**Evidence Based Library and Information Practice**

*Commentary*

# How to Develop a Validated Geographic Search Filter: Five Key Steps

Lynda Ayiku
Information Specialist
National Institute for Health and Care Excellence (NICE)
Manchester, United Kingdom
Email: lynda.ayiku@nice.org.uk

Jenny Craven
Information Specialist
National Institute for Health and Care Excellence (NICE)
Manchester, United Kingdom
Email: jenny.craven@nice.org.uk

Thomas Hudson
Information Specialist
National Institute for Health and Care Excellence (NICE)
Manchester, United Kingdom
Email: thomas.hudson@nice.org.uk

Paul Levay
Information Specialist
National Institute for Health and Care Excellence (NICE)
Manchester, United Kingdom
Email: paul.levay@nice.org.uk

**Received:** 5 Sept. 2019 **Accepted:** 24 Jan. 2020



DOI: 10.18438/eblip29633

**Introduction**

The purpose of this commentary is to increase awareness of the existing validated geographic search filters and to encourage the creation of new filters for additional places in the world.

Search filters are collections of search terms that are designed to find evidence with a common feature (Glanville et al., 2008). They differ from search strategies because their retrieval ability has been tested (validated) against a set of relevant references (Glanville et al., 2008). This provides users with an indication of how successfully filters work for retrieving the type of evidence that they wish to identify.

Most filters aim to retrieve evidence with a specific study design (Damarell, May, Hammond, Sladek & Tieman, 2019). Information professionals will probably be most familiar with those for systematic reviews or randomized controlled trials. However, an increasing number of "topic search filters" have been developed for clinical conditions, demography, health care delivery issues, and geographic locations (Damarell et al., 2019).

Geographic search filters are applied to literature searches with the aim of retrieving evidence about geographic locations such as continents or countries. As of 2020, only three validated geographic filters are available in published literature (Glanville, Lefebvre & Wright, 2020):

1. Spain: PubMed (Valderas, Mendivil, Parada, Losada-Yáñez,& Alonso, 2006)
2. Africa: PubMed and Embase (Pienaar, Grobler, Busgeeth, Eisinga, & Siegfried, 2011)
3. UK: MEDLINE and Embase, OVID platform (Ayiku et al., 2017, 2019)

There are search strategies for other geographic locations that are labelled as "search filters", but these have not been created and validated using recognized filter development methods (Ayiku et al., 2017).

Geographic restrictions are not always applied to searches with a geographic focus when validated geographic filters are unavailable. For instance, in a post-development study for the National Institute for Health and Care Excellence (NICE) UK filters, 100 UK-focused systematic reviews were identified that had no geographic restrictions in their searches (the searches were conducted before the UK filters were available publicly) (Ayiku & Finnegan, 2019). A potential reason for this is that information professionals may have concerns about excluding relevant geographic evidence by accident through the use of untested search approaches. However, when restrictions are not applied, references about a specific location need to be identified from a larger set of irrelevant geographic literature. This approach is time-consuming and inefficient.

Geographic filters enable effective and efficient literature searches for topics with a geographic focus. They can retrieve most of the evidence about a geographic region while limiting the retrieval of irrelevant references about other geographic regions (Ayiku et al., 2017, 2019). Geographic filters therefore save time and associated resource costs spent on selecting evidence for topics about specific regions.

**Developing and Validating Geographic Search Filters: Five Key Steps**

The following steps are based on filter development methodologies (Jenkins, 2004; Sampson et al., 2006; Glanville et al., 2008) in addition to the authors' knowledge gained during the creation of the NICE UK filters for MEDLINE and Embase (Ayiku et al., 2017, 2019). The process for developing geographic filters is outlined in Figure 1.

