# Artificial intelligence in systematic reviews: Uncharted waters for librarians

Elizabeth Moreton, MLS

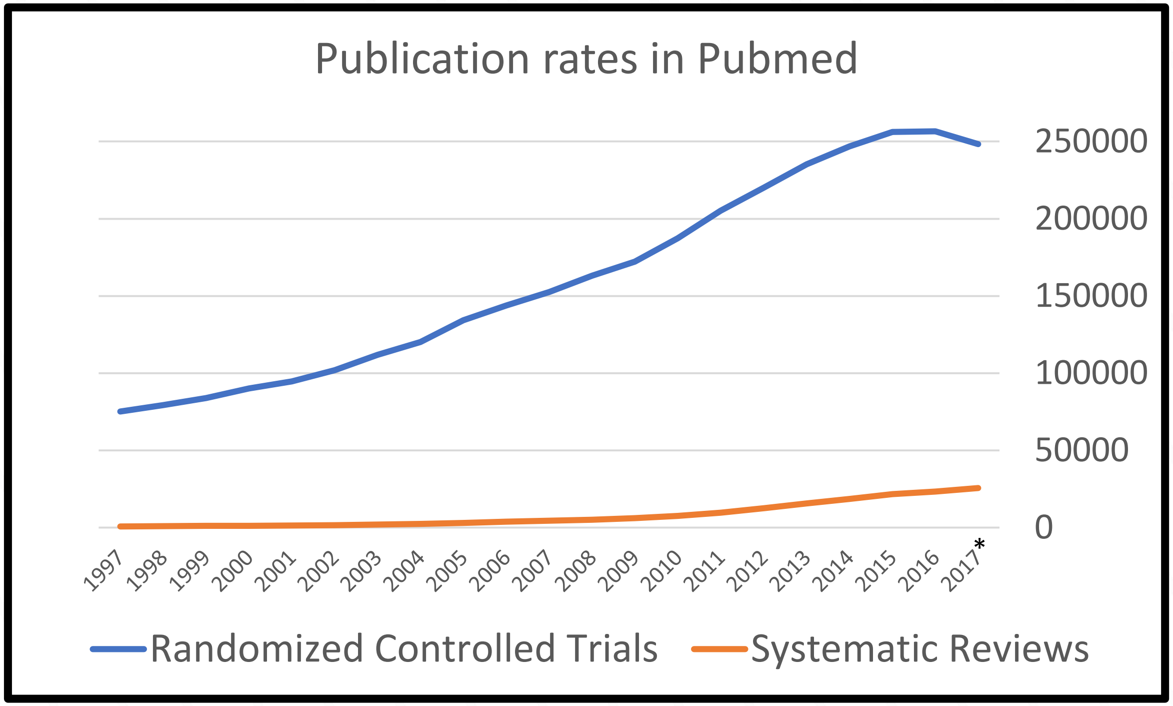
Health Sciences Library
University of North Carolina at Chapel Hill

# Project team

- Michelle Cawley
- Adam Dodd
- Elizabeth Moreton
- Jennifer Walker
- Fei Yu

# The problem with systematic reviews

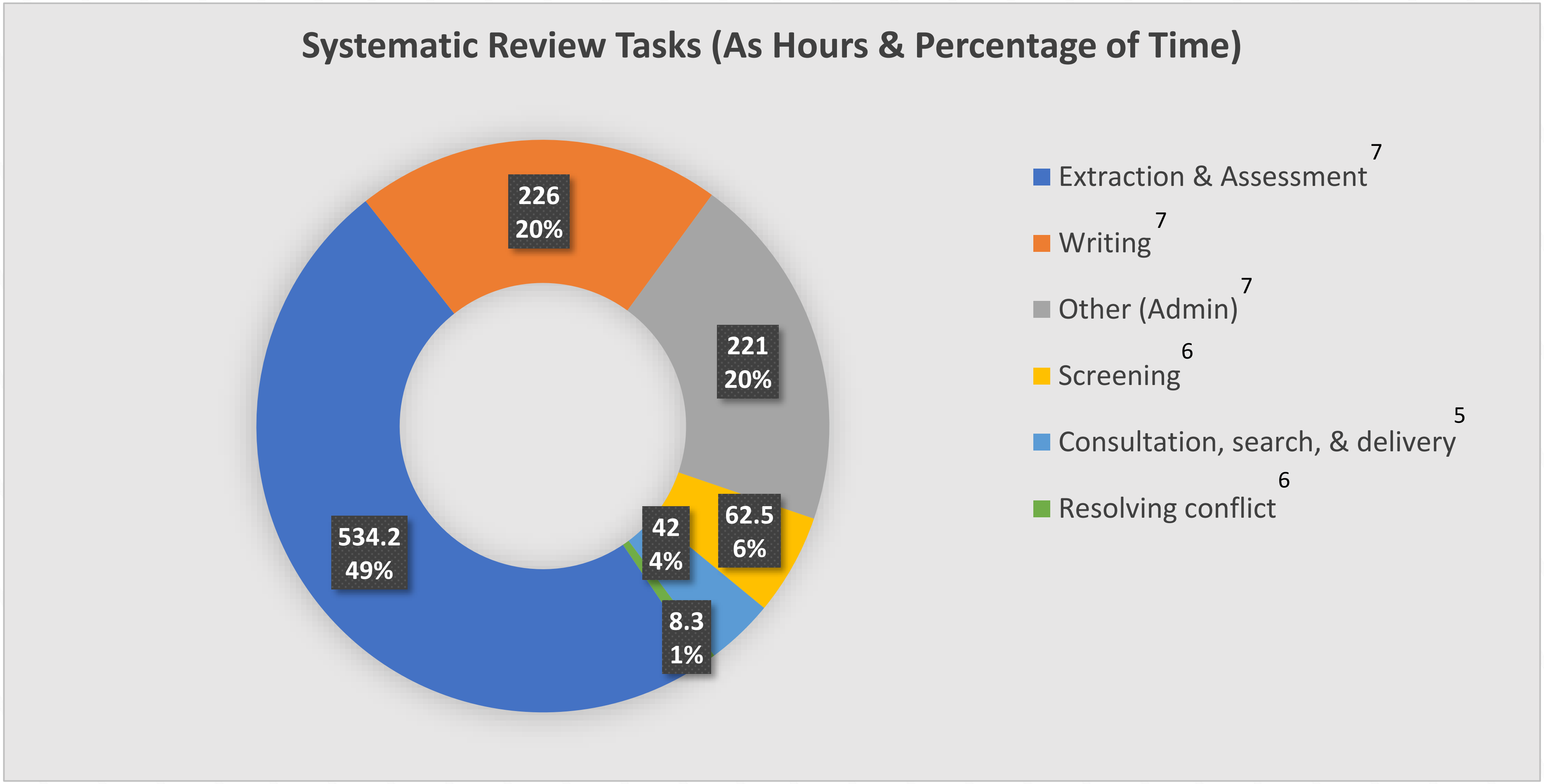
- Significant growth of trial publications, but systematic review publication rates aren't keeping up[1]
- Complex review methods[2]
- Limited resources/time
- Reviews need updating[3,4]



