PERFORMANCE OF OVID MEDLINE SEARCH FILTERS TO IDENTIFY HEALTH STATE UTILITY STUDIES

Mick Arber  
York Health Economics Consortium  
mick.arber@york.ac.uk  

Sonia Garcia  
Thomas Veale  
Mary Edwards  
Alison Shaw  
Julie M. Glanville  
York Health Economics Consortium

Objectives: This study was designed to assess the sensitivity of three Ovid MEDLINE search filters developed to identify studies reporting health state utility values (HSUVs), to improve the performance of the best performing filter, and to validate resulting search filters.

Methods: Three quasi-gold standard sets (QGS1, QGS2, QGS3) of relevant studies were harvested from reviews of studies reporting HSUVs. The performance of three initial filters was assessed by measuring their relative recall of studies in QGS1. The best performing filter was then developed further using QGS2. This resulted in three final search filters (FSF1, FSF2, and FSF3), which were validated using QGS3.

Results: FSF1 (sensitivity maximizing) retrieved 132/139 records (sensitivity: 95 percent) in the QGS3 validation set. FSF1 had a number needed to read (NNR) of 842. FSF2 (balancing sensitivity and precision) retrieved 128/139 records (sensitivity: 92 percent) with a NNR of 502. FSF3 (precision maximizing) retrieved 123/139 records (sensitivity: 88 percent) with a NNR of 383.

Conclusions: We have developed and validated a search filter (FSF1) to identify studies reporting HSUVs with high sensitivity (95 percent) and two other search filters (FSF2 and FSF3) with reasonably high sensitivity (92 percent and 88 percent) but greater precision, resulting in a lower NNR. These seem to be the first validated filters available for HSUVs. The availability of filters with a range of sensitivity and precision options enables researchers to choose the filter which is most appropriate to the resources available for their specific research.

Keywords: Quality of life, Quality-adjusted life-years, Cost-Benefit Analysis, Information storage and retrieval, Databases, Bibliographic

Measuring quality of life (QoL) in people with disease is a key information component to healthcare decision-making and priority setting. Quality-adjusted life-years (QALYs) are an important technique to allow common assessment of different treatments (1). One component of the QALY is a health state utility value (HSUV). When preparing health economic models to estimate costs and benefits of comparative treatments, the HSUV is required to calculate QALYs relating to the treatments being compared (2). Researchers working in health technology assessment (HTA) and model production need to effectively and efficiently identify studies reporting HSUVs.

Some databases currently include (ScHARRHUD (3), Cost-Effectiveness Analysis Registry (4)), or have included (NHS Economic Evaluation Database (5)) studies reporting HSUVs. Although these are valuable resources, they are not comprehensive in their coverage of studies reporting HSUVs. National Institute for Health and Care Excellence (NICE) Decision Support Unit (DSU) guidance recommends that searches to identify HSUVs for reviews should be systematic, transparent and sensitive, including a broad electronic database search using an extensive list of search terms for HSUV concepts (6). Recent research recommends MEDLINE should be searched as a minimum (7). Creating an effective strategy that identifies HSUVs with both high sensitivity and acceptable precision is a recognized search challenge (6–8). This is due to a range of factors, including the lack of Medical Subject Headings (MeSH) specific to HSUVs or utility measures, the number of studies reporting HSUVs which do not explicitly refer to HSUVs or utility measures in the title or abstract, and the number of irrelevant records retrieved when useful yet poorly discriminating terms such as “utility” or “quality of life” are included in the strategy.

Although a range of potentially useful terms has been suggested for possible use in searches for HSUVs (6;8), we have not identified any published, validated search filters designed for this purpose in key search filter resources (9–11). Search filters have played a significant role in the effective and efficient retrieval of other study types such as randomized controlled trials and economic evaluations. The limitations in indexing and abstract description pose a recognized challenge for filter use in identifying HSUVs (6–8). Given this challenge, it is important...
that researchers have evidence on how a group of search terms used for this purpose is likely to perform. A validated filter designed to identify studies reporting HSUVs would, therefore, be a useful asset in HTA and economic model production.

Three different versions of a filter (Current Search Filter 1, CSF2, CSF3) to identify studies reporting HSUVs in Ovid MEDLINE have been developed using pragmatic methods from 2008 onward for use in-house by York Health Economic Consortium (YHEC). The filters were subjectively derived and based on the information specialists’ experience in finding health state utility values to inform reviews and economic models. CSF1 is the most sensitive and CSF3 is the most precise. However, all three filters aim to maximize precision and, therefore, intentionally omit some search terms that are recognized as likely to increase sensitivity at the expense of precision (such as terms relating to QoL) (6–8).

Although the three CSFs seemed to perform at reasonable levels of sensitivity, we had never formally tested them. In anticipation that a validated search filter would be useful beyond YHEC, we report on the testing of and further development of the three filters. In particular, we were interested to learn how the filters would perform given the emphasis on precision and the intentional omission of broader search terms such as QoL. We were also interested in the extent to which we could develop the sensitivity of the filters further, without introducing a significant impact on precision.