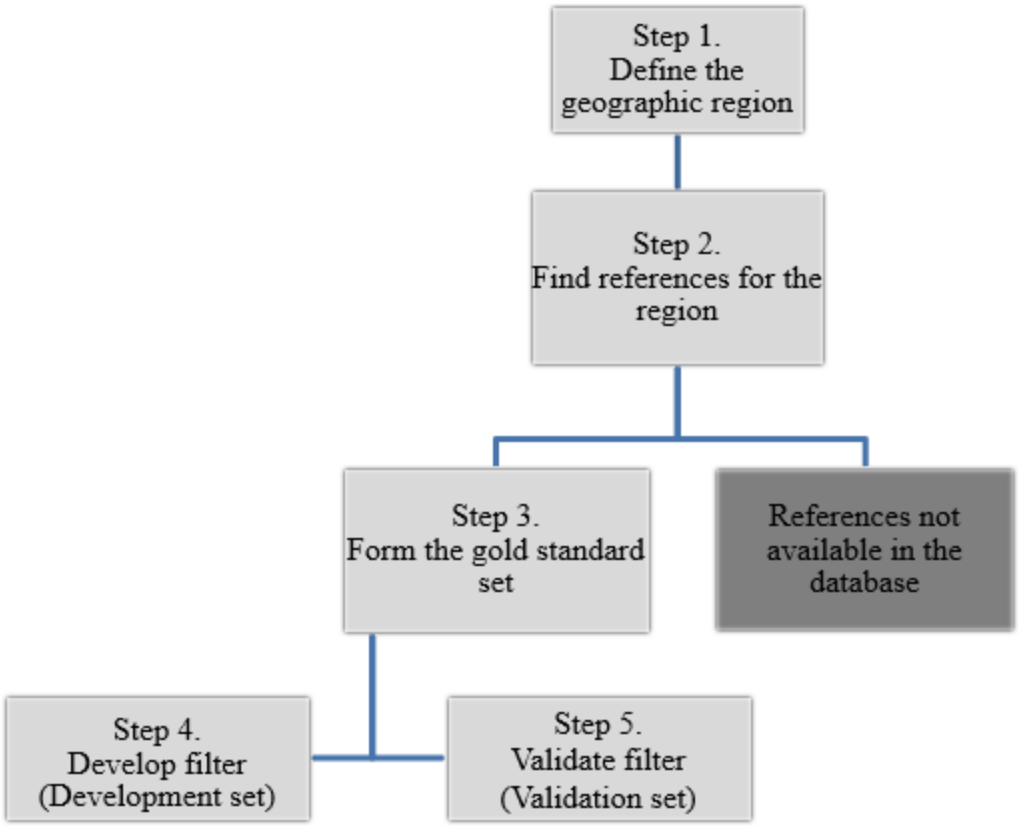


Figure 1
Process for developing a geographic search filter.

***Step 1. Define the Geographic Region***

Official definitions can help to specify the geographic region for the filter if required.

***Step 2. Find References for the Region***

*2a. Identifying References*

A set of references about the geographic region for the filter is required to develop and validate geographic search filters. This set is called a "gold standard" (also known as a "reference set") (Jenkins, 2004). Evidence based sources such as systematic reviews or guidelines usually provide descriptions about the geographic setting of the references that informed them. The gold standard set can be created by pooling relevant references that have informed evidence-based sources (Sampson et al., 2006). The aim is to enable the pragmatic collection of references that have been previously identified for the topic of the filter. This method of reference identification is used to validate filters via the "relative recall" approach and it is quicker than finding relevant references by hand searching journals (Sampson et al., 2006). However, hand searching can be used to create a gold standard set for geographic filters if preferred.

The authors identified references with a UK setting for the gold standard set from NICE guidance documents to develop the NICE UK filters (Ayiku et al., 2017, 2019).

*2b. How Many References for the Gold Standard Set are Needed?*

The authors advise that at least 300 references about a geographic location should be identified for the gold standard set. This is because it is possible that some references will not be available in the bibliographic database for the filter. In addition, the references will need to be divided into the following sets:

1. Development set: used to create filters
2. Validation set: used to validate filters

Sampson et al. (2006) suggest that at least 100 references are required to validate filters because this sample size will provide a reasonable confidence interval (assuming that the filter retrieves 90% of the validation set references). Finding a minimum of 300 references will help to ensure that there are at least 100 references for the validation set.

***Step 3. Form the Gold Standard Set***

*3a. Locating References in the Bibliographic Database for the Filter*

When 300 or more references have been identified, their availability in the database for the filter needs to be checked. To locate the references in the database, enter key bibliographic details (such as title and author) for each reference into the database. The references that are available will form the gold standard set.

The existing geographic filters have been designed for the PubMed, MEDLINE, and Embase bibliographic databases (Valderas et al., 2006; Pienaar et al., 2011; Ayiku et al., 2017, 2019). However, it may be appropriate to design a filter for another database if it is relevant to do so.

*3b. Creating the Development and Validation Sets*

Next, the references in the gold standard set need to be split into a development set and a validation set. For rigor, the references should be randomized prior to their division. To do this, assign each of the references a number (this could simply be their number order). A free online randomizer tool can be used to randomize the numbers. The authors used RANDOM.ORG (Randomness and Integrity Services Ltd, 2020) for the NICE UK filters (Ayiku et al., 2017, 2019).

