Original Article

Research Synthesis Methods

Received 4 January 2013, Revised 10 June 2013, Accepted 29 June 2013 Published online 23 August 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1093

Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews

Ian Shemilt,[a]*† Antonia Simon,[b] Gareth J. Hollands,[a] Theresa M. Marteau,[a] David Ogilvie,[a] Alison O'Mara-Eves,[c] Michael P. Kelly[d] and James Thomas[c]

In scoping reviews, boundaries of relevant evidence may be initially fuzzy, with refined conceptual understanding of interventions and their proposed mechanisms of action an intended output of the scoping process rather than its starting point. Electronic searches are therefore sensitive, often retrieving very large record sets that are impractical to screen in their entirety. This paper describes methods for applying and evaluating the use of text mining (TM) technologies to reduce impractical screening workload in reviews, using examples of two extremely large-scale scoping reviews of public health evidence (choice architecture (CA) and economic environment (EE)).

Electronic searches retrieved >800,000 (CA) and >1 million (EE) records. TM technologies were used to prioritise records for manual screening. TM performance was measured prospectively.

TM reduced manual screening workload by 90% (CA) and 88% (EE) compared with conventional screening (absolute reductions of ≈430 000 (CA) and ≈378 000 (EE) records). This study expands an emerging corpus of empirical evidence for the use of TM to expedite study selection in reviews. By reducing screening workload to manageable levels, TM made it possible to assemble and configure large, complex evidence bases that crossed research discipline boundaries. These methods are transferable to other scoping and systematic reviews incorporating conceptual development or explanatory dimensions. © 2013 The Authors. Research Synthesis Methods published by John Wiley & Sons, Ltd.

Keywords: text mining; scoping review methods; systematic review methods; study selection

1. Introduction

This paper describes the use of text mining (TM) technologies to expedite study screening and selection in reviews by reference to an evaluation of their application in two extremely large-scale scoping reviews of public health evidence, to help identify and configure large and heterogeneous evidence bases.

1.1. Scoping review methods

Scoping reviews (systematic maps) are often conducted to explore, delimit and describe broad evidence bases as a preliminary stage to help inform the design of conventional systematic reviews (Oliver and Sutcliffe, 2012;

[a]*Behaviour and Health Research Unit, University of Cambridge, Cambridge, UK*
[b]*Thomas Coram Research Unit, Department of Children and Health, Institute of Education, London, UK*
[c]*Evidence for Policy and Practice Information and Co-ordinating Centre, Department of Children and Health, Institute of Education, London, UK*
[d]*Centre for Public Health, National Institute for Health and Care Excellence, London, UK*
**Correspondence to: Ian Shemilt, Senior Research Associate, Behaviour and Health Research Unit, Institute of Public Health, University of Cambridge, Forvie Site, Robinson Way, Cambridge CB2 0SR, UK.*
[†]*E-mail: ian.shemilt@medschl.cam.ac.uk*

Arksey and O'Malley 2005; Valaitis *et al*., 2012). Their methods diverge from those of conventional systematic reviews in several key dimensions (see also Figure 1).

First, the boundaries of relevant evidence explored by scoping reviews (and characteristics of studies that fall within those boundaries) are typically unclear at the outset. Prespecified study eligibility criteria are therefore inevitably provisional, and it is accepted that these will be refined and re-applied iteratively during the review process, based on emergent knowledge of the studies and evidence encountered. In these circumstances, refined conceptual understanding of interventions, their proposed mechanisms of action and related phenomena of interest becomes an intended output of the scoping process rather than its starting point, to be derived through an *a posteriori* interpretive process (Kelly and Moore, 2012). Second, the obligation to identify every eligible study may be relaxed to some extent, as scoping reviews typically prioritise conceptual breadth (the aim to assemble a range and distribution of eligible studies that are representative of the target evidence base in terms of key study characteristics) over depth (the aim to assemble all eligible studies) (Brunton *et al*., 2012). Third, scoping reviews typically employ evidence synthesis strategies that focus on configuring or mapping evidence and generating or exploring intervention theory (Gough and Thomas, 2012; Hammersley 2002), rather than on aggregating evidence and testing intervention theory, as exemplified by the use of meta-analysis to combine evidence for intervention effects (Green *et al*., 2008; Deeks *et al*., 2008). As such, many scoping reviews include a conceptual or theoretical development dimension. The broad, initially fuzzy scope and configurative approach of scoping reviews means that electronic search strategies, designed to locate records of potentially eligible studies, are necessarily sensitive and can retrieve very large numbers of study records.