The objective of this study was to test the sensitivity, precision, and number needed to read (NNR) of the three CSFs, to develop the best performing filter with the aim of improving sensitivity, precision, and NNR, and to validate the resulting final search filters (FSFs) using a set of relevant studies reporting HSUVs.

**METHODS**

The study was conducted in three phases. In Phase 1, the performance of the three CSFs was assessed, and the best performing CSF was identified. In Phase 2, the best performing CSF was developed to produce three FSFs. In Phase 3, the performance of the three FSFs was validated. The process of filter assessment, development and validation is shown in Figure 1.

**Building the Quasi-Gold Standards**

In each phase, filter performance was assessed by measuring the relative recall (RR) of the filters of a quasi-gold standard (QGS) set of relevant studies. The QGS sets were studies harvested from recent systematic reviews (SRs) of studies reporting HSUVs, and from ten manufacturers’ submissions (MSs) for NICE single technology appraisals (STAs) which had conducted reviews of studies reporting HSUVs. To ensure the work was manageable within project resources, the search for SRs was limited to studies published from 2004 and the number of MSs included was restricted to ten. To identify the sample of SRs, a search was carried out on Ovid MEDLINE (Supplementary Table 1). The titles and abstracts of retrieved reports of SRs were double screened by two researchers using pragmatic screening criteria (Supplementary Table 2). Full text documents were sought for eligible SRs.

RR gold standards built from included studies from SRs are only as good as the searches conducted in those SRs. The search methodology of each SR was therefore quality assessed using a pragmatic full text assessment checklist (Supplementary Table 3). Disagreement over eligibility at each stage was resolved by discussion and/or by seeking input from a third researcher. SRs with search strategies that fulfilled the quality checklist criteria (i.e., which achieved “YES” for each of the quality checklist items) were selected and citations for HSUVs studies were extracted from those SRs. To identify the 10 MSs, a search was made on the NICE Web site to identify MSs for NICE STAs, beginning with those published in April 2015 and working backward until ten utility reviews had been identified. Results were assessed to identify submissions which included reviews of studies reporting HSUVs. Ten such submissions across a range of healthcare conditions were selected and citations for HSUV studies were extracted. Citation details for all HSUV studies extracted from the SRs and MSs were added to an EndNote library and de-duplicated. The HSUV studies with records in MEDLINE formed the QGSs.

The potential SRs were assessed in author alphabetical order. The first 60 percent of the potential SRs plus the 10 MSs were used to source the first QGS set of studies reporting HSUVs (QGS1, for strategy testing in Phase 1 and strategy development use in Phase 2). The remaining 40 percent of the potential SRs were used to source QGS2 (for strategy testing and development in Phase 2) and QGS3 (for strategy validation in Phase 3), with records randomly allocated to QGS2 and QGS3 using a random number generator (12).

**Phase 1: Testing the Current Filters**

The PubMed id codes for records in QGS1 were used to search Ovid MEDLINE and the resulting set of records were combined with CSF1, CSF2, and CSF3.

The most sensitive CSF was identified. Where filters had the same sensitivity, the filter with the lowest NNR was nominated as the best performing filter.

In addition to calculating sensitivity and NNR in relation to QGS1, we identified the number of records retrieved when each filter was combined with terms for three example health conditions (diabetes, chronic obstructive pulmonary disease (COPD), and constipation). This analysis was undertaken to illustrate the impact of using the different filters on the volume of retrieved records in those three conditions.

**Phase 2: Filter Development**

To improve the sensitivity of the best performing CSF from Phase 1, QGS1 database records that were not retrieved by
the filter in Phase 1 were analyzed to identify candidate search terms that could be added to the filter to increase retrieval. The records were analyzed manually and using Ovid search functionality, PubMed PubReMiner (13), and TerMine (14). Index terms and free text terms that appeared relatively frequently in the missed records were identified. Different combinations of candidate terms were tested iteratively to identify those that, when added to the search filter, would retrieve the largest number of records previously missed from QGS1, while minimizing the increase in total record numbers retrieved. In addition, some extra EQ-5D related terms not previously included in the CSFs were added to the filter. The EQ-5D is a measure of health-related QoL and we added these terms to assess their impact on sensitivity. For some organizations (such as NICE) the EQ-5D is the preferred instrument for generating HSUVs (6).

To improve the precision of the best performing CSF while maintaining sensitivity, the impact of individual search lines in the filter on retrieval of studies from QGS1 was examined. Lines that did not retrieve any records in QGS1, and lines that did not retrieve unique records, were removed. Free-text terms
in one search line that did not retrieve unique records were also removed.