Once the references have been randomized and divided, create two search strategies in the database for the filter; one for the development set references and another for the validation set references. For both search strategies, combine the references at the end using the OR Boolean operator. As an example, the NICE UK filter search strategies for the development set and validation set references were structured as follows:

1. Langford I (author) AND "The potential effects of climate change on winter mortality in England and Wales" (title) AND 1995 (year)
2. Chahal R (author) AND "A study of the morbidity, mortality and long-term survival following radical cystectomy and radical radiotherapy in the treatment of invasive bladder cancer in Yorkshire" (title) AND 2003 (year)
3. Saka O (author) AND "Cost of stroke in the United Kingdom" (title) AND 2009 (year)
4. Etc…
5. 1 OR 2 OR 3 OR 4…

Save both search strategies in the database account so that they can be re-run to test the retrieval ability of the filter during steps four and five.

***Step 4. Develop Filter***

*4a. Development Set*

The purpose of the development set references is to identify the most relevant search fields and search terms to create the geographic filter. Creating filters using fields and terms from the development set references will help to ensure that the most relevant details for the filter are identified (Hausner, Waffenschmidt, Kaiser, & Simon, 2012). Filters that are created in this way are known as "objectively-derived" filters (Jenkins, 2004). The authors used this approach to create the NICE UK filters (Ayiku et al., 2017, 2019).

*Identifying Relevant Search Fields*

An Excel spreadsheet can be used to identify relevant search fields from the development set references. If the filter is for an Ovid database, the "Excel sheet" export option can be used to transfer the database records for development set references into Excel. Using the "CSV" export option will work in a similar way to transfer database records into Excel if the filter is for PubMed.

In the Excel spreadsheet, the content for each search field from the development set database records is displayed in separate columns. The search fields that contain geographic setting details about your region of interest will be the relevant fields for your filter.

The most relevant fields found in Excel for the NICE UK filters (Ayiku et al., 2017, 2019) were:

- Subject heading
- Title
- Abstract
- Journal name
- Institution

UK setting terms also appeared in the 'country of publication' field but it was not included in the final version of the filter. This is because several UK-based publishing companies produce journals that contain international content. However, it may be useful to add the 'country of publication' field if your filter is for a country in which publishing companies are more likely to publish geographic-specific content.

*Identifying Search Terms*

Once the relevant search fields have been identified, word frequency analysis can be conducted to find candidate geographic setting search terms for the filter. The authors used the WriteWords (2020) word and phrase counter tool to conduct the frequency analysis for the NICE UK filters (Ayiku et al., 2017, 2019). WriteWords (2020) is available for free online. Other free online counters are available such as DataBasic (Bhargava & D'Ignazio, 2020) and commercial counters can be used too.

For the NICE UK filters, the authors copied the content contained in each relevant search field from Excel and pasted it into WriteWords (2020) one field at a time. The frequency of single words up to phrases containing four words was then recorded for each field. Next, the high frequency words and phrases used to describe UK settings were examined. The most frequent UK settings identified from the development set references were:

- Countries
- Nationalities
- Cities
- UK National Health Service (NHS)

*4b. Constructing the Filter*

A geographic filter can be drafted once the relevant search fields and geographic setting terms have been identified. Save the draft filter in the database account so that it can be easily re-run to test its retrieval ability.

1. Search terms for country name (e.g., Britain OR GB OR United Kingdom OR UK OR England OR Scotland OR Northern Ireland OR Wales)
2. Search terms for nationalities (e.g., British OR English OR Scottish OR Northern Irish OR Welsh)
3. Search terms for city names (e.g., London OR Edinburgh OR Belfast OR Cardiff)
4. Search terms for national health systems (e.g., National Health Service OR NHS)
5. 1 OR 2 OR 3 OR 4

Figure 2
Example structure for a geographic filter to retrieve evidence about a country.

As an example of a geographic filter structure, an outline of the NICE UK filters is provided in Figure 2. The full NICE UK filters for MEDLINE and Embase can be found in published journal articles (Ayiku et al., 2017, 2019) and in the InterTASC Information Specialists' Sub-Group (ISSG) Search Filter Resource section on geographic search filters (Glanville et al., 2020).

