MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Developing Optimal Search Strategies for Detecting Clinically Relevant Qualitative Studies in MEDLINE

Sharon S.-L. Wong, Nancy L. Wilczynski, R. Brian Haynes, for the Hedges Team

Health Information Research Unit, McMaster University, Hamilton, Ontario, Canada

Abstract

Background: The growing interest in qualitative research within the evidence based practice framework highlights the need for accurate search strategies to enhance the retrieval of qualitative studies. To date, little work has been done on developing optimal search filters for retrieving qualitative studies. The current study extends our earlier work, on developing optimal search strategies, to include qualitative studies.

Objective: To develop optimal search strategies for detecting clinically relevant qualitative studies in MEDLINE in the publishing year 2000.

Design: Comparison of the retrieval performance of methodologic search strategies in MEDLINE with a manual review ("gold standard") of each article for each issue of 161 core health care journals for the year 2000.

Methods: 6 experienced research assistants who had been trained and intensively calibrated reviewed all issues of 161 journals for the publishing year 2000. Each article was systematically classified for "format" (whether it was an original study, review article, general article, or case report), "interest" (whether or not it was of interest to the health care of humans), and "purpose" (whether it pertained to therapy, diagnosis, prognosis, causation, economics, costs, or clinical prediction; was of a qualitative nature; or was about something else). Search strategies were developed for all purpose categories, including qualitative studies.

Main outcome measures: The sensitivity (recall), specificity, precision, and accuracy of single and combinations of search terms.

Results: 49,028 articles were identified after matching the hand search records with the data downloaded from MEDLINE, of which 366 (0.75%) were classified as qualitative. Combinations of search terms reached peak sensitivities of 95%. Compared with the best single term, a three-term strategy increased sensitivity for qualitative studies by 23.6% (absolute increase), but with some loss of specificity when sensitivity was maximized. When search terms were combined to optimize sensitivity and specificity, both these values peaked above 90%.

Conclusion: *Several search strategies can achieve high performance in retrieving qualitative studies from MEDLINE.*

Keywords:

Information storage and retrieval, Qualitative research, Medical informatics.

Introduction

In the past decade, qualitative research has been increasing in volume and interest for exploring health care problems [1]. Qualitative research seeks to understand and interpret personal experiences, behaviors, interactions, and social contexts to explain the phenomena of interest [2], such as the attitudes, beliefs, and perspectives of patients and clinicians; the interpersonal nature of caregiver and patient relationships; the illness experience; or the impact of human suffering. The term, "qualitative", represents various research methodologies including ethnography, phenomenology, grounded theory, and narrative analysis [2]. Instead of quantifying or statistically portraying the data, qualitative research focuses on the narrative account of the individual and in so doing gives voice to the patient or provider in the health care decision making process.

As the body of qualitative research grows, the task of identifying clinically relevant qualitative studies is becoming more difficult, for several reasons. First, qualitative research is a very small portion of the over two million new articles published every year in the biomedical research literature [3, 4], all of which are mixed in electronic databases such as MEDLINE. Second, inconsistent indexing can hinder the ability to successfully search the literature. Qualitative studies that use creative titles or provide inadequate information in their abstracts have a greater risk of not being indexed appropriately [5], especially if persons involved in maintaining bibliographic databases are not familiar with the spectrum of qualitative terminology. Finally, because bibliographic databases vary in their indexing practices, an effective search may necessitate using more than one database and different search filters. For example, researchers suggest that the indexing of qualitative studies in MEDLINE makes use of fewer methodologic index terms that describe the qualitative design than the indexing of the same studies in CINAHL, possibly because CINAHL focuses on professions that commonly use qualitative methodologies [5, 6]. Therefore, searching MEDLINE plus CINAHL may improve the yield for relevant qualitative studies.

Methodologic search filters have been developed for improving the accuracy of searching for clinically relevant and sound quantitative studies [7]. For qualitative research, however, only preliminary work has been done in developing search filters for Ovid CINAHL by Nesbit [8].

Using a gold standard based on a set of 161 journals for the year 2000, we evaluated as part of a larger study the retrieval properties of search filters for identifying clinically relevant qualitative studies in MEDLINE.