This resulted in five potential filters. A search string was created using PubMed id codes for QGS2 records in Ovid MEDLINE and the result set was combined with each of the five potential filters. The sensitivity, precision, and NNR for each potential filter was calculated. The five potential filters had differing balances of sensitivity and precision. From the five potential filters, we chose three potential filters for further development: the strategy with the highest sensitivity, the strategy with the highest precision which still achieved above 85 percent sensitivity using QGS2, and the strategy which offered the best compromise between sensitivity and precision.

Database records from QGS2 that were not retrieved by the three potential filters were analyzed with the aim of identifying search terms that could be added to the strategies to retrieve the missing QGS2 studies. The same approach was used as described in Phase 1. This resulted in three final search filters:

- FSF1 - sensitivity maximizing;
- FSF2 - precision and sensitivity balanced;
- FSF3 - precision maximizing.

Phase 3: Filter Validation, Real-World Volume Illustration, and Comparison with Previously Published Search Strategy

Filter Validation. The three FSFs were validated using the QGS3 set of records. A search string was created using the PubMed id codes for the QGS3 records and the results were combined with each FSF in Ovid MEDLINE. The sensitivity, precision, and NNR for each filter was calculated.

Real-World Volume Illustration. In addition to calculating sensitivity, precision, and NNR in relation to QGS3, we identified the number of records retrieved when each filter was combined with terms for three example health conditions (diabetes, COPD, and constipation). This analysis was undertaken to illustrate the impact of using the different filters on the volume of retrieved records in those three conditions.

Comparison with previously published search strategy. As previously stated, including QoL terms when searching for studies reporting HSUVs is recognized as being likely to increase sensitivity at the expense of precision (6–8). To increase precision, the three FSFs intentionally did not include “broad” QoL terms (i.e., QoL terms searched alone). To assess the relative performance of a set of search terms used in a published search strategy that did include broad QoL terms, we calculated the sensitivity, precision, and NNR for the set of search terms used in the case study included in the NICE DSU Technical Support Document on the identification, review, and synthesis of HSUVs from the literature (6) (NICE case study [CS] terms, Supplementary Table 4). We compared the results of the NICE CS terms with those from the FSF testing. It is important to note that the NICE CS terms were only used as an example of a set of terms used in a published strategy; we acknowledge that the authors of the NICE DSU document do not propose that the CS terms are a formal tested search filter.

RESULTS

Phase 1; Testing the Current Filters

Creation of QGS1. The MEDLINE search identified 1,485 potential SRs (search date February 27, 2015). For Phase 1, a total of 885/1485 SRs were used as the source for QGS1 and were double screened (with 600 being reserved for QGS2 and QGS3). After screening and full text assessment, ten SRs had search strategies that fulfilled the quality checklist criteria. From these ten SRs and the ten MSs, 346 citations for studies reporting HSUVs were extracted. After de-duplication, 294 of these records could be identified in MEDLINE and formed QGS1.

Performance of CSF1, CSF2 and CSF3. All three filters retrieved 268/294 (sensitivity: 91 percent) QGS1 records in MEDLINE (Table 1). CSF3 (Supplementary Table 5) had the lowest NNR (365) and was nominated the best performing CSF. CSF3 retrieved the lowest volume of results in the three example search scenarios.

Phase 2: Filter Development

Creation of QGS2 and QGS3. For Phase 2, 600 potential SRs were double screened. Thirty SRs had search strategies that fulfilled the quality checklist criteria. Two hundred and seventy eight citations for studies reporting HSUVs were extracted and all could be identified in MEDLINE. After de-duplication and randomization, 139 records formed the QGS2 set used for filter development, and 139 records formed the QGS3 set used for filter validation.

The three FSFs resulting from Phase 2 filter development work are shown in Table 2. Lines 1–12 in all three FSFs are identical. Lines 13–16 in FSF1 and FSF2 are identical. Lines 17–22 are unique to FSF1.

Phase 3: Filter Validation, Real-World Volume Illustration, and Comparison with Previously Published Search Strategy

Filter Validation: Performance of FSF1, FSF2, and FSF3. Results are shown in Table 3. FSF1 (sensitivity maximizing) retrieved 132/139 records (sensitivity: 95 percent) in QGS3. FSF1 had a NNR of 842. FSF2 (balance of sensitivity and precision) retrieved 128/139 records (sensitivity: 92 percent) with a NNR of 502. FSF3 (precision maximizing) retrieved 123/139 records (sensitivity: 88 percent) with a NNR of 383.