*4c. Internal Validity Test*

When the geographic filter is drafted, the next step is to test how successfully it retrieves the references that were used to create it. This is known as an "internal validity" test (Jenkins, 2004). To do this, run the saved search strategy for the development set references. Next, run the saved search strategy for the draft filter and apply it to the development set search strategy using the AND Boolean operator. For example, the search strategy structure used to test the retrieval ability of the NICE UK filters was as follows:

1. Langford I (author) AND "The potential effects of climate change on winter mortality in England and Wales" (title) AND 1995 (year)
2. Chahal R (author) AND "A study of the morbidity, mortality and long-term survival following radical cystectomy and radical radiotherapy in the treatment of invasive bladder cancer in Yorkshire" (title) AND 2003 (year)
3. Saka O (author) AND "Cost of stroke in the United Kingdom" (title) AND 2009 (year)
4. Etc…
5. 1 OR 2 OR 3 OR 4…
6. Draft UK geographic search filter
7. 5 AND 6

It is unlikely that the draft filter will retrieve all of the development set references because it is rare for search filters to have a 100% retrieval rate. For instance, some references will contain no details about their geographic setting in their database records (Ayiku et al., 2017, 2019).

If the draft filter retrieves all of the references in the development set, it can be validated using the instructions in step five. If the draft filter does not retrieve all of the references, the reasons why the missing references were not retrieved must be investigated. Carefully look through the database records for the missing references to see if any geographic setting details are contained within them. Consider making modifications to the filter to retrieve missing references that contain setting details for the region. Ensure that you record any changes you make to the filter and provide explanations about why the changes were made.

Also make a record of any references that cannot retrieved by the draft filter and explain why the references were not retrieved. Save the final version of the filter in the database account so that it can be easily re-run to validate the filter (see step five).

**Step 5. Validate Filter**

Validation is the final process for filter development. The validation set contains references that have not been used previously to develop the filter and it is used to assess the filter's "external validity" (Glanville et al., 2008). Validating filters using an independent set of references provides an indication of how well filters perform in retrieving relevant evidence in any search (Glanville et al., 2008).

To validate the filter, run the saved search strategy for the validation set references. Next, run the saved search strategy for the final version of the filter. Apply the filter to the validation set search strategy using the AND Boolean operator following the same example structure shown above in step four.

The filter's recall can now be calculated. "Recall", also known as "sensitivity", is used to measure a filter's ability to retrieve a set of known relevant references and it is calculated as follows (Jenkins, 2004):

- Number of relevant records retrieved by filter/Total number of relevant records (× 100 to express as a percentage)

The term "relative recall" is more accurate than "recall" when the relative recall approach has been used to identify references pooled from multiple evidence based sources for the validation set (Sampson et al., 2006), however, in practice both terms are used.

It is unlikely that the filter will achieve 100% recall and the reasons why missing references were not retrieved should be investigated and recorded. There is no standard definition of "high" recall. However, 90% or above has been used as a threshold in previous studies (Beynon et al., 2013). The existing geographic filters performed as follows:

- Spain filter: PubMed: 88.1% recall (Valderas et al., 2006)
- Africa filters: PubMed: 74% recall, Embase: 73% recall (Pienaar et al., 2011)
- NICE UK filters: MEDLINE UK filter: 99.5% recall, Embase UK filter: 99.8% recall (for references with UK identifiers) (Ayiku et al., 2017, 2019)

Note that no changes can be made to the filter once its recall against the validation set has been calculated. Another validation set containing at least 100 previously unused references will need to be created if filter modifications are required to increase recall. In this case, the former validation set becomes a "test set" that was used to inform the filter's development.

*Tips for Creating Filters*

*Seek Advice*

It may be helpful to seek advice from a professional peer with relevant experience if needed.

*Limiting Retrieval of Irrelevant Results*

Some setting names for the geographic region of the filter may be found elsewhere in the world. Using the NOT Boolean operator can help to minimize the retrieval of irrelevant geographic references. For example, the NICE UK filters included the following strategy to help minimize the retrieval of irrelevant geographic references about the US: York NOT "New York" (Ayiku et al., 2017, 2019).

*Language Variations*

If relevant, use language variations for the geographic region. For instance, the Spain filter

included the following language variations for the country: Spain, Espagne, Espana, and Spagna (Valderas et al., 2006).

*Retrieving References by Language*

Consider retrieving references by language if the filter is for a region with a language that is uncommon in other geographic locations. The search strategy to retrieve references by language is: "language.lg" for OVID databases or "language.la" for PubMed (e.g., welsh.lg or welsh.la). Add the language search strategy to the rest of the filter using the OR Boolean operator.

*Share the Filter*

The filter should be published along with the accompanying filter development processes to make it widely available. It will be added to the ISSG Search Filter Resource section on geographic search filters when it is published which will increase its dissemination (Glanville et al., 2020). In addition, the filter could be promoted at conferences and on social media.

*Acknowledge Limitations*

No filter is perfect, it is unlikely that the filter will achieve 100% recall. Make sure to explain why the filter does not retrieve certain geographic references so that users understand its limitations.