Methods

The operating characteristics of methodologic search strategies in MEDLINE (accessed using Ovid) were compared with a manual review of all articles in each issue of 161 core health care journals for the year 2000. To evaluate MEDLINE strategies designed to retrieve qualitative studies, MeSH terms and textwords related to qualitative research design features were run as search strategies. These search strategies were treated as diagnostic tests for qualitative studies, and the manual review of the literature was treated as the "gold standard." The sensitivity (recall), specificity, precision, and accuracy of MEDLINE searches were determined. For example, for each MEDLINE search strategy designed to retrieve qualitative studies, sensitivity (recall) was calculated as the proportion of qualitative articles that were retrieved; specificity as the proportion of non-qualitative articles (those articles not classified as qualitative) that were not retrieved; precision as the proportion of retrieved articles that were classified as qualitative studies; and accuracy as the proportion of all citations that were correctly classified.

Six research assistants assessed all articles for each issue of 161 journals for the year 2000. For articles in 7 purpose categories (causation, prognosis, diagnosis, treatment, economics, clinical prediction, and reviews), methodologic criteria were applied to determine if the article was scientifically sound. Purpose category definitions were used to classify qualitative and cost studies, but methodologic criteria were not applied to these types of studies. Purpose category definitions and methodologic rigor criteria have previously been published [9]. Original studies or review articles that were of interest to the health care of humans were classified as qualitative if the content related to how people experience certain situations and if the data collection and analytical methods used were appropriate for qualitative data.

The 161 journals reviewed in the year 2000 were selected using an iterative process based on recommendations of clinicians and librarians, Science Citation Index Impact Factors, and their ongoing yield of sound and clinically relevant studies and reviews for the disciplines of internal medicine, general medical practice, mental health, and general nursing practice (a list of reviewed journals is available upon request from authors). Research staff underwent training and intensive calibration; inter-rater reliability (assessed by the kappa statistic) for classifying articles according to methodologic criteria was greater than 80% for all purpose categories [9].

To construct a comprehensive set of search terms, we generated a list of MeSH terms and textwords from qualitative research studies, and sought input from clinicians and librarians in the United States and Canada through interviews with known searchers, requests at meetings and conferences, and requests to the National Library of Medicine. These experts were asked which terms or phrases they used when searching for studies of causation, prognosis, diagnosis, treatment, economics, clinical prediction, reviews, costs, and studies of a qualitative nature. Terms could be MeSH terms, including publication types (pt), check tags, and subheadings (sh), or textwords (tw) denoting methodology in titles and abstracts of articles. We compiled a list of 5,345 terms for all types of studies (a list of tested terms is available upon request from authors) and tested all of these for each category, including qualitative studies. Search strategies were tested and developed in a random 60% of the entire database (the development data set) and validated in the remaining 40% of the database (the validation data set).

Results

49,028 articles were identified after matching the hand search records with the data downloaded from MEDLINE. 366 articles (0.75%) were classified as qualitative (358 original studies and 8 reviews).

The single terms having the best sensitivity, best specificity, and best optimization of sensitivity and specificity for detecting qualitative articles in MEDLINE in 2000 were derived in the development data set (29,397 articles) and are displayed in Table 1; their operating characteristics tested in the validation data set (19,631 articles) are also displayed in Table 1. The top performing single terms had high specificities (> 97%). When specificity was maximized, a pronounced but expected reduction in sensitivity and increase in precision (absolute increase of 13% in both development and validation data sets) were seen.

The operating characteristics of top-performing two- or three-term strategies are displayed in Table 2. Compared with single terms, combined terms reached higher peak sensitivities albeit with reduced specificity.

The three-term strategy, "interview:.tw. OR px.fs. OR exp health services administration", yielded the best sensitivity, 95.05%, and had a specificity of 69.98% (development data set). Compared with the best sensitivity single term, "interview:.mp." (71.43% sensitivity, 97.05% specificity; development data set), the best three-term strategy yielded an absolute increase in sensitivity of 23.62%, but with an absolute loss in specificity of 27.07%. Also as expected, when sensitivity was increased, precision was compromised (falling from 13.10% [best sensitivity single term] to 1.93% [best sensitivity combined term] in the development data set).

The two-term strategy, "qualitative.tw. OR themes.tw.", yielded the best specificity (superior to any of the three-term strategies), 99.36% (development data set), but with a definite trade-off in sensitivity, which lowered to 60.99% (development data set). Yet when specificity was maximized, a comparatively remarkable rise was also seen with precision, which reached 37.37% (development data set). Compared with the most sensitive three-term strategy, this represents an absolute increase in precision of 35.44% (comparing within the development data set). When search terms were combined to optimize sensitivity and specificity, these values were >90% in the development data set.