Real-World Volume Illustration. Results are shown in Table 3. When combined with diabetes terms, FSF1 retrieved 5,025 records, FSF2 retrieved 3,222 records, and FSF3 retrieved 2,476 records. When combined with the COPD terms, FSF1 retrieved 2,530 records, FSF2 retrieved 1,363 records, and FSF3 retrieved 759 records. When combined with the
Table 1. Performance of CSFs Using QGS1

<table>
<thead>
<tr>
<th>Current YHEC HSUVs search filters</th>
<th>CSF1</th>
<th>CSF2</th>
<th>CSF3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of records retrieved by the filter in MEDLINE</td>
<td>255198</td>
<td>159724</td>
<td>97946</td>
</tr>
<tr>
<td>Sensitivity % (QGS1)\textsuperscript{a}</td>
<td>91 (268/294)</td>
<td>91 (268/294)</td>
<td>91 (268/294)</td>
</tr>
<tr>
<td>Precision % (QGS1)\textsuperscript{a}</td>
<td>0.1</td>
<td>0.16</td>
<td>0.27</td>
</tr>
<tr>
<td>NNR (QGS1)\textsuperscript{a}</td>
<td>952</td>
<td>596</td>
<td>365</td>
</tr>
<tr>
<td>Number of records retrieved when combined with diabetes terms</td>
<td>7481</td>
<td>5191</td>
<td>4194</td>
</tr>
<tr>
<td>Number of records retrieved when combined with COPD terms</td>
<td>2051</td>
<td>1602</td>
<td>1125</td>
</tr>
<tr>
<td>Number of records retrieved when combined with constipation terms</td>
<td>1477</td>
<td>1249</td>
<td>617</td>
</tr>
</tbody>
</table>

\textsuperscript{a}The sensitivity, precision and NNR for each filter was calculated using the following formulae: (i) Sensitivity % = (number of records in QGS retrieved / number of records in QGS) x 100; (ii) Precision % = (number of records in QGS retrieved / total number of records retrieved) x 100; (iii) NNR = (total number of records retrieved / number of records in QGS retrieved).

COPD, chronic obstructive pulmonary disease; CSF, current search filter; HSUV, health state utility value; NNR, number needed to read; YHEC, York Health Economics Consortium; QGS, quasi-gold standard.

constipation terms, FSF1 retrieved 1,445 records, FSF2 retrieved 667 records, and FSF3 retrieved 334 records.

Comparison with Previously Published Search Strategy: Results are shown in Table 3. The NICE CS terms retrieved 133/139 records (sensitivity: 96 percent) in QGS3 and had a NNR of 2033. When combined with diabetes terms, the NICE CS terms retrieved 11,300 records. When combined with COPD terms, the NICE CS terms retrieved 6,071 records. When combined with constipation terms, the NICE CS terms retrieved 3,933 records.

DISCUSSION
We have developed and validated a search filter (FSF1) to identify studies reporting HSUVs that performs with high sensitivity (95 percent). We have also developed and validated two other search filters (FSF2 and FSF3) with reasonably high sensitivity (92 percent and 88 percent) but greater precision, resulting in a lower NNR.

Comparison of FSF1, FSF2, and FSF3, and Real-World Volume Illustration
FSF1 seems the most appropriate filter to use in projects where sensitivity is a priority or where the health condition of interest yields small numbers of results. In project contexts where there is increased emphasis on precision and on minimizing numbers of records to screen, FSF2 or FSF3 could be considered as alternatives. Although less sensitive than FSF1, both filters still have reasonably high sensitivity and both would reduce the number of records retrieved from MEDLINE for screening. We have not measured the performance of the filters in the context of gold standard sets of records for specific conditions. However, our assessment of the volume of retrieved records using the filters in three example conditions suggests that using the more precise filters could have a clear impact on this volume.

When combined with the COPD terms for example, FSF1 retrieved 2,530 records, FSF2 retrieved 1,363 records, and FSF3 retrieved 759 records. Researchers working with more limited resources or who do not require a search which aims to be “exhaustive,” might be happy to accept the trade-off of a relatively small reduction in sensitivity (for example, 3–7 percent) for a reduction in number of records for screening of 46–70 percent in the case of COPD. Previous guidance has suggested that, where project context means highly sensitive search strategies for HSUVs are not possible, it is reasonable to use a less sensitive search strategy, because relevant citations missed by the main search are likely be retrieved by supplementary search techniques such as citation searching and reference list checking (8).

Comparison of FSF1 with Previously Published Search Strategy
It is already known that including QoL search terms on their own will retrieve HSUV studies, but will also reduce already low precision (6–8). These terms were intentionally omitted from our filters to increase precision. We were, therefore, interested to see how our filters would compare with a set of HSUV search terms where the QoL terms were included, and we used the NICE CS terms to make this comparison. FSF1 compares well with the NICE CS terms in terms of sensitivity and NNR balance. FSF1 had a slightly lower sensitivity (95 percent versus 96 percent) but much lower NNR (842 versus 2,033) than the NICE CS terms. As can be seen by the results retrieved in the three search scenarios in Table 3, this difference has the potential to translate to a reasonably large decrease in record numbers for screening (with benefits for project timelines / resources). Our findings suggest that using FSF1 can achieve similar sensitivity to a strategy including broad QoL terms, but with a lower NNR.
### Table 2. FSFs for Identifying HSUVs

