Query Automation for Systematic Reviews

Harry Scells
Leipzig University
https://scells.me

Glasgow IR Seminar · July 10, 2023

How do clinicians become informed about how to treat their patients?

How do governments and institutions make health policy decisions?

How do clinicians become informed about how to treat their patients?

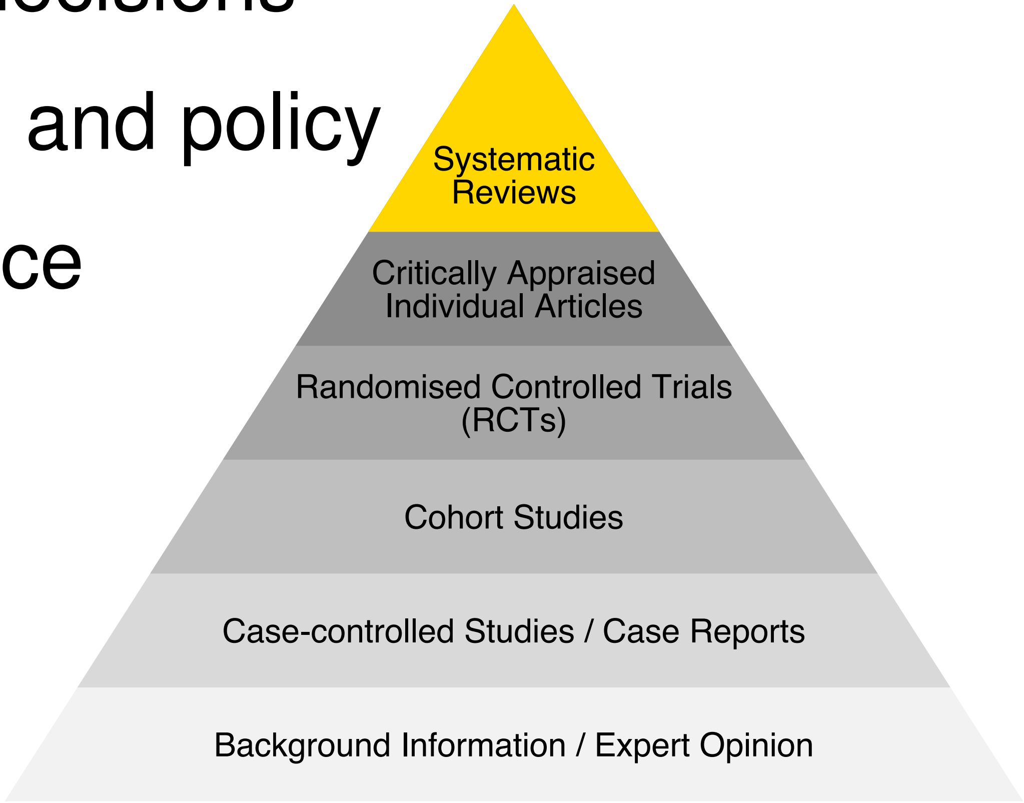
Systematic Reviews

How do governments and institutions make health policy decisions?

Systematic Reviews

Overview

- **Guide** clinical decisions
- **Inform** practice and policy
- **Provide** evidence



Systematic Reviews

Critically Appraised Individual Articles

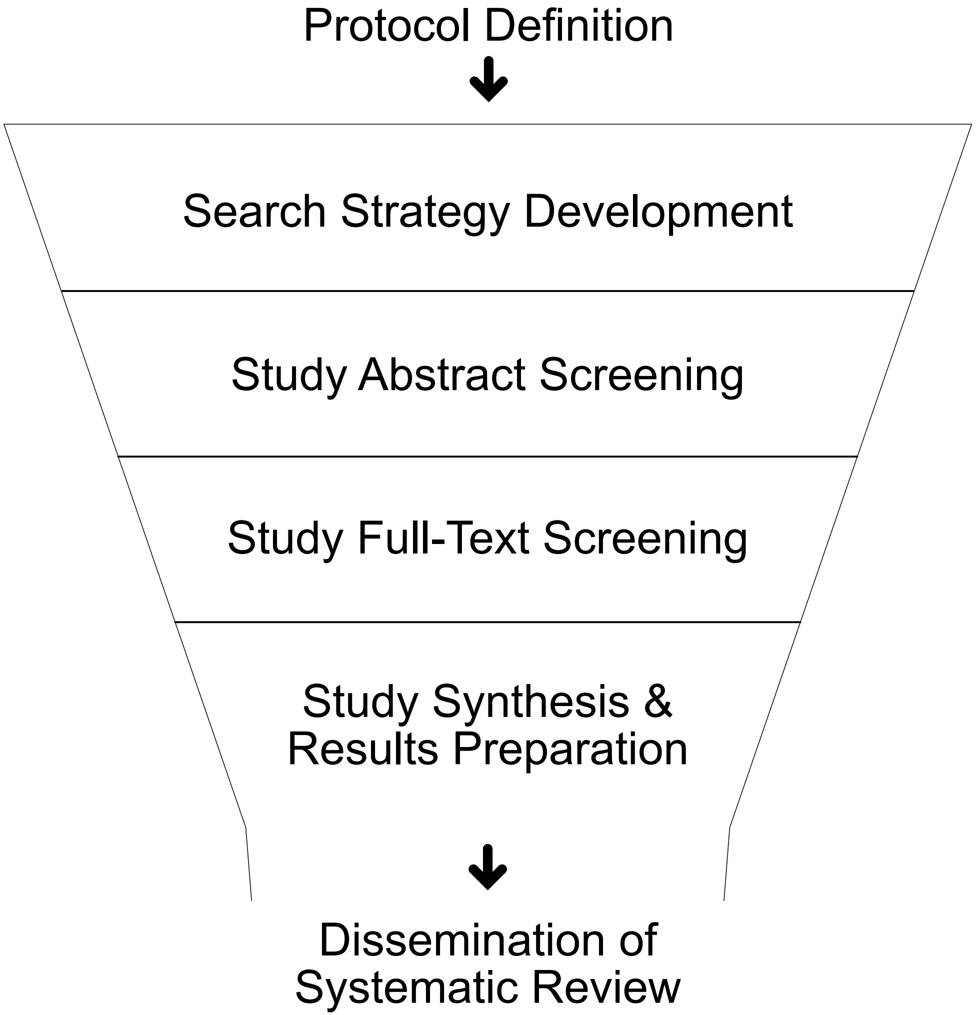
Randomised Controlled Trials (RCTs)

Cohort Studies

Case-controlled Studies / Case Reports

Background Information / Expert Opinion

Systematic review creation is hard!



Protocol Definition

Search Strategy Development

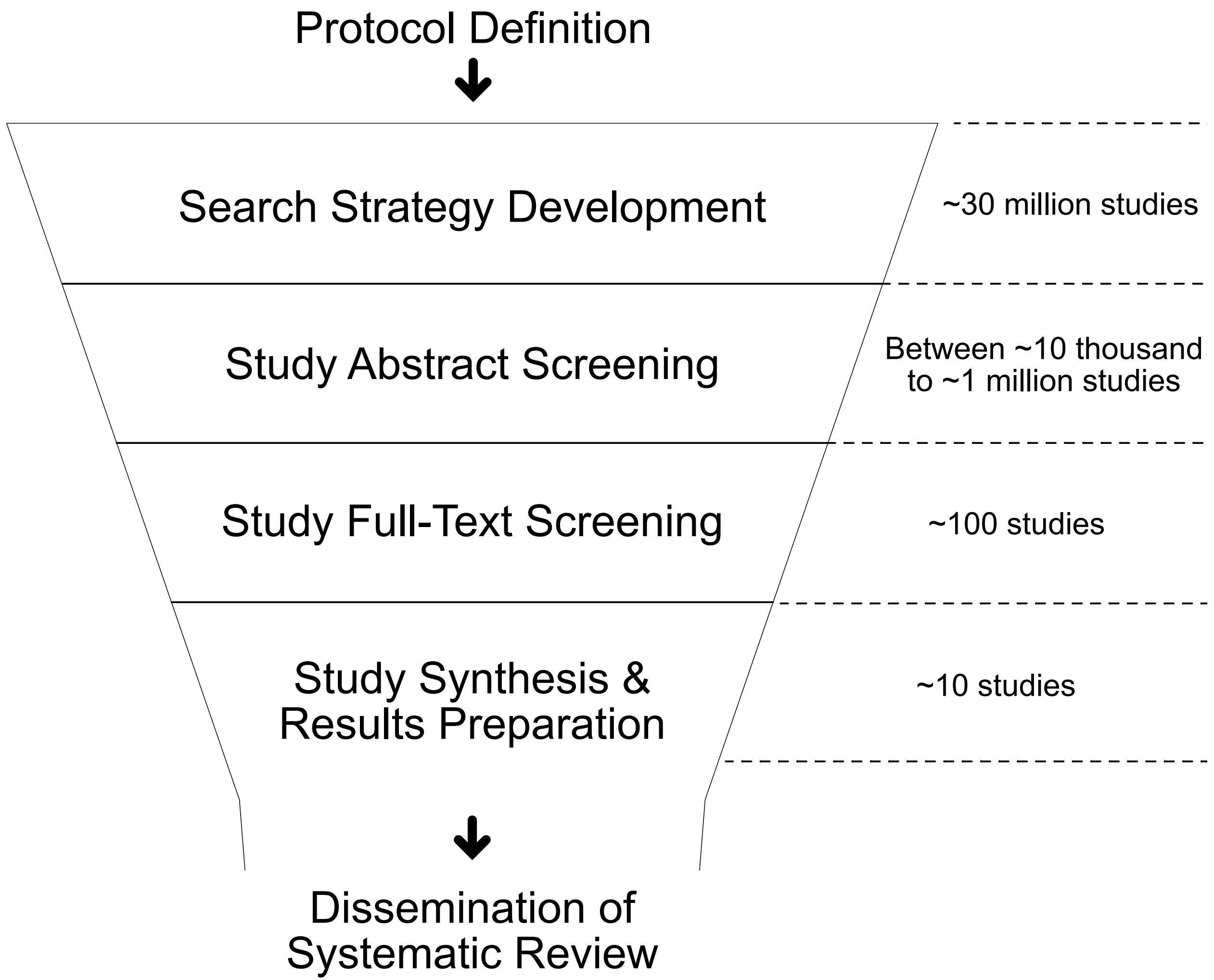
Study Abstract Screening

Study Full-Text Screening

Study Synthesis &
Results Preparation

Dissemination of
Systematic Review

Why is systematic review creation hard?



Protocol Definition

Search Strategy Development

~30 million studies

Study Abstract Screening

Between ~10 thousand to ~1 million studies

Study Full-Text Screening

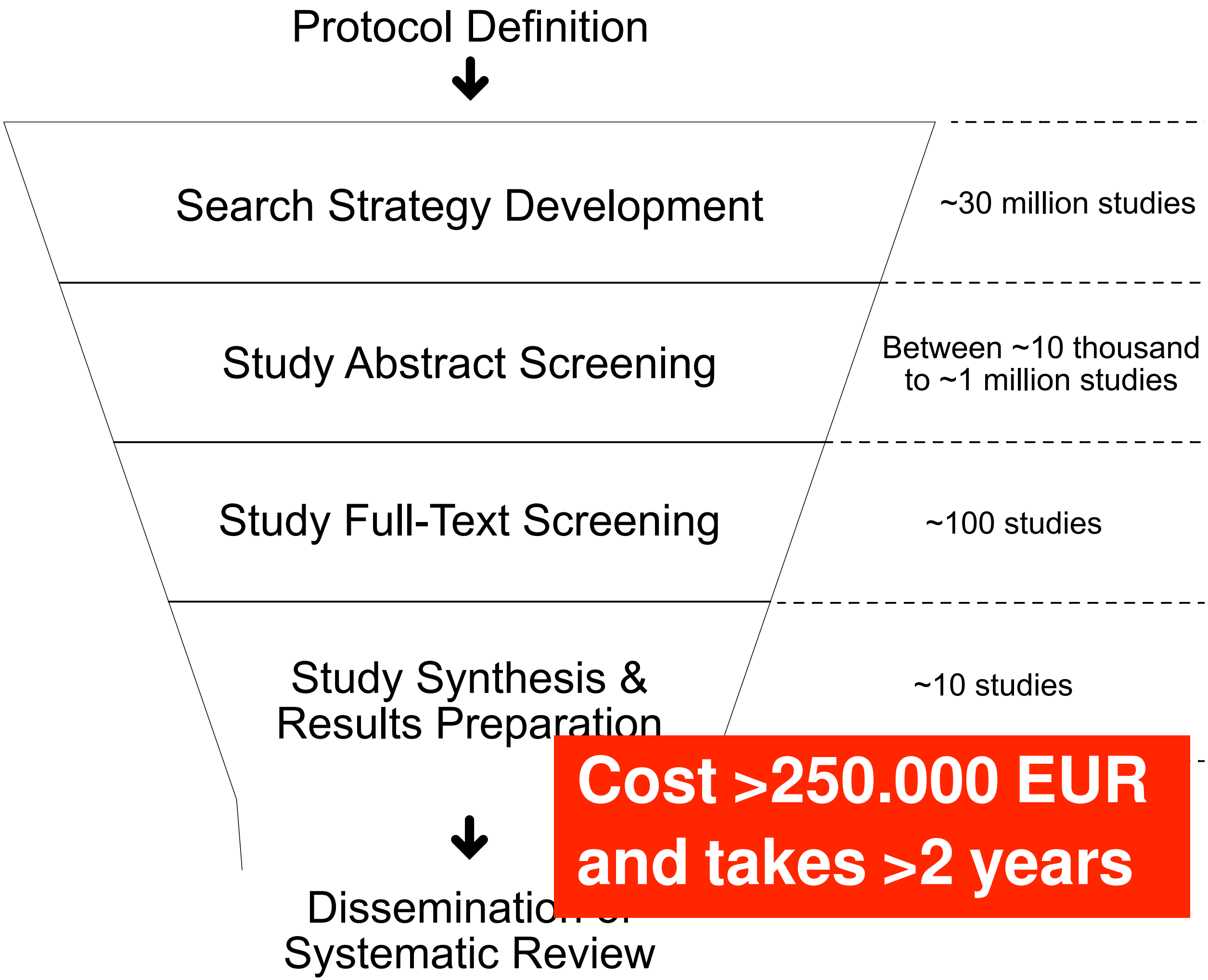
~100 studies

Study Synthesis & Results Preparation

~10 studies

Dissemination of Systematic Review

Why is systematic review creation hard?



Protocol Definition

Search Strategy Development

~30 million studies

Study Abstract Screening

Between ~10 thousand
to ~1 million studies

Study Full-Text Screening

~100 studies

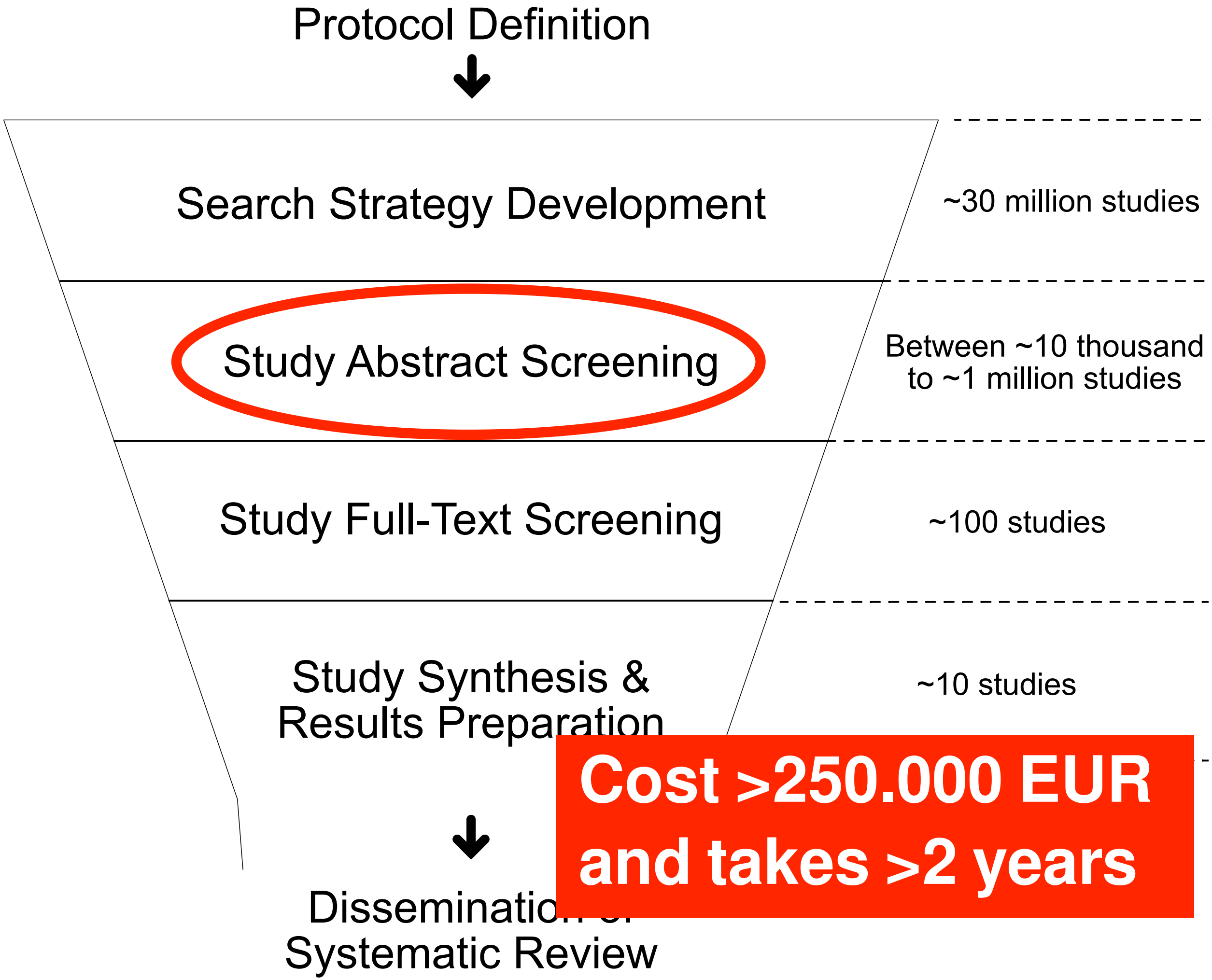
Study Synthesis &
Results Preparation

~10 studies

Disseminatio
Systematic Review

Cost >250.000 EUR
and takes >2 years

Why is systematic review creation hard?



Protocol Definition

Search Strategy Development

~30 million studies

Study Abstract Screening

Between ~10 thousand
to ~1 million studies

Study Full-Text Screening

~100 studies

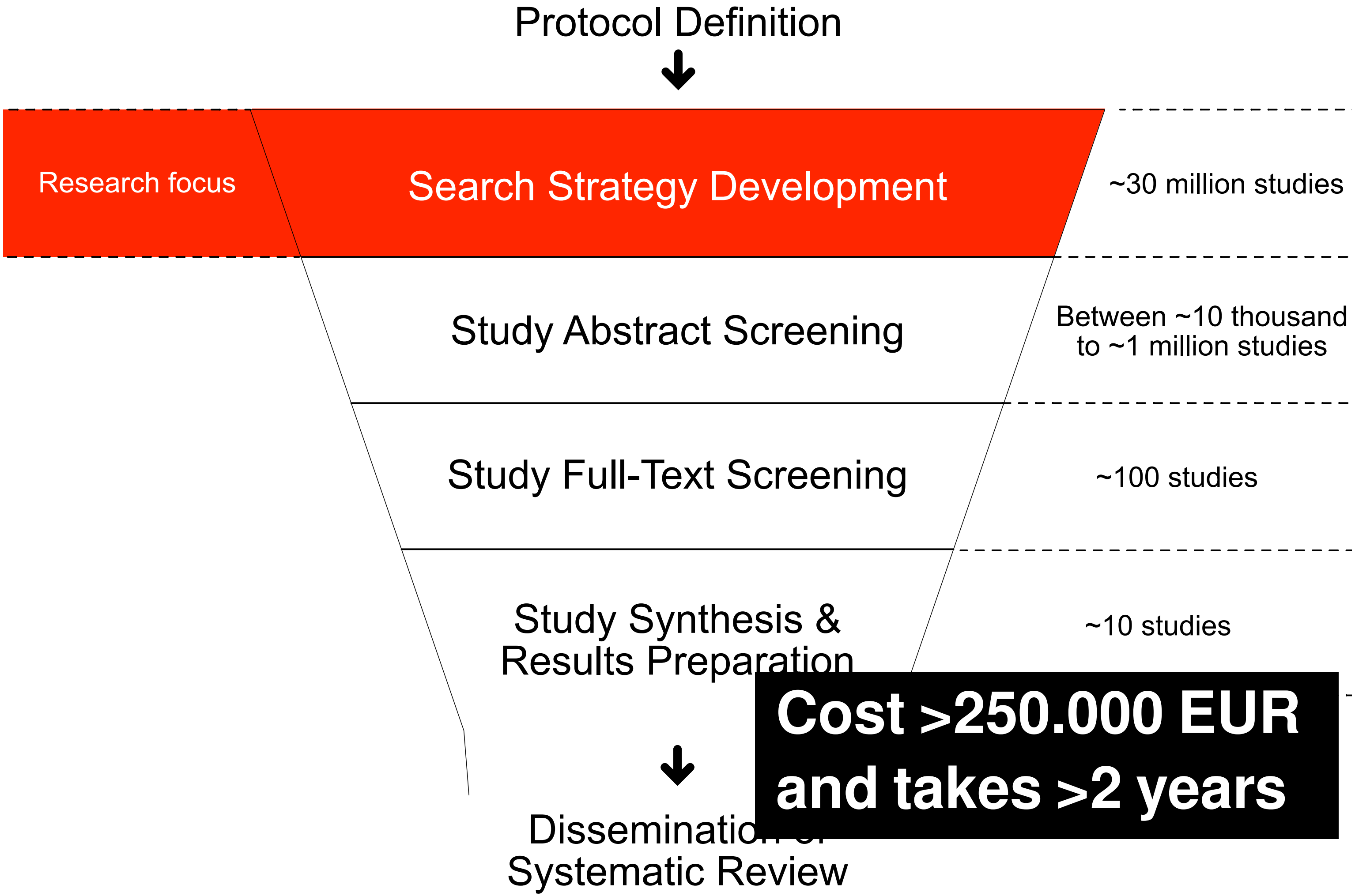
Study Synthesis &
Results Preparation

~10 studies

Disseminatio
Systematic Review

**Cost >250.000 EUR
and takes >2 years**

Why is systematic review creation hard?



Protocol Definition

Research focus

Search Strategy Development

~30 million studies

Study Abstract Screening

Between ~10 thousand to ~1 million studies

Study Full-Text Screening

~100 studies

Study Synthesis & Results Preparation

~10 studies

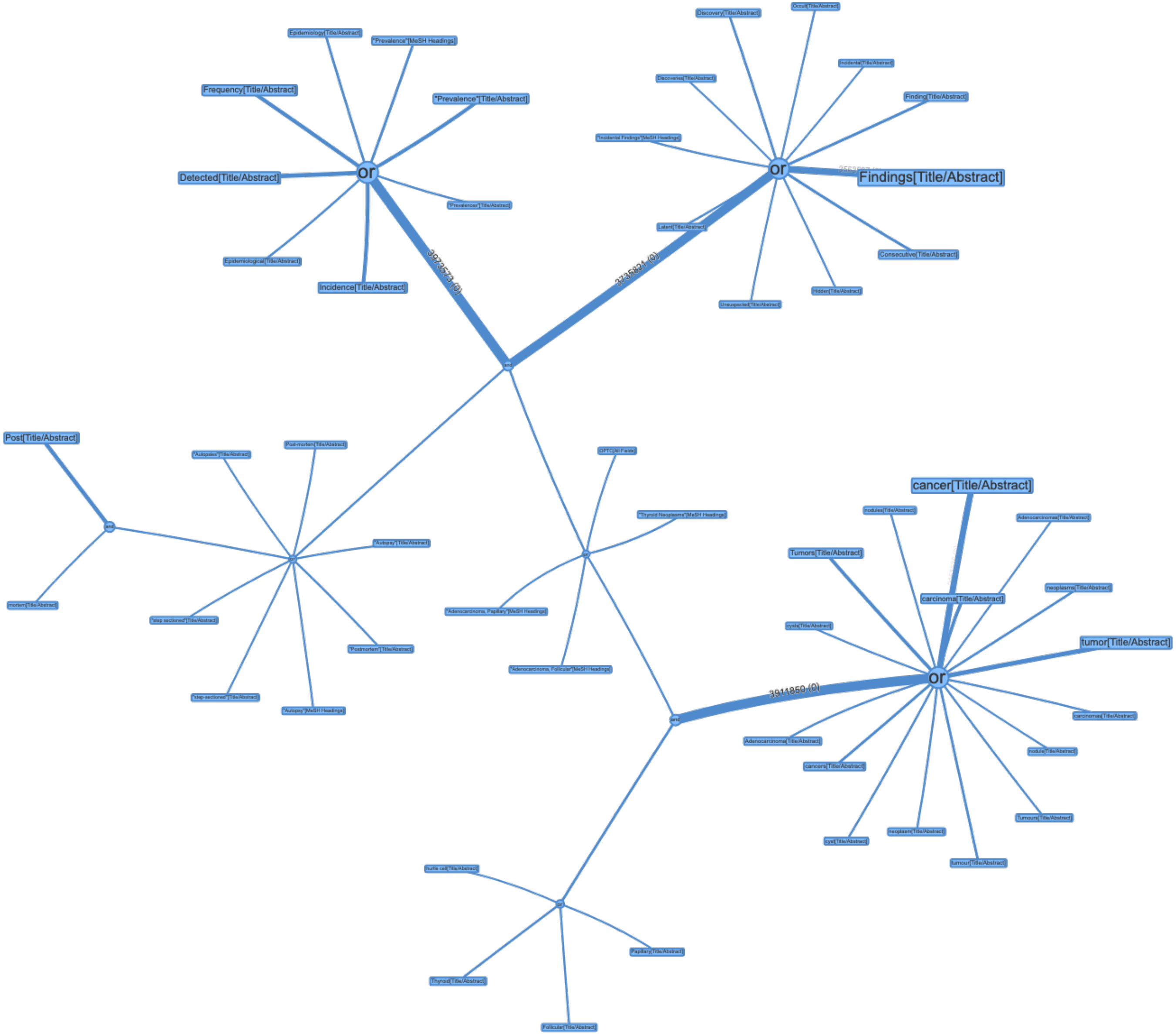
Dissemination of Systematic Review

Cost >250.000 EUR and takes >2 years

Why such little research on queries?

```
 (("Thyroid Neoplasms"[MeSH] OR "Adenocarcinoma, Follicular"[MeSH]
OR "Adenocarcinoma, Papillary"[MeSH] OR OPTC OR ((Thyroid[tiab]
OR Follicular[tiab] OR Papillary[tiab] OR hurtle cell[tiab]) AND
(cancer[tiab] OR cancers[tiab] OR carcinoma[tiab] OR carcinomas[tiab]
OR Adenocarcinoma[tiab] OR Adenocarcinomas[tiab] OR neoplasm[tiab]
OR neoplasms[tiab] OR nodule[tiab] OR nodules[tiab] OR tumor[tiab]
OR tumour[tiab] OR Tumors[tiab] OR Tumours[tiab] OR cyst[tiab]
OR cysts[tiab]))) AND ("Autopsy"[MeSH] OR "Autopsy"[tiab] OR
"Autopsies"[tiab] OR "Postmortem"[tiab] OR Post-mortem[tiab] OR
"step-sectioned"[tiab] OR "step sectioned"[tiab] OR (Post[tiab]
AND mortem[tiab])) AND (Prevalence"[MeSH] OR Prevalence"[tiab] OR
Prevalences"[tiab] OR Incidence[tiab] OR Epidemiology[tiab] OR
Epidemiological[tiab] OR Frequency[tiab] OR Detected[tiab]) AND
("Incidental Findings"[MeSH] OR Incidental[tiab] OR Unsuspected[tiab] OR
Discovery[tiab] OR Discoveries[tiab] OR Findings[tiab] OR Finding[tiab]
OR Occult[tiab] OR Hidden[tiab] OR Latent[tiab] OR Consecutive[tiab]))
```

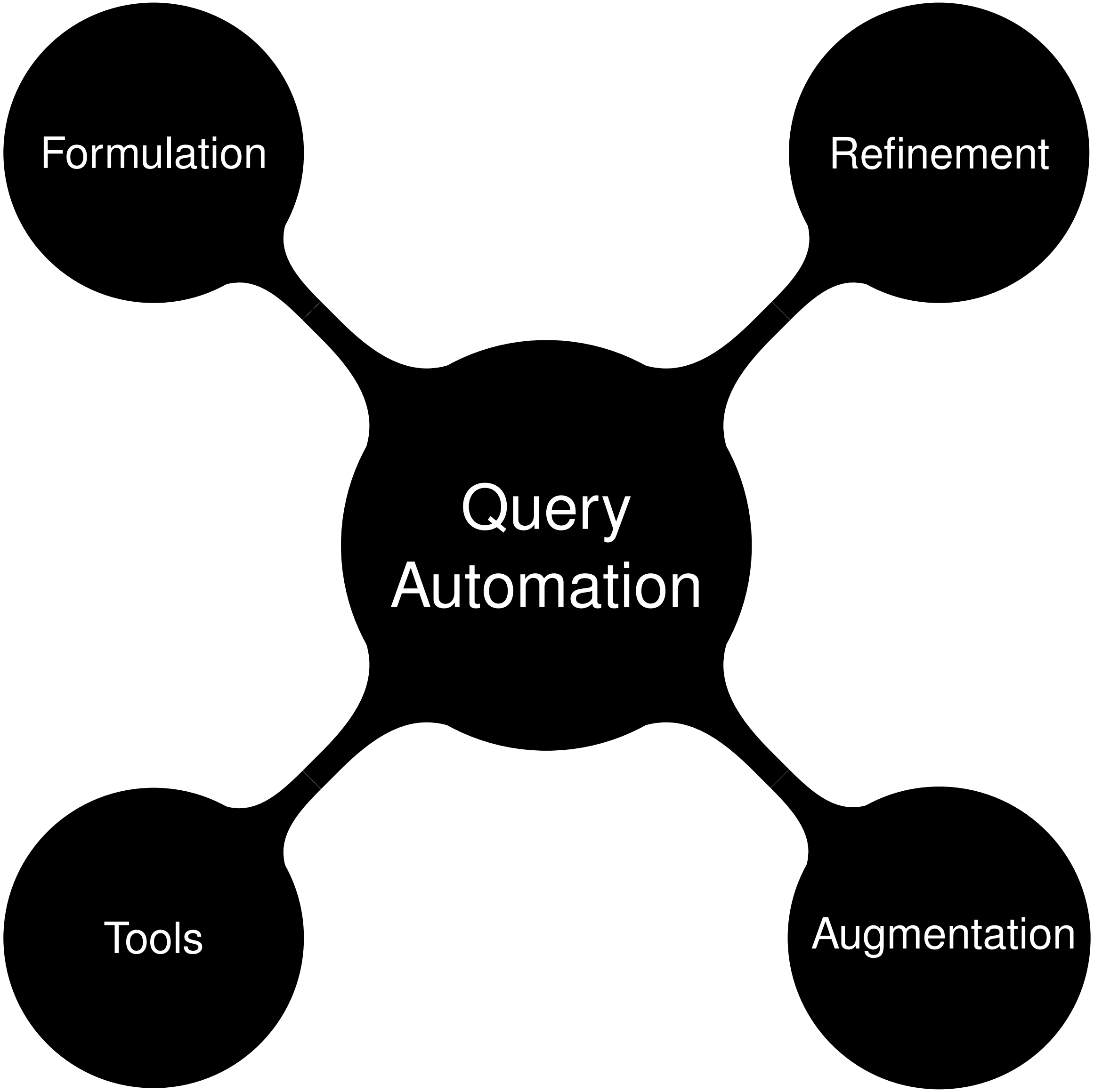
Why such little research on queries?


Epidemiology[Title/Abstract]
"Prevalence"[MeSH Headings]
Frequency[Title/Abstract]
"Prevalence"[Title/Abstract]
Detected[Title/Abstract]
or
"Prevalences"[Title/Abstract]
Epidemiological[Title/Abstract]
Incidence[Title/Abstract]
3973573 (0)
Discovery[Title/Abstract]
Occult[Title/Abstract]
Incidental[Title/Abstract]
Discoveries[Title/Abstract]
Finding[Title/Abstract]
"Incidental Findings"[MeSH Headings]
or
Findings[Title/Abstract]
Latent[Title/Abstract]
Consecutive[Title/Abstract]
3735821 (0)
Hidden[Title/Abstract]
Unsuspected[Title/Abstract]
and
Post[Title/Abstract]
"Autopsies"[Title/Abstract]
Post-mortem[Title/Abstract]
and
"Autopsy"[Title/Abstract]
or
mortem[Title/Abstract]
"step sectioned"[Title/Abstract]
"Postmortem"[Title/Abstract]
"step-sectioned"[Title/Abstract]
"Autopsy"[MeSH Headings]
OPTC[All Fields]
"Thyroid Neoplasms"[MeSH Headings]
or
"Adenocarcinoma, Papillary"[MeSH Headings]
"Adenocarcinoma, Follicular"[MeSH Headings]
cancer[Title/Abstract]
nodules[Title/Abstract]
Adenocarcinomas[Title/Abstract]
Tumors[Title/Abstract]
carcinoma[Title/Abstract]
neoplasms[Title/Abstract]
cysts[Title/Abstract]
tumor[Title/Abstract]
or
3911850 (0)
and
carcinomas[Title/Abstract]
Adenocarcinoma[Title/Abstract]
nodule[Title/Abstract]
cancers[Title/Abstract]
Tumours[Title/Abstract]
neoplasm[Title/Abstract]
cyst[Title/Abstract]
tumour[Title/Abstract]
hurtle cell[Title/Abstract]
or
Papillary[Title/Abstract]
Thyroid[Title/Abstract]
Follicular[Title/Abstract]

Why are Boolean queries used?