1.2. Text mining: a potential solution to the 'too many records' problem

A potential solution to the problem of impractically large search yields (the 'too many records' problem) entails use of TM technologies to prioritise retrieved study records for manual screening. TM is 'an automated process that can assist with the identification and structuring of patterns in the text of individual documents and across multiple documents' (Gough *et al*., 2012). Its application in the field of systematic reviews is relatively new (Thomas *et al*., 2011; Ananiadou *et al*., 2009) but the technique is being recognised for its ability to classify text and to enable more sensitive searching without increasing manual screening workload (Brunton *et al*., 2012). Critically, for study screening and selection applications, it can change the distribution of retrieved study records so that those records most likely to meet eligibility criteria are placed at the top of the search results list. This means that study screening and selection can be better focused and, potentially, take less time to complete compared with conventional manual screening.

However, TM technologies are highly dependent on the records available to train from (i.e. screened records from which TM automatically 'learns rules' for use to classify or prioritise further unscreened records). This raises concern that using TM may result in a bias towards finding 'more of what is already known'—the problem of 'hasty generalisation' (Wallace *et al*., 2010a). In theory, deployment of multiple TM technologies, which may individually be susceptible to introducing this form of bias into a review, in combination would ameliorate the risk of such bias (assuming that the different techniques make different, uncorrelated errors).

In the remainder of this introduction, we describe challenges presented by two scoping reviews of public health evidence in which systematic searches yielded extremely large numbers of study records, and describe the TM methods we used to address them. Similar methods could reasonably be applied in other scoping or

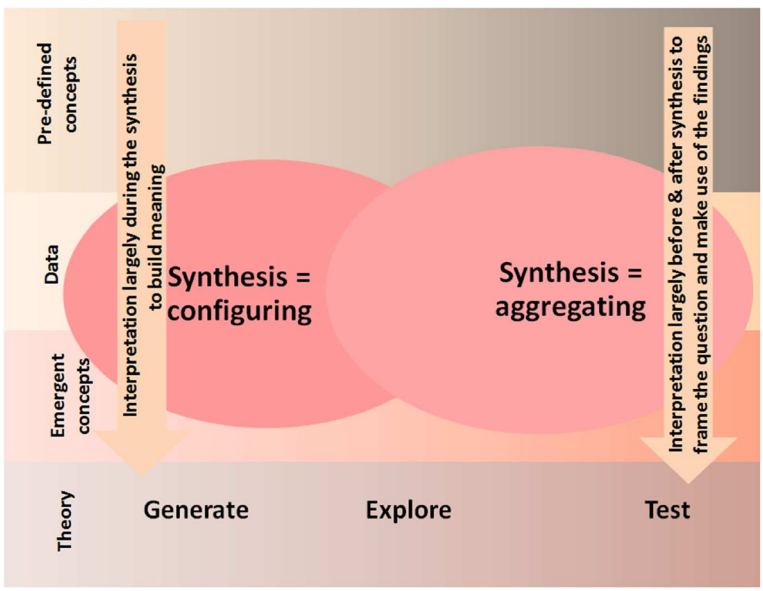

Figure 1. A spectrum of approaches to evidence synthesis.

systematic reviews with moderately large initial record sets (i.e. application of the techniques is not limited to the extreme quantities of records retrieved in these two examples).