*Keep the Filter Up to Date*

Make sure that the filter is kept updated with any changes to the geographic setting terms. The updated filter may not be validated but the original recall level can still be considered as a baseline for this type of change.

## Conclusion

Geographic search filters enable effective and efficient systematic literature searches for topics with a geographic focus. There are currently only three validated filters identified in the published literature for Spain, Africa and the UK (Glanville et al., 2020). The authors hope that this commentary has increased awareness of the existing filters and encourages the creation of new geographic filters for additional places in the world.

## References

Ayiku, L., & Finnegan, A. (2019). OP23 smart searches for context-sensitive topics: Geographic search filters. *International Journal of Technology Assessment in Health Care, 35*(S1), 5. https://doi.org/10.1017/S0266462319000953

Ayiku, L., Levay, P., Hudson, T., Craven, J., Barrett, E., Finnegan, A., & Adams, R. (2017). The MEDLINE UK filter: Development and validation of a geographic search filter to retrieve research about the UK from OVID MEDLINE. *Health Information and Libraries Journal, 34*(3), 200–216. https://doi.org/10.1111/hir.12187

Ayiku, L., Levay, P., Hudson, T., Craven, J., Finnegan, A., Adams, R., & Barrett, E. (2019). The Embase UK filter: Validation of a geographic search filter to retrieve research about the UK from OVID Embase. *Health Information and Libraries Journal, 36*(2), 121-133. https://doi.org/10.1111/hir.12252

Beynon, R., Leeflang, M. M., McDonald, S., Eisinga, A., Mitchell, R. L., Whiting, P., & Glanville, J. M. (2013). Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. *Cochrane Database of Systematic Reviews* (9), MR000022. https://doi.org/10.1002/14651858.mr000022.pub3

Bhargava, R., & D'Ignazio, C. (2020). *DataBasic Word Counter*. Emerson College and University of Massachusetts, Massachusetts, USA. Retrieved from https://DataBasic.io/en/wordcounter/

Damarell, R. A., May, N., Hammond, S., Sladek R. M., & Tieman, J. J. (2019). Topic search filters: A systematic scoping review. *Health Information and Libraries Journal, 36*(1), 4-40. https://doi.org/10.1111/hir.12244

Glanville, J., Bayliss, S., Booth, A., Dunda, Y., Fernandes, H., Fleeman, N. D., Foster, L., Fraser, C., Fry-Smith, A., Golder, S., Lefebvre, C., Miller, C., Paisley, S., Payne, L., Price, A., Welch, K. (2008). So many filters, so little time: The development of a search filter appraisal checklist. *Journal of the Medical Library Association, 96*(4), 356–361. https://doi.org/10.3163/1536-5050.96.4.011.

Glanville, J., Lefebvre, C., & Wright, K. (2020). The InterTASC information specialists' sub-group search filter resource: Filters to find studies of geographic locations. ISSG Filters Resource. University of York, York, UK. Retrieved from https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/other-filters/filters-to-find-studies-of-geographic-locations

Hausner E., Waffenschmidt, S., Kaiser, T., & Simon, M. (2012). Routine development of objectively derived search strategies. *Systematic Reviews, 1*(19). https://doi.org/10.1186/2046-4053-1-19

Jenkins, M. (2004). Evaluation of methodological search filters – A review. *Health Information and Libraries Journal, 21*(3), 148–163. https://doi.org/10.1111/j.1471-1842.2004.00511.x

Pienaar, E., Grobler, L., Busgeeth, K., Eisinga, A., & Siegfried, N. (2011). Developing a geographic search filter to identify randomised controlled trials in Africa: Finding the optimal balance between sensitivity and precision. *Health Information and Libraries Journal, 28*(3), 210–215. https://doi.org/10.1111/j.1471-1842.2011.00936.x

Randomness and Integrity Services Ltd. (2020). *RANDOM.ORG.* Retrieved from https://www.random.org/

Sampson, M., Zhang, L., Morrison, A., Barrowman, N. J., Clifford, T. J., Platt, R. W., Klassen, T. P., & Moher, D. (2006). An alternative to the hand searching gold standard: Validating methodological search filters using relative recall. *BMC Medical Research Methodology, 6*(33). https://doi.org/10.1186/1471-2288-6-33

Valderas, J., Mendivil, J., Parada, A., Losada-Yáñez, M. & Alonso, J. (2006). Development of a geographic filter for PubMed to identify studies performed in Spain. Revista *Española de Cardiología, 59*(12), 1244–1251. https://doi.org/10.1016/S1885-5857(07)60080-2

WriteWords. (2020). *WriteWords Frequency Counters*. Retrieved from http://www.writewords.org.uk/word_count.asp