Reproducibility ➜ double check screening

Understandability ➜ control set size



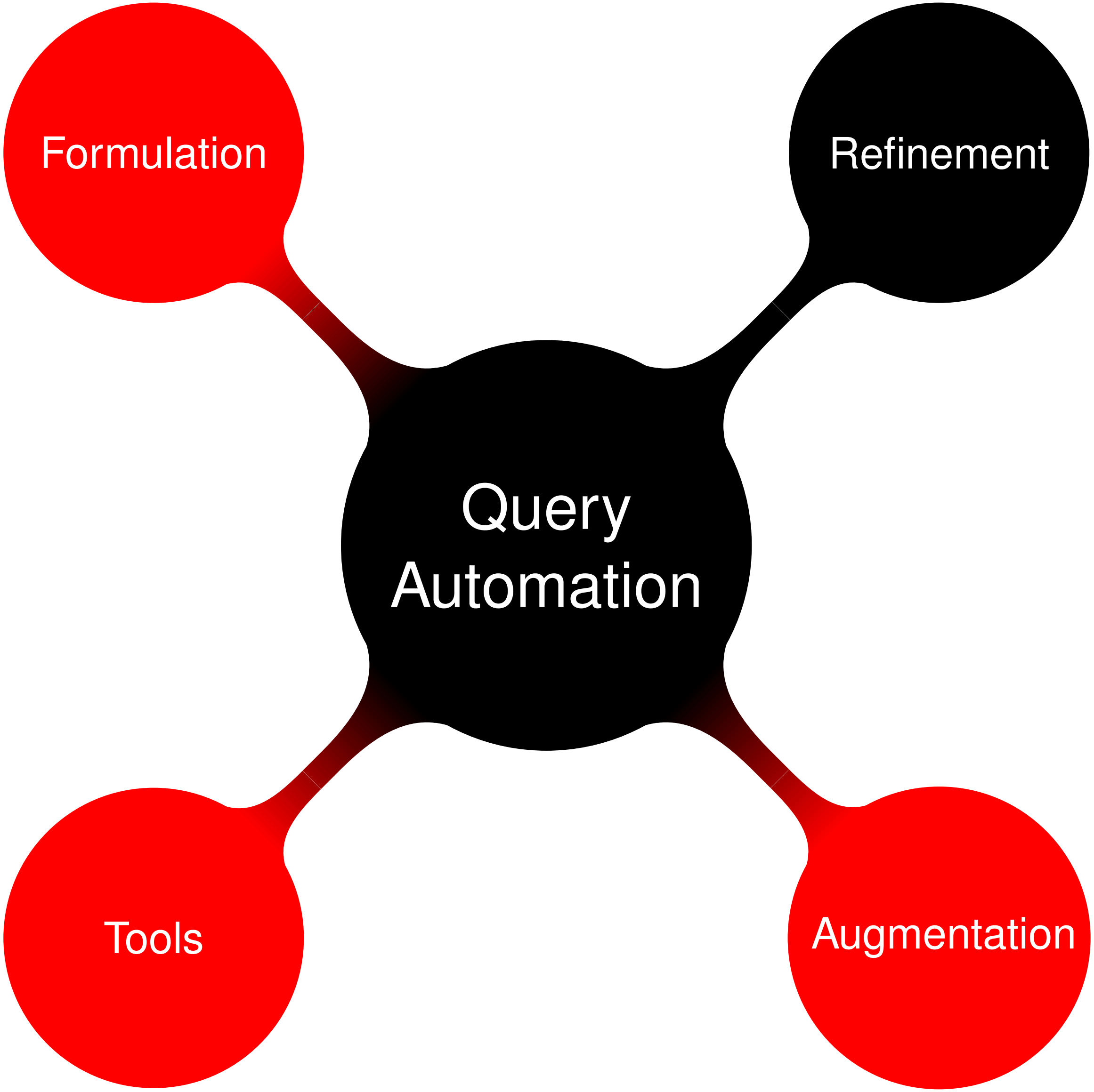
Formulation

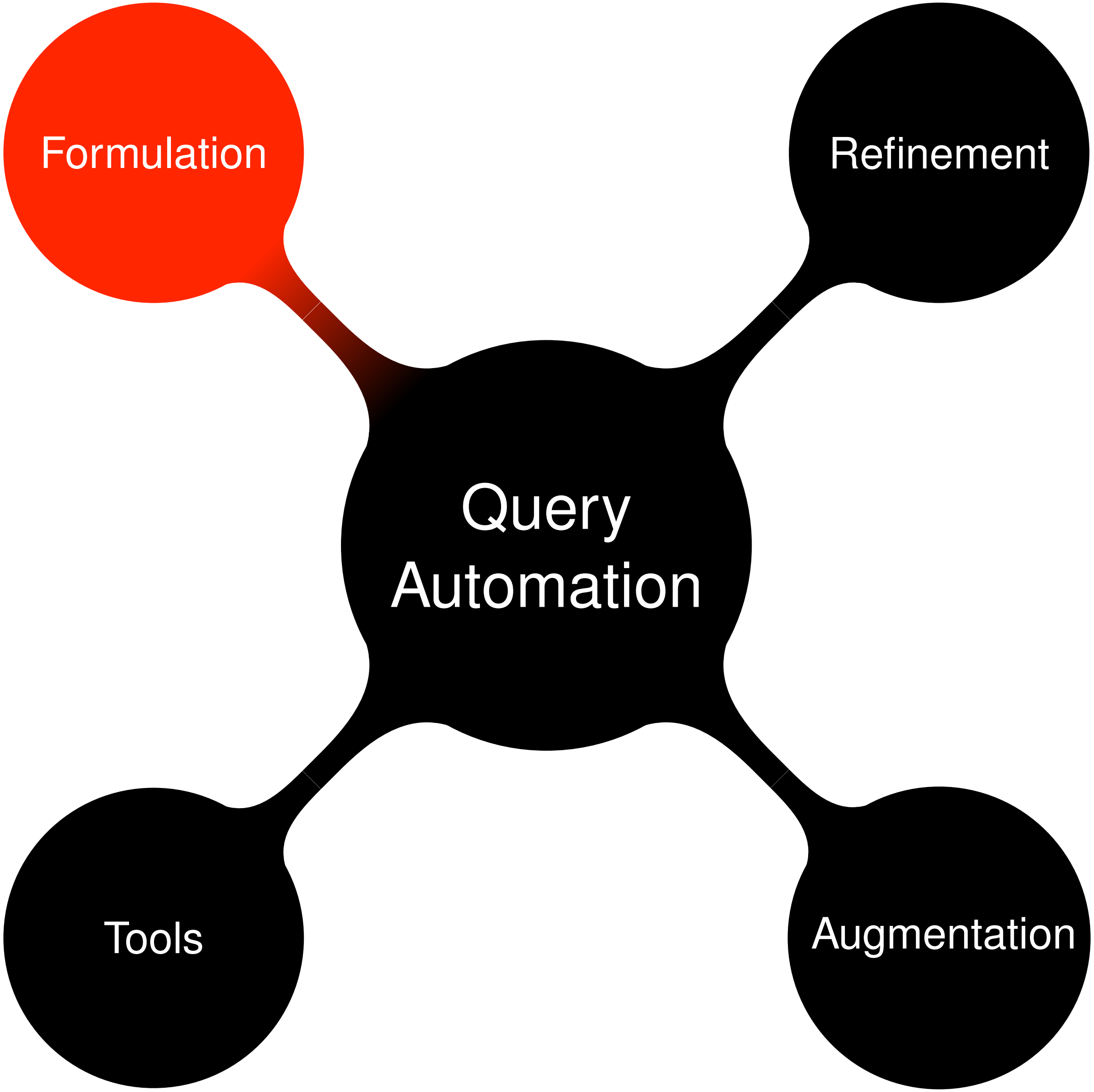
Refinement

Query Automation

Tools

Augmentation


Formulation
Refinement
Query Automation
Tools
Augmentation


Formulation
Refinement
Query Automation
Tools
Augmentation

Query Formulation

The automatic creation of complex queries for the task of systematic review literature search

- ❑ Content covered
  - – Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023 **(to appear at SIGIR'23)**
- ❑ Further reading
  - – Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. Automatic boolean query formulation for systematic review literature search. In *Proceedings of the 29th World Wide Web Conference*, pages 1071–1081, 2020
  - – Harrisen Scells, Guido Zuccon, and Bevan Koopman. A computational approach for objectively derived systematic review search strategies. In *Proceedings of the 42nd European Conference on Information Retrieval*, pages 385–398, 2020
  - – Harrisen Scells, Guido Zuccon, and Bevan Koopman. A comparison of automatic boolean query formulation for systematic reviews. *Information Retrieval Journal*, pages 1–26, 2020

How humans formulate queries

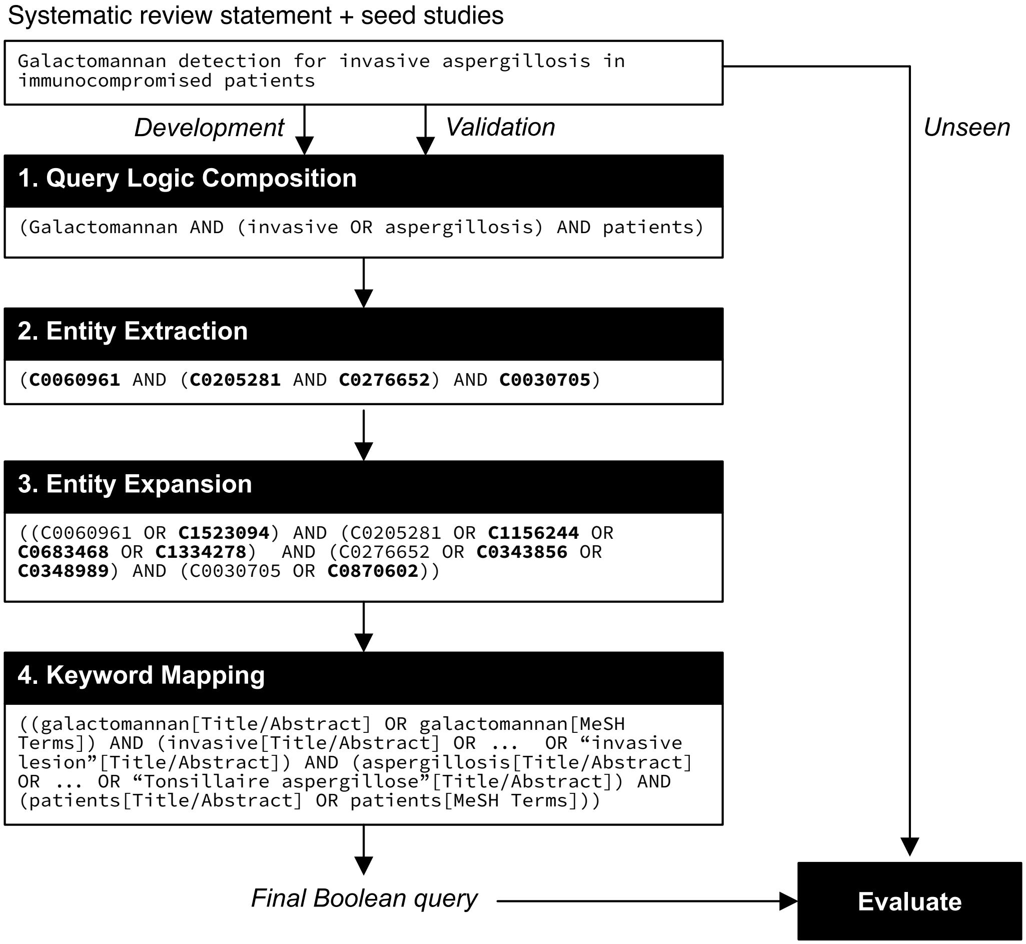
Overview

Conceptual method [Clark 2013] ➜ Human expertise

Objective method [Hausner et al. 2012] ➜ More algorithmic

Both methods ➜ Seed studies

Automating the conceptual method



Systematic review statement + seed studies

Galactomannan detection for invasive aspergillosis in immunocompromised patients

Development

Validation

Unseen

1. Query Logic Composition

(Galactomannan AND (invasive OR aspergillosis) AND patients)

2. Entity Extraction

(**C0060961** AND (**C0205281** AND **C0276652**) AND **C0030705**)

3. Entity Expansion

((C0060961 OR **C1523094**) AND (C0205281 OR **C1156244** OR **C0683468** OR **C1334278**) AND (C0276652 OR **C0343856** OR **C0348989**) AND (C0030705 OR **C0870602**))

4. Keyword Mapping

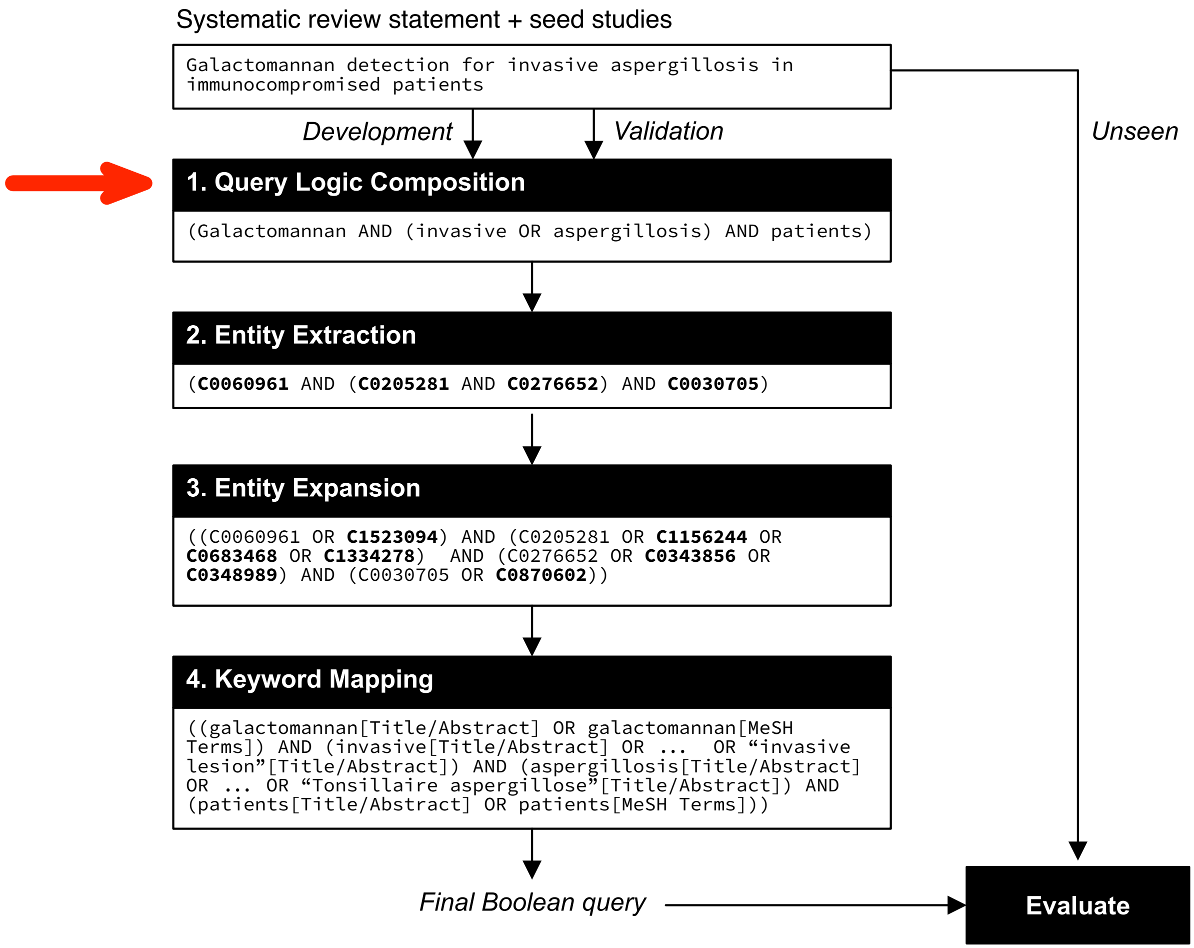
((galactomannan[Title/Abstract] OR galactomannan[MeSH Terms]) AND (invasive[Title/Abstract] OR ... OR “invasive lesion”[Title/Abstract]) AND (aspergillosis[Title/Abstract] OR ... OR “Tonsillaire aspergillose”[Title/Abstract]) AND (patients[Title/Abstract] OR patients[MeSH Terms]))

Final Boolean query

Evaluate

High level concepts ➜ broaden search ➜ iterate until satisfied

Automating the conceptual method



Systematic review statement + seed studies

Galactomannan detection for invasive aspergillosis in immunocompromised patients

Development

Validation

Unseen

1. Query Logic Composition

(Galactomannan AND (invasive OR aspergillosis) AND patients)

2. Entity Extraction

(**C0060961** AND (**C0205281** AND **C0276652**) AND **C0030705**)

3. Entity Expansion

((C0060961 OR **C1523094**) AND (C0205281 OR **C1156244** OR **C0683468** OR **C1334278**) AND (C0276652 OR **C0343856** OR **C0348989**) AND (C0030705 OR **C0870602**))

4. Keyword Mapping

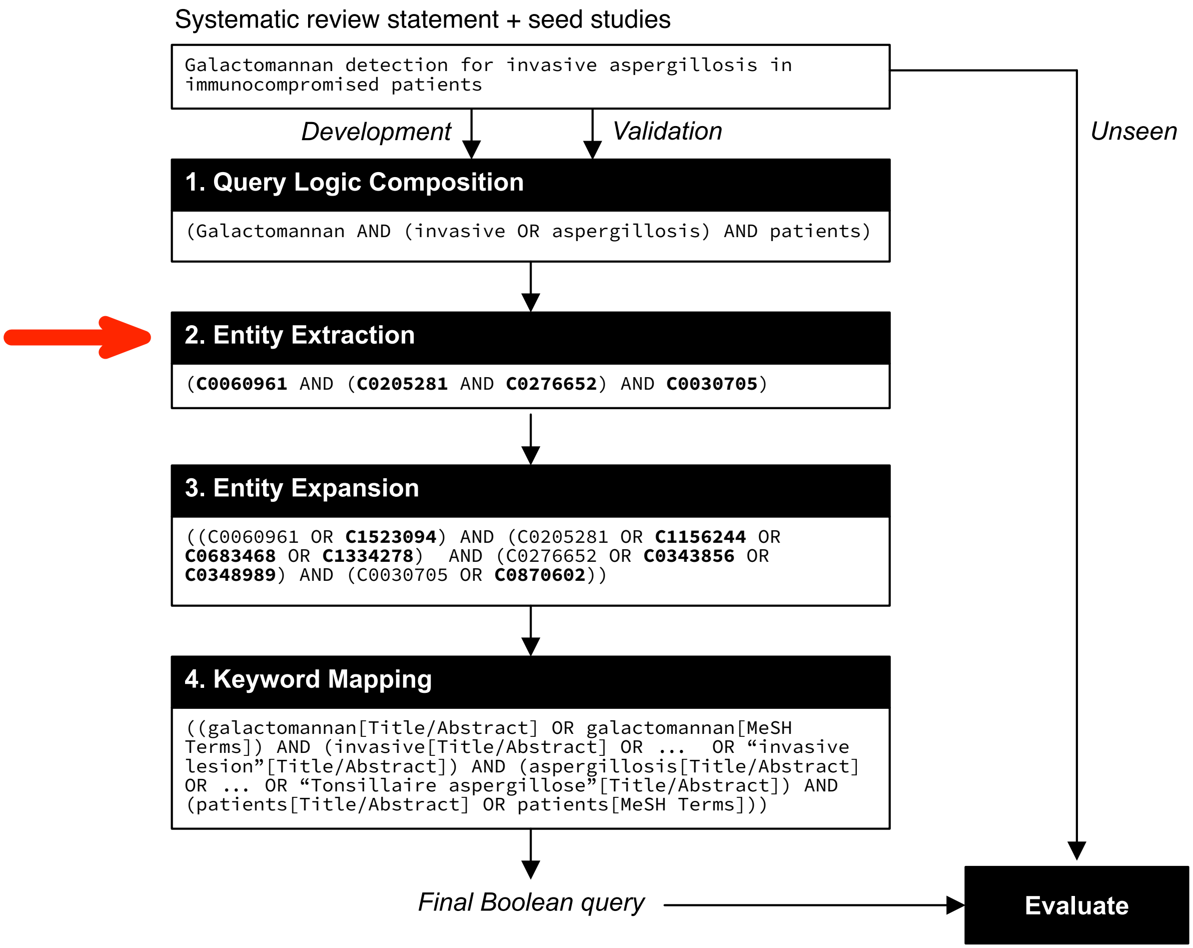
((galactomannan[Title/Abstract] OR galactomannan[MeSH Terms]) AND (invasive[Title/Abstract] OR … OR “invasive lesion”[Title/Abstract]) AND (aspergillosis[Title/Abstract] OR … OR “Tonsillaire aspergillose”[Title/Abstract]) AND (patients[Title/Abstract] OR patients[MeSH Terms]))

Final Boolean query

Evaluate

POS tagger ➜ parse grammar & segment words into noun phrases

Automating the conceptual method



Systematic review statement + seed studies

Galactomannan detection for invasive aspergillosis in immunocompromised patients

Development

Validation

Unseen

1. Query Logic Composition

(Galactomannan AND (invasive OR aspergillosis) AND patients)

2. Entity Extraction

(**C0060961** AND (**C0205281** AND **C0276652**) AND **C0030705**)

3. Entity Expansion

((C0060961 OR **C1523094**) AND (C0205281 OR **C1156244** OR **C0683468** OR **C1334278**) AND (C0276652 OR **C0343856** OR **C0348989**) AND (C0030705 OR **C0870602**))

4. Keyword Mapping

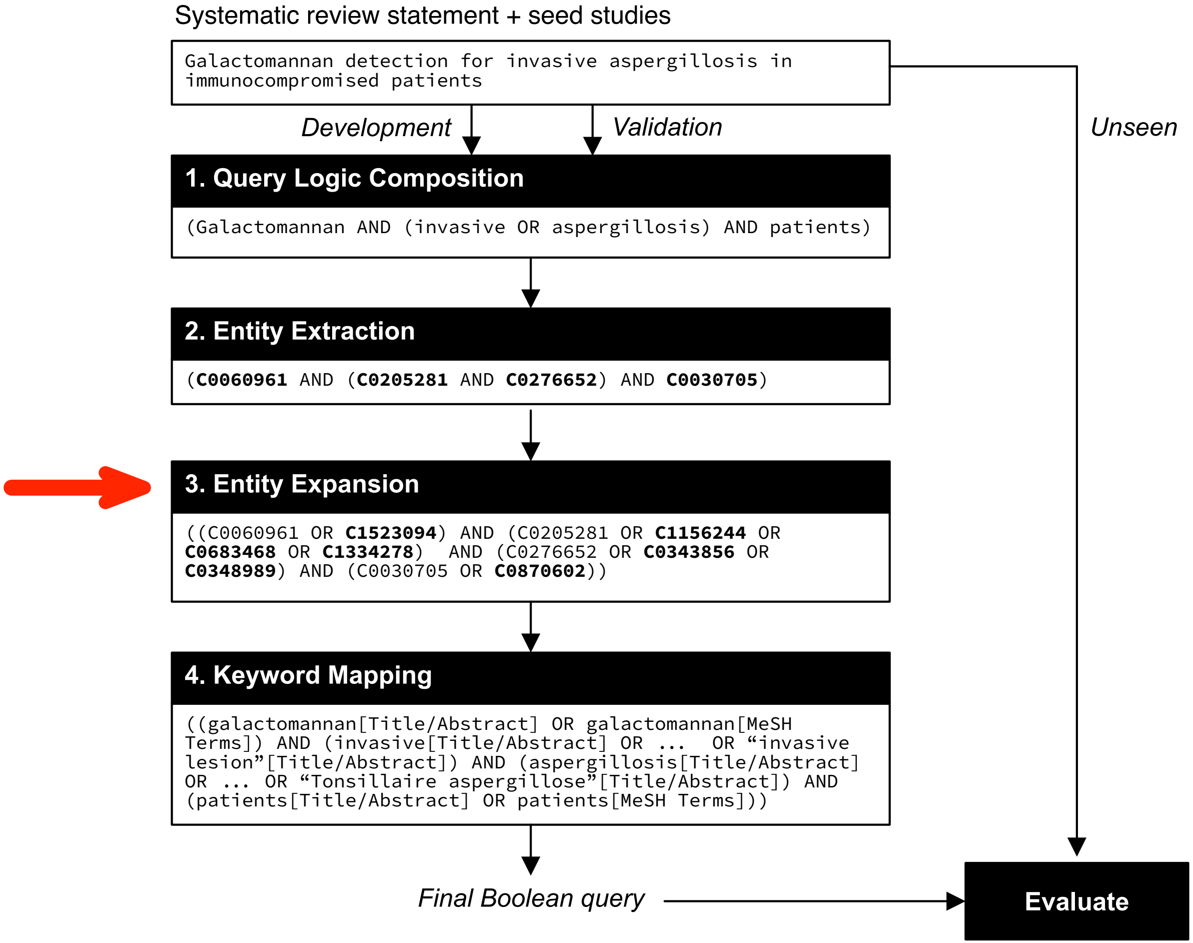
((galactomannan[Title/Abstract] OR galactomannan[MeSH Terms]) AND (invasive[Title/Abstract] OR ... OR “invasive lesion”[Title/Abstract]) AND (aspergillosis[Title/Abstract] OR ... OR “Tonsillaire aspergillose”[Title/Abstract]) AND (patients[Title/Abstract] OR patients[MeSH Terms]))

Final Boolean query

Evaluate

MetaMap ➜ extract CUIs from UMLS ontology

Automating the conceptual method



Systematic review statement + seed studies

`Galactomannan detection for invasive aspergillosis in immunocompromised patients`

Development

Validation

Unseen

1. Query Logic Composition

`(Galactomannan AND (invasive OR aspergillosis) AND patients)`

2. Entity Extraction

(**C0060961** AND (**C0205281** AND **C0276652**) AND **C0030705**)

3. Entity Expansion

((C0060961 OR **C1523094**) AND (C0205281 OR **C1156244** OR **C0683468** OR **C1334278**) AND (C0276652 OR **C0343856** OR **C0348989**) AND (C0030705 OR **C0870602**))

4. Keyword Mapping

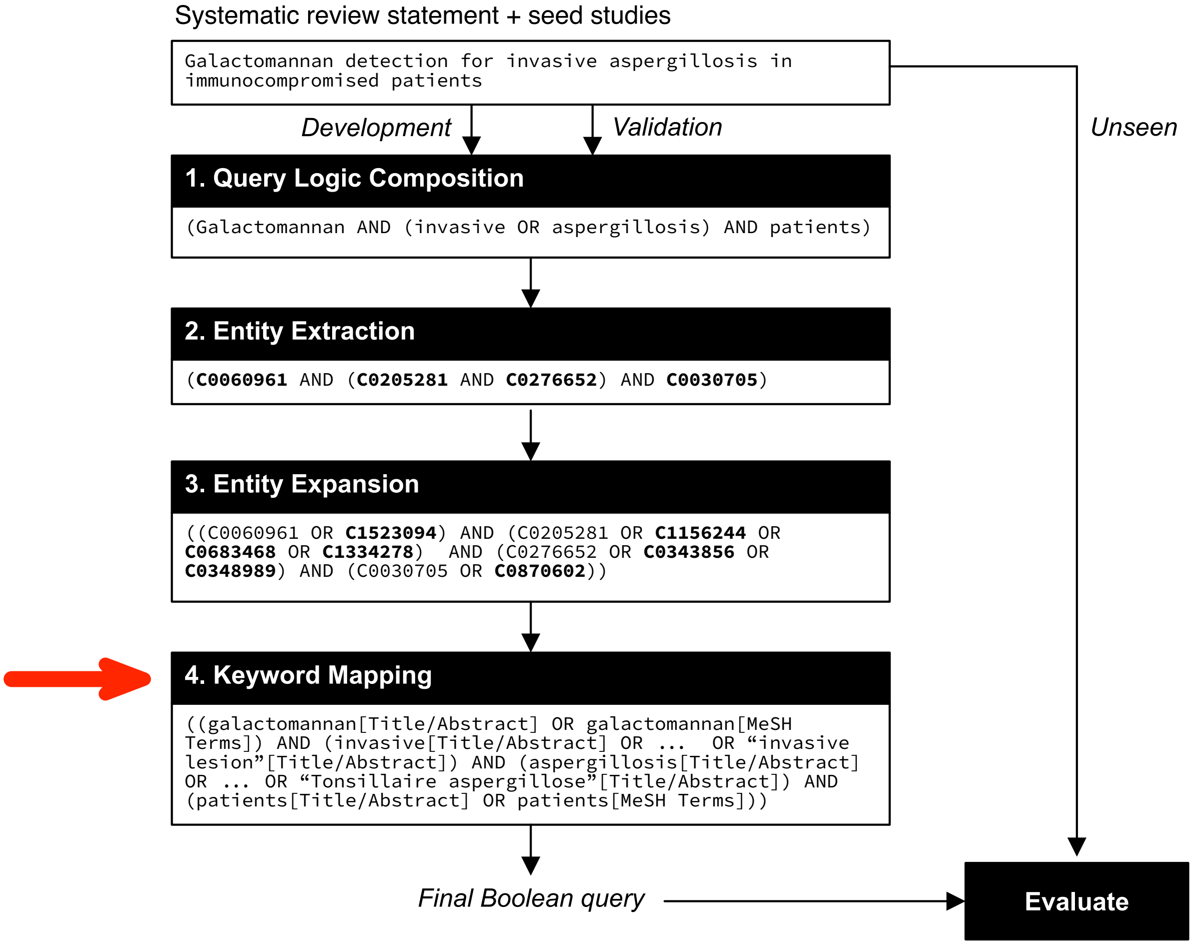
`((galactomannan[Title/Abstract] OR galactomannan[MeSH Terms]) AND (invasive[Title/Abstract] OR ... OR “invasive lesion”[Title/Abstract]) AND (aspergillosis[Title/Abstract] OR ... OR “Tonsillaire aspergillose”[Title/Abstract]) AND (patients[Title/Abstract] OR patients[MeSH Terms]))`

Final Boolean query

Evaluate

Skipgram model ➜ broaden scope

Automating the conceptual method



Systematic review statement + seed studies

Galactomannan detection for invasive aspergillosis in immunocompromised patients

Development

Validation

Unseen

1. Query Logic Composition

(Galactomannan AND (invasive OR aspergillosis) AND patients)

2. Entity Extraction

(**C0060961** AND (**C0205281** AND **C0276652**) AND **C0030705**)

3. Entity Expansion

((C0060961 OR **C1523094**) AND (C0205281 OR **C1156244** OR **C0683468** OR **C1334278**) AND (C0276652 OR **C0343856** OR **C0348989**) AND (C0030705 OR **C0870602**))

4. Keyword Mapping

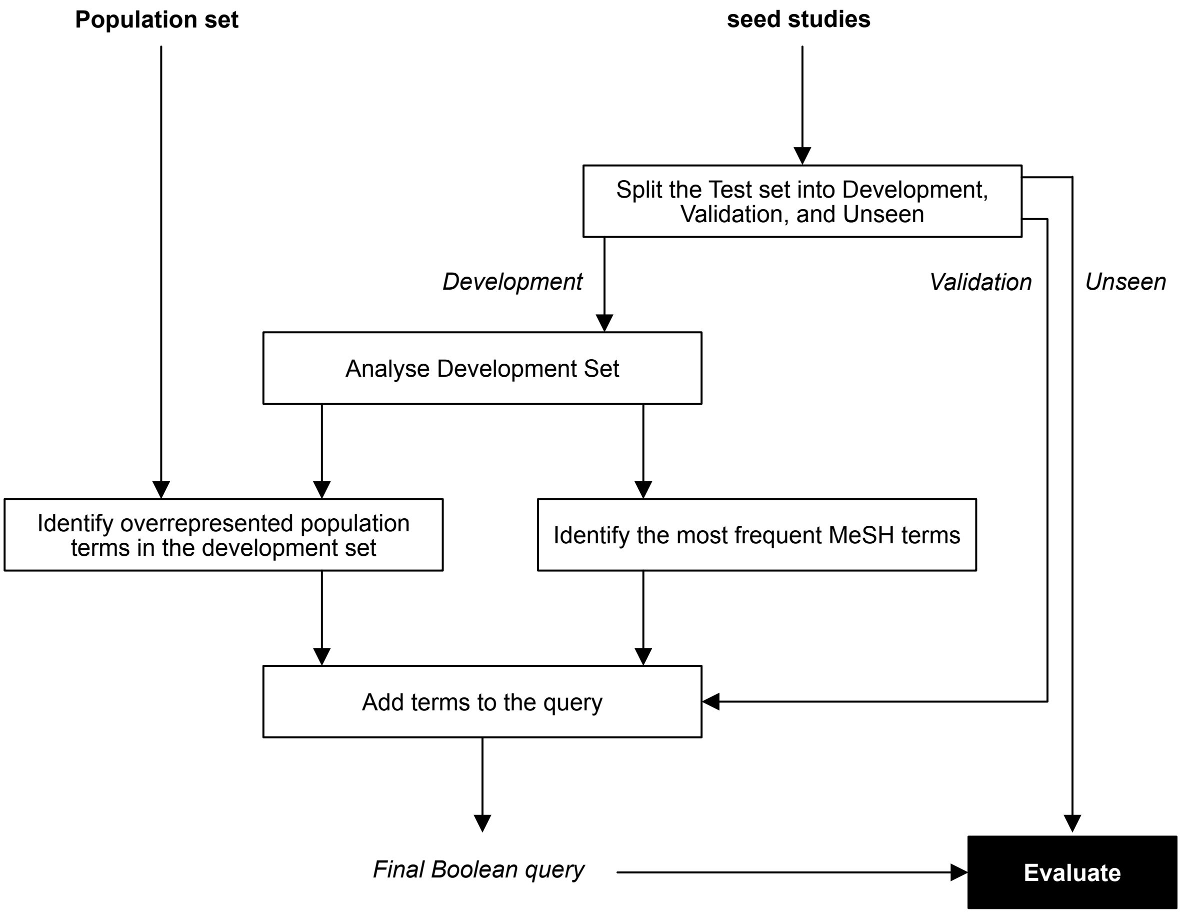
((galactomannan[Title/Abstract] OR galactomannan[MeSH Terms]) AND (invasive[Title/Abstract] OR ... OR “invasive lesion”[Title/Abstract]) AND (aspergillosis[Title/Abstract] OR ... OR “Tonsillaire aspergillose”[Title/Abstract]) AND (patients[Title/Abstract] OR patients[MeSH Terms]))

Final Boolean query

Evaluate

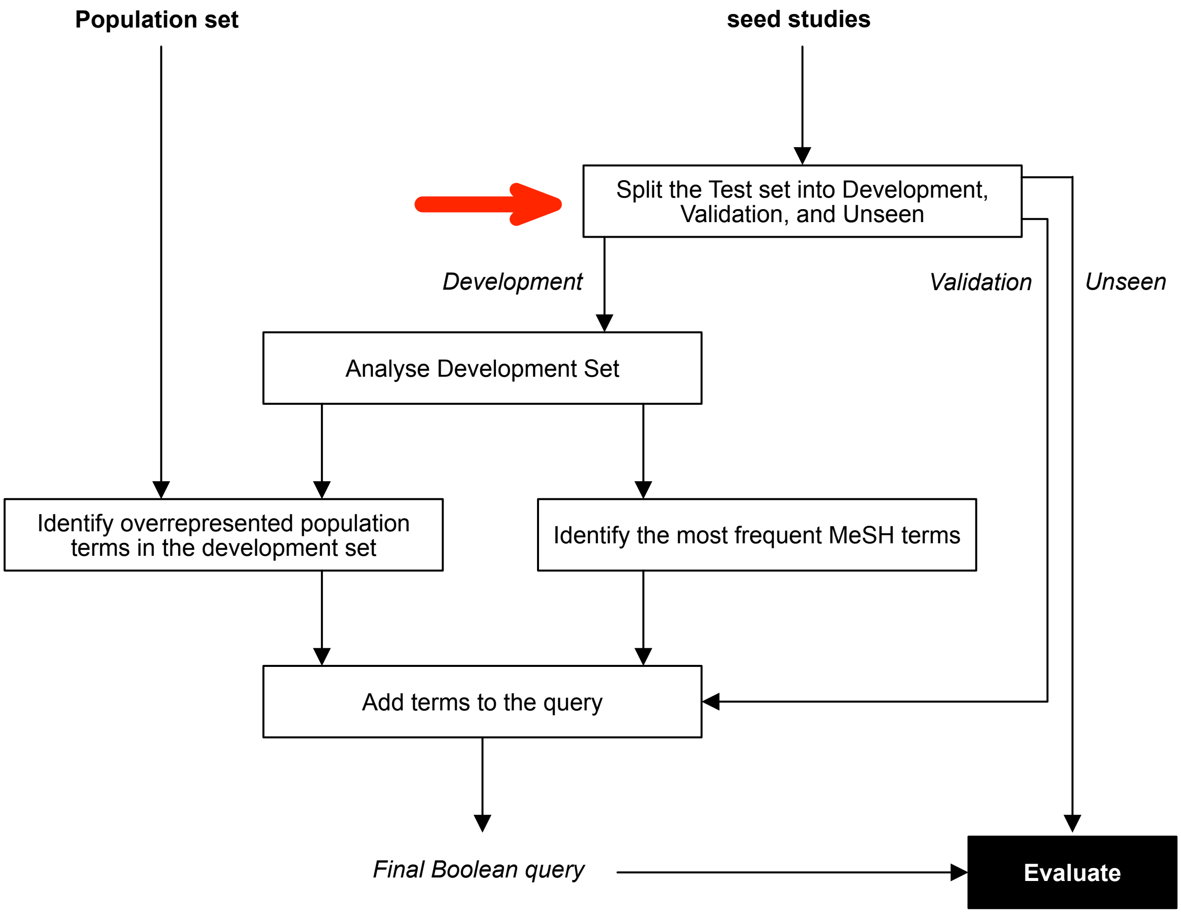
Map concepts (CUIs) to terms

Automating the objective method


Population set
seed studies
Split the Test set into Development, Validation, and Unseen
Development
Validation
Unseen
Analyse Development Set
Identify overrepresented population terms in the development set
Identify the most frequent MeSH terms
Add terms to the query
Final Boolean query
Evaluate

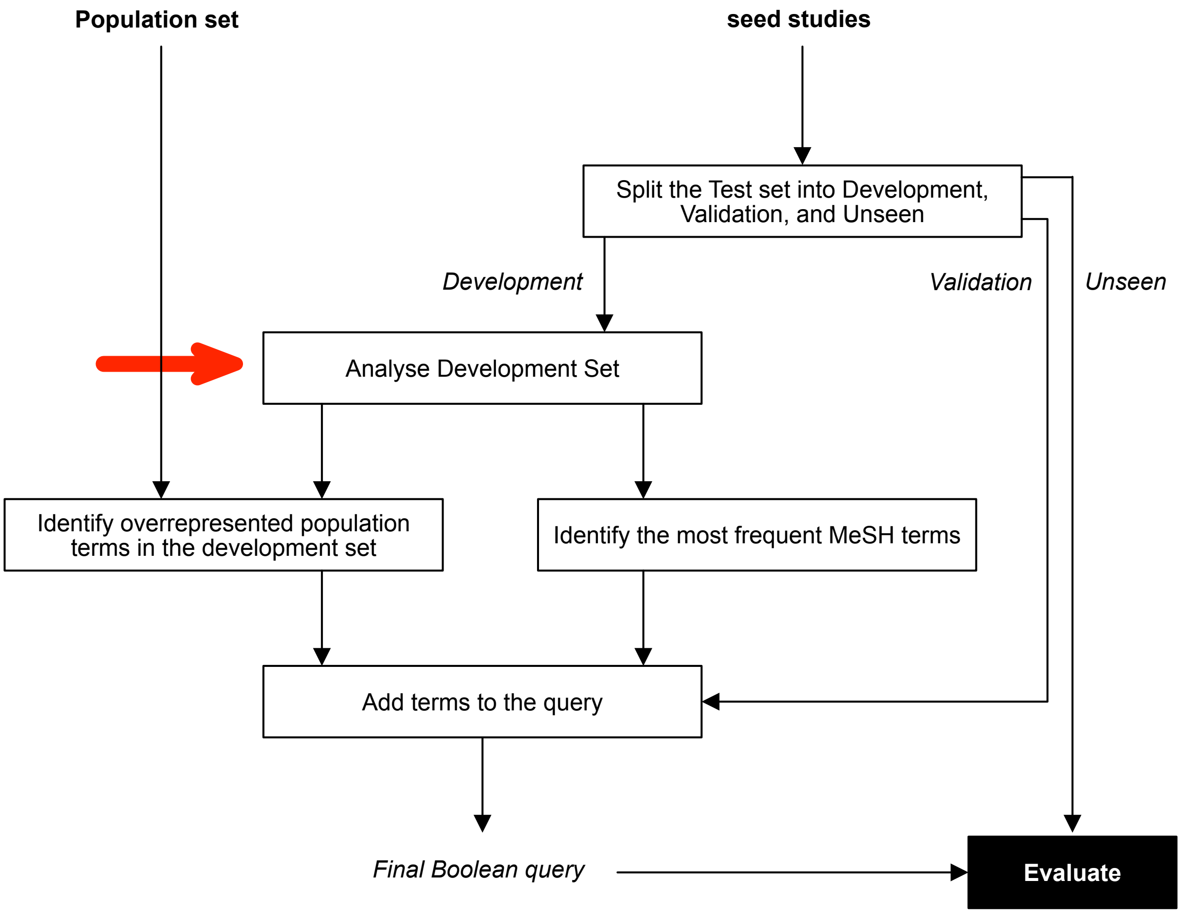
Find prominent terms from docs ➜ Add these terms to query

Automating the objective method


Population set
seed studies
Split the Test set into Development, Validation, and Unseen
Development
Validation
Unseen
Analyse Development Set
Identify overrepresented population terms in the development set
Identify the most frequent MeSH terms
Add terms to the query
Final Boolean query
Evaluate

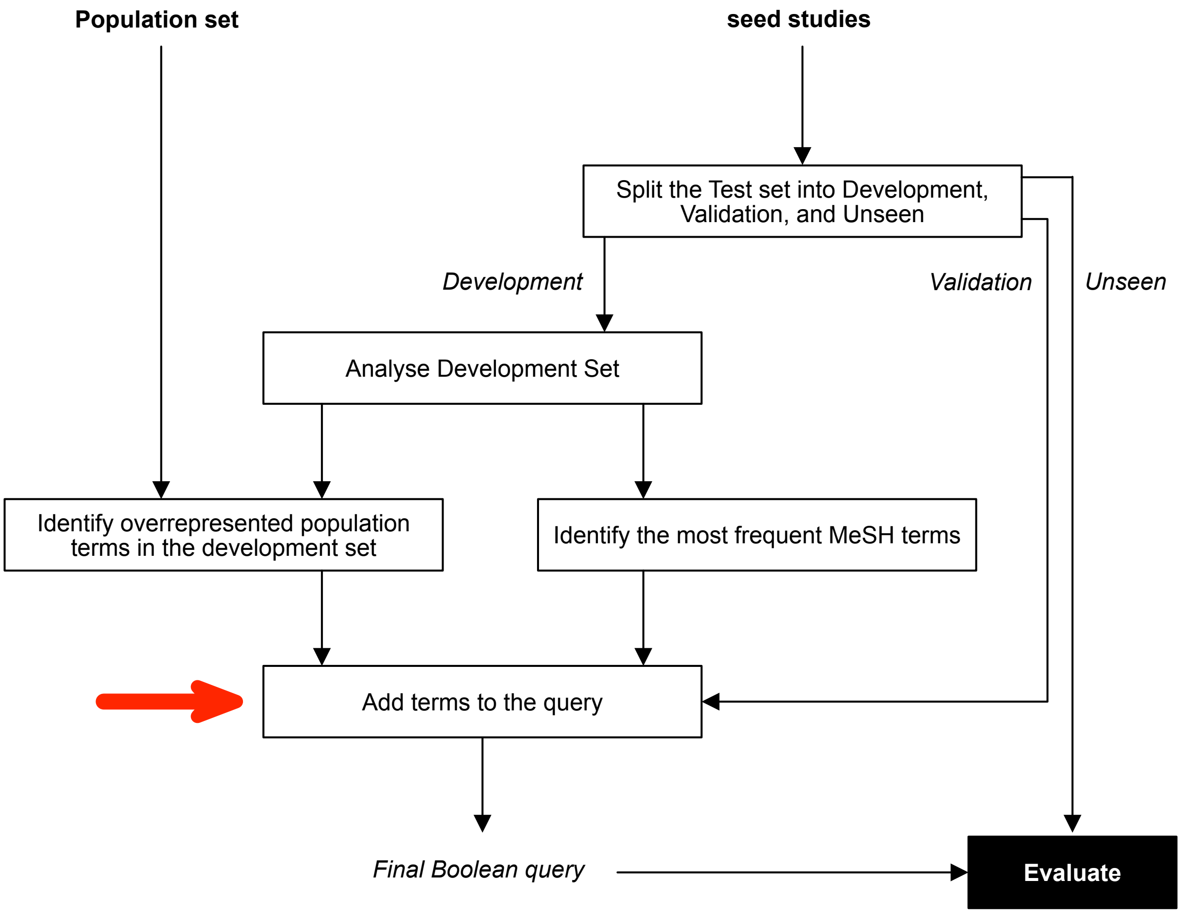
Extract list of keywords from seed studies

Automating the objective method


Population set
seed studies
Split the Test set into Development, Validation, and Unseen
Development
Validation
Unseen
Analyse Development Set
Identify overrepresented population terms in the development set
Identify the most frequent MeSH terms
Add terms to the query
Final Boolean query
Evaluate

Rank documents using term frequency

Automating the objective method



Population set

seed studies

Split the Test set into Development, Validation, and Unseen

Development

Validation

Unseen

Analyse Development Set

Identify overrepresented population terms in the development set

Identify the most frequent MeSH terms

Add terms to the query

Final Boolean query

Evaluate

Add keywords from documents to query

Conceptual versus objective results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	F1	Recall
Conceptual	0.0014	0.0027	0.6996
Objective	0.0002	0.0005	0.9128

num. queries=72

Seed study collection [Wang et al. 2022]

Precision	F1	Recall
0.0018	0.0036	0.4138
0.0057	0.0113	0.5192

num. queries=40

Conceptual versus objective results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	F1	Recall
Human	0.0207	0.0290	0.8317
Conceptual	0.0014	0.0027	0.6996
Objective	0.0002	0.0005	0.9128

num. queries=72

Seed study collection [Wang et al. 2022]

Precision	F1	Recall
0.0367	0.0651	0.7366
0.0018	0.0036	0.4138
0.0057	0.0113	0.5192

num. queries=40

Conceptual versus objective results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	F1	Recall
Human	0.0207	0.0290	0.8317
Conceptual	0.0014	0.0027	0.6996
Objective	0.0002	0.0005	0.9128

num. queries=72

Seed study collection [Wang et al. 2022]

	Precision	F1	Recall
Human	0.0367	0.0651	0.7366
Conceptual	0.0018	0.0036	0.4138
Objective	0.0057	0.0113	0.5192

num. queries=40

Objective provides highest recall, but larger trade-offs than conceptual

Conceptual versus objective results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	F1	Recall
Human	0.0207	0.0290	0.8317
Conceptual	0.0014	0.0027	0.6996
Objective	0.0002	0.0005	0.9128

num. queries=72

Seed study collection [Wang et al. 2022]

	Precision	F1	Recall
Human	0.0367	0.0651	0.7366
Conceptual	0.0018	0.0036	0.4138
Objective	0.0057	0.0113	0.5192

num. queries=40

Objective provides highest recall, but larger trade-offs than conceptual

Both methods are highly sensitive to initial seed studies

Conceptual versus objective results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	F1	Recall
Human	0.0207	0.0290	0.8317
Conceptual	0.0014	0.0027	0.6996
Objective	0.0002	0.0005	0.9128

num. queries=72

Seed study collection [Wang et al. 2022]

Precision	F1	Recall
0.0367	0.0651	0.7366
0.0018	0.0036	0.4138
0.0057	0.0113	0.5192

num. queries=40

Objective provides highest recall, but larger trade-offs than conceptual

Both methods are highly sensitive to initial seed studies

Conceptual method better starting point for precision, and objective for recall

Can ChatGPT formulate effective queries?

Hey ChatGPT, formulate a query for me!

Sure, no problem! ;-)

- Translation task: prompt ➜ query
- Two methods to formulate queries:
  - Unguided prompt *one shot to generate query*
  - Guided prompt *follows objective approach*
- Prevalence of Differentiated Thyroid Cancer in Autopsy Studies Over Six Decades: A Meta-Analysis

Unguided prompt examples

Simple

For a systematic review titled "`{review_title}`", can you **generate a systematic review Boolean query** to find all included studies on PubMed for the review topic?

(**"Prevalence of Differentiated Thyroid Cancer"** AND "Autopsy Studies"AND **"Six Decades"**) AND (systematic review OR meta-analysis)

Unguided prompt examples

Detailed

You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. **Your specialty is developing queries** that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. Now you have your information need to conduct research on "`{review_title}`". Please construct a highly effective systematic review Boolean query that can best serve your information need.

(**"Differentiated Thyroid Cancer"** OR "Papillary Thyroid Cancer" OR "Follicular Thyroid Cancer") AND "Autopsy Studies" AND "Prevalence" AND **(("1960/01/01"[PDAT] : "2023/01/30"[PDAT]) OR ("1960"[PDAT]:"2023"[PDAT]))**

Unguided prompt examples

With Example

You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. **You are able to take an information need** such as: "`{example_review_title}`" and **generate valid pubmed queries** such as: "`{example_review_query}`". Now you have the information need to conduct research on "`{review_title}`", please generate a highly effective systematic review Boolean query for the information need.

(**"Differentiated Thyroid Cancer"[Mesh]** OR (differentiated[All Fields] AND thyroid[All Fields] AND cancer[All Fields]) OR Papillary Thyroid Cancer"[Mesh] OR (papillary[All Fields] AND thyroid[All Fields] AND cancer[All Fields]) OR Follicular Thyroid Cancer"[Mesh] OR (follicular[All Fields] AND thyroid[All Fields] AND cancer[All Fields])) AND ((prevalence[All Fields] OR incidence[All Fields]) AND (autopsy[All Fields] OR post-mortem[All Fields] OR mortuary[All Fields]) **AND (study[All Fields] OR studies[All Fields]) AND (six[All Fields] OR 60[All Fields] OR decades[All Fields])**)

Previous query formulation results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	F1	Recall
Human	0.0207	0.0290	0.8317
Conceptual	0.0014	0.0027	0.6996
Objective	0.0002	0.0005	0.9128

Seed study collection [Wang et al. 2022]

Precision	F1	Recall
0.0367	0.0651	0.7366
0.0018	0.0036	0.4138
0.0057	0.0113	0.5192

Unguided query formulation results

	CLEF TAR [Kanoulas et al. 2017, 2018]			Seed study collection [Wang et al. 2022]		
	Precision	F1	Recall	Precision	F1	Recall
Human	0.0207	0.0290	0.8317	0.0367	0.0651	0.7366
Conceptual	0.0014	0.0027	0.6996	0.0018	0.0036	0.4138
Objective	0.0002	0.0005	0.9128	0.0057	0.0113	0.5192
Simple	0.0543	0.0500	0.1293	0.0501	0.0274	0.0528
Detailed	0.1166	0.0654	0.1310	0.0983	0.0310	0.0394
Example	0.0752	0.0642	0.5035	0.0283	0.0274	0.1290

Unguided query formulation results

CLEF TAR [Kanoulas et al. 2017, 2018]

	Precision	F1	Recall
Human	0.0207	0.0290	0.8317
Conceptual	0.0014	0.0027	0.6996
Objective	0.0002	0.0005	0.9128
Simple	0.0543	0.0500	0.1293
Detailed	0.1166	0.0654	0.1310
Example	0.0752	0.0642	0.5035

Seed study collection [Wang et al. 2022]

Precision	F1	Recall
0.0367	0.0651	0.7366
0.0018	0.0036	0.4138
0.0057	0.0113	0.5192
0.0501	0.0274	0.0528
0.0983	0.0310	0.0394
0.0283	0.0274	0.1290

Collection contains **seed studies**
➜ Use ChatGPT for objective method

Guided prompt examples

Step 1

Follow my instructions precisely to develop a highly effective Boolean query for a medical systematic review literature search. Do not explain or elaborate. First, Given the following statement and text from a relevant study, **identify 50 terms or phrases that are relevant**. The terms you identify should be used to retrieve more relevant studies, so be careful that the terms you choose are not too broad. You are not allowed to have duplicates in your list. Statement: `{statement}`. Text: `{text}`

1. **Differentiated thyroid cancer**
2. Prevalence
3. Autopsy studies
4. Meta-analysis
5. Occult carcinomas

...

50. Parameters.

Guided prompt examples

Step 2

...
50. Parameters.

For each item in the list you created in Step 1, **classify it into three categories**: terms relating to health conditions (A), terms relating to a treatment (B), terms relating to types of study design (C). When an item does not fit one of these categories, mark it as (N/A). Each item needs to be categorised into (A), (B), (C), or (N/A).

1. **(A) Differentiated thyroid cancer**
2. (N/A) Prevalence
3. (C) Autopsy studies
4. (C) Meta-analysis
5. (A) Occult carcinomas

...
50. (N/A) Parameters.

Guided prompt examples

Step 3

...
50. (N/A) Parameters.

Using the categorised list you created in step 2, create a Boolean query that can be submitted to PubMed which groups together items from each category. For example: ((**itemA1** OR **itemA2** or **itemA3**) AND (itemB1 OR itemB2 OR itemB3) AND (itemC1 OR itemC2 OR itemC3))

(((**differentiated**[Title/Abstract] OR **thyroid**[Title/Abstract] OR **carcinoma**[Title/Abstract] OR papillary[Title/Abstract] OR microcarcinoma[Title/Abstract]) AND (cancer[Title/Abstract] OR incidence[Title/Abstract] OR etiology[Title/Abstract] OR risk[Title/Abstract] OR gender[Title/Abstract] OR hormonal[Title/Abstract] OR nodular[Title/Abstract] OR goiter[Title/Abstract] OR Hashimoto's[Title/Abstract] OR malignancy[Title/Abstract] OR concomitant[Title/Abstract] OR tumor[Title/Abstract] OR infiltrate[Title/Abstract] OR fibrosis[Title/Abstract] OR development[Title/Abstract] OR frequency[Title/Abstract])) AND (autopsy[Title/Abstract] OR surgical[Title/Abstract] OR material[Title/Abstract] OR series[Title/Abstract] OR specimens[Title/Abstract] OR cases[Title/Abstract]))

Guided prompt examples

Step 4

...AND (autopsy[Title/Abstract] OR surgical[Title/Abstract] OR material[Title/Abstract] OR series[Title/Abstract] OR specimens[Title/Abstract] OR cases[Title/Abstract]))

Use your expert knowledge to **refine the query**, making it retrieve as many relevant documents as possible while minimising the total number of documents retrieved. Also add relevant MeSH terms into the query where necessary, e.g., MeSHTerm[MeSH]. Retain the general structure of the query, however, with each main clause of the query corresponding to a PICO element. The final query still needs to be executable on PubMed, so it should be a valid query.

(((**differentiated thyroid cancer[MeSH]** OR **"differentiated thyroid"**[All Fields] OR **"thyroid carcinoma"**[All Fields] OR "papillary microcarcinoma"[All Fields]) AND (prevalence[All Fields] OR incidence[MeSH] OR "etiology of"[All Fields] OR "risk factors"[All Fields] OR gender[All Fields] OR hormonal[All Fields] OR "nodular goiter"[All Fields] OR **"Hashimoto's thyroiditis"[MeSH]** OR malignancy[MeSH] OR "concomitant lesion"[All Fields] OR tumor[All Fields] OR infiltrate[All Fields] OR fibrosis[All Fields] OR "early stages of development"[All Fields] OR frequency[All Fields])) AND (autopsy[MeSH] OR surgical[All Fields] OR material[All Fields] OR series[All Fields] OR specimens[All Fields] OR cases[All Fields]))

Unguided query formulation results

Seed study collection [Wang et al. 2022]

	Precision	F1	Recall
Human	0.0367	0.0651	0.7366
Conceptual	0.0018	0.0036	0.4138
Objective	0.0057	0.0113	0.5192
Simple	0.0501	0.0274	0.0528
Detailed	0.0983	0.0310	0.0394
Example	0.0283	0.0274	0.1290

Guided query formulation results

Seed study collection [Wang et al. 2022]

	Precision	F1	Recall
Human	0.0367	0.0651	0.7366
Conceptual	0.0018	0.0036	0.4138
Objective	0.0057	0.0113	0.5192
Simple	0.0501	0.0274	0.0528
Detailed	0.0983	0.0310	0.0394
Example	0.0283	0.0274	0.1290
Guided	0.0993	0.0492	0.5171

Guided query formulation results

Seed study collection [Wang et al. 2022]

	Precision	F1	Recall
Human	0.0367	0.0651	0.7366
Conceptual	0.0018	0.0036	0.4138
Objective	0.0057	0.0113	0.5192
Simple	0.0501	0.0274	0.0528
Detailed	0.0983	0.0310	0.0394
Example	0.0283	0.0274	0.1290
Guided	0.0993	0.0492	0.5171

ChatGPT is more effective than automatic conceptual and objective methods

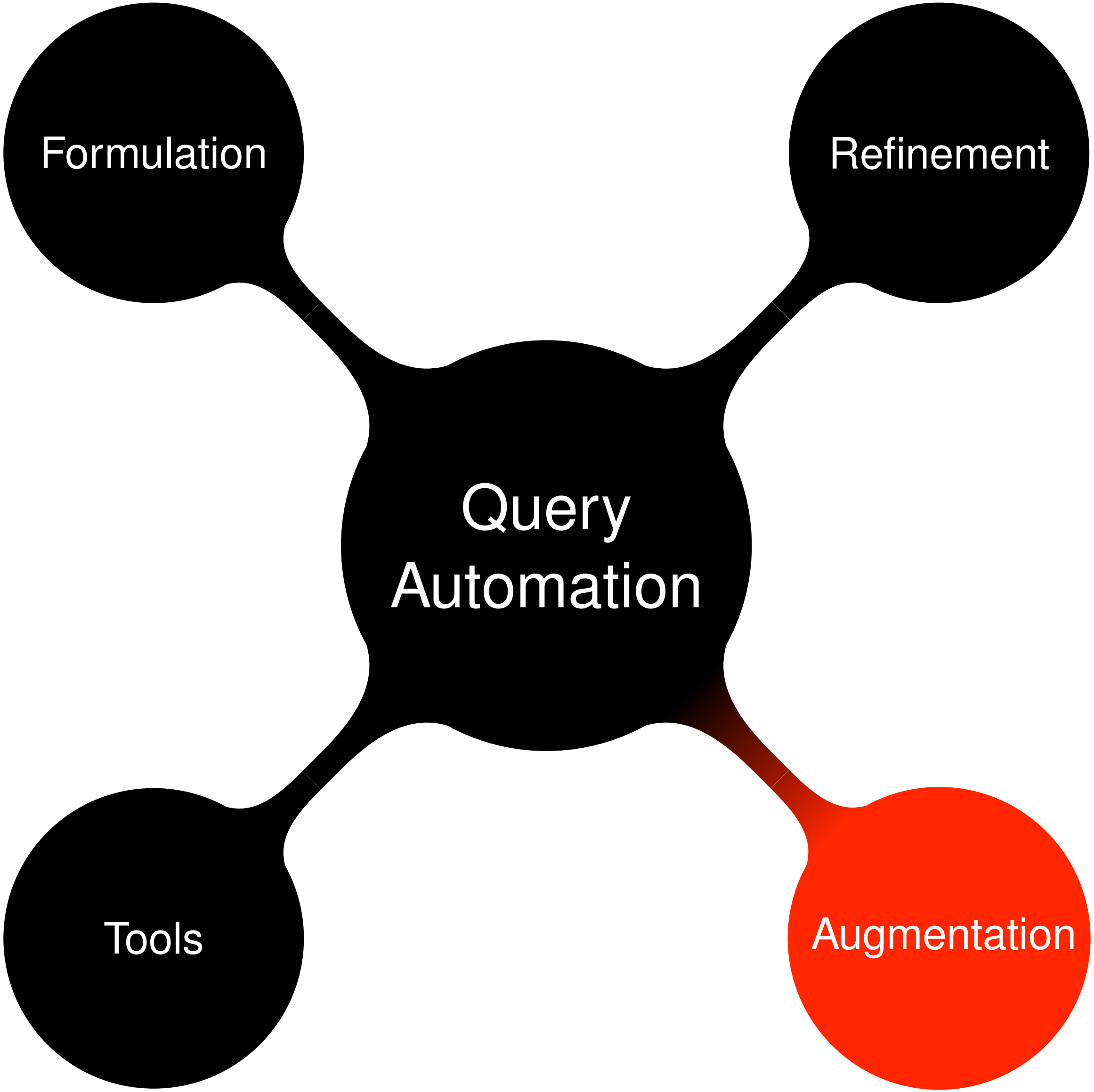
Guided query formulation results

Seed study collection [Wang et al. 2022]

	Precision	F1	Recall
Human	0.0367	0.0651	0.7366
Conceptual	0.0018	0.0036	0.4138
Objective	0.0057	0.0113	0.5192
Simple	0.0501	0.0274	0.0528
Detailed	0.0983	0.0310	0.0394
Example	0.0283	0.0274	0.1290
Guided	0.0993	0.0492	0.5171

ChatGPT is more effective than automatic conceptual and objective methods

ChatGPT is highly dependent on prompt and prone to hallucination



Formulation

Refinement

Query Automation

Tools

Augmentation

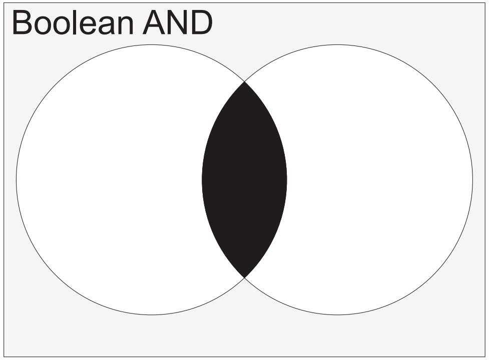
Query Augmentation

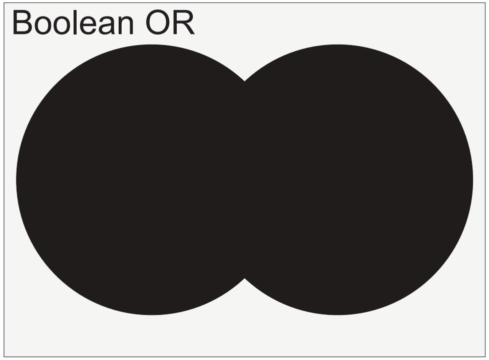
The modification or extension of complex queries in order to improve their effectiveness at the task of systematic review literature search

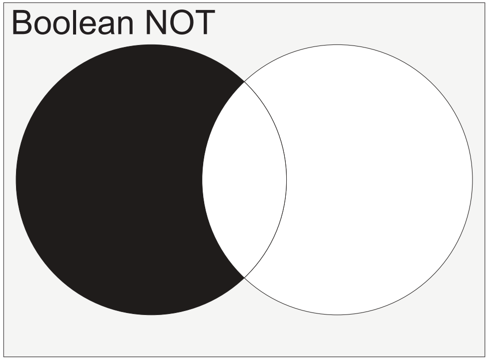
- ❑ Content covered
  - Harrisen Scells, Ferdinand Schlatt, and Martin Potthast. Smooth Operators for Effective Systematic Review Queries. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023 **(to appear at SIGIR'23)**
- ❑ Further reading
  - Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. Integrating the framing of clinical questions via PICO into the retrieval of medical literature for systematic reviews. In *Proceedings of the 26th International Conference on Information and Knowledge Management*, pages 2291–2294, 2017
  - Harrisen Scells and Guido Zuccon. You can teach an old dog new tricks: Rank fusion applied to coordination level matching for ranking in systematic reviews. In *Proceedings of the 42nd European Conference on Information Retrieval*, pages 399–414, 2020

Smooth operators

Intuition

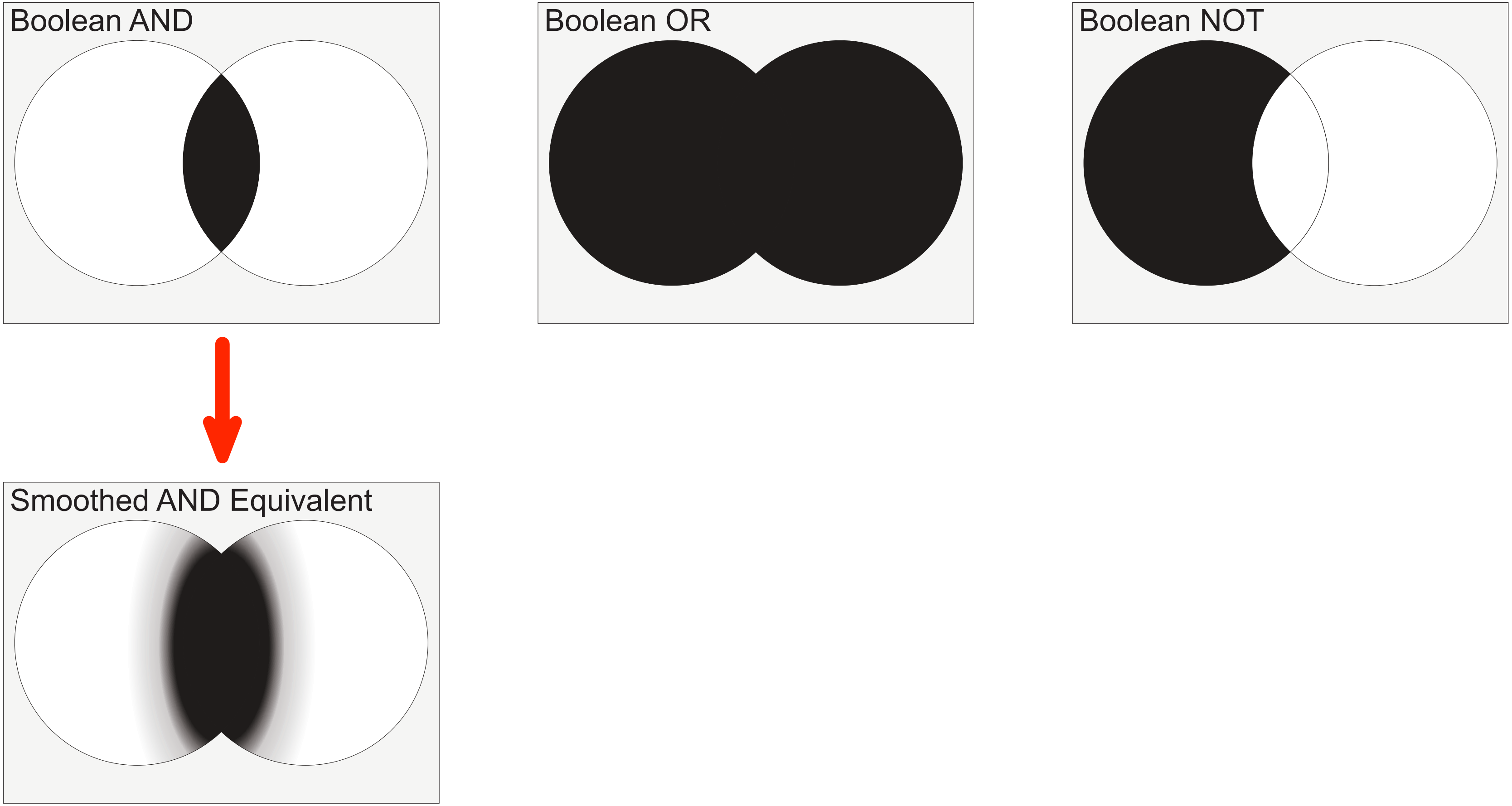






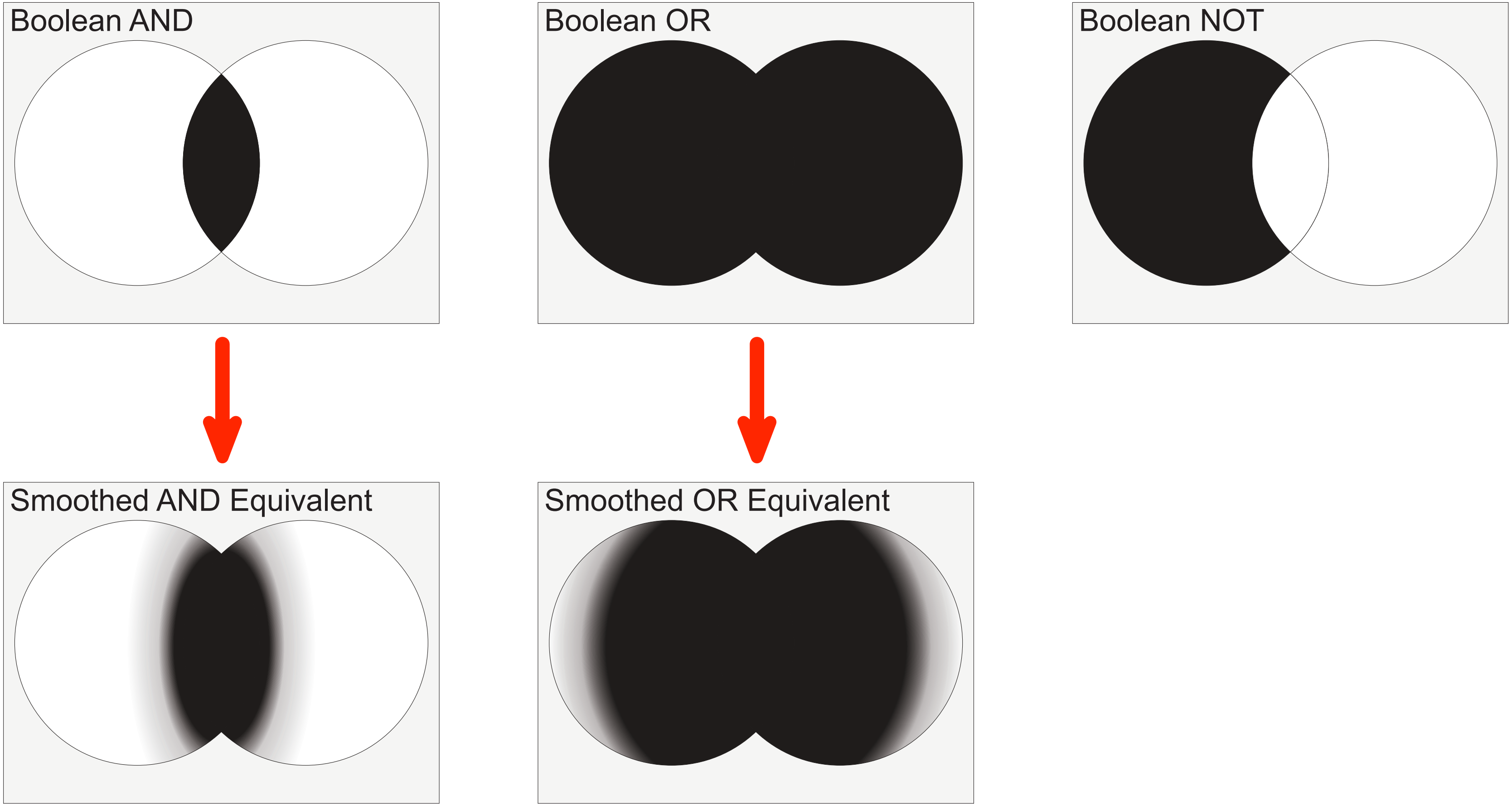
Smooth operators

Intuition



Smooth operators

Intuition


Boolean AND

Boolean OR

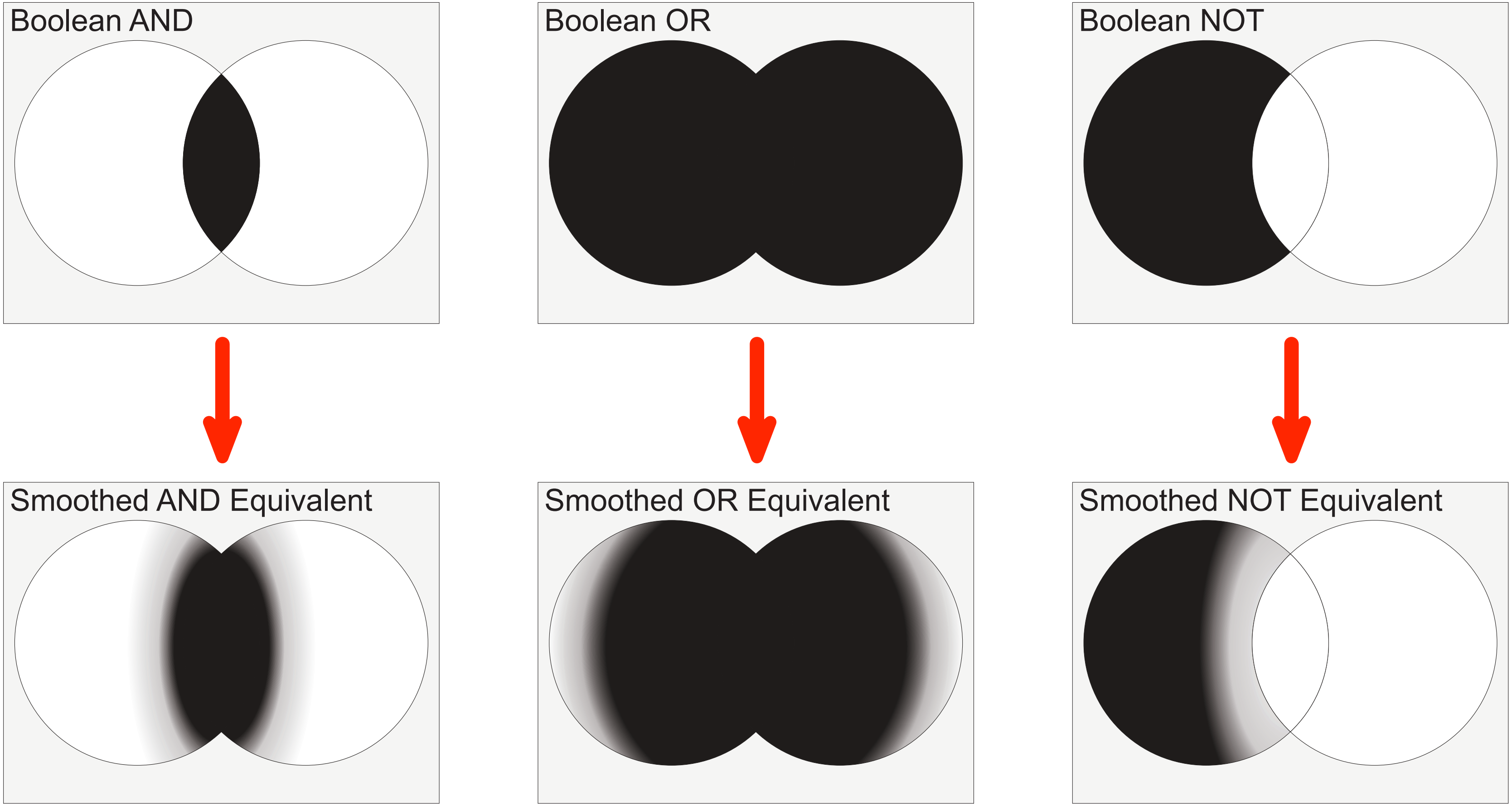
Boolean NOT

Smoothed AND Equivalent

Smoothed OR Equivalent

Smooth operators

Intuition



Boolean AND

Boolean OR

Boolean NOT

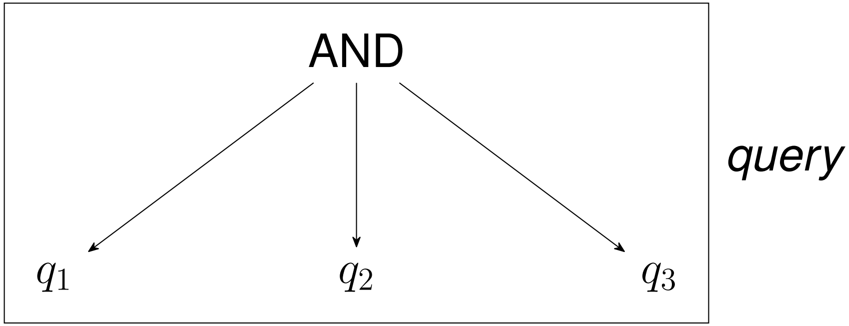
Smoothed AND Equivalent

Smoothed OR Equivalent

Smoothed NOT Equivalent

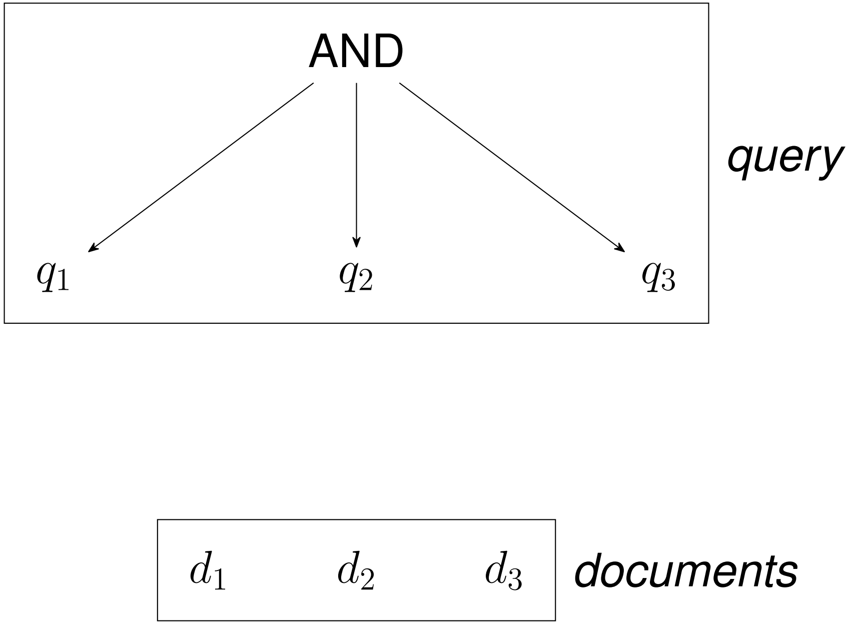
Smooth operators

Intuitive example



Smooth operators

Intuitive example



AND

q_1 q_2 q_3

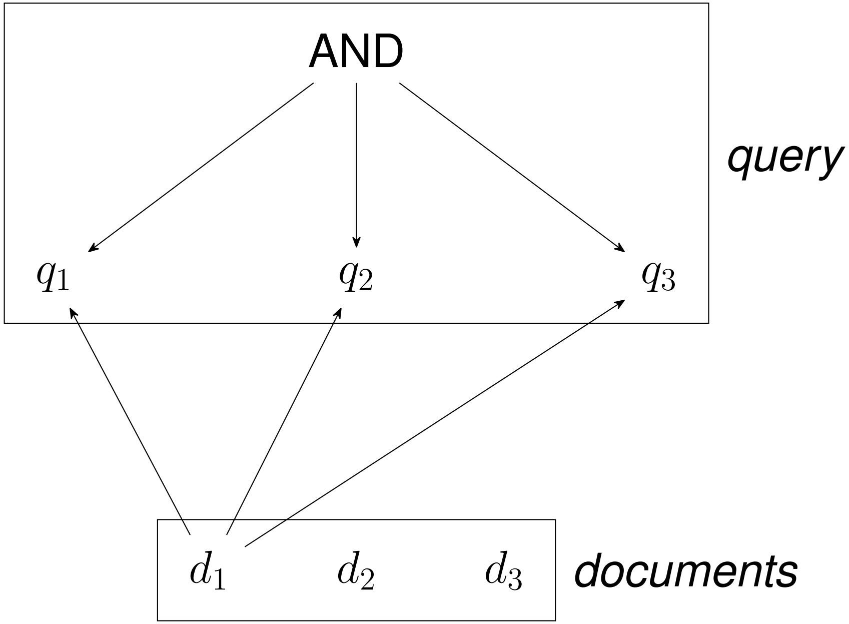
query

d_1 d_2 d_3

documents

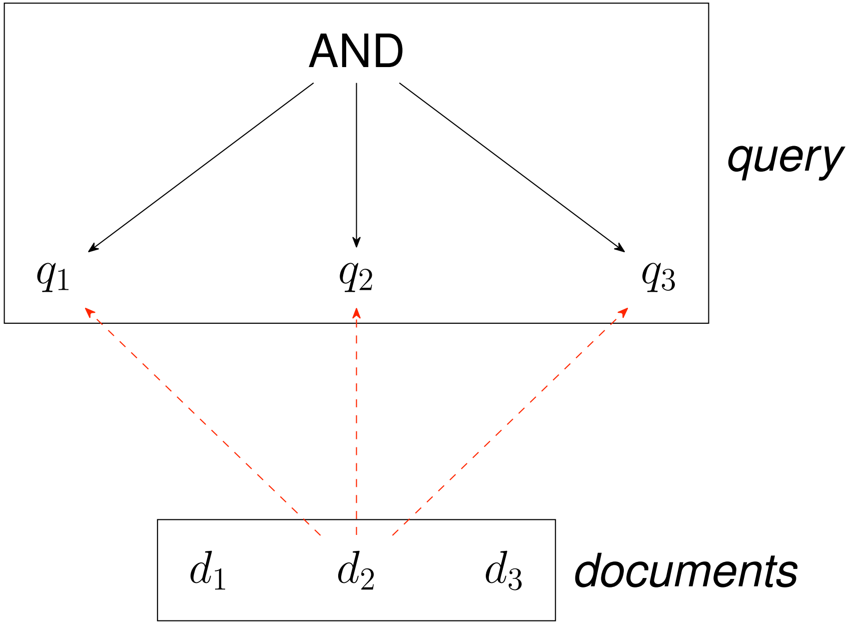
Smooth operators

Intuitive example



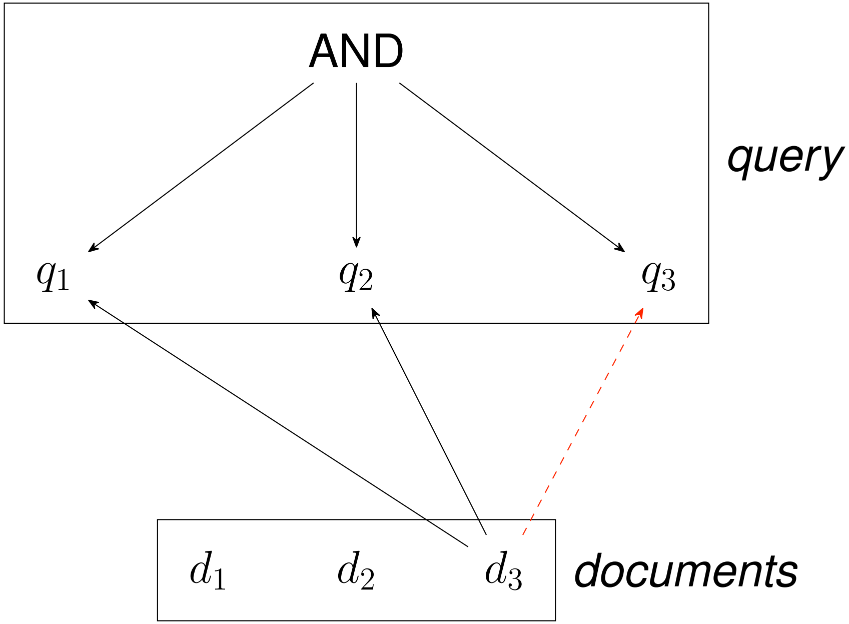
Smooth operators

Intuitive example



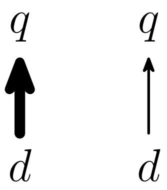
Smooth operators

Intuitive example



Theory

Smoothing result sets

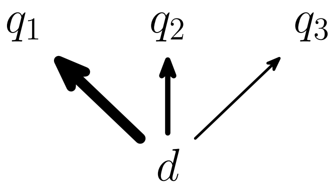


$P(d|q)$ ➜ extent to which d should belong to q

$$P(d|q) = \frac{P(d)P(q|d)}{P(q)}$$

Theory

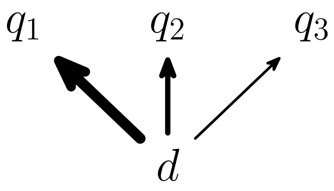
Smoothing result sets



$$P(d|q_1, ..., q_k) = \frac{P(d)\prod P(q_i|d)}{P(d)\prod P(q_i|d) + P(\bar{d})\prod P(q_i|\bar{d})}$$

Theory

Smoothing result sets



$$P(d|q_1, ..., q_k) = \frac{P(d)\prod P(q_i|d)}{P(d)\prod P(q_i|d) + P(\bar{d})\prod P(q_i|\bar{d})}$$

Leaves two estimations:

- $P(d)$ ➜ Probability of a document
- $P(q_i|d)$ ➜ Probability of a query given a document

Theory

Smoothing result sets

$P(d)$ ➜ ratio of queries that retrieve d

$$P(d) = \frac{|\{\forall q_i \in q \ : \ d \in D_{q_i}\}|}{|q|}$$

Theory

Smoothing result sets

$P(d)$ ➜ ratio of queries that retrieve d

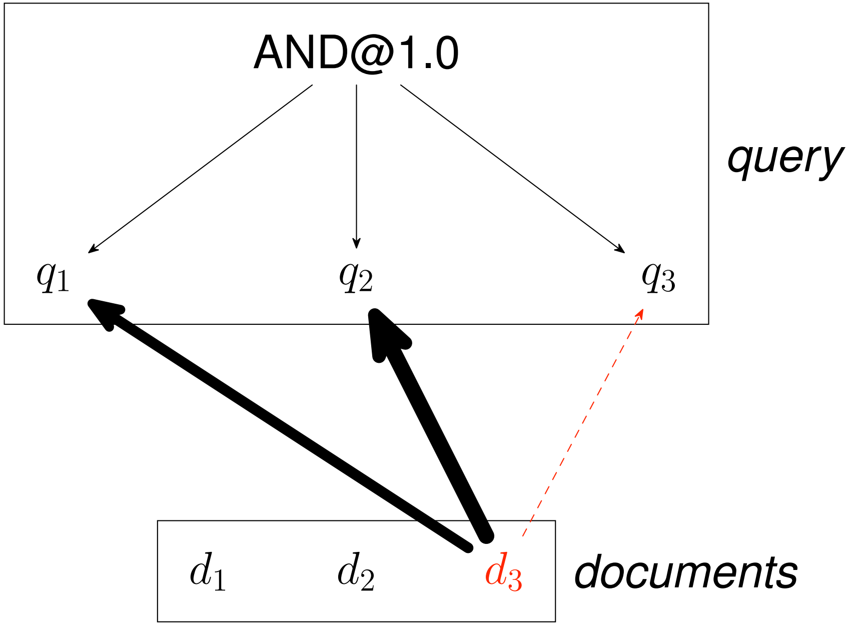
$$P(d) = \frac{|\{\forall q_i \in q \ : \ d \in D_{q_i}\}|}{|q|}$$

$P(q_i|d)$ ➜ relevance between q_i and d

$$P(q_i|d) = 1 - \frac{pos(q_i, d)}{|D_{q_i}|}$$

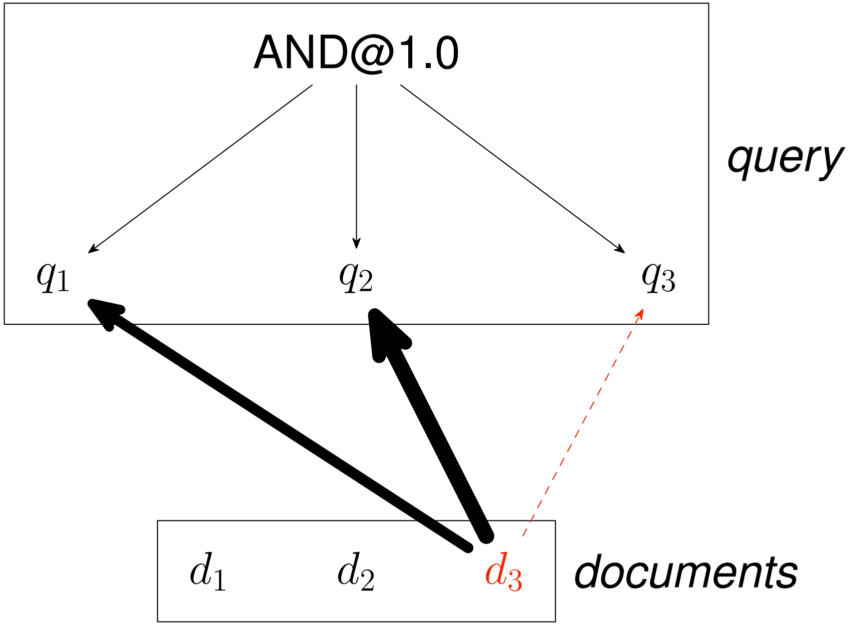
Theory

Implementing smooth operators



Theory

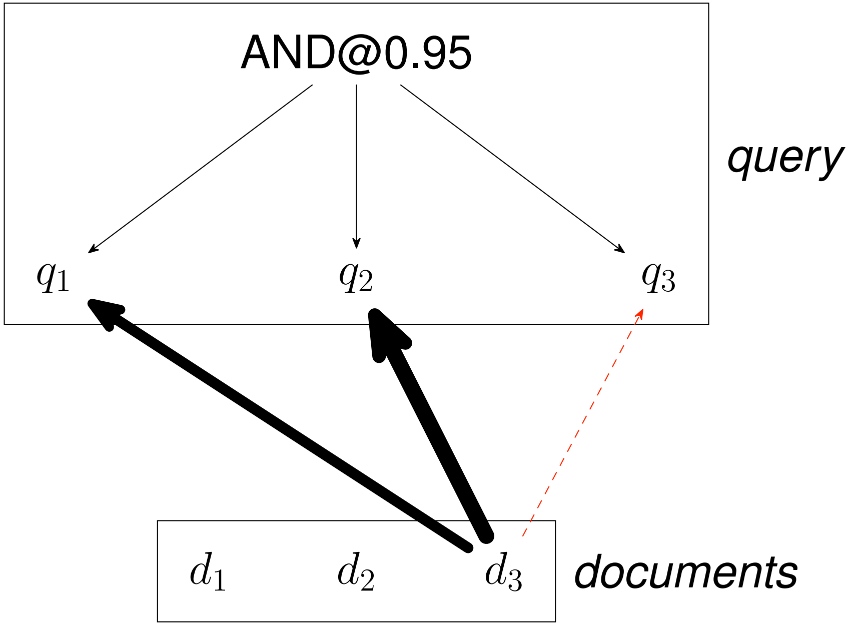
Implementing smooth operators



$$P(d_3|q_1, q_2, q_3) = 0.95$$

Theory

Implementing smooth operators



$$P(d_3|q_1, q_2, q_3) = 0.95$$

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486

Only using smooth operators for ranking is already considerably better than using BM25

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486
AND@0.99	0.7206	0.0038	0.0019	0.2331
AND@0.9	0.7658	0.0003	0.0002	0.2180

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486
AND@0.99	0.7206	0.0038	0.0019	0.2331
AND@0.9	0.7658	0.0003	0.0002	0.2180

Smoothing AND operators increases recall at the cost of precision and ranking effectiveness

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486
AND@0.99	0.7206	0.0038	0.0019	0.2331
AND@0.9	0.7658	0.0003	0.0002	0.2180
OR@0.01	0.3078	0.0582	0.0366	0.1604
OR@0.1	0.0612	0.0417	0.0486	0.0620

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486
AND@0.99	0.7206	0.0038	0.0019	0.2331
AND@0.9	0.7658	0.0003	0.0002	0.2180
OR@0.01	0.3078	0.0582	0.0366	0.1604
OR@0.1	0.0612	0.0417	0.0486	0.0620

Smoothing OR operators increases precision at the cost of recall and ranking effectiveness

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486
AND@0.99	0.7206	0.0038	0.0019	0.2331
AND@0.9	0.7658	0.0003	0.0002	0.2180
OR@0.01	0.3078	0.0582	0.0366	0.1604
OR@0.1	0.0612	0.0417	0.0486	0.0620
Predictor	0.3876	0.0566	0.0364	0.1651

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486
AND@0.99	0.7206	0.0038	0.0019	0.2331
AND@0.9	0.7658	0.0003	0.0002	0.2180
OR@0.01	0.3078	0.0582	0.0366	0.1604
OR@0.1	0.0612	0.0417	0.0486	0.0620
Predictor	0.3876	0.0566	0.0364	0.1651

Predicting the smoothness using features struck middleground between smooth OR and AND

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486
AND@0.99	0.7206	0.0038	0.0019	0.2331
AND@0.9	0.7658	0.0003	0.0002	0.2180
OR@0.01	0.3078	0.0582	0.0366	0.1604
OR@0.1	0.0612	0.0417	0.0486	0.0620
Predictor	0.3876	0.0566	0.0364	0.1651
Oracle	0.7437	0.0769	0.0440	0.2885

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486
AND@0.99	0.7206	0.0038	0.0019	0.2331
AND@0.9	0.7658	0.0003	0.0002	0.2180
OR@0.01	0.3078	0.0582	0.0366	0.1604
OR@0.1	0.0612	0.0417	0.0486	0.0620
Predictor	0.3876	0.0566	0.0364	0.1651
Oracle	0.7437	0.0769	0.0440	0.2885

However: using ground truth information, more effective queries are possible

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486
AND@0.99	0.7206	0.0038	0.0019	0.2331
AND@0.9	0.7658	0.0003	0.0002	0.2180
OR@0.01	0.3078	0.0582	0.0366	0.1604
OR@0.1	0.0612	0.0417	0.0486	0.0620
Predictor	0.3876	0.0566	0.0364	0.1651
Oracle	0.7437	0.0769	0.0440	0.2885
PubmedBERT	0.7148	0.0643	0.0363	0.2252
BERT	0.7149	0.0644	0.0363	0.2447
DistilBERT	0.7118	0.0641	0.0362	0.2449

Results

Seed study collection [Wang et al. 2022]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.7149	0.0642	0.0362	-
BM25 Title	0.7149	0.0642	0.0362	0.0972
smooth Boolean equivalents	0.7149	0.0642	0.0362	0.2486
AND@0.99	0.7206	0.0038	0.0019	0.2331
AND@0.9	0.7658	0.0003	0.0002	0.2180
OR@0.01	0.3078	0.0582	0.0366	0.1604
OR@0.1	0.0612	0.0417	0.0486	0.0620
Predictor	0.3876	0.0566	0.0364	0.1651
Oracle	0.7437	0.0769	0.0440	0.2885
PubmedBERT	0.7148	0.0643	0.0363	0.2252
BERT	0.7149	0.0644	0.0363	0.2447
DistilBERT	0.7118	0.0641	0.0362	0.2449

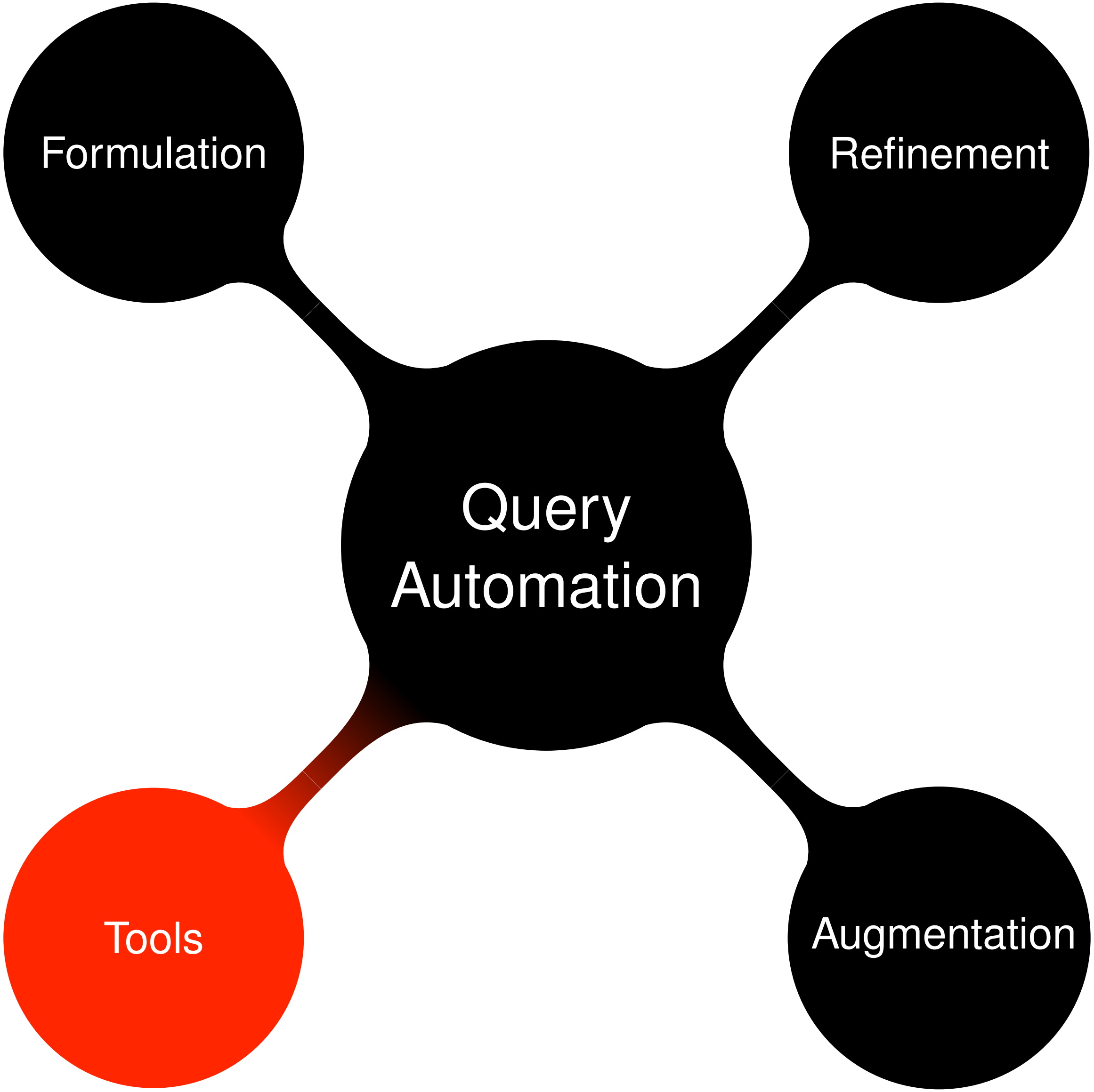
More advanced neural ranking models did not have any effect on ranking effectiveness

Results

CLEF TAR [Kanoulas et al. 2018]

	Recall	F_1	Precision	nDCG@100
Boolean operators	0.8344	0.0385	0.0204	-
BM25 Title	0.8344	0.0385	0.0204	0.0232
smooth Boolean equivalents	0.8344	0.0385	0.0204	0.1995
Predictor	0.6205	0.0372	0.0206	0.1698
Oracle	0.8487	0.0397	0.0211	0.2125
BERT	0.8344	0.0385	0.0204	0.2191
ECNU_RUN1	0.5147	0.0806	0.0490	0.2440
ECNU_RUN2	0.3831	0.0823	0.0539	0.1368
ECNU_RUN3	0.5147	0.0806	0.0490	0.2438
sheffield-bm25	0.4525	0.0180	0.0095	0.1197
sheffield-boolean	0.3048	0.0116	0.0061	0.0562
sheffield-tfidf	0.2572	0.0112	0.0059	0.0154

[Wu et al. 2018, Alharbi et al. 2018]



Formulation
Refinement
Query Automation
Tools
Augmentation

Tools

Software to support systematic review and information retrieval practitioners in undertaking query automation

- ❑ Content covered
  - Harrisen Scells and Martin Potthast. pybool_ir: A Toolkit for Domain-Specific Search Experiments. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023 **(to appear at SIGIR'23)**
- ❑ Further reading
  - Harrisen Scells, Daniel Locke, and Guido Zuccon. An information retrieval experiment framework for domain specific applications. In *Proceedings of the 41st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1281–1284, 2018
  - Harrisen Scells and Guido Zuccon. Searchrefiner: A query visualisation and understanding tool for systematic reviews. In *Proceedings of the 27th International Conference on Information and Knowledge Management*, pages 1939–1942, 2018
  - Hang Li, Harrisen Scells, and Guido Zuccon. Systematic review automation tools for end-to-end query formulation. In *Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–30, 2020

pybool_ir

Domain-specific search has high barrier to entry

- Slow/limited search APIs
- Specific indexing and document processing
- Complex query languages

pybool_ir

Domain-specific search has high barrier to entry

- Slow/limited search APIs
- Specific indexing and document processing
- Complex query languages

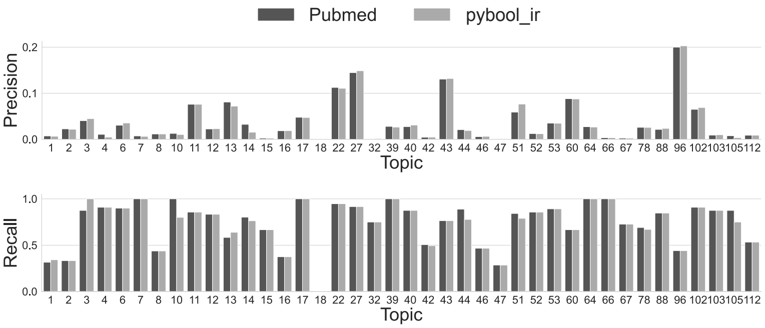
```python
from pybool_ir.experiments.collections import load_collection
from pybool_ir.experiments.retrieval import RetrievalExperiment
from ir_measures import *
import ir_measures

# Automatically downloads, then loads this collection.
col = load_collection("ielab/sysrev-seed-collection")

# Point the experiment to your index, your collection.
with RetrievalExperiment(indexer=PubmedIndexer("./pubmed"),
                         collection=col) as experiment:
        # Get the run of the experiment.
        # This automatically executes the queries.
        run = experiment.run

# Evaluate the run using ir_measures.
ir_measures.calc_aggregate([SetP, SetR, SetF], col.qrels, run)
```

pybool_ir

Domain-specific search has high barrier to entry

- ❑ Slow/limited search APIs
- ❑ Specific indexing and document processing
- ❑ Complex query languages

```
from pybool_ir.experiments.collections import load_collection
from pybool_ir.experiments.retrieval import RetrievalExperiment
from ir_measures import *
import ir_measures

# Automatically downloads, then loads this collection.
col = load_collection("ielab/sysrev-seed-collection")

# Point the experiment to your index, your collection.
with RetrievalExperiment(indexer=PubmedIndexer("./pubmed"),
                         collection=col) as experiment:
        # Get the run of the experiment.
        # This automatically executes the queries.
        run = experiment.run

# Evaluate the run using ir_measures.
ir_measures.calc_aggregate([SetP, SetR, SetF], col.qrels, run)
```

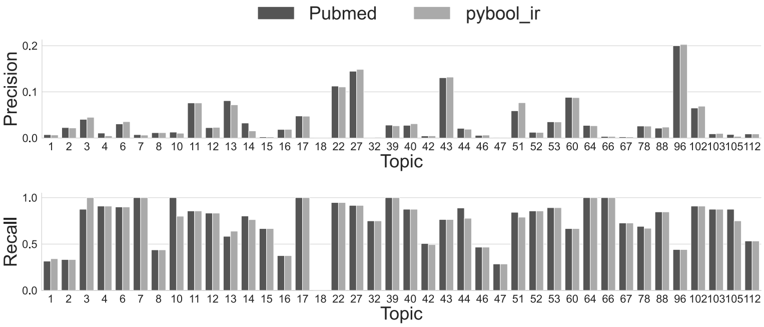


pybool_ir

Domain-specific search has high barrier to entry

- Slow/limited search APIs
- Specific indexing and document processing
- Complex query languages

```
from pybool_ir.experiments.collections import load_collection
from pybool_ir.experiments.retrieval import RetrievalExperiment
from ir_measures import *
import ir_measures

# Automatically downloads, then loads this collection.
col = load_collection("ielab/sysrev-seed-collection")

# Point the experiment to your index, your collection.
with RetrievalExperiment(indexer=PubmedIndexer("./pubmed"),
                         collection=col) as experiment:
        # Get the run of the experiment.
        # This automatically executes the queries.
        run = experiment.run

# Evaluate the run using ir_measures.
ir_measures.calc_aggregate([SetP, SetR, SetF], col.qrels, run)
```



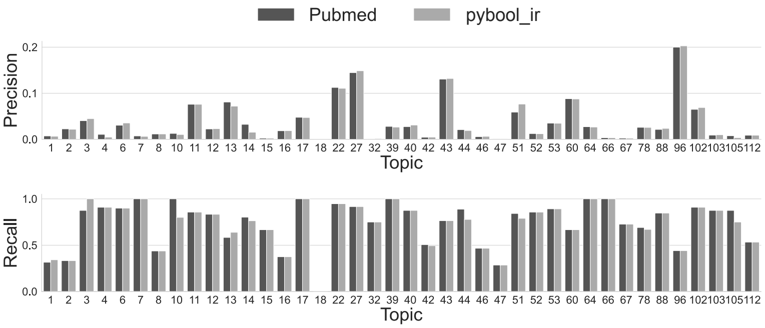
- Domain-specific indexing
  - ir_datasets
  - Arbitrary JSON
- Extend query syntaxes
  - Smooth operators
  - Faster demo prototyping
- Compatibility with pyserini
  - index ➜ search

pybool_ir

Domain-specific search has high barrier to entry

- Slow/limited search APIs
- Specific indexing and document processing
- Complex query languages

```python
from pybool_ir.experiments.collections import load_collection
from pybool_ir.experiments.retrieval import RetrievalExperiment
from ir_measures import *
import ir_measures

# Automatically downloads, then loads this collection.
col = load_collection("ielab/sysrev-seed-collection")

# Point the experiment to your index, your collection.
with RetrievalExperiment(indexer=PubmedIndexer("./pubmed"),
                         collection=col) as experiment:
        # Get the run of the experiment.
        # This automatically executes the queries.
        run = experiment.run

# Evaluate the run using ir_measures.
ir_measures.calc_aggregate([SetP, SetR, SetF], col.qrels, run)
```



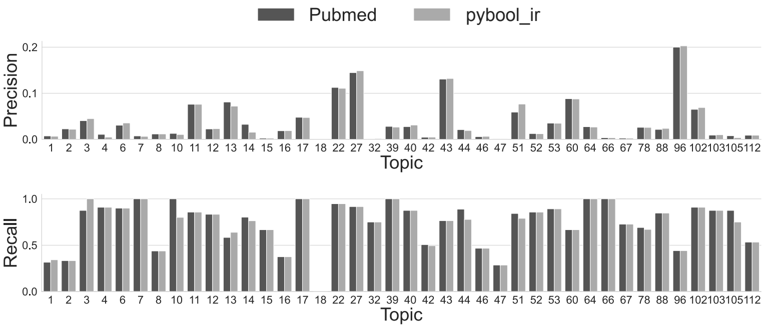
- Domain-specific indexing
  - ir_datasets
  - Arbitrary JSON
- Extend query syntaxes
  - Smooth operators
  - Faster demo prototyping
- Compatibility with pyserini
  - index ➜ search

pybool_ir

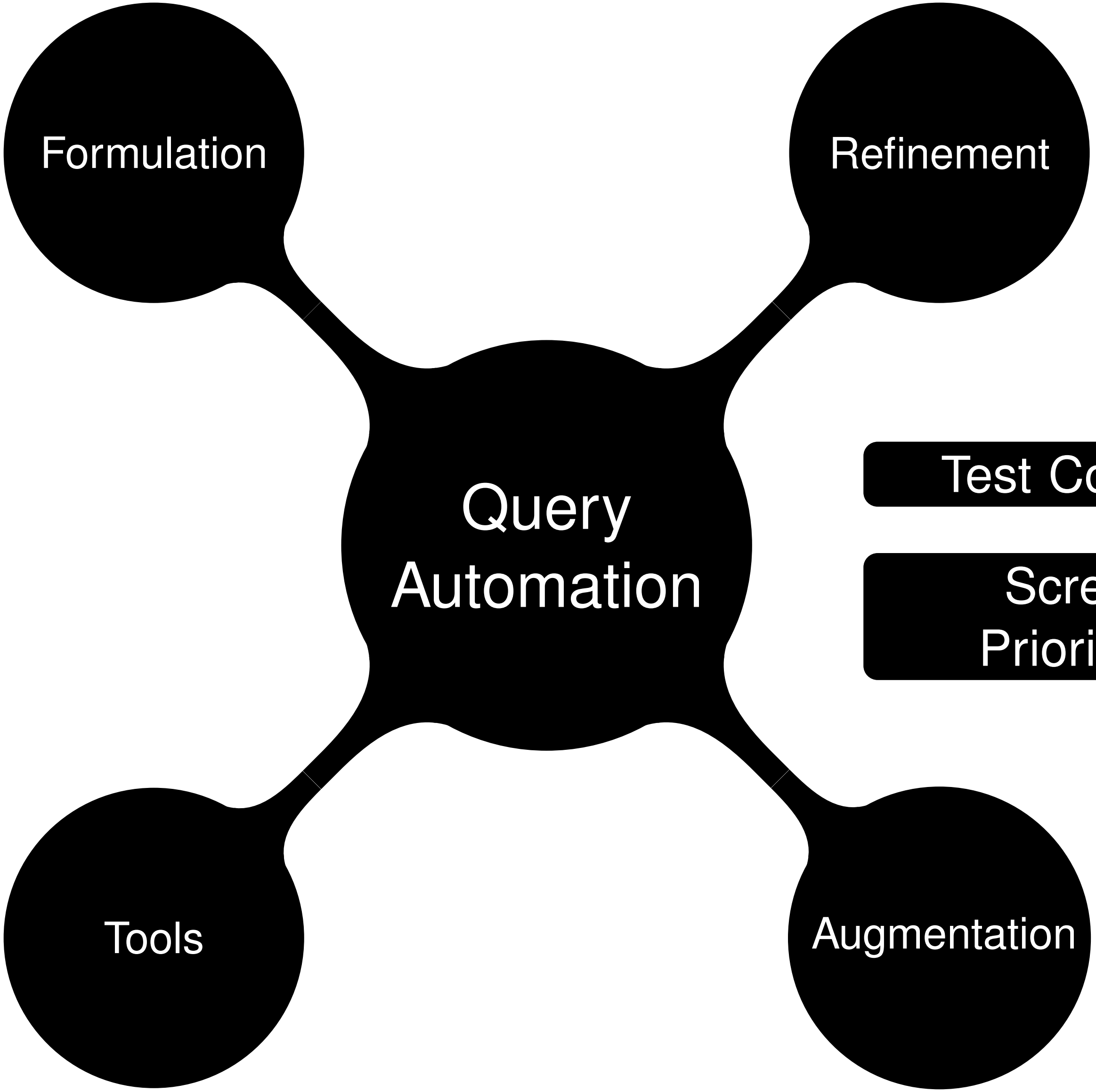
Domain-specific search has high barrier to entry

- Slow/limited search APIs
- Specific indexing and document processing
- Complex query languages

```python
from pybool_ir.experiments.collections import load_collection
from pybool_ir.experiments.retrieval import RetrievalExperiment
from ir_measures import *
import ir_measures

# Automatically downloads, then loads this collection.
col = load_collection("ielab/sysrev-seed-collection")

# Point the experiment to your index, your collection.
with RetrievalExperiment(indexer=PubmedIndexer("./pubmed"),
                         collection=col) as experiment:
        # Get the run of the experiment.
        # This automatically executes the queries.
        run = experiment.run

# Evaluate the run using ir_measures.
ir_measures.calc_aggregate([SetP, SetR, SetF], col.qrels, run)
```



- Domain-specific indexing
  - ir_datasets
  - Arbitrary JSON
- Extend query syntaxes
  - Smooth operators
  - Faster demo prototyping
- Compatibility with pyserini
  - index ➜ search



Query
Automation

Formulation

Refinement

Tools

Augmentation

Test Collections

Screening
Prioritisation

Outlook

- **Money** ➜ Can cost upwards of 250,000 Euros
- **Time** ➜ Can take over a year

[McGowan and Sampson, 2005]

Outlook

- **Money** ➜ Can cost upwards of 250,000 Euros
- **Time** ➜ Can take over a year

[McGowan and Sampson, 2005]

Does hydrocloroquin treat COVID-19?

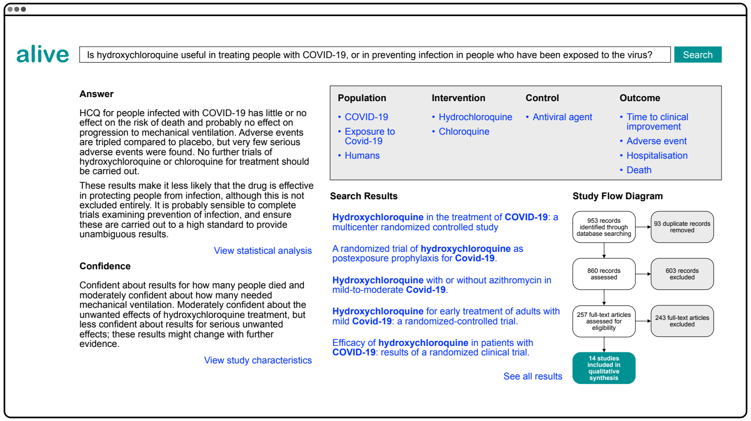
Should everyone wear a mask during the COVID-19 pandemic?

- Priority for decision making?
- Level of uncertainty in the literature?
- Frequency of new studies is high?

Conclusion

Next steps

- Evaluating generative IR
  Is information relevant/correct/readable?
- Making query development easier
  Formulation, refinement, augmentation
- Furthering tools to enable research
  Reproduction with pybool_ir



Is hydroxychloroquine useful in treating people with COVID-19, or in preventing infection in people who have been exposed to the virus?

Answer

HCQ for people infected with COVID-19 has little or no effect on the risk of death and probably no effect on progression to mechanical ventilation. Adverse events are tripled compared to placebo, but very few serious adverse events were found. No further trials of hydroxychloroquine or chloroquine for treatment should be carried out.

These results make it less likely that the drug is effective in protecting people from infection, although this is not excluded entirely. It is probably sensible to complete trials examining prevention of infection, and ensure these are carried out to a high standard to provide unambiguous results.

View statistical analysis

Confidence

Confident about results for how many people died and moderately confident about how many needed mechanical ventilation. Moderately confident about the unwanted effects of hydroxychloroquine treatment, but less confident about results for serious unwanted effects; these results might change with further evidence.

View study characteristics

Population	Intervention	Control	Outcome
• COVID-19 • Exposure to Covid-19 • Humans	• Hydrochloroquine • Chloroquine	• Antiviral agent	• Time to clinical improvement • Adverse event • Hospitalisation • Death

Search Results

Hydroxychloroquine in the treatment of **COVID-19**: a multicenter randomized controlled study

A randomized trial of **hydroxychloroquine** as postexposure prophylaxis for **Covid-19**.

Hydroxychloroquine with or without azithromycin in mild-to-moderate **Covid-19**.

Hydroxychloroquine for early treatment of adults with mild **Covid-19**: a randomized-controlled trial.

Efficacy of **hydroxychloroquine** in patients with **COVID-19**: results of a randomized clinical trial.

See all results

Study Flow Diagram

953 records identified through database searching
93 duplicate records removed
860 records assessed
603 records excluded
257 full-text articles assessed for eligibility
243 full-text articles excluded
14 studies included in qualitative synthesis

Envisioned outcomes

- Faster and less expensive systematic reviews
- Fully automated evidence synthesis
- Tools for librarians and researchers to automate evidence creation

Stay in touch

- @hscells
- https://scells.me
- harry.scells@uni-leipzig.de