1.3. The example scoping reviews

We conducted two scoping reviews to explore, delimit and describe broad evidence bases for two classes of interventions that involve altering environments with the aim of encouraging change in health behaviours at population level (Marteau *et al*., 2012; House of Lords Science and Technology Select Committee, 2011): choice architecture ('nudge') interventions (Hollands *et al*., 2013) and economic environment interventions (Shemilt *et al*., 2013; Shemilt *et al*., in press). Both had objectives that align with the characteristics of scoping reviews outlined above. For example, the objectives of the choice architecture (CA) review were

- to develop an operational definition of choice architecture applicable to public health interventions;
- to develop a provisional typology of a set of such interventions that involve altering micro-environments to change people's health behaviour;
- to map empirical evidence for the effects of the interventions on tobacco, alcohol, diet, and physical activity-related behaviours and
- to identify next steps for the development and evaluation of choice architecture interventions to change health behaviour at population level.

These objectives reflected our observation that previous conceptual work had not produced a clear definition of choice architecture applicable to public health intervention. Although extensive policy interest had stimulated efforts to clarify the concept and highlight example choice architecture interventions with public health objectives (Marteau *et al*., 2011; Dolan *et al*., 2010; Cabinet Office Behavioural Insights Team, 2010), terminology used to describe the concept has remained inconsistent (House of Lords Science and Technology Select Committee, 2011). Moreover, because the concept of choice architecture was proposed relatively recently (Thaler and Sunstein, 2008), we could expect empirical study reports to describe relevant interventions using terms and theoretical constructs that predated it, with variant terminology between research disciplines. In the economic environment (EE) review, though we had a clearer initial understanding of the likely range of eligible interventions, search terms based on target intervention concepts were unlikely to be specific to titles, abstracts or index terms of the target set of empirical studies. These issues presented thorny challenges for selection of search terms based on target intervention concepts for incorporation into electronic search strategies.

1.4. Search methods for locating studies

A range of search methods are available to locate studies for scoping and systematic reviews and related evidence synthesis activities (Greenhalgh and Peacock, 2005; Lefebvre *et al*., 2008; Booth, 2008; Papaioannou *et al*., 2010; Kaltenthaler *et al*., 2011). In the initial stages of each scoping review (CA and EE), we consulted topic experts and conducted preliminary, non-systematic searches to assemble an initial corpus of empirical studies closely aligned with, or close to the boundaries of, provisional eligibility criteria for interventions and outcomes. Snowball searches aim to locate further eligible studies by checking references lists of eligible study reports that have already been located, coupled with forward citation tracking from those reports, using online platforms such as PubMed and Google Scholar (Lefebvre *et al*., 2008). We intended to conduct snowball searches in both the CA and EE reviews, building on our initial corpora of eligible studies. However, using this technique as the principal method for locating studies might introduce bias into the reviews if it failed to locate clusters of eligible studies unconnected to our initial corpora via chains of linked citations. This concern was particularly germane to these scoping reviews (CA and EE) because eligible studies would likely be distributed across (and clustered within) different fields of applied research. We therefore decided to invest major effort in conducting systematic searches of electronic literature databases in parallel with snowball searches.

Designing electronic searches of bibliographic databases for published reports of intervention studies involves selecting discrete sets of keyword and index terms based on target concepts derived from a 'structured participants, interventions, comparators, outcomes and study designs' (PICOS) framework (or alternative framework for translating research questions into study eligibility criteria) (O'Connor, 2008; Booth, 2003). These sets of terms are then combined using Boolean operators so that they locate records of studies that meet all eligibility criteria (Lefebvre *et al*., 2008). Alongside challenges in the selection of search terms based on intervention concepts in the CA and EE scoping reviews, it was not appropriate—for the purposes of scoping each broad evidence base—to impose exclusion criteria relating to participants, comparators or study designs. Further, although we did prespecify eligibility criteria for outcomes, these encompassed multiple sets of health behaviours (CA: tobacco and alcohol use, diet and physical activity; EE: diet and physical activity), proximal consequences of these behaviours (e.g. energy, nutrient, alcohol and tobacco intake and energy expenditure) and, in the EE review, their more distal consequences—modifiable physiological and metabolic risk factors for non-communicable diseases (e.g. overweight and obesity, raised blood pressure, raised blood glucose and raised blood cholesterol). These design specifications ensured electronic search strategies would be highly sensitive.