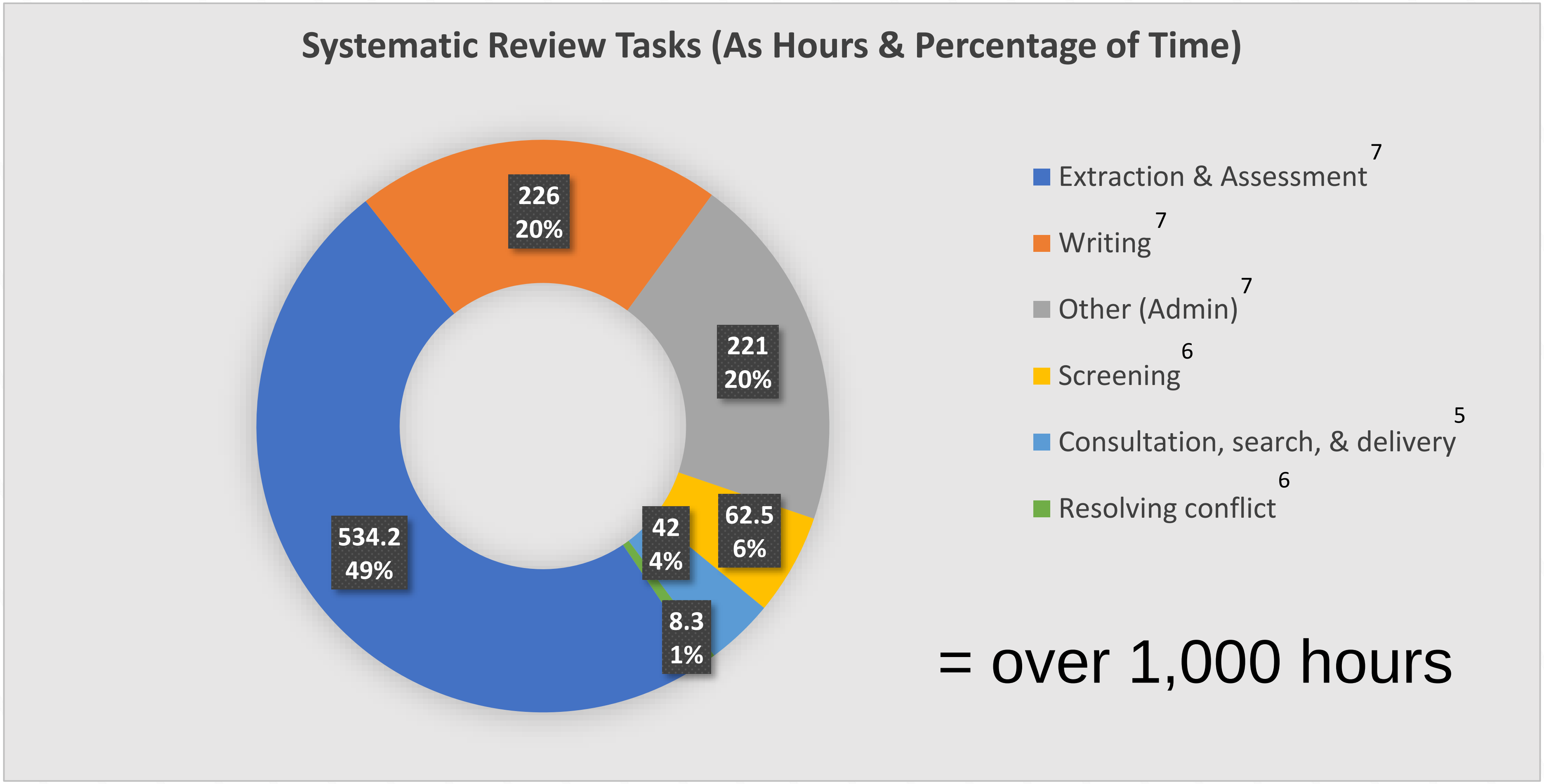

*may represent incomplete data

Data from http://dan.corlan.net/medline-trend.html

# The problem with systematic reviews

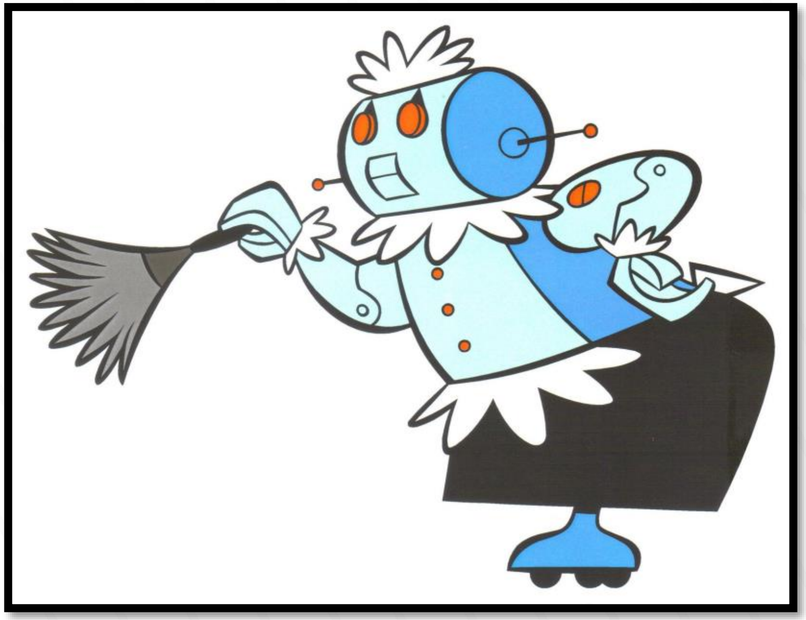


# The problem with systematic reviews



# What you may think of as Artificial Intelligence (AI)

- Robots
- Roomba
- Self-driving cars
- Netflix & Amazon recommendations
- IBM Watson



http://thejetsons.wikia.com/wiki/Rosey

# AI in the systematic review context

- **Artificial intelligence:** Artificial intelligence (AI) makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks.[8]
- **Machine learning:** Machine learning (ML) is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.[9]
- **Natural language processing:** Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language.[10]