No statistically significant differences were seen for any of the operating characteristics in the development compared with the validation data sets for any of the single or combined terms.

Table 1: Single Terms with the Best Sensitivity (keeping Specificity ≥ 50%), Best Specificity (keeping Sensitivity ≥ 50%), and Best Optimization of Sensitivity and Specificity (based on absolute [sensitivity-specificity] < 26%) for Detecting Qualitative Articles in MEDLINE in 2000

OVID Search Terms	Sensitivity (%) Development Validation Difference (95% CI)*	Specificity (%) Development Validation Difference (95% CI)*	Precision (%) Development Validation Difference (95% CI)*	Accuracy (%) Development Validation Difference (95% CI)*
Best Sensitivity†	71.43	97.05	13.10	96.89
interview:.mp.	66.44	97.17	14.95	96.94
	-5.0 (-15 to 5.1)	0.12 (-0.19 to 0.42)	1.8 (-1.6 to 5.3)	0.05 (-0.26 to 0.36)
Best Specificity	55.49	99.01	25.83	98.74
interviews.tw.	52.74	98.99	28.21	98.65
	-2.8 (-13.6 to 8.1)	-0.01 (-0.19 to 0.17)	2.4 (-4.5 to 9.3)	-0.09 (-0.29 to 0.12)
Best Optimization of Sensitivity and Specificity†	71.43	97.05	13.10	96.89
	66.44	97.17	14.95	96.94
interview:.mp.	-5.0 (-15 to 5.1)	0.12 (-0.19 to 0.42)	1.8 (-1.6 to 5.3)	0.05 (-0.26 to 0.36)

*None of the differences between the development and validation data sets were statistically significant.
†The same single term gives the best sensitivity and the best optimization of sensitivity and specificity.

Discussion

We have developed search filters that can enhance the retrieval of qualitative studies. Searchers should examine our filters and determine the most desirable trade-off between sensitivity and specificity for their needs. Those willing to spend the time to sort out irrelevant articles to avoid missing key articles would benefit most from using a highly sensitive strategy. If an exhaustive collection of relevant articles were not needed, a highly specific strategy would be best.

In our strategies, precision was generally low. Maximizing specificity improved precision, which was particularly evident with the best specificity combined term where precision reached 37%. This trend was not surprising considering that specificity is a key determinant of precision. Low precision in our study was inevitable because MEDLINE is a large multi-purpose database, containing few qualitative articles. Precision would likely have been greater had we combined our strategies with content terms, or had we tested "and" and "and not" combinations, but this would predictably lead to losses in sensitivity.

In contrast to search filters for quantitative research, our strategies for retrieving qualitative studies did not include criteria for methodologic merit for qualitative studies, and thus do not filter the higher quality from the lower quality articles. This is because we could not find agreement in the literature about criteria for methodologic merit for this type of research. To date, there has been considerable debate over how to appraise the merit of qualitative studies [10, 11]. Because qualitative and quantitative research are so distinct, both philosophically and methodologically, criteria used to determine scientific rigor of quantitative studies are not necessarily fitting for qualitative studies.

Further work is needed to develop and validate more sophisticated search strategies and to determine how well our strategies perform when combined with content and age terms. Qualitative researchers should also help by ensuring that their study methods are clearly described in their publications, particularly in the title and abstract sections. This is especially important because qualitative research encompasses a variety of methodologies (e.g., ethnography, phenomenology, grounded theory). Explicit use of qualitative design terminology would hopefully lead to better indexing of qualitative articles in MEDLINE and other bibliographic databases, and a greater capacity for detecting these types of articles. Because qualitative studies are gaining momentum within the evidence based practice framework, indexing of these studies at least in MEDLINE is evolving. Testing the robustness of our MEDLINE qualitative filters for searching in future publication years may be warranted if indexing changes are made.

Our study focused on developing qualitative search filters for searching in MEDLINE. Given that there is a lack of standardization across bibliographic databases with indexing practices, and that evidence suggests that CINAHL may index qualitative articles more thoroughly than MEDLINE [5], searchers requiring an exhaustive search of qualitative papers may very well benefit from searching more than one database (e.g., searching MEDLINE plus CINAHL). To this end, further work is needed to develop search filters for other databases including for CINAHL, which will be a focus of some of our future work on this project.