<table>
<thead>
<tr>
<th>FSF1 — sensitivity maximizing</th>
<th>FSF2 — balance of sensitivity and precision</th>
<th>FSF3 — precision maximizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Quality-Adjusted Life Years/</td>
<td>1 Quality-Adjusted Life Years/</td>
<td>1 Quality-Adjusted Life Years/</td>
</tr>
<tr>
<td>2 (quality adjusted or adjusted life year$).ti,ab,kf.</td>
<td>2 (quality adjusted or adjusted life year$).ti,ab,kf.</td>
<td>2 (quality adjusted or adjusted life year$).ti,ab,kf.</td>
</tr>
<tr>
<td>3 (qaly$ or qald$ or qale$ or qtime$).ti,ab,kf.</td>
<td>3 (qaly$ or qald$ or qale$ or qtime$).ti,ab,kf.</td>
<td>3 (qaly$ or qald$ or qale$ or qtime$).ti,ab,kf.</td>
</tr>
<tr>
<td>4 (illness state$1 or health state$1).ti,ab,kf.</td>
<td>4 (illness state$1 or health state$1).ti,ab,kf.</td>
<td>4 (illness state$1 or health state$1).ti,ab,kf.</td>
</tr>
<tr>
<td>5 (hui or hui1 or hui2 or hui3).ti,ab,kf.</td>
<td>5 (hui or hui1 or hui2 or hui3).ti,ab,kf.</td>
<td>5 (hui or hui1 or hui2 or hui3).ti,ab,kf.</td>
</tr>
<tr>
<td>6 (multiattribute$ or multi attribute$).ti,ab,kf.</td>
<td>6 (multiattribute$ or multi attribute$).ti,ab,kf.</td>
<td>6 (multiattribute$ or multi attribute$).ti,ab,kf.</td>
</tr>
<tr>
<td>7 (utility adj3 (score$1 or valu$ or health$ or cost$ or measur$ or disease$ or mean or gain or gains or index$)).ti,ab,kf.</td>
<td>7 (utility adj3 (score$1 or valu$ or health$ or cost$ or measur$ or disease$ or mean or gain or gains or index$)).ti,ab,kf.</td>
<td>7 (utility adj3 (score$1 or valu$ or health$ or cost$ or measur$ or disease$ or mean or gain or gains or index$)).ti,ab,kf.</td>
</tr>
<tr>
<td>8 utilities.ti,ab,kf.</td>
<td>8 utilities.ti,ab,kf.</td>
<td>8 utilities.ti,ab,kf.</td>
</tr>
<tr>
<td>9 (eq-5d or eq5d or eq-5 or eq5 or euro qual or euroqual or euroQUAL5d or euro qol or euroqol or euro qol5d or euroqol5d or euro qual or euroqual or euroQUAL5d or euro qol or euroqol or euro qol5d or euroqol5d).ti,ab,kf.</td>
<td>9 (eq-5d or eq5d or eq-5 or eq5 or euro qual or euroqual or euroQUAL5d or euro qol or euroqol or euro qol5d or euroqol5d).ti,ab,kf.</td>
<td>9 (eq-5d or eq5d or eq-5 or eq5 or euro qual or euroqual or euroQUAL5d or euro qol or euroqol or euro qol5d or euroqol5d).ti,ab,kf.</td>
</tr>
<tr>
<td>10 (euro$ adj3 (5 d or 5d or 5 dimension$ or 5dimension$ or 5 domain$ or 5domain$)).ti,ab,kf.</td>
<td>10 (euro$ adj3 (5 d or 5d or 5 dimension$ or 5dimension$ or 5 domain$ or 5domain$)).ti,ab,kf.</td>
<td>10 (euro$ adj3 (5 d or 5d or 5 dimension$ or 5dimension$ or 5 domain$ or 5domain$)).ti,ab,kf.</td>
</tr>
<tr>
<td>11 (sf36$ or sf 36$ or sf thirtysix or sf thirty six).ti,ab,kf.</td>
<td>11 (sf36$ or sf 36$ or sf thirtysix or sf thirty six).ti,ab,kf.</td>
<td>11 (sf36$ or sf 36$ or sf thirtysix or sf thirty six).ti,ab,kf.</td>
</tr>
<tr>
<td>12 (time trade off$1 or time tradeoff$1 or tto or timetradeoff$1).ti,ab,kf.</td>
<td>12 (time trade off$1 or time tradeoff$1 or tto or timetradeoff$1).ti,ab,kf.</td>
<td>12 (time trade off$1 or time tradeoff$1 or tto or timetradeoff$1).ti,ab,kf.</td>
</tr>
<tr>
<td>13 quality of life/ and ((quality of life or qol) adj (score$1 or measure$1)).ti,ab,kf.</td>
<td>13 quality of life/ and ((quality of life or qol) adj (score$1 or measure$1)).ti,ab,kf.</td>
<td>13 quality of life/ and ((quality of life or qol) adj (score$1 or measure$1)).ti,ab,kf.</td>
</tr>
<tr>
<td>14 quality of life/ and ec.fs.</td>
<td>14 quality of life/ and ec.fs.</td>
<td>14 quality of life/ and ec.fs.</td>
</tr>
<tr>
<td>15 quality of life/ and (health adj3 status).ti,ab,kf.</td>
<td>15 quality of life/ and (health adj3 status).ti,ab,kf.</td>
<td>15 quality of life/ and (health adj3 status).ti,ab,kf.</td>
</tr>
<tr>
<td>16 (quality of life or qol).ti,ab,kf. and Cost-Benefit Analysis/</td>
<td>16 (quality of life or qol).ti,ab,kf. and Cost-Benefit Analysis/</td>
<td>16 (quality of life or qol).ti,ab,kf. and Cost-Benefit Analysis/</td>
</tr>
<tr>
<td>17 ((qol or hqol or quality of life).ti,kf. or &quot;quality of life&quot;)/ and ((qol or hqol$ or quality of life) adj2 (increase$ or decrease$ or improv$ or declin$ or reduc$ or high$ or low$ or effect or effects or worse or score or scores or change$1 or impact$1 or impacted or deteriorat$1)).ab.</td>
<td>17 ((qol or hqol or quality of life).ti,kf. or &quot;quality of life&quot;)/ and ((qol or hqol$ or quality of life) adj2 (increase$ or decrease$ or improv$ or declin$ or reduc$ or high$ or low$ or effect or effects or worse or score or scores or change$1 or impact$1 or impacted or deteriorat$1)).ab.</td>
<td>17 ((qol or hqol or quality of life).ti,kf. or &quot;quality of life&quot;)/ and ((qol or hqol$ or quality of life) adj2 (increase$ or decrease$ or improv$ or declin$ or reduc$ or high$ or low$ or effect or effects or worse or score or scores or change$1 or impact$1 or impacted or deteriorat$1)).ab.</td>
</tr>
<tr>
<td>18 Cost-Benefit Analysis/ and (cost-effectiveness ratio$ and (perspective$ or life expectancy$)).ti,ab,kf.</td>
<td>18 Cost-Benefit Analysis/ and (cost-effectiveness ratio$ and (perspective$ or life expectancy$)).ti,ab,kf.</td>
<td>18 Cost-Benefit Analysis/ and (cost-effectiveness ratio$ and (perspective$ or life expectancy$)).ti,ab,kf.</td>
</tr>
<tr>
<td>20 quality of life/ and ((quality of life or qol) adj3 (improv$ or chang$)).ti,ab,kf.</td>
<td>20 quality of life/ and ((quality of life or qol) adj3 (improv$ or chang$)).ti,ab,kf.</td>
<td>20 quality of life/ and ((quality of life or qol) adj3 (improv$ or chang$)).ti,ab,kf.</td>
</tr>
<tr>
<td>21 quality of life/ and health-related quality of life.ti,ab,kf.</td>
<td>21 quality of life/ and health-related quality of life.ti,ab,kf.</td>
<td>21 quality of life/ and health-related quality of life.ti,ab,kf.</td>
</tr>
<tr>
<td>22 models,economic/</td>
<td>22 models,economic/</td>
<td>22 models,economic/</td>
</tr>
<tr>
<td>23 or/1-22</td>
<td>23 or/1-22</td>
<td>23 or/1-22</td>
</tr>
</tbody>
</table>

FSF, final search filter; HSUV, health state utility value.
for example, we had included in our filters the terms strategy, several of these studies would have been retrieved. If, include terms closely relevant to the concept of HSUVs. As majority of the studies (9/12) were missed because they did not
database records of the missed studies were examined to establish the reasons for nonretrieval. The
tary Table 5, lines 9, 12, 20, and 23), but were then removed
Studies Missed by FSF1
FSF1 achieved high sensitivity (Table 3), but failed to retrieve
seven records in the QGS3 validation set and twelve records
across all three QGSs. The database records of the missed studies
were examined to establish the reasons for nonretrieval. The
majority of the studies (9/12) were missed because they did not
include terms closely relevant to the concept of HSUVs. As anticipated, if we had included QoL terms on their own in the
strategy, several of these studies would have been retrieved. If, for example, we had included in our filters the terms quality of life/ or (quality of life or qol).ti.ab,kf., the sensitivity of FSF1
using the QGS3 validation set would have increased from 95 percent to 98 percent. The NNR would have also increased,
small number of relevant records may not be generalizable to
LIMITATIONS OF THIS STUDY
It is not possible to know what proportion of the relevant literature a QGS set actually represents. The more representative studies in the QGS set are of all relevant studies, the more robust will be research findings based on RR. In part, this depends on the quality of the searches undertaken in the SRs used to source the studies which form the QGS set. Despite passing the quality checklist, the search methods of some SRs did use supplementary search approaches in addition to bibliographic database searches. Published advice has indicated the key role that supplementary searches can have in identification of studies reporting HSUVs. However, some of these searches using the QGS3 validation set would have increased from 97.9 percent to 98.4 percent. The small size of the increase confirms that these are not key terms for identifying unique studies reporting HSUVs. However, reinsertion of these terms would have minimal adverse impact on precision; therefore, we recommend researchers maximizing sensitivity consider adding these terms to the FSFs.

FSF1 achieved high sensitivity (Table 3), but failed to retrieve
seven records in the QGS3 validation set and twelve records
across all three QGSs. The database records of the missed studies
were examined to establish the reasons for nonretrieval. The
majority of the studies (9/12) were missed because they did not
include terms closely relevant to the concept of HSUVs. As anticipated, if we had included QoL terms on their own in the
strategy, several of these studies would have been retrieved. If, for example, we had included in our filters the terms quality of life/ or (quality of life or qol).ti.ab,kf., the sensitivity of FSF1
using the QGS3 validation set would have increased from 95 percent to 98 percent. The NNR would have also increased,
small number of relevant records may not be generalizable to
LIMITATIONS OF THIS STUDY
It is not possible to know what proportion of the relevant literature a QGS set actually represents. The more representative studies in the QGS set are of all relevant studies, the more robust will be research findings based on RR. In part, this depends on the quality of the searches undertaken in the SRs used to source the studies which form the QGS set. Despite passing the quality checklist, the search methods of some SRs did use supplementary search approaches in addition to bibliographic database searches. Published advice has indicated the key role that supplementary searches can have in identification of studies reporting HSUVs. However, some of these searches using the QGS3 validation set would have increased from 97.9 percent to 98.4 percent. The small size of the increase confirms that these are not key terms for identifying unique studies reporting HSUVs. However, reinsertion of these terms would have minimal adverse impact on precision; therefore, we recommend researchers maximizing sensitivity consider adding these terms to the FSFs.

### Table 3. Performance of FSFs and NICE CS Terms Using QGS3 (Validation Set) and Across All QGSs

<table>
<thead>
<tr>
<th>Revised YHEC HSUVs search filters</th>
<th>FSF1 – sensitivity maximizing</th>
<th>FSF2 – balance of sensitivity and precision</th>
<th>FSF3 – precision maximizing</th>
<th>NICE CS terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of records retrieved by the filter in MEDLINE</td>
<td>111205</td>
<td>64288</td>
<td>47060</td>
<td>270406</td>
</tr>
<tr>
<td>Sensitivity % (QGS3 - validation set)</td>
<td>95 (132/139)</td>
<td>92 (128/139)</td>
<td>88 (123/139)</td>
<td>96 (133/139)</td>
</tr>
<tr>
<td>Precision % (QGS3 - validation set)</td>
<td>0.12</td>
<td>0.2</td>
<td>0.26</td>
<td>0.05</td>
</tr>
<tr>
<td>NNR (QGS3 - validation set)</td>
<td>842</td>
<td>502</td>
<td>383</td>
<td>2033</td>
</tr>
<tr>
<td>Sensitivity % (across all QGSs)</td>
<td>98 (555/567)</td>
<td>94 (531/567)</td>
<td>89 (503/567)</td>
<td>96 (544/567)</td>
</tr>
<tr>
<td>Precision % (across all QGSs)</td>
<td>0.5</td>
<td>0.83</td>
<td>1.07</td>
<td>0.2</td>
</tr>
<tr>
<td>NNR (across all QGSs)</td>
<td>200</td>
<td>121</td>
<td>94</td>
<td>497</td>
</tr>
<tr>
<td>Number of records retrieved when combined with diabetes terms</td>
<td>5025</td>
<td>3222</td>
<td>2476</td>
<td>11300</td>
</tr>
<tr>
<td>Number of records retrieved when combined with COPD terms</td>
<td>2530</td>
<td>1363</td>
<td>759</td>
<td>6071</td>
</tr>
<tr>
<td>Number of records retrieved when combined with constipation terms</td>
<td>1445</td>
<td>667</td>
<td>334</td>
<td>3933</td>
</tr>
</tbody>
</table>

*The sensitivity, precision and NNR for each filter was calculated using the following formulae: (i) Sensitivity % = (number of records in QGS retrieved / number of records in QGS) x 100; (ii) Precision % = (number of records in QGS retrieved / total number of records retrieved) x 100; (iii) NNR = (total number of records retrieved / number of records in QGS retrieved). COPD, chronic obstructive pulmonary disease; FSF, final search filter; HSUV, health state utility value; NICE CS, National Institute for Health and Care Excellence case study; NNR: number needed to read; YHEC, York Health Economics Consortium; QGS, quasi-gold standard.

Of the missed studies, three of twelve included terms in the database record that we would regard as closely relevant to the concept of HSUVs (utility loss, disutility, short form, SF-12). These terms were included in our original CSFs (Supplementary Table 5, lines 9, 12, 20, and 23), but were then removed during strategy development phase to increase precision, as they were not responsible for retrieving any additional unique records in QGS1 or QGS2. In QGS3, however, these terms would have retrieved three additional unique records. If these terms had been included in FSF1, sensitivity across all QGS would have increased from 97.9 percent to 98.4 percent. The small size of the increase confirms that these are not key terms for identifying unique studies reporting HSUVs. However, reinsertion of these terms would have minimal adverse impact on precision; therefore, we recommend researchers maximizing sensitivity consider adding these terms to the FSFs.

### STUDIES MISSED BY FSF1

Studies Missed by FSF1
FSF1 achieved high sensitivity (Table 3), but failed to retrieve
seven records in the QGS3 validation set and twelve records
across all three QGSs. The database records of the missed studies
were examined to establish the reasons for nonretrieval. The
majority of the studies (9/12) were missed because they did not
include terms closely relevant to the concept of HSUVs. As anticipated, if we had included QoL terms on their own in the
strategy, several of these studies would have been retrieved. If, for example, we had included in our filters the terms quality of life/ or (quality of life or qol).ti.ab,kf., the sensitivity of FSF1
using the QGS3 validation set would have increased from 95 percent to 98 percent. The NNR would have also increased,
small number of relevant records may not be generalizable to
LIMITATIONS OF THIS STUDY
It is not possible to know what proportion of the relevant literature a QGS set actually represents. The more representative studies in the QGS set are of all relevant studies, the more robust will be research findings based on RR. In part, this depends on the quality of the searches undertaken in the SRs used to source the studies which form the QGS set. Despite passing the quality checklist, the search methods of some SRs did use supplementary search approaches in addition to bibliographic database searches. Published advice has indicated the key role that supplementary searches can have in identification of studies reporting HSUVs. However, some of these searches using the QGS3 validation set would have increased from 97.9 percent to 98.4 percent. The small size of the increase confirms that these are not key terms for identifying unique studies reporting HSUVs. However, reinsertion of these terms would have minimal adverse impact on precision; therefore, we recommend researchers maximizing sensitivity consider adding these terms to the FSFs.
the wider group of relevant records. The smaller the group of records, the more likely this will be. So, for example, during the filter development phase, the decision was made to remove several terms (utility loss, disutility, short form, SF-12) to increase precision, based on an analysis of the records in QGS1 and QGS2. However, when validating the final filters using QGS3, we found that inclusion of these terms would have retrieved an additional three records; sensitivity would, therefore, have been increased if these terms had been retained. Similarly, the terms that were added to increase strategy sensitivity in Phase 2 resulted from an analysis of a very small number of records that were not being retrieved by the CSFs. These missed records were outliers, which are less likely to be representative of a wide body of relevant records. Revisions made to a strategy to increase sensitivity based on retrieval of such records risks the revisions relating only to that small number of specific studies, with little or no relevance generally to the type of studies being sought.

The initial candidate filters that we wished to test (CSFs) were developed previously through pragmatic and traditional subjective methods, rather than from a gold standard set using objective filter design methods (15;16). We have, therefore, not conducted a search filter design, test, and validation project from start to finish (15;16). However, despite the pragmatic development approach used to design our initial candidate strategies, these strategies performed well in their first test, and we were able to use recommended approaches using separate QGS sets to test and improve the CSFs’ performance and to validate the FSFs.

The FSFs are designed to find reviews, cost-utility analyses, and utility elicitation studies that report HSUVs. They also aim to find utility mapping studies where disease-specific QoL instruments have been mapped to generic instruments. The filters do not attempt to distinguish between these different types of research. This means that precision for researchers looking only for utility elicitation studies, for example, will be lower than results reported here.

IMPLICATIONS FOR PRACTICE
Three validated search filters designed to identify studies reporting HSUVs in Ovid MEDLINE are now available for researchers to use. To the best of our knowledge, these are the first validated filters to be designed for this purpose. Providing three filters with a range of sensitivity and precision enables researchers to choose the filter that is most appropriate to the resource context of their specific research.

IMPLICATIONS FOR RESEARCH
These filters are designed for Ovid MEDLINE. Their performance when adapted to other interfaces (such as PubMed) and databases needs to be explored. HSUVs are often published in gray literature such as health technology assessment reports that are not indexed in the major bibliographic databases such as MEDLINE and Embase. The optimal use of search terms in retrieving documents from gray literature sources remains to be explored. Case studies have demonstrated the importance of supplementary search techniques in the identification of studies in HSUV reviews (6;8). Further exploration of the relative utility of the various supplementary search approaches would be valuable.

SUPPLEMENTARY MATERIAL
Supplementary Table 1: https://doi.org/10.1017/S0266462317000897
Supplementary Table 2: https://doi.org/10.1017/S0266462317000897
Supplementary Table 3: https://doi.org/10.1017/S0266462317000897
Supplementary Table 4: https://doi.org/10.1017/S0266462317000897
Supplementary Table 5: https://doi.org/10.1017/S0266462317000897

CONFLICTS OF INTEREST
The authors have nothing to disclose.

REFERENCES