We developed a draft MEDLINE search strategy for each review (see Appendix A), adapted these for Embase and PsycINFO, and tested all three for their sensitivity to retrieve study records contained in our early initial corpora of eligible or borderline studies. We refined draft search strategies until they retrieved 100% of initial corpus records indexed in each database, with a concurrent aim to minimise search yields (so far as possible). The process of testing and refining draft search strategies confirmed that we could not achieve greater specificity without sacrificing sensitivity to retrieve eligible records. We adapted final search strategies for 12 (CA) and 11 (EE) relevant electronic literature databases and executed these between 11 July and 11 August 2011 (Table 1).

Electronic searches retrieved totals of 1 207 611 (CA) and 1 426 032 (EE) title-abstract records, prior to removal of duplicates (Table 1). All phases of study selection were managed using EPPI REVIEWER 4 (ER4) systematic review software (Social Science Research Unit, Institute of Education, London, UK). We exported all retrieved records from source databases to ENDNOTE X4 databases (Thomson Reuters, San Fransisco CA, USA), exported to RIS file format and then bulk imported all into two ER4 review databases (CA and EE). We next ran automatic deduplication software to remove the majority of duplicate records present in each record set. Following automatic duplicate removal, 804 919 (CA) and 1 053,908 (EE) title-abstract records remained (Table 1). To our knowledge, the size of these initial record sets makes these the two largest scale scoping (or systematic) reviews ever attempted.

1.5. Use of text mining to support study screening and selection

Conventional methods for screening title-abstract records in reviews involve researchers manually inspecting all records and provisionally selecting those judged likely to meet eligibility criteria (Higgins and Deeks, 2008). In addition to the sheer size of record sets our searches had retrieved, the magnitude of the screening challenge was exacerbated because provisional eligibility criteria were intended to be refined and re-applied iteratively during the study screening and selection process. This meant provisional selection and exclusion decisions needed to be reviewed periodically as our eligibility criteria evolved. Given these factors, application of conventional screening methods was beyond the time and resources available for title-abstract screening.

We therefore used TM technologies to prioritise title-abstract records for manual screening. In both scoping reviews (CA and EE), the principal aim of using TM technologies was to improve the efficiency of the screening process, by identifying as many provisionally eligible records as possible whilst reducing manual screening workload to practicable levels. Alongside the two scoping reviews (CA and EE), we conducted a methodological study to evaluate the use of TM technologies to support title-abstract screening and selection. This study is novel because TM technologies have not previously been applied or evaluated in scoping or systematic reviews of this size. The remainder of this paper describes the methods and results of our evaluation.

2. Study objective

Our study objective was to assess the performance of TM technologies to support title-abstract screening in two extremely large-scale scoping reviews of public health evidence (CA and EE), compared with conventional screening (an unobserved counterfactual).

Table 1. Yields from systematic searches of electronic literature databases.

Database	CA (N records)	EE (N records)
Applied Social Sciences Index and Abstracts (CSA illumina)	28 358	56 523
Cochrane Database of Systematic Reviews (Wiley Online Library)	1193	2427
Database of Abstracts of Reviews of Effects (Wiley Online Library)	127	171
Database of Promoting Health Effectiveness Reviews (EPPI-Centre)	1854	1604
EconLit (EBSCO)	31 660	152 188
Embase (Ovid)	402 410	619 990
Health Technology Assessment Database (Wiley Online Library)	15	139
MEDLINE (Ovid)	418 040	432 641
NHS Economic Evaluation Database (Wiley Online Library)	45	1164
PsycINFO (Ovid)	150 325	68 659
SPORTDiscus with full text	5334	90 526
Web of Science (Thomson Reuters)	168 250	—
Subtotals (prior to automatic de-duplication)	1 207 611	1 426 032
Totals (after automatic de-duplication)	804 919	1 053 908

CA, choice architecture review; EE, economic environment review.

3. Methods

3.1. Application of TM technologies

Three TM technologies were applied in both reviews: *automatