go	<input type="checkbox"/>	I McGhan S	A children's asthma education program: Roaring Adventures of Puff (RAP), improves quality of life	2010	99
Go	<input type="checkbox"/>	I Bruzzese	Feasibility and impact of a school-based intervention for families of urban adolescents with asthma: results from a randomized pil	2008	98
Go	<input type="checkbox"/>	I Horner S	Improvement of rural children's asthma self-management by lay health educators	2008	98
Go	<input type="checkbox"/>	I Cicutto L	Breaking the access barrier: evaluating an asthma center's efforts to provide education to children with asthma in schools	2005	97
Go	<input type="checkbox"/>	I Bruzzese	Reducing Morbidity And Urgent Health Care Utilization In Urban Pre-adolescents With Asthma: Results Of A Randomized Control T	2010	96
Go	<input type="checkbox"/>	I Kintner E	Randomized clinical trial of a school-based academic and counseling program for older school-age students	2009	96
Go	<input type="checkbox"/>	I McWhirte	Can schools promote the health of children with asthma?	2008	96
Go	<input type="checkbox"/>	I Patterson	A cluster randomised intervention trial of asthma clubs to improve quality of life in primary school children: the School Care and A	2005	96
Go	<input type="checkbox"/>	I Shah S ;	Effect of peer led programme for asthma education in adolescents: cluster randomised controlled trial	2001	95
Go	<input type="checkbox"/>	I Clark NM	An evaluation of asthma interventions for preteen students	2010	93
Go	<input type="checkbox"/>	I McGhan S	Evaluation of an education program for elementary school children with asthma	2003	86
Go	<input type="checkbox"/>	I Bruzzese	Effects of a school-based intervention for urban adolescents with asthma. A controlled trial	2011	84
Go	<input type="checkbox"/>	I Gerald LB	Outcomes for a comprehensive school-based asthma management program	2006	81
Go	<input type="checkbox"/>	I Clark NM	Effects of a comprehensive school-based asthma program on symptoms, parent management, grades, and absenteeism	2004	78
Go	<input type="checkbox"/>	I Butz A ; F	Rural children with asthma: Impact of a parent and child asthma education program	2005	77
Go	<input type="checkbox"/>	I Joseph Cl	A web-based, tailored asthma management program for urban African-American high school students	2007	76
Go	<input type="checkbox"/>	I Cicutto L	A randomized controlled trial of a public health nurse-delivered asthma program to elementary schools	2013	75
Go	<input type="checkbox"/>	I Guglani L	Exploring the impact of elevated depressive symptoms on the ability of a tailored asthma intervention to improve medication adh	2013	74
Go	<input type="checkbox"/>	I Levy M ; I	The efficacy of asthma case management in an urban school district in reducing school absences and hospitalizations for asthma	2006	74
Go	<input type="checkbox"/>	I Mandhan	A child's asthma quality of life rating does not significantly influence management of their asthma	2010	74
Go	<input type="checkbox"/>	I Clark NM	A trial of asthma self-management in Beijing schools	2005	70
Go	<input type="checkbox"/>	I Joseph Cl	Feasibility Of Web-based Asthma Education For Urban Teenagers: Intervention Compliance And Computer Access [Abstract]	2010	70
Go	<input type="checkbox"/>	I Henry RL	Randomized controlled trial of a teacher-led asthma education program	2004	63
Go	<input type="checkbox"/>	I McCann C	A controlled trial of a school-based intervention to improve asthma management	2006	60
Go	<input type="checkbox"/>	I Kintner E	Effectiveness of a school-based academic asthma health education and counseling program on fostering acceptance of asthma in	2014	57

Page 1 of 1

Codes

- Source
- James's codes
- Screening Criteria
- Admin codes
- EPPI-Centre Health prom
- Mapping Coding Tool
- Mapping tool allocations
- Surveys n=100
- Jeff 3.1 check
- YP tobacco sources - Da

Codes Sources Review statistics

Status: Normal. The Online Shop is now active in the Account Manager pages. It... [Show More] | User: James Thomas | Review: Young people's access to tobacco: a mixed-method systematic review.



# Developing search strategies





# Text recognition/text analytics

- Text analysis (frequency counts, phrases or nearby terms in text, or statistical frequency across corpus)
- Term extraction and automatic clustering (ranked list of words or phrases from a combination of linguistic and statistical analyses)

## Applications:

- Provides a rapid overview of words/phrases or controlled terms in sample, depending on tools used.
- Identifying words or phrases that not considered
- Identify unwanted search terms

# Termine: Automatic Term Identification

<http://www.nactem.ac.uk/software/termine/>

- Termine automatically identifies and ranks terms according to their importance to the citation list

The software tools and services which NaCTeM supplies allow researchers to apply text mining techniques to problems within their specific areas of interest - examples of these tools are highlighted below. In addition to providing services, the Centre is also involved in, and makes significant contributions to, the text mining research community both nationally and internationally in initiatives such as Europe PubMed Central.

Rank	Term	Score
1	text mining	3
2	text mining research community	2
3	datum mining system	1.584962
3	natural language processing	1.584962

Frantzi, Katerina, Sophia Ananiadou, and Hideki Mima. "Automatic recognition of multi-word terms: the C-value/NC-value method." International Journal on Digital Libraries 3.2 (2000): 115-130.

# Text recognition/text analytics

Sample N=52

Using TD\*IDF

Term	Score
ID	76.61
people	44.67
resident	39.81
adult	38.26
woman	37.08
older people	35.73
menopause	32.58
participant	29.33
client	28.07
life	27.35
older adult	26.21
death	25.76
older person	25.76
age	23.62
active ageing	19.76
retirement	19.55
person	18.72
population	18.14
end-of-life care	17.95
future planning	17.12
ageing adult	15.8

Using Termine

intellectual disability	148.896545
aged care	14.3125
end-of-life care	10
intellectual disability nurse	6.33985
community-based aged care	6.33985
active ageing	6
moderate id	6
down syndrome	5
future planning	5
residential aged care	4.754888
residential aged care facility	4
learning disability	4
moderate intellectual disability	3.169925
future plan	3
developmental disability	3
id group home staff	2
mild id	2
future care	2
death method	2
hospice care	2
elderly people	2
profound id	2
late life	2
age-related illness	2
retirement option	2
community-based aged care results	2
nursing home	2
normal ageing	2

# Identifying search terms for: Research on care and support needs for ageing populations of adults with learning disabilities

52 items relevant to topic



Analysed with both Termine and TD\*IDF



Selected population terms related to older people, ageing, aged care.  
Re-ran search to find 44/52 items.



Scanned remaining 8 records manually,



Revisited term lists to identify additional terms:

Future planning, planning for the future, Aged related  
illness, menopausal

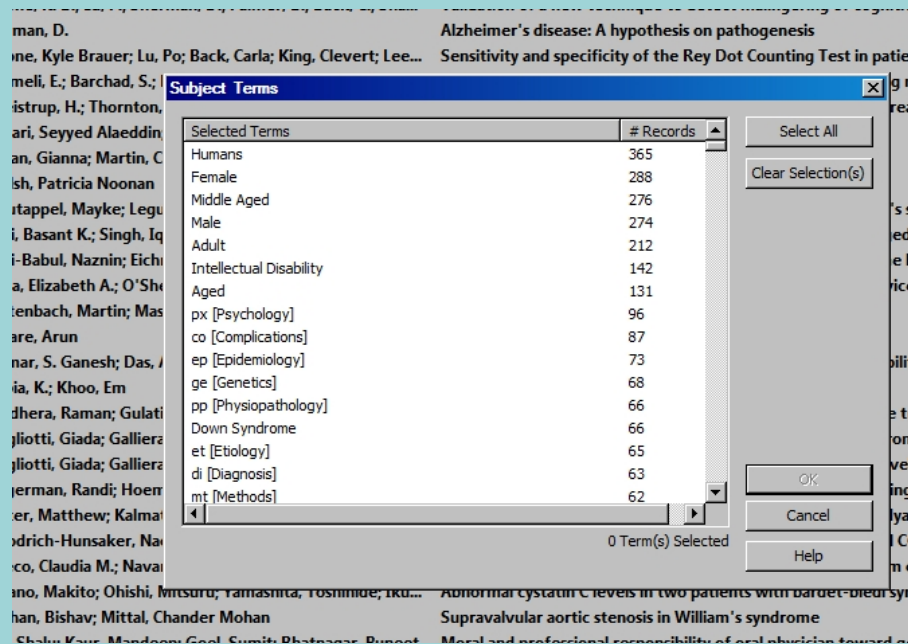


Incorporated into existing search strategy

# Rapid overview of controlled terms

## Analysing controlled terms in Endnote:

N=500 items  
to analyse  
search subset



The screenshot shows the 'Subject Terms' dialog box in Endnote. The dialog box has a title bar 'Subject Terms' and a close button. It contains a table with two columns: 'Selected Terms' and '# Records'. The table lists various terms and their corresponding record counts. Below the table, there are buttons for 'Select All', 'Clear Selection(s)', 'OK', 'Cancel', and 'Help'. At the bottom of the dialog, it says '0 Term(s) Selected'.

Selected Terms	# Records
Humans	365
Female	288
Middle Aged	276
Male	274
Adult	212
Intellectual Disability	142
Aged	131
px [Psychology]	96
co [Complications]	87
ep [Epidemiology]	73
ge [Genetics]	68
pp [Physiopathology]	66
Down Syndrome	66
et [Etiology]	65
di [Diagnosis]	63
mt [Methods]	62

Method in: Hayman, S and Shaheem, Y (2014). Smart Searching: Logical Steps to Building and Testing Your Literature Search. CareSearch Palliative Care Knowledge Network.<http://sites.google.com/site/smartsearchinglogical/home>

# Rapid overview of words per citation

Word frequency per citation:

In Bibexcel analysed titles/abstracts from 1,000 items in test search - count per citation, found:

Gene (96),  
Premutation (59),  
protein (57), FMR1  
(57), mutation  
(51), FXTAS (44)

The screenshot shows the BibExcel software interface. The main window is titled "BibExcel - A toolbox for bibliometricians, by Ole Persson, Version 2014-03-25". The interface is divided into several sections:

- Select file here:** A file explorer showing a directory structure with files like "500temsti.cit", "500temsti.doc", "500temsti.out", "500temsti.oux", "500temsT1.txt", "342 from line16.ris", "age search string.doc", and "Alborz 2004.pdf".
- Select field to be analysed, view file to get info about which fields are available:** A dropdown menu set to "Blank-separated words (e.g. title)" and a "Prep" button.
- Select documents:** A "Start" button.
- Select rows:** A "Start" button.
- The Box:** A text area showing the view file path: "View file: s:\nice nccsc\topic 9 older people w le learning difficulties\search strategy development\500temsti.cit Sec:0 Units:2206 Unique:2206".
- Type new file name here:** An empty text input field.
- Frequency distribution:** A section with a "Select type of unit" dropdown set to "Whole string", a "Start" button, and checkboxes for "Sort descending" (checked), "Remove duplicates", "Make new out-file", and "Fractionalize". There are also input fields for "Min number", "Max number", and "Any number".
- Old Tag:** A text input field containing "ER" and an "Add field to units" button.
- New Tag:** An empty text input field and an "Add new field to docs" button.
- The List:** A list of words with their frequencies. The word "intellectual" is highlighted with a frequency of 32. Other words include "Alternate" (499), "adults" (99), "Down" (65), "syndrome" (60), "disabilities" (48), "disabilitiesTitle" (47), "syndromeTitle" (47), "disability" (46), "people" (44), "older" (39), "care" (37), "studyTitle" (37), "patients" (30), and "clinical" (28).
- Row no: 2** and **Search** and **View whole file** buttons are at the bottom.

Directory of tools:  
<http://www.tapor.ca/>

# Query expansion

<http://nactem.ac.uk/hom>

- Augment query with synonyms, related terms, orthographic variations etc.

The screenshot shows a search interface with a search bar containing the term "tobacco". Below the search bar are two buttons: "New Search" (red) and "Refine Search" (green). Below the search bar is a section titled "Related Terms" which displays a word cloud of related terms. The terms are connected by lines, indicating relationships. The terms include: ban, billboard, sponsor, grit, sponsorship, hookah, cigar, brand, carcinogen, smoke, pharmaceutical, habit, cigarette, smoker, ets, abatement, nicotine, cigarillo, and chimney, benzpyrene.

Term ✕  
tobacco

New Search Refine Search

Related Terms

ban billboard sponsor grit sponsorship hookah  
cigar brand carcinogen smoke  
pharmaceutical habit  
cigarette smoker ets  
abatement nicotine  
cigarillo  
chimney benzpyrene

# Query expansion

<http://nactem.ac.uk/hom>

The screenshot shows the History Of Medicine website interface. On the left, a 'Search Query' panel lists the terms 'intellect', 'disabled', and 'nursing home'. Below it, a 'Related Terms' panel lists: handicapped, residential care home, geriatric nursing home, voluntary home, private nursing home, home nursing equipment, ordinary nursing home, psychiatric nursing home, and residential home. The main area displays 'Search Results' for 'Showing page 1 of 15242 results'. A 'Term Frequencies' section contains a line graph titled 'Term frequency over time' showing the frequency of 'disabled' (blue line) and 'nursing home' (red line) from 1840 to 2008. The 'disabled' line shows a significant peak around 1996, while the 'nursing home' line shows a peak around 1960. Below the graph, a list of search results is shown, including 'Nursing-homes', 'Community care: the first year', 'Staff problem in old people's homes', 'NHS must pay nursing costs for dependent nursing home residents', and 'Private nursing home care: the middle way'.

Search Query

Term  
intellect

Term  
disabled

Term  
nursing home

New Search Refine Search

Related Terms

handicapped

residential care home

geriatric nursing home

voluntary home

private nursing home

home nursing equipment

ordinary nursing home

psychiatric nursing home

residential home

Search Results <sup>0</sup> Showing page 1 of 15242 results

Term Frequencies

Term frequency over time

Frequency

Year

disabled

nursing home

Nursing-homes

Nursing-homes SIR -The last sentence of the final paragraph of your annotation on nursing-homes (Journal, June 12, p. 1368) provokes thought and comment. The long-stay annexes of newly established geriatric units still house large numbers of crippled and enfeebled patients who once existed in pub... [bmj\\_2079276](#)

Community care: the first year

Community care: the first year EDITOR -Trish Groves reports on mental health care in Bassetlaw one year into the Community Care Act, referring to people with enduring mental health problems and the importance of rehabilitation.' There is no mention of finding community placement for the most disa... [bmj\\_2539906](#)

Staff problem in old people's homes

BRITISH MEDICAL JOURNAL VOLUME 282 3 JANUARY 1981 77 Staff problem in old people's homes SIR -Local authority residential homes for the elderly ("Part III accommodation") have been notoriously understaffed. At any given time there is only one ... [bmj\\_1503746](#)

NHS must pay nursing costs for dependent nursing home residents

NHS must pay nursing costs for dependent nursing home residents Clare Dyer - legal correspondent BMJ The Court of Appeal in London last week overturned a High Court judge's decision which would have forced the NHS to find an extra £220m (\$352m) a year for the costs of caring for patients in nurs... [bmj\\_1116322](#)

Private nursing home care: the middle way

732 BRITISH MEDICAL JOURNAL VOLUME 296 12 MARCH 1988 Private nursing home care: the middle way The place of private nursing homes in the care of ... [bmj\\_1116322](#)

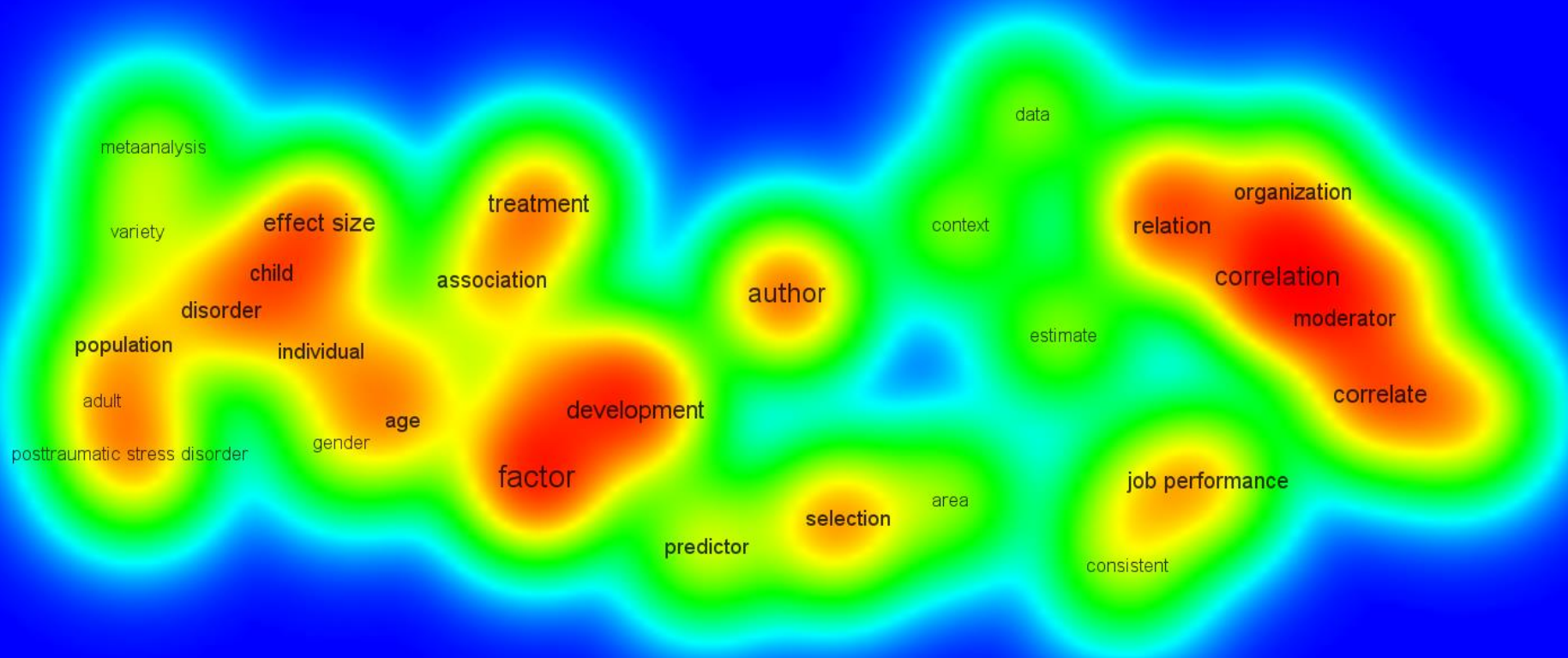


# Mapping research activity



# Technologies for identifying subsets of citations

- Different families of techniques
  - Fairly simple approaches which examine term frequencies to group similar citations
  - More complex approaches, such as Latent Dirichlet Allocation (LDA)
- Identifies groups of documents which use similar combinations of terms
- The difficult part is finding good labels to describe the clusters
  - But are labels always needed?
- Visualisations are often incorporated into tools



e.g. Carrot

The screenshot shows a search engine interface with a circular topic map. The map is divided into segments representing related concepts: SEARCH, UNSTRUCTURED DATA, INFORMATION RETRIEVAL, CUSTOMER, COMPUTER, TOOLS, TEXT PROCESSING, TEXT MINING AND ANALYTICS, RESEARCH, ANALYSIS AND TEXT MINING, WORDSTAT, OTHER TOPICS, SEMANTIC SEARCH, RESPONSE PREDICTION, ORACLE, NATIONAL CENTRE FOR, LAUNCHES, DISTRIBUTED, UNIVERSITY, UPDATED, BIOMEDICAL TEXT MINING, WIKIPEDIA, TM PACKAGE, COMBINING, and TEXT MINING TECHNOLOGY. The search results on the right list various articles and resources related to text mining.

The screenshot shows a search engine interface with a hexagonal topic map. The map is divided into segments representing related concepts: Unstructured Data, Computer, Text Mining and Analytics, Information Retrieval, Tools, Search, Oracle, Other Topics, Tm Package, Semantic Search, and University. The search results on the right list various articles and resources related to text mining.

The screenshot shows a search engine interface with a list of search results. The results include links to various articles and resources related to text mining, such as 'What is Text Mining?' and 'What is text mining (text analytics)?'. The search results on the right list various articles and resources related to text mining.

# “Topic modelling”



```

CRD_Ida_topics0_AB - Notepad
File Edit Format View Help
0.148*use + 0.050*drug + 0.038*substance + 0.036*alcohol + 0.022*drugs + 0.020*users +
0.016*marijuana + 0.016*among + 0.015*cocaine + 0.014*cannabis0.052*sexual + 0.038*hiv +
0.031*sex + 0.029*risk + 0.020*use + 0.017*drug + 0.015*partners + 0.015*among + 0.015*men +
0.014*infection0.069*intake + 0.050*dietary + 0.034*food + 0.025*diet + 0.018*energy +
0.015*vitamin + 0.014*consumption + 0.014*fat + 0.014*fruit + 0.013*vegetables0.101*men +
0.071*women + 0.026*body + 0.019*mass + 0.018*fat + 0.017*age + 0.015*associated +
0.012*index + 0.012*association + 0.009*associations0.013*compliance + 0.011*volume +
0.008*immigrants + 0.007*salivary + 0.007*epilepsy + 0.006*volumes + 0.006*immigrant +
0.006*doctors + 0.006*conventional + 0.006*acamprosate0.094*blood + 0.078*pressure +
0.054*hypertension + 0.025*systolic + 0.020*bp + 0.015*diastolic + 0.013*diabetic + 0.012*mm
+ 0.010*hypertensive + 0.010*mmhg0.110*risk + 0.072*disease + 0.057*factors + 0.040*heart +
0.039*cardiovascular + 0.033*coronary + 0.029*stroke + 0.022*chd + 0.019*smoking +
0.014*hypertension0.019*markers + 0.016*stress + 0.015*cortisol + 0.014*tissue + 0.014*muscle
+ 0.013*inflammation + 0.012*growth + 0.012*inflammatory + 0.011*oxidative + 0.010*visceral
0.019*media + 0.016*internet + 0.012*information + 0.011*ra + 0.011*messages + 0.010*online +
0.009*content + 0.009*mdma + 0.008*advertising + 0.008*incontinence0.084*suicide +
0.042*suicidal + 0.030*athletes + 0.030*attempts + 0.026*sports + 0.021*ideation +
0.017*sport + 0.017*attempt + 0.014*attempted + 0.010*team0.142*patients + 0.013*treatment +
0.011*clinical + 0.011*patient + 0.010*hospital + 0.009*therapy + 0.007*cases + 0.007*pain +
0.006*medication + 0.006*diagnosis0.076*95 + 0.065*ci + 0.038*risk + 0.025*odds +
0.021*confidence + 0.020*associated + 0.018*ratio + 0.018*interval + 0.015*association +
0.013*adjusted0.072*gambling + 0.023*problem + 0.021*gamblers + 0.020*pathological + 0.010*pd
    
```

	A	B	C	D	E	F	G	H	I	J	
1	Included?	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9
2	No	7.395112	5.946818	5.040004	4.744282	4.253583	3.83581	2.417316	3.087711	4.860644	3.087711
3	Yes	14.07189	2.137577	4.443276	5.034399	0.7431	2.299016	1.26992	1.574956	7.010434	4.600000

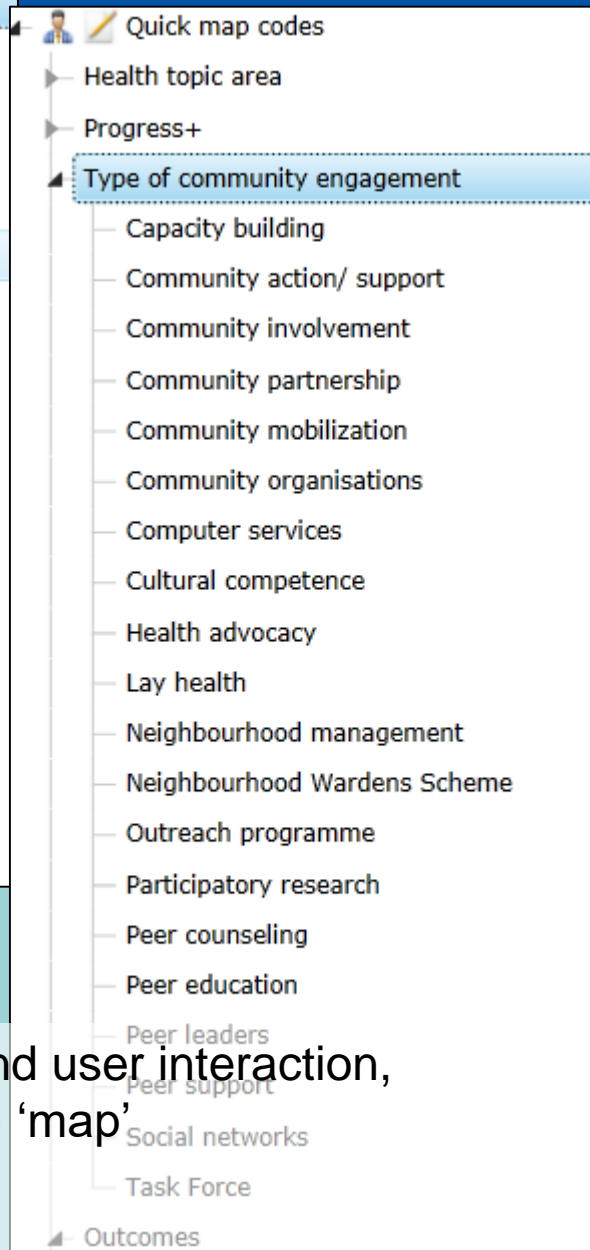
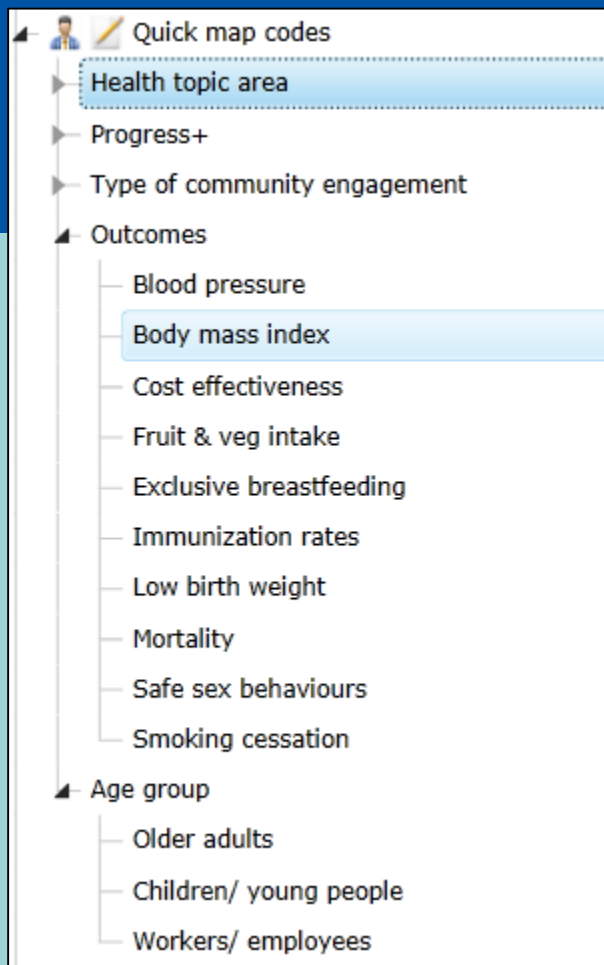
# Used in a review about community engagement to improve health amongst disadvantaged groups

**Lingo3G clusters**

- ▶ African American
- ▶ Health Promotion
- ▶ High Risk
- ▶ Low Income
- ▶ Physical Activity
- ▶ **Lay Health**
- ▶ Peer Education
- ▶ Cost Effectiveness
- ▶ Substance Abuse
- ▶ National Evaluation
- ▶ HIV Risk Reduction
- ▶ Mental Health
- ▶ Injury Prevention
- ▶ Pregnant Women
- ▶ Randomised Controlled Trial
- ▶ Cigarette Smoking
- ▶ Months Later
- ▶ Community Action

The screenshot shows the ePPI Reviewer software interface. On the left, a table lists included documents with columns for Authors, Title, and Year. On the right, a 'Codes' panel displays a hierarchical list of Lingo3G clusters, with 'Lay Health' highlighted. The status bar at the bottom indicates the user is James Thomas reviewing 'CHERI: Community engagement to reduce inequalities in health'.

Go	Authors	Title	Year
Go	I Aaron S J; Jenk	Postponing sexual intercourse among urban junior high school students-a randomized controlled evaluation	2000
Go	I Ahmed NU ; Hab	Randomized controlled trial of mammography intervention in insured very low-income women.	2010
Go	I Allen JP ; Philli	Preventing Teen Pregnancy and Academic Failure: Experimental Evaluation of a Developmentally Based Approach	1997
Go	I Allen P ; Thomp	Impact of periodic follow-up testing among urban American Indian women with impaired fasting glucose.	2008
Go	I Anand S S; Davl	A family-based intervention to promote healthy lifestyles in an aboriginal community in Canada	2007
Go	I Andersen M R; Y	The effectiveness of mammography promotion by volunteers in rural communities	2000
Go	I Andersen M R; H	Recruitment, Retention, and Activity of Volunteers Promoting Mammography Use in Rural Communities	2000
Go	I Anderson AK ; D	A randomized trial assessing the efficacy of peer counseling on exclusive breastfeeding in a predominantly Latina low-income	2005
Go	I Andrews J O; Be	Using community-based participatory research to develop a culturally sensitive smoking cessation intervention with public ho	2007
Go	I Andrews JO ; Fel	The effect of a multi-component smoking cessation intervention in African American women residing in public housing	2007
Go	I Arlotti J P; Cott	Breastfeeding among low-income women with and without peer support	1998
Go	I Asepline Robert	Mentoring as a Drug Prevention Strategy: An Evaluation of "Across Ages."	2000
Go	I Auld GW ; Roma	Outcomes from a school-based nutrition education program using resource teachers and cross-disciplinary models	1998
Go	I Auslander W ; H	A controlled evaluation of staging dietary patterns to reduce the risk of diabetes in African-American women	2002
Go	I Avila P ; Hovell	Physical activity training for weight loss in Latinas: a controlled trial	1994
Go	I Ayala G X; Elder	Longitudinal intervention effects on parenting of the Aventuras para Niños study	2010
Go	I Baker K ; Pollack	Violence prevention through informal socialization: An evaluation of the South Baltimore Youth Center	1995
Go	I Baker E A; Bouk	The Latino Health Advocacy Program: a collaborative lay health advisor approach	1997
Go	I Balcazar HG ; de	A promotores de salud intervention to reduce cardiovascular disease risk in a high-risk Hispanic border population, 2005-200	2010
Go	I Banks Erin Rash	Being Healthy Counts To H.I.M.: An Examination of Health Behavior Among Participants in a Diabetes Prevention and Health I	2009
Go	I Baranowski T ; S	A center-based program for exercise change among black-American families	1990
Go	I Barnes K ; Fried	Impact of community volunteers on immunization rates of children younger than 2 years	1999
Go	I Barnes-Boyd C ;	Promoting infant health through home visiting by a nurse-managed community worker team	2001
Go	I Baruth Meghan ;	Psychosocial mediators of a faith-based physical activity intervention: Implications and lessons learned from null findings	2010



With a little organisation and user interaction, we constructed a workable 'map'



Creating databases of  
research  
prospectively

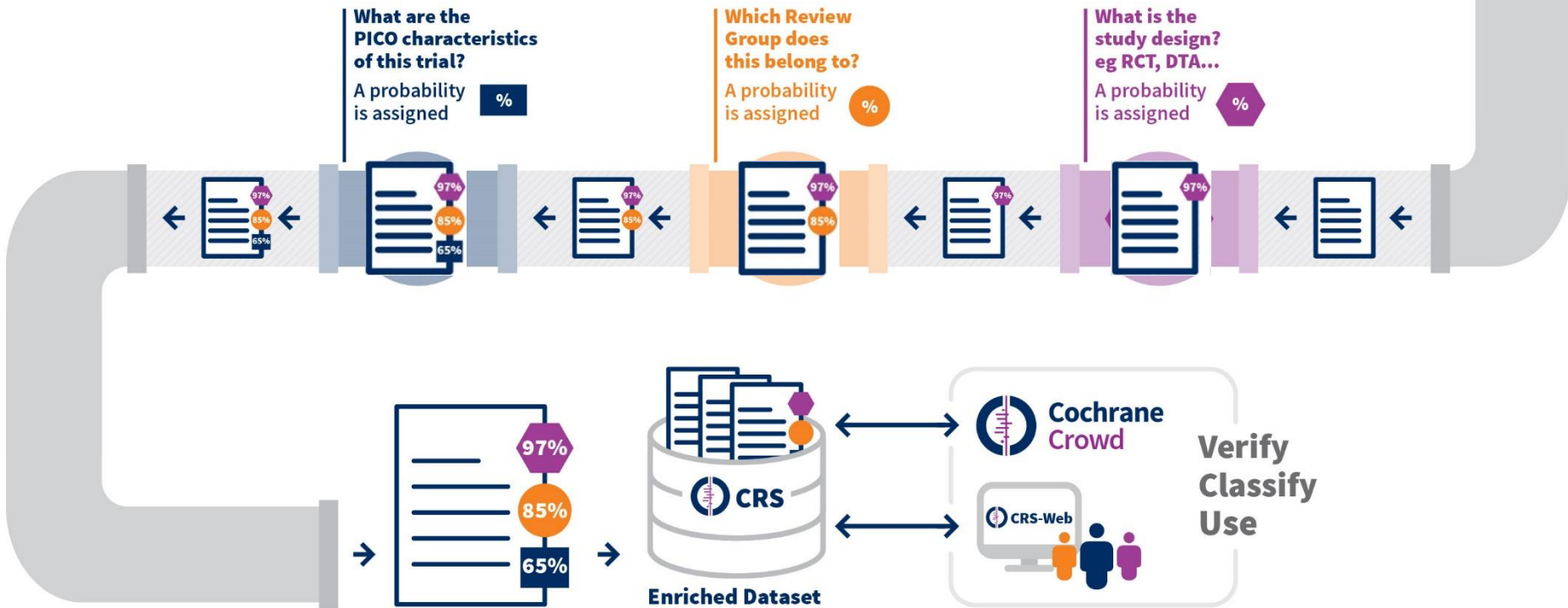
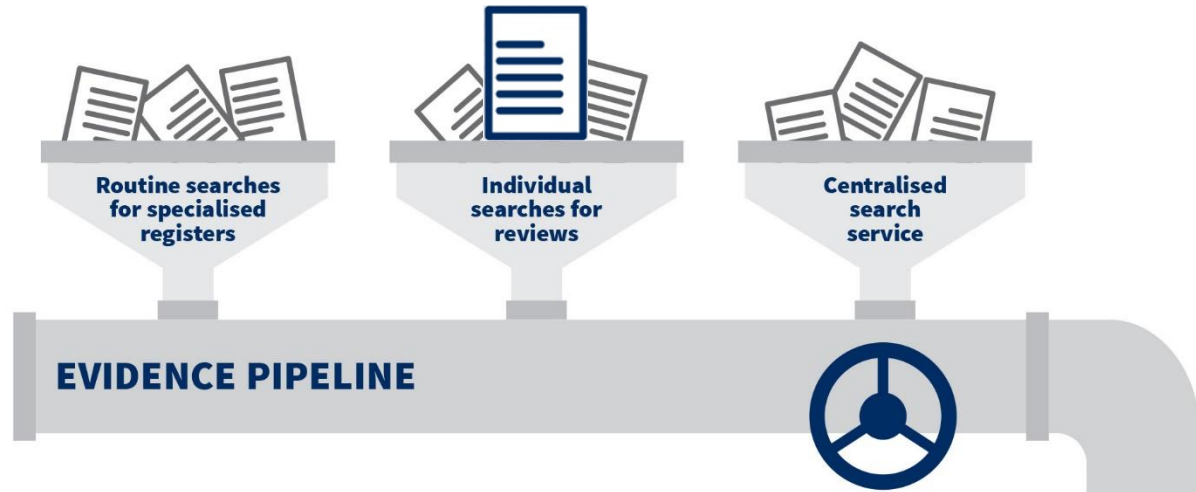




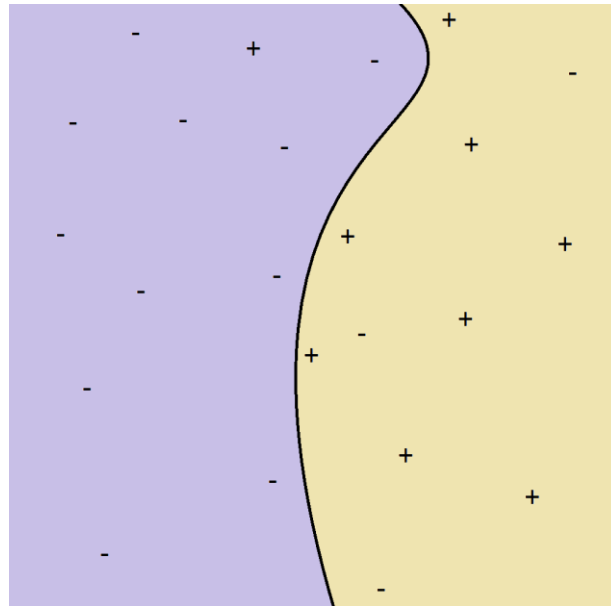


## Evidence Pipeline

Finding and classifying relevant research



# How does it work?



Building machine classifiers



# 1. A dictionary and index are created

- First, the key terms in the studies are listed (ignoring very common words)
- Second, the studies are indexed against the list of terms
  - (the resulting matrix can be quite large)
- Next...

**e.g. We have two studies – one is an RCT, and one isn't an RCT**

Study 1 Effectiveness of asthma self-care interventions: a systematic review (not an RCT)

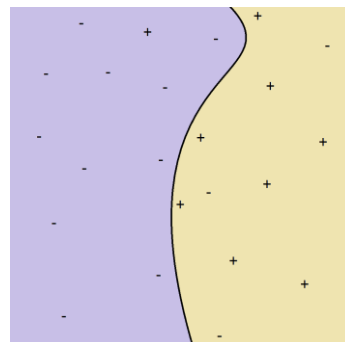
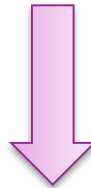
Study 2 Effectiveness of a self-monitoring asthma intervention: an RCT (an RCT)

RCT?										
0	1	1	1	1	1	1	1	0	0	0
1	1	1	1	0	0	0	0	1	1	1

## 2. A statistical model is built

The matrix is used to create a statistical model which is able to distinguish between the two classes of document (e.g. between RCTs and non-RCTs)

RCT?	Effectiveness	asthma	self care	interventions	systematic	review	monitoring	intervention	RCT
0	1	1	1	1	1	1	0	0	0
1	1	1	1	0	0	0	1	1	1

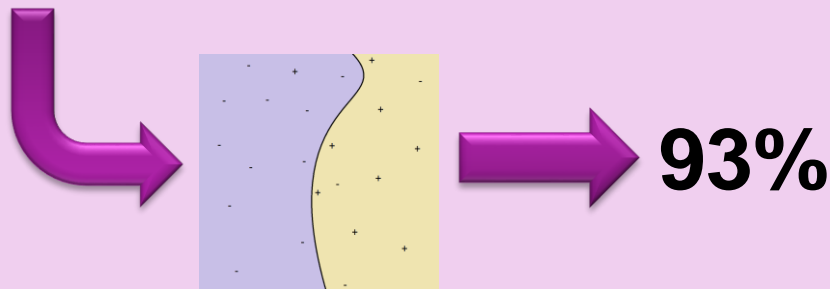


### 3. The model is applied to new documents

- New citations are indexed against the previously generated list of terms
- The resulting matrix is fed into the previously generated model
- And the model will assign a probability that the new document is, or is not a member of the class in question

e.g. The effectiveness of a school-based asthma management programme: an RCT

Effectiveness	asthma	self	care	interventions	systematic	review	monitoring	intervention	RCT
1	1	0	0	0	0	0	0	0	1





# The new CRS-Web

CRS Web (online) | crsdemo.metaxis.com/index.php#Search

Search Simple MeSH Classifier Saved Tracking

## Classifier search

Records that have been through the classifier have probabilities assigned to them to indicate how likely they are to have certain properties, like whether they are of interest to a review group, or whether they are likely to be an RCT. Choose the classifier model you are interested in, set the model parameters and click Search to find the records

RCT

Number of references

Score

Approximately 32129 records that are between 99 and 100 percent likely to be of interest

[Search](#)

You can find your records that are currently being processed by the classifier by searching for INPROCESS:CLASSIFIER

[Find those records](#)

## Cochrane Register of Studies

Anna Noel-Storr [DEMENTIA] | logout

Dashboard Records Import Journals CT.GOV Reports To do

Search Layout1 Layout2 Layout3 Layout4

Search results (399 records) Page 1 of 8

#	Title	Author
1	Cognitive effects of treating obstructive sleep apnea in Alzheimer's disease: a randomized controlled study	Ancoli-Israel S // Palmer BW // Cooke
2	Efficacy of galantamine in probable vascular dementia and Alzheimer's disease combined with cerebrovascular disease: a random...	Erkinjuntti T // Kurz A // Gauthier S //
3	Donepezil improved memory in multiple sclerosis in a randomized clinical trial	Krupp LB // Christodoulou C // Melvil
4	A randomized, 26-week, double-blind, placebo-controlled trial to evaluate the safety and efficacy of galantamine in the treatme...	Auchus A
5	A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. Donepezil Study Group	Rogers SL // Farlow MR // Doody RS /
6	A Controlled, Double-Blind, Randomized Pilot Clinical Trial of Hydroxysafflor Yellow a on Cognitive Function in Patients With Vas...	Tian J

Record

Fields Duplicates Links Reviews Classifier Files Audit CENTRAL REGISTER

The bar chart below shows the classifier scores for this record. Scores are presented in the range 0 - 100 where higher scores mean a higher likelihood that the record is of interest to the group. You can tell a group about this record if it doesn't already have it in its segment by clicking the bar for that group.

In register 
  In segment 
  Not in segment 
  Not relevant to my group

There is a 99% likelihood that this record is an RCT [Confirm this is not an RCT] [Confirm this is an RCT]

99 73 37 29 23 16 15 15 11 11 11 10 9

DEMENTIA AIRWAYS MOVEMENT ENDOC SYMPT VASC ANAESTH COMMON EPOC EPILEPSY COMPRED OCCHEALTH MS

## Collection: Direct oral anticoagulants for thromboprophylaxis after orthopedic surgery

Owner: dperezrada@gmail.com

Update recommendations

Delete Collection

Related (26) Included (119) Excluded (59) Machine Recommendations (24)

### Primary Study

**Single intravenous administration of TB-402 for the prophylaxis of venous thromboembolism after total knee replacement: a dose-escalating, randomized, controlled trial.**

**AUTHORS** » Verhamme P, Tangelder M, Verhaeghe R, Agenc W, Glazer S, Prins M, Jacquemin M, Büller H, TB-402 Study Group

**JOURNAL** » Journal of thrombosis and haemostasis : JTH

**YEAR** » 2011

**LINKS** » Pubmed, DOI

This article is included in 1 Systematic Review

Include Exclude

### Systematic Review

**Rivaroxaban for thromboprophylaxis after orthopaedic surgery: pooled analysis of two studies.**

**AUTHORS** » Fisher WD, Eriksson BI, Bauer KA, Borris L, Dahl OE, Gent M, Haas S, Homering M, Huisman MV, Kakkar AK, Kålebo P, Kwong LM, Misselwitz F, Turpie AG

**JOURNAL** » Thrombosis and haemostasis

**YEAR** » 2007

**LINKS** » Pubmed

Without references

Include Exclude

### Unclassified

**Fondaparinux for prevention of venous thromboembolism in major orthopedic surgery.**

**AUTHORS** » Tran AH, Lee G

**JOURNAL** » The Annals of pharmacotherapy

**YEAR** » 2003

**LINKS** » Pubmed, DOI

Include Exclude

### Primary Study

**Evidence of venous thromboembolism after lower limb**

### Abstract

### About this article

Despite current guidelines recommendations about anticoagulant prophylaxis, many studies have shown an high venous thromboembolism (VTE) incidence in patients undergoing total hip and knee arthroplasty. A number of anticoagulants are currently available, but they have some limitations that affect their applicability and consequently their effectiveness. Several new oral anticoagulants (NOACs) have been developed in an attempt to overcome these limitations. Apixaban is a NOAC that selectively inhibits the coagulation factor Xa; it is approved for the prevention of VTE after total hip replacement and total knee replacement surgery. This review examines the results of main trials designed to test efficacy and safety of apixaban in major elective orthopedic surgery.



Project Transform

# Today



## You can make a difference

Become a Cochrane citizen scientist. Anyone can join our collaborative volunteer effort to help categorise and summarise healthcare evidence so that we can make better healthcare decisions.

Give it a try

Cochrane Crowd  
[crowd.cochrane.org](https://crowd.cochrane.org)





# Risk of bias assessment



# Risk of Bias assessment

- Emerging area; e.g.
  - RobotReviewer
  - Millard, Flach and Higgins
- Tools can accomplish two purposes:
  - 1. identify relevant text in the document
  - 2. automatically assess risk of bias
- Can perform very well though authors do not yet suggest well enough to replace humans

Int. J. Epidemiol. Advance Access published December 8, 2015

International Journal of Epidemiology, 2015, 1–12  
doi: 10.1093/ije/dyv306  
Original article

---

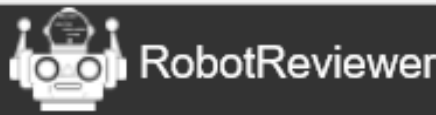
Original article

## Machine learning to assist risk-of-bias assessments in systematic reviews

Louise A.C. Millard,<sup>1,2,3\*</sup> Peter A. Flach<sup>1,3</sup> and Julian P.T. Higgins<sup>1,2</sup>

<sup>1</sup>MRC Integrative Epidemiology Unit, <sup>2</sup>School of Social and Community Medicine and <sup>3</sup>Intelligent Systems Laboratory, University of Bristol, Bristol, UK

The screenshot shows the RobotReviewer web interface. On the left, a document titled 'robot-reviewer.vortex.systems' is open, displaying text from a research paper. The text includes sections for 'Methods', 'Study population', 'Randomisation', 'Data collection', and 'Study protocol and interventions'. On the right, a 'Risk of Bias' assessment panel is visible, featuring a 'Random Sequence Generation' section with a 'Allocation Concealment' sub-section. The 'Allocation Concealment' section shows a prediction of 'low' bias, with a list of reasons: 'At enrollment the sequence was concealed from researchers...', 'Randomisation An independent statistician generated a random size stratified by recruitment centre, in a 1:1 ratio. At enrollment the sequence was concealed from researchers who confirmed consent and eligibility on an online database before allocation was revealed. It was not feasible to mask participants or researchers to group allocation.', 'Blinding Of Participants And Personnel', 'Blinding Of Outcome Assessment', 'Incomplete Outcome Data', and 'Selective Reporting'. Below this, the 'PICO' section is partially visible, showing 'Population' and 'Intervention'.



# Data extraction



# Data extraction

- RobotReviewer can identify phrases relating to study PICO characteristics
- ExaCT extracts trial characteristics (e.g. eligibility criteria)
- Systematic review found that no unified framework yet exists
- More evaluative work is needed on larger datasets
- Further challenges include extraction of data from tables and graphs

http://www.biomedcentral.com/1472-6947/10/56

BMC Medical Informatics & Decision Making

**TECHNICAL ADVANCE** **Open Access**

## ExaCT: automatic extraction of clinical trial characteristics from journal publications

Jonnalagadda *et al.* *Systematic Reviews* (2015) 4:78  
DOI 10.1186/s13643-015-0066-7

**RESEARCH** **Open Access**

## Automating data extraction in systematic reviews: a systematic review

Siddhartha R. Jonnalagadda<sup>1\*</sup>, Pawan Goyal<sup>2</sup> and Mark D. Huffman<sup>3</sup>

**SYSTEMATIC REVIEWS** CrossMark

The screenshot shows the RobotReviewer web interface. The main content area displays a systematic review article titled "Physical activity for smoking cessation in pregnancy: randomised controlled trial" by Michael Usher, Sarah Lewis, Paul Aveyard, Isaac Manyonda, Robert West, Beth Lewis, Bess Marcus, Muhammad Riaz, Adrian Taylor, Ananda Daley, and Tim Coleman. The article includes an abstract, objectives, design, setting, participants, interventions, and main outcome measures. On the right side, a sidebar titled "PICO" shows the extracted characteristics for each element:
 

- Population:** 789 pregnant smokers, aged 16-50 years and at 10-24 weeks' gestation, who smoked at least one cigarette daily and were prepared to quit smoking one week after enrollment were randomised (1,1); 785 were included in the intention to treat analyses, with 392 assigned to the physical activity group.
- Intervention:** Adding a physical activity intervention to behavioural smoking cessation support for pregnant women did not increase cessation rates at end of pregnancy. During pregnancy, physical activity is not recommended for smoking cessation but remains indicated for general health benefits.
- Comparison:** Parallel group, randomised controlled, multicentre.
- Outcome:** At enrollment the sequence was concealed from research...

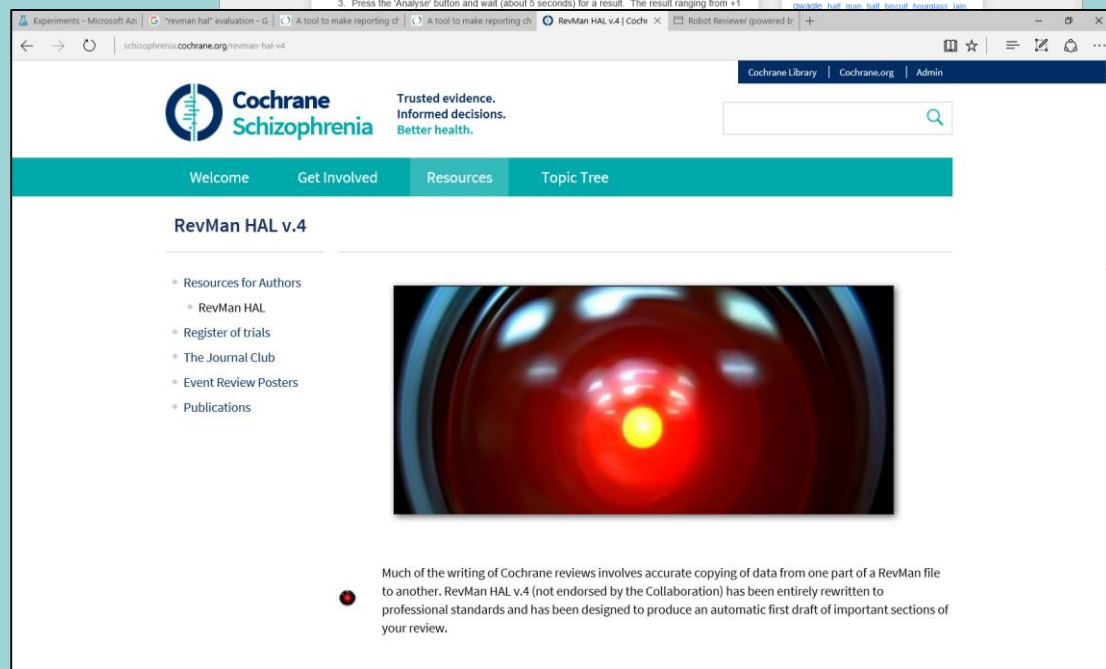
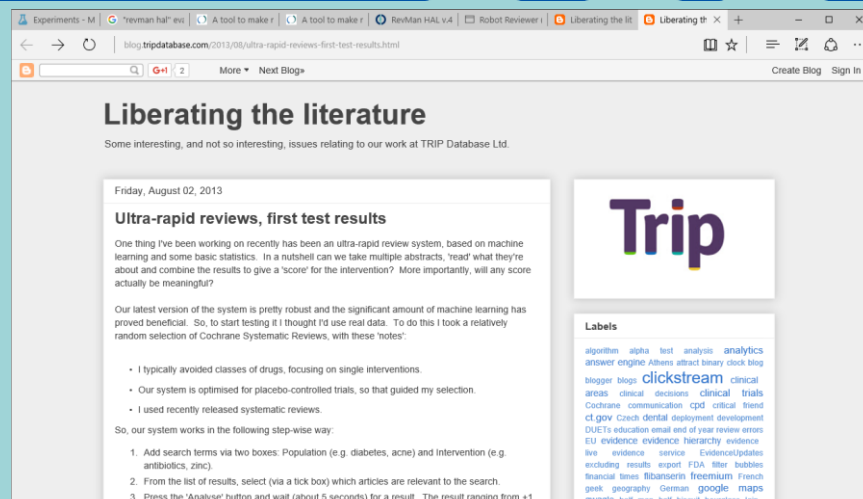


# Synthesis



# Synthesis and conclusions

- Summarisation and synthesis of text is an active area for development in computer science
- Many hurdles to overcome before this technology can be used routinely
- Some systems automate parts of the process



# Tools to try

This page: <http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3677>

Classification: RCT Classifier; and 'custom' classifier

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

Workload reduction during citation screening (using 'active learning')

- Rayyan systematic reviews tool: <http://rayyan.qcri.org/reviews/5>
- Microsoft Azure Machine Learning (James can demo for those interested)

Identifying sub-sets of citations (clustering)

- Carrot2 search: <http://search.carrot2.org/stable/search>
- Topic modelling:  
<http://nbviewer.jupyter.org/github/bmabey/pyLDavis/blob/master/notebooks/GraphLab.ipynb#topic=12&lambd=0.84&term=>

Search strategy development

- Termine: <http://www.nactem.ac.uk/software/termine/>
- NaCTeM History of Medicine: <http://nactem.ac.uk/hom>

## **Discussion: in small groups**

**Discuss methodological issues that the use of these technologies raise**

- which tools do you already use?
- what challenges do you face using existing tools?
  - Is recall of 99.879% for a workload reduction of 60% reasonable?
- what challenges do you think the new technologies would raise?
- what do you think would help you to adopt the new technologies?
- are there situations when using these tools might threaten the reliability of the review?



## Thank you

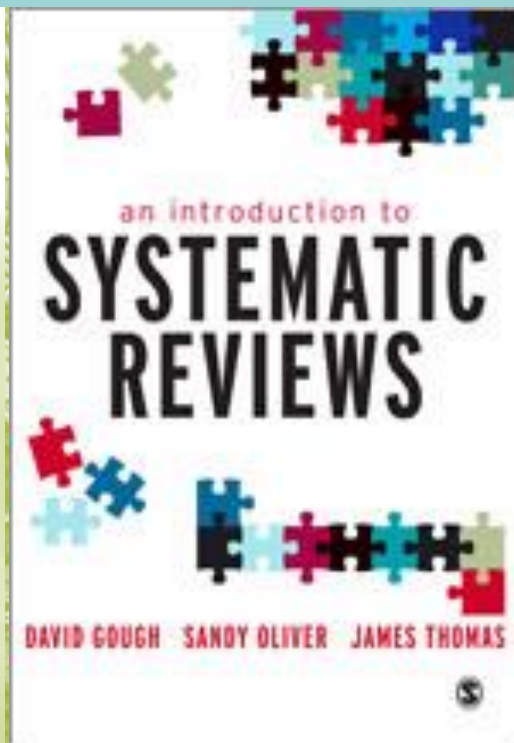
SSRU website: <http://www.ioe.ac.uk/ssru/>  
SSRU's EPPI website: <http://eppi.ioe.ac.uk>

Email

[j.thomas@ioe.ac.uk](mailto:j.thomas@ioe.ac.uk)

[c.stansfield@ioe.ac.uk](mailto:c.stansfield@ioe.ac.uk)

[a.omara-eves@ioe.ac.uk](mailto:a.omara-eves@ioe.ac.uk)



**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](mailto:eppi@ioe.ac.uk)  
Web [eppi.ioe.ac.uk/](http://eppi.ioe.ac.uk/)