Erasmus MC
University Medical Center Rotterdam

Removing duplicates in retrieval sets from electronic databases

comparing the efficiency and accuracy of the Bramer-method with other methods and software packages

Wichor Bramer – Erasmus MC – Medical Library

Leslie Holland, Jurgen Mollema, Todd Hannon, Tanja Bekhuis (USA / NL)

What are duplicate referencess?

Referering to the same bibliographic entity

Unique identifiers?

DOI / PMID

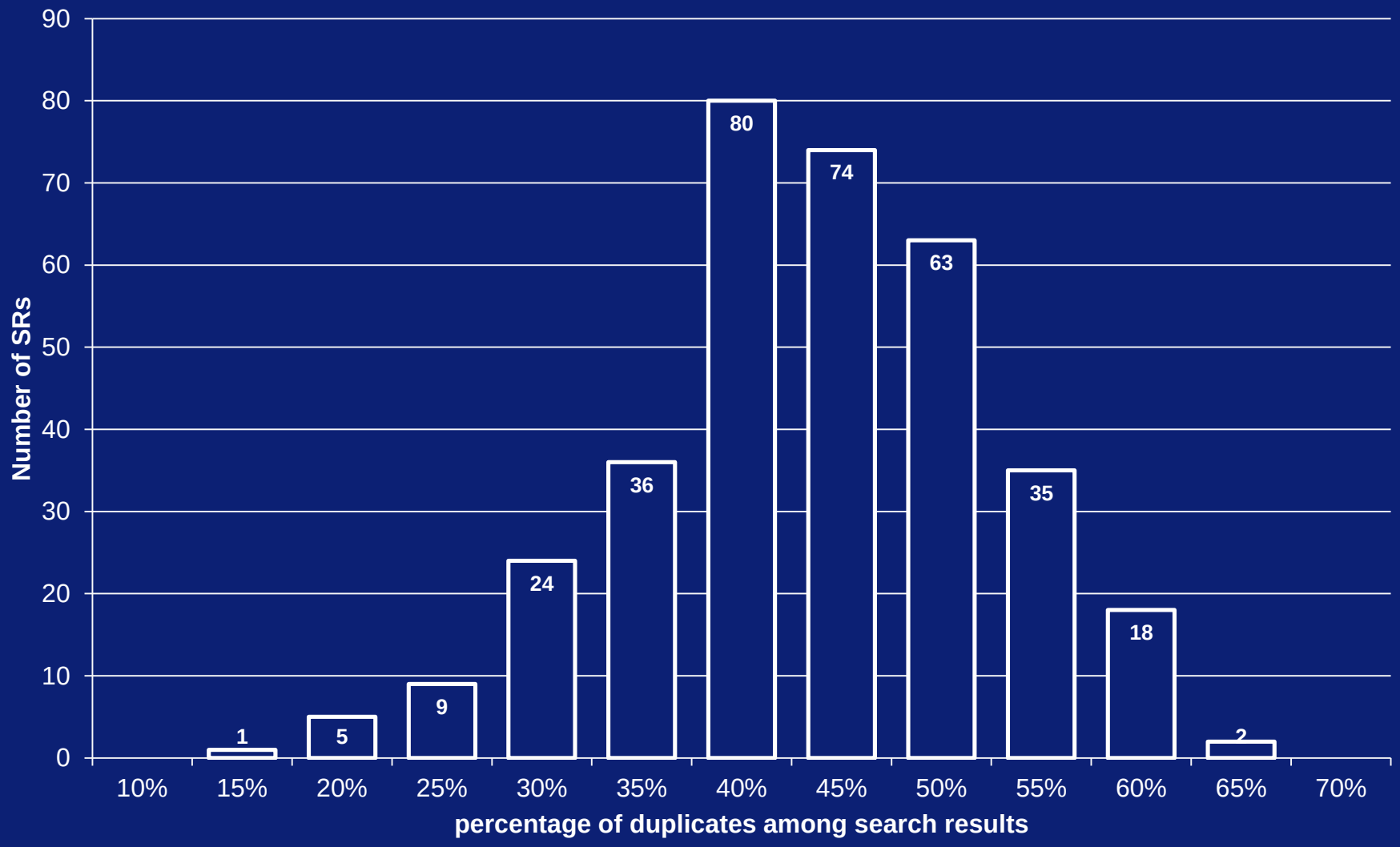
→ Not always present in database or in export files