Conclusion

Several search strategies can achieve high performance in retrieving qualitative studies from MEDLINE, and the most beneficial trade-off between sensitivity and specificity should be weighed according to the searcher's needs.

Acknowledgments

This research was funded by the National Library of Medicine, USA. The Hedges Team includes Angela Eady, Brian Haynes, Susan Marks, Ann McKibbon, Doug Morgan, Cindy Walker-Dilks, Stephen Walter, Nancy Wilczynski, and Sharon Wong.

References

[1] Green J, Britten N. Qualitative research and evidence based medicine. *BMJ* 1998;316:1230-2.

Table 2: Two- or Three-Term Strategies with the Best with the Best Sensitivity (keeping Specificity ≥ 50%), Best Specificity (keeping Sensitivity ≥ 50%), and Best Optimization of Sensitivity and Specificity (based on absolute [sensitivity-specificity] < 40%) for Detecting Qualitative Articles in MEDLINE in 2000

OVID Search Strategies	Sensitivity (%) Development Validation Diff (95% CI)*	Specificity (%) Development Validation Diff (95% CI)*	Precision (%) Development Validation Diff (95% CI)*	Accuracy (%) Development Validation Diff (95% CI)*
Best Sensitivity				
interview:.tw. OR px.fs. OR exp health services administration	95.05 92.47 -2.6 (-7.9 to 2.7)	69.98 69.94 -0.04 (-0.87 to 0.79)	1.93 2.25 0.32 (-0.15 to 0.79)	70.13 70.10 -0.03 (-0.86 to 0.08)
Best Sensitivity – Small decrease in Sensitivity with large increase in Specificity				
interview:.mp. OR px.fs. OR qualitative.tw.	93.96 91.10 -2.9 (-8.6 to 2.9)	88.78 89.17 0.39 (-0.18 to 0.96)	4.96 5.93 0.97 (-0.25 to 2.2)	88.81 89.18 0.37 (-0.19 to 0.94)
Best Specificity	60.99	99.36	37.37	99.13
qualitative.tw. OR themes.tw.	53.42 -7.6 (-18 to 3.2)	99.39 0.03 (-0.12 to 0.17)	39.59 2.2 (-6.5 to 11)	99.05 -0.08 (-0.25 to 0.09)
Best Specificity – Small decrease in Specificity with large increase in Sensitivity				
interviews.mp,pt. OR qualitative.mp. OR experiences.tw.	84.07 78.77 -5.3 (-14 to 3.2)	97.83 97.86 0.03 (-0.23 to 0.30)	19.42 21.62 2.2 (-2.2 to 6.7)	97.74 97.72 -0.02 (-0.29 to 0.25)
Best Optimization of Sensitivity and Specificity	92.31	91.98	6.69	91.98
interview:.mp. OR experience:.mp. OR qualitative.tw.	86.99 -5.3 (-12 to 1.4)	92.00 0.02 (-0.47 to 0.51)	7.53 0.84 (-0.75 to 2.4)	91.96 -0.02 (-0.51 to 0.47)

*None of the differences between the development and validation data sets were statistically significant.

[2] Ploeg J. Identifying the best research design to fit the question. Part 2: qualitative designs. *Evid Based Nurs* 1999;2:36-7.

[3] Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. *JAMA* 2002;287:2805-8.

[4] Haynes RB, Sackett DL, Tugwell P. Problems in handling of clinical and research evidence by medical practitioners. *Arch Intern Med* 1983;143:1971-5.

[5] Evans D. Database searches for qualitative research. *J Med Libr Assoc* 2002;90:290-3.

[6] Barroso J, Gollop CJ, Sandelowski M, Meynell J, Pearce PF, Collins LJ. The challenges of searching for and retrieving qualitative studies. *W J Nurs Res* 2003;25:153-78.

[7] Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. *Int J Epidemiol.* 2002;31:150-3.

[8] University of Rochester Medical Center. Edward G. Miner Library. Evidence-based filters for Ovid CINAHL. http://www.urmc.rochester.edu/hslt/miner/digital_library/tip_sheets/cinahl_eb_filters.pdf Accessed September 5, 2003.

[9] Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search literature. *Medinfo 2001*;10(Pt 1): 390-3.

[10]Barbour RS. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *BMJ* 2001;322:1115-7.

[11]Morse JM, Barrett M, Mayan M, Olson K, Spiers J. Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods 1* (2). 2002. Article 2.