- **Text mining:** Text mining (TM) is the process of analyzing collections of textual materials in order to capture key concepts and themes and uncover hidden relationships and trends without requiring that you know the precise words or terms that authors have used to express those concepts.[11]

# What we <u>don't</u> mean:

- Robots doing all the work by themselves
- Removing the librarian from the systematic review process
- Robots taking our jobs

https://www.fiercehealthcare.com/practices/ehr-involve-physicians-development-artificial-intelligence-stanford-university

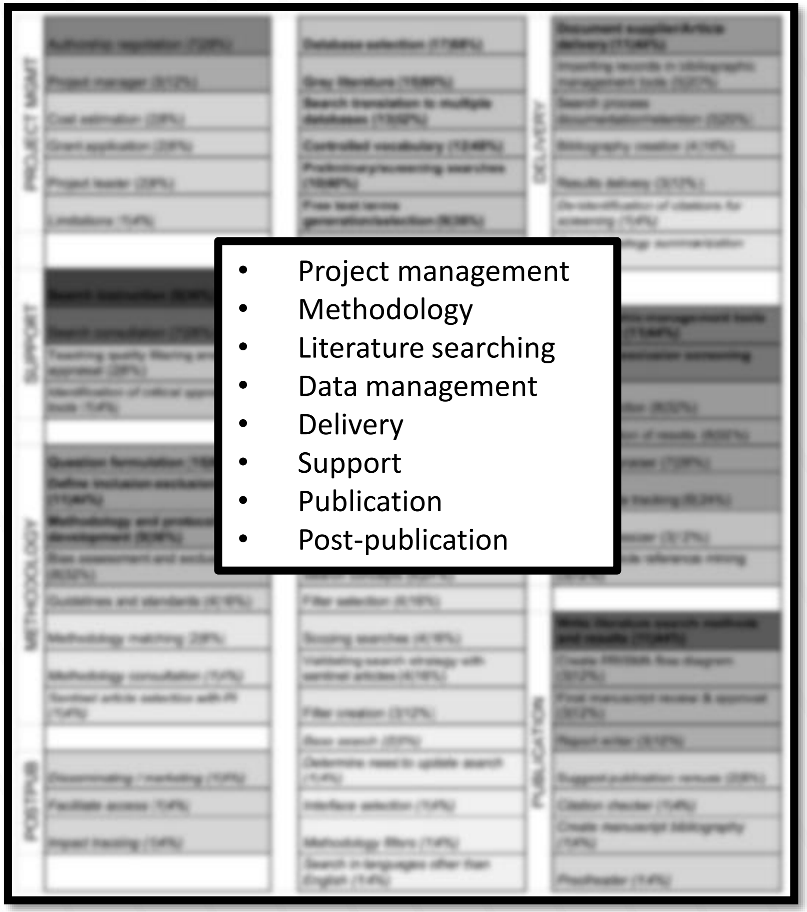
# What we do mean:

- Machines assisting with tasks- automating & predicting
- The team works faster and more efficiently
- The librarian becomes an expert consultant

https://www.fiercehealthcare.com/practices/ehr-involve-physicians-development-artificial-intelligence-stanford-university

# Librarian roles in systematic reviews

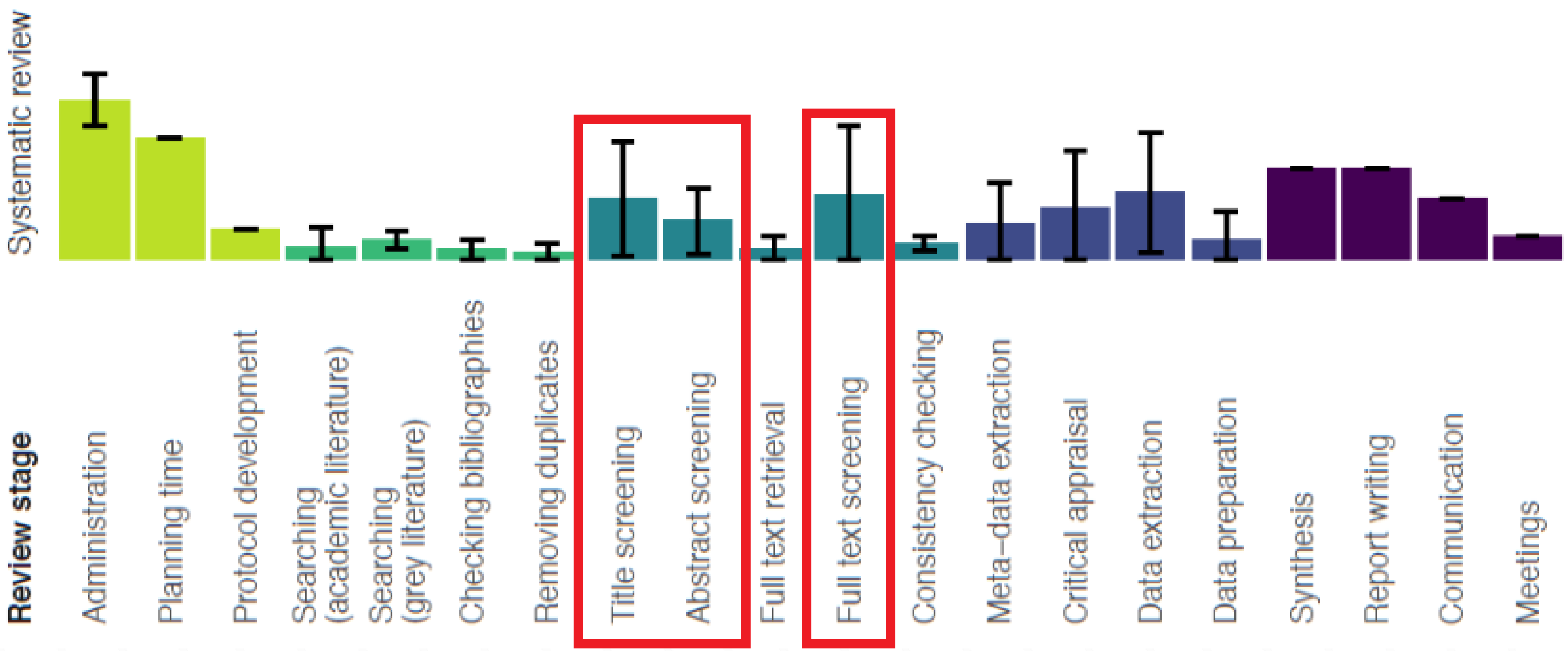
- 2018 JMLA scoping review by Spencer & Eldredge[12] found **18** roles for librarians in systematic review process
- 2018 MLA presentation by Ginier & Anderson[13] itemized each part of the SR process & found **69** roles librarians can perform




(Ginier & Anderson 2018)

# Where AI can accelerate SR process

# Our first automation challenge: screening

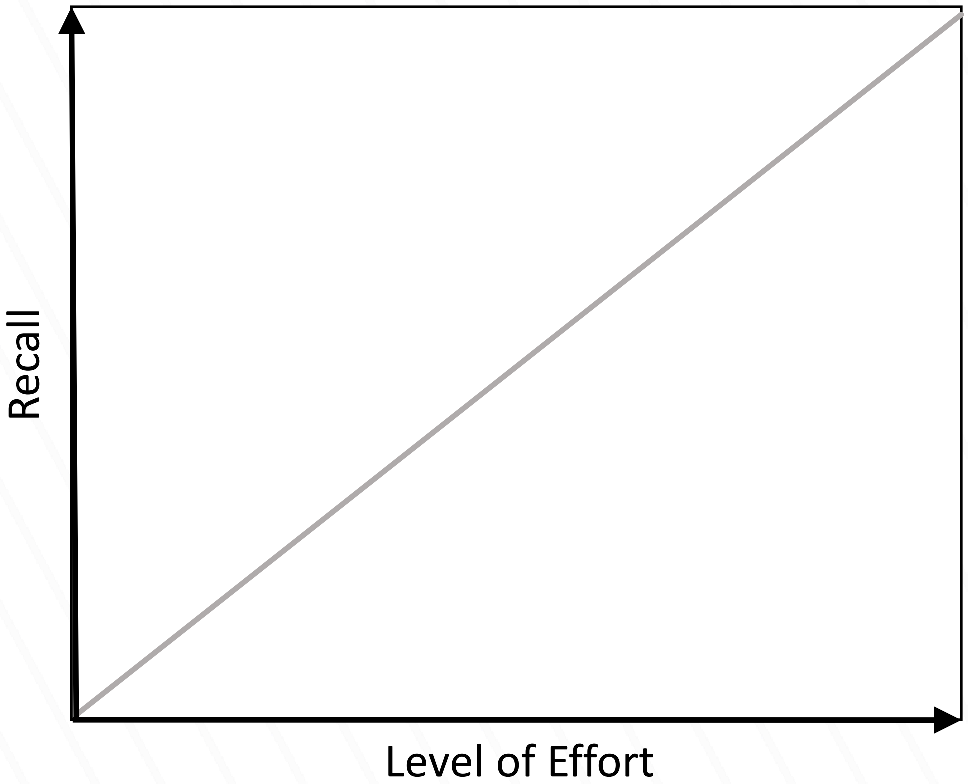



(Haddaway 2018)

# How is screening automation measured?

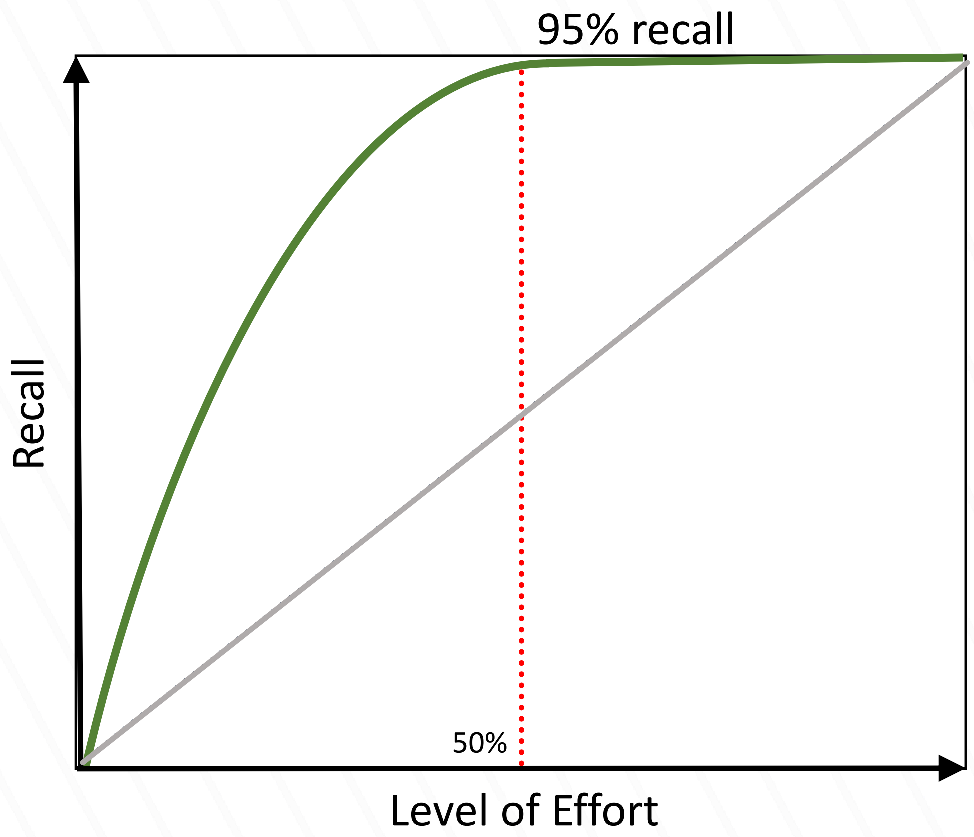
- **Recall/Sensitivity:** number of relevant reports identified divided by the total number of relevant reports in existence[15]
- **Precision/ Specificity:** number of relevant reports identified divided by the total number of reports identified[15]
- **F1 Score:** a weighted average of precision and recall[16]
- **WSS:** the reduction in workload in systematic review preparation when using a classifier[6]
- **AUR:** average workload across all recall levels[6]

# What would good automation performance look like?




Standard human performance where every article is screened by 2 reviewers

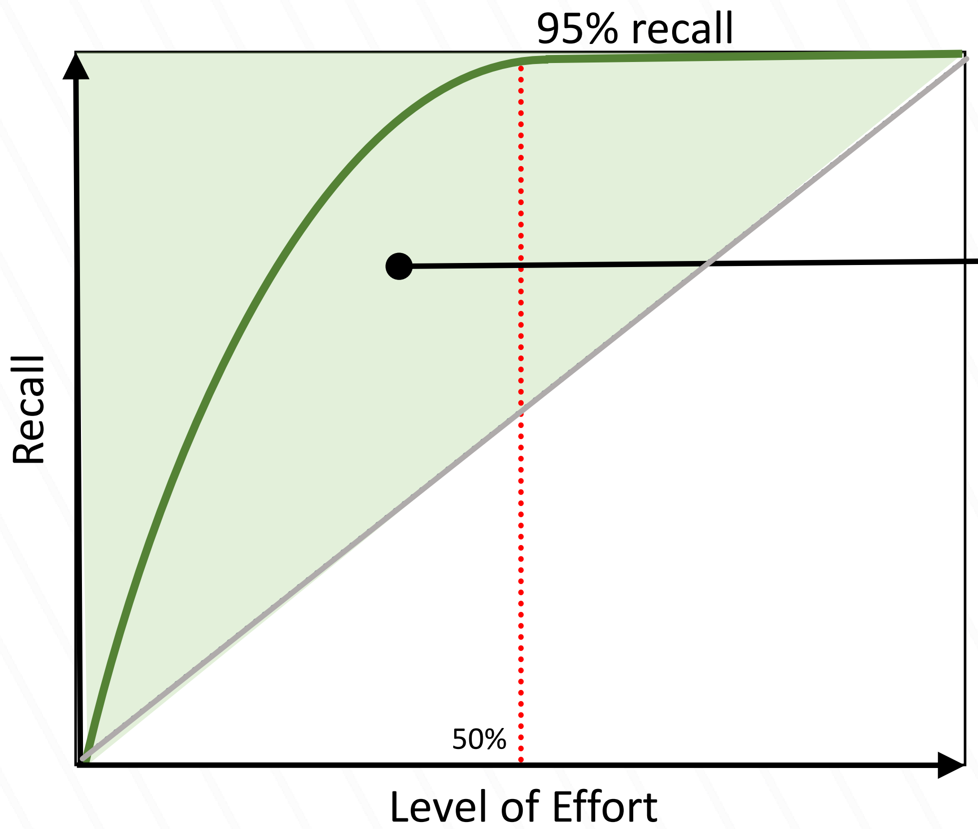
# What would good automation performance look like?




Ideally, adding automation would look something like:

approx. 95% of the relevant studies screened in 50% of the time

# What would good automation performance look like?




These approaches allow us to work in this spectrum, achieving high recall while minimizing level of effort.

# What would good automation performance look like?

- Wallace B, Noel-Storr A, Marshall I, Cohen A, Smalheiser N, Thomas J. (2017) [18]

- Retrospective simulation identifying randomized controlled trials using crowdsourcing (manual) vs. hybrid (manual + machine learning)

- **Manual approach**: combination of novices, experts, resolvers screen all citations
- **Hybrid**: computer screens out obvious non-RCTs, then novices, experts, resolvers screen

- 61,365 citations screened

# What would good automation performance look like? [18]

| | Novice Screener Decisions (cost x1) | Expert Screener Decisions (cost x2) | Conflict Resolver Decisions (cost x4) | Total cost units (cu) |
|---|---|---|---|---|
| Manual | 29,376 | 97,512 | 1,895 | 231,980 cu |
| Hybrid | | | | |
| Change in # of decisions | | | | |

# What would good automation performance look like?[18]

| | Novice Screener Decisions (cost x1) | Expert Screener Decisions (cost x2) | Conflict Resolver Decisions (cost x4) | Total cost units (cu) |
|---|---|---|---|---|
| Manual | 29,376 | 97,512 | 1,895 | 231,980 cu |
| Hybrid | 3,884 | 12,218 | 4,175 | 45, 020 cu |
| Change in # of decisions | | | | |

# What would good automation performance look like?[18]

| | Novice Screener Decisions (cost x1) | Expert Screener Decisions (cost x2) | Conflict Resolver Decisions (cost x4) | Total cost units (cu) |
|---|---|---|---|---|
| Manual | 29,376 | 97,512 | 1,895 | 231,980 cu |
| Hybrid | 3,884 | 12,218 | 4,175 | 45,020 cu |
| Change in # of decisions | ↓ 25,492 | ↓ 85,294 | ↑ 2,280 | |

# What would good automation performance look like? [18]

| | Novice Screener Decisions (cost x1) | Expert Screener Decisions (cost x2) | Conflict Resolver Decisions (cost x4) | Total cost units (cu) |
|---|---|---|---|---|
| Manual | 29,376 | 97,512 | 1,895 | 231,980 cu |
| Hybrid | 3,884 | 12,218 | 4,175 | 45,020 cu |
| Change in # of decisions | 25,492 | 85,294 | 2,280 | 186,960 cu |

# What have previous studies found?

- “Most suggested that a saving in workload of between **30%** and **70%** might be possible (with some a little higher or a little lower than this), though sometimes the saving in workload is accompanied by the loss of 5% of relevant studies (i.e., a 95% recall).” [34]

- **Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan**. (Olofsson H, Brolund A, Hellberg C, et al. 2017) [19]
- **Machine Learning Versus Standard Techniques for Updating Searches for Systematic Reviews: A Diagnostic Accuracy Study**. (Shekelle PG, Shetty K, Newberry S, Maglione M, Motala A.  2017) [20]
- **Towards automating the initial screening phase of a systematic review.** (Bekhuis T, Demner-Fushman D. 2010) [21]
- **Towards Automatic Recognition of Scientifically Rigorous Clinical Research Evidence**. (Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes BR. 2009) [22]
- **Reducing workload in systematic review preparation using automated citation classification.** (Cohen AM, Hersh WR, Peterson K, Yen P-Y. 2006) [23]
- **Text categorization models for high-quality article retrieval in internal medicine.** (Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. 2005)[24]

# So where is it?

- First success with automation in SR in 2006- over 10 years ago![25]
- Reducing workload in systematic review preparation using automated citation classification. Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006).[23]

# So where is it?

- First success with automation in SR in 2006- over 10 years ago![25]
- Reducing workload in systematic review preparation using automated citation classification. Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006).[23]

- Rogers' Diffusion of Innovation model provides insight[25, 26, 27]

1. ***Perceived relative advantage*** (does it appear to have benefits to the user?)
2. ***Compatibility*** (is it consistent with past experiences and the needs/values of the user?)
3. ***Trialability*** (can the user try it out in their own work?)
4. ***Observability*** (are the results of the innovation visible to others?)
5. ***Complexity*** (is it perceived as easy to understand and use?)

(Stansfield et al 2015) [27]

# So where is it?

- First success with automation in SR in 2006- over 10 years ago! [25]
  - Reducing workload in systematic review preparation using automated citation classification. Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006).[23]

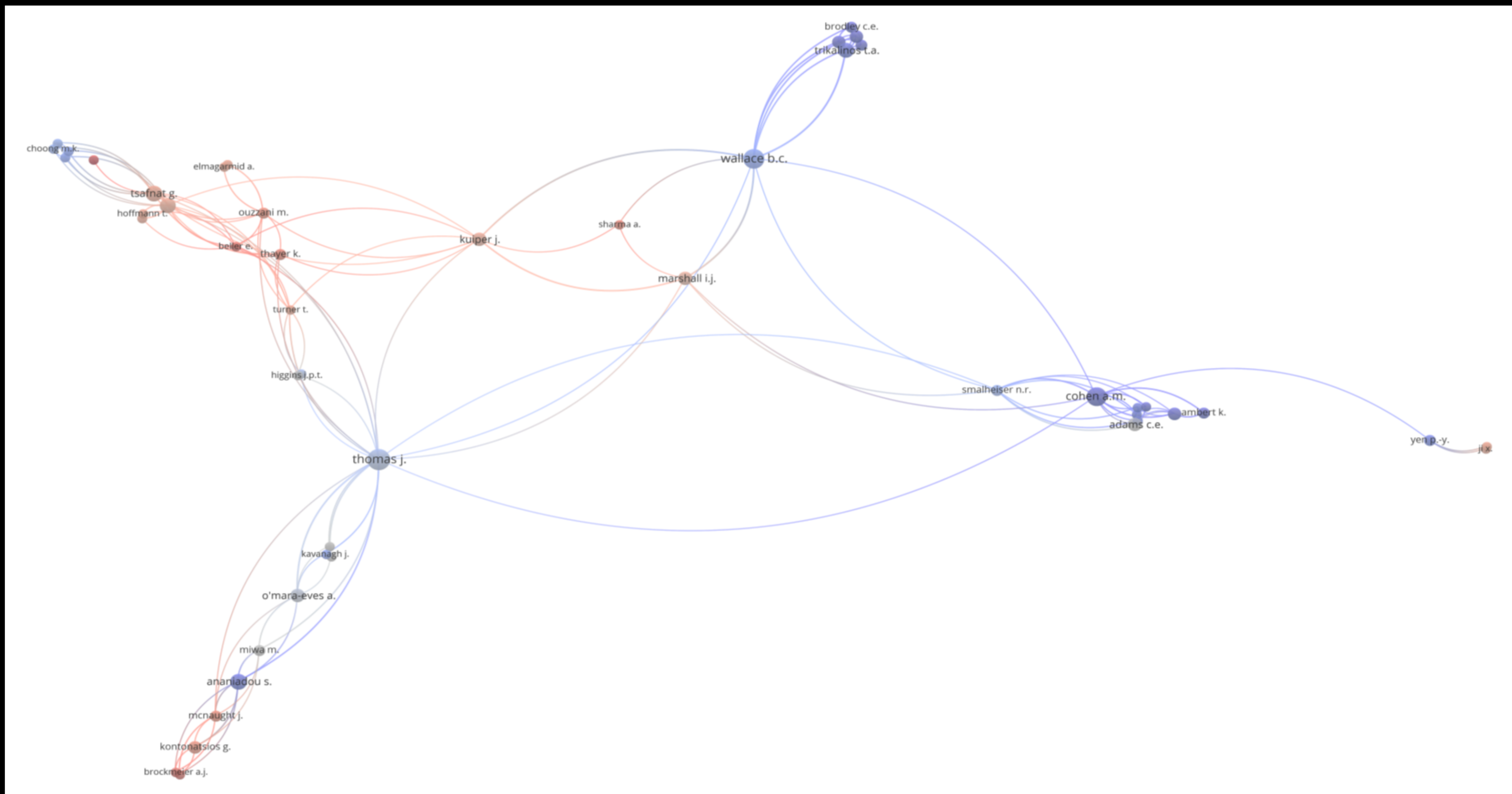
- Rogers' Diffusion of Innovation model provides insight[25, 26, 27]

- The nature of our field- we're busy!

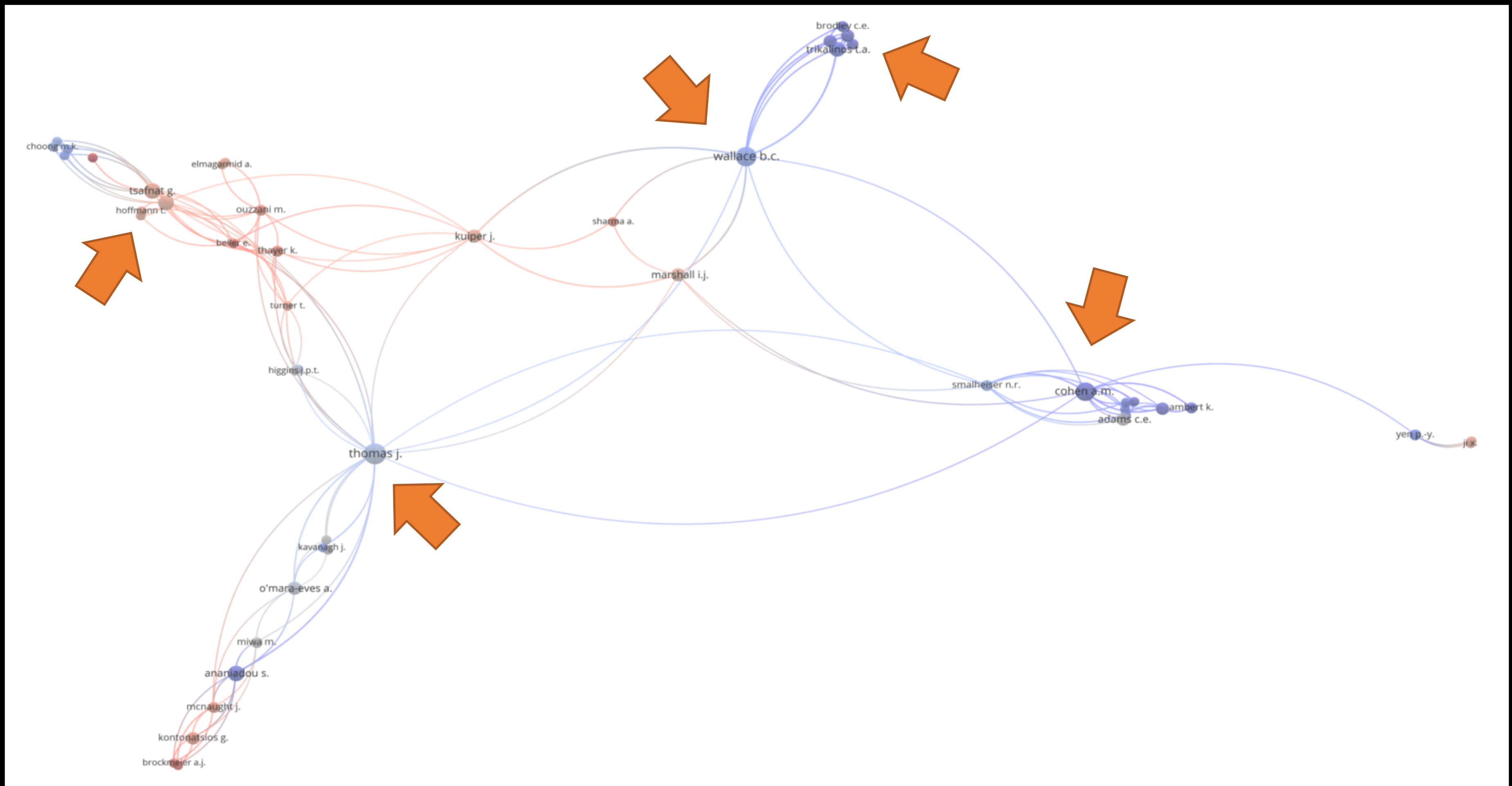
1. ***Perceived relative advantage*** (does it appear to have benefits to the user?)
2. ***Compatibility*** (is it consistent with past experiences and the needs/values of the user?)
3. ***Trialability*** (can the user try it out in their own work?)
4. ***Observability*** (are the results of the innovation visible to others?)
5. ***Complexity*** (is it perceived as easy to understand and use?)

(Stansfield et al 2015) [27]

# Who publishes the literature?



# Who publishes the literature?



# Who publishes the literature?

- Thomas
- O'Mara-Eves
- Glasziou
- Adams
- C. Marshall
- Trikalinos

- Wallace
- Cohen
- Ananiadou
- Brereton
- Felizardo
- Jonnalagadda
- Brodley
- Tsafnat

- I. Marshall
- Elliott

# Who publishes the literature?

Evidence-Based Practice

- Thomas: EPPI-Centre
- O'Mara-Eves: EPPI-Centre
- Glasziou: Centre for Research in Evidence-Based Practice
- Adams: Cochrane
- C. Marshall: York Health Economics Consortium
- Trikalinos: Health Services, Policy and Practice

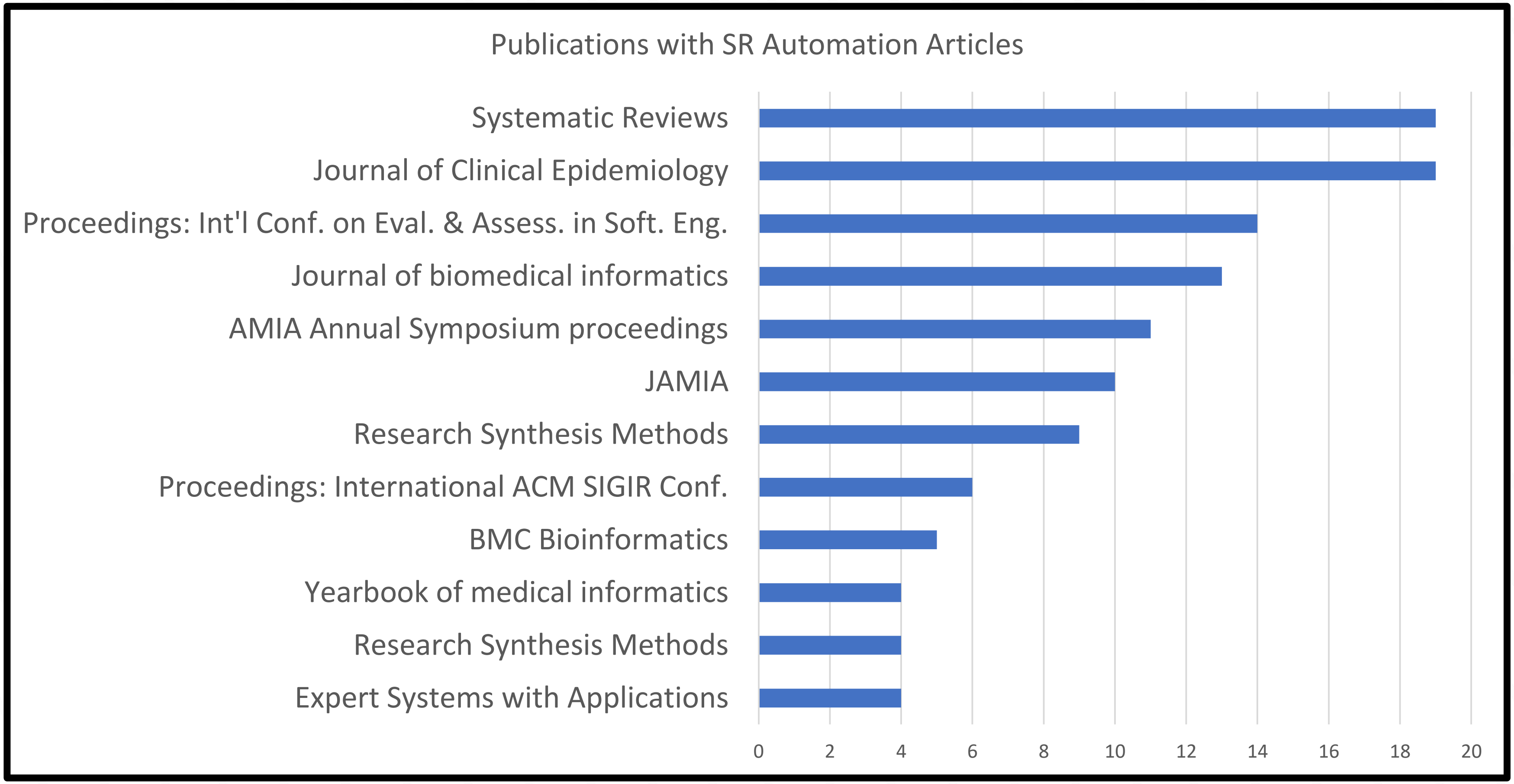
Public Health

- I. Marshall: primary care & public health
- Elliott: public health

Computer Science

- Wallace: computer science
- Cohen: medical informatics
- Ananiadou: National Centre for Text Mining
- Brereton: computing & mathematics
- Felizardo: computer science
- Jonnalagadda: Microsoft
- Brodley: computer science
- Tsafnat: Australian Institute of Health Innovation

# Where are they publishing?



# Challenges with the technology

- Incorrect classifications[28]
  - False negatives
  - Hasty generalization
- More confidence: less effort, better precision, worse recall[13]

- Cost/benefit
- Buy-in from review team, publishers, others
- Limited ability to observe tools in action
- Limited validation studies
- Not sure how the tools work
- Learning curve/ Requires coding experience[25]

# Tools that are free & ready/easy to implement

- Abstrakr[28]
- Colandr
- Cadima[29]
- Rayyan[19]
- RobotAnalyst[5]
- Swift Review[30]

SR TOOL BOX

systematicreviewtools.com

# Comparisons of SR automation tools

1. **Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools**. Kohl C, McIntosh EJ, Unger S, et al. 2018.[29]

2. **EPC Methods: An Exploration of the Use of Text-Mining Software in Systematic Reviews.** Paynter R, Banez LL, Berliner E, et al. 2016.[28]

3. **Tool support for systematic reviews in software engineering**. (Dissertation) Marshall C. 2016. [31]

   - **Tools to support systematic reviews in software engineering: a feature analysis**. Marshall C, Brereton P, Kitchenham B. 2014. [32]
   - **Tools to support systematic reviews in software engineering: a cross-domain survey using semi-structured interviews**. Marshall C, Brereton P, Kitchenham B. 2015. [33]

4. **Using text mining for study identification in systematic reviews: a systematic review of current approaches**. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. 2015. [34]

5. **Systematic literature review (SLR) automation: A systematic literature review**. Hamad Z, Salim N. 2014. [35]

# Future challenges

- Encouraging potential reviewers that other types of reviews may be more appropriate for their needs[25]
- Building partnerships across disciplines to test tools[34] including clinical and technical literature, as well as social science and theoretical[25]
  - Testing a variety of review topics from many disciplines
  - Comparing non-automated reviews to automated reviews
  - Testing a variety of levels of automation integration
- Developing or prompting the development of tool improvements[25]

# Questions?

For reference list, visit:

go.unc.edu/ai-refs

Elizabeth (Beth) Moreton

emoreton@email.unc.edu