→ Limited use in software

Equal author, title, journal, volume, issue, pages

→ Data can vary between databases or in time

Removing duplicates is important (median 43%)



Removing duplicates is cumbersome

- Do you deduplicate for your patrons?

- … Does not use default settings because of abbreviated and long forms of journal names.
- … Several iterations with different settings. Ends with manual scan.
- … Manually checks author names and page numbers to de-dupe.
- … Manually de-dupes in reverse chronological order.

Removing duplicates is problematic

- “Missed duplicates despite best efforts”
- “Authors who publish similar titles at various conferences”
- “Having to manually eyeball exact matches”
- “De-duping can take forever”

Removing duplicates is time consuming

Number of references	Average time needed
500	30 minutes
2000	1.5 hours
10000	6 hours

Sources: non-published questionnaires by Bekhuis, and by Bramer

Challenges for deduplication methods

- Reduce the number of hits substantially
- Without deleting false duplicates

  Not not any or too much?
- Without taking hours to perform

Methods for deduplication

Software programs

Endnote	Reference Manager	Refworks
Papers	Mendeley	Zotero
Jabref	Paperpile	*and?*

Published algorithms

- Qi, Yang et al, 2013 – PLoS One
- Jiang, Lin et al, 2014 – Database

Own algorithm

Bramer method

Methods

- Three gold standard sets
  - Around 1000 records each
  - 4 databases (embase.com, medline OvidSP, Web-of-Science, Scopus)
  - Deduplicated manually (author sorted, title sorted, manual comparison)
- Golden standard sets deduplicated using the standard methods of the software

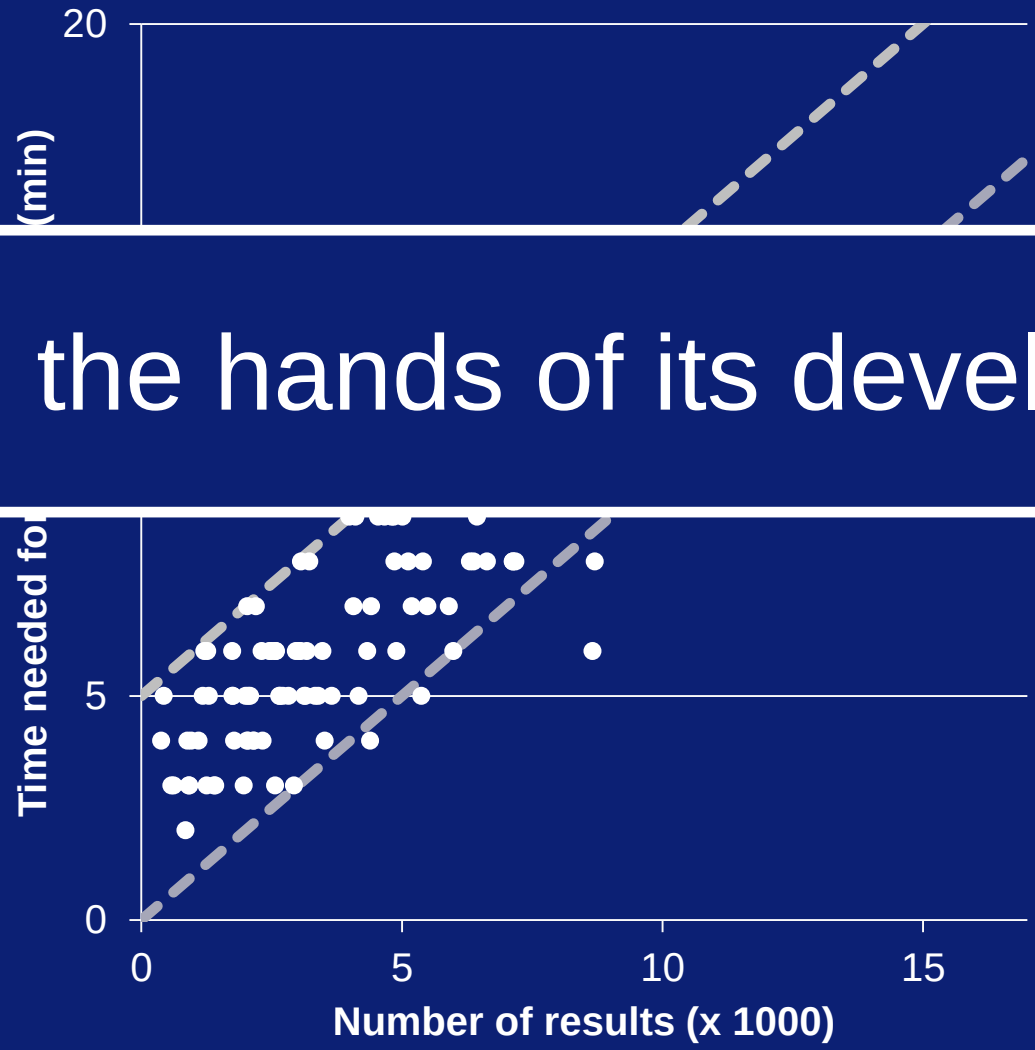
  → recording effort (time and clicks)
- Results compared to hand deduplicated results

  → # of records en # false duplicates

For now by one person, but plans are to repeat the experiments

	hits after	# false	time	# of clicks	score
Bramer algorithm	101%	1	6	157	8,9
endnote qi	100%	8	10	285	6,6
manual	100%	4	15	992	6,3
paperpile	115%	0	8	5	5,5
mendeley import	103%	13	1	0	5,1
jabref	115%	4	1	0	4,5
mendeley check	100%	18	1	13	4,4
refworks close	102%	15	8	109	4,4
zotero	102%	13	15	337	4,2
endnote standaard	121%	0	1	5	4,0
refman author	119%	5	9	17	2,6
endnote web	142%	0	3	2	-1,9
refman standard	136%	6	7	9	-2,5
refman algorithm	98%	44	13	64	-3,8
refworks exact	150%	0	7	59	-4,6

The Bramer method is fast



In the hands of its developer

Is the Bramer method accurate?

Golden standard:	1 error in 3423 records	→ 0,03%
Qi reference set:	2 errors in 22339 records	→ 0,01%
Jiang reference set:	14 errors in 6265 records	→ 0,22%

Two equal conference proceedings	4
Updated Cochrane review	4
Conference proceedings kept full text dropped	4
Truly false duplicates removed	2

10?	→	0,16%
6?	→	0,10%
2?	→	0,03%

Discussion

What is a problematic false duplicate (what is a valueable bibliographic entity)	Librarians (N=7)	Researchers (N=27)
Conf – Conf	71%	7%
Full – Conf	57%	2%
Conf – Full	86%	93%
When you consider that for relevant conference papers you try to find the published article		29%
Version 2 – Version 1	64%	20%

Discussion

Is it problematic to falsely delete 0.2% unique references?

With on average 2-3% of the results included

0.2% deduplication errors means 0.5 include missed in 10,000 references

(How sure are you that the search did not miss any relevant articles)

Limitations of the Bramer method

- Bound to EndNote software package
- Data restructuring helpfull (required for speed) :

  embase, WoS, Scopus: abbreviated journal titles

  medline / cochrane: full page numbers
- Possibly rather steep learning curve

Ongoing research

You are invited to use the Bramer method for your own deduplication process

- Please share your experiences about its speed and accuracy
- We will continue comparing other (new) methods
- And replicate the experiments already performed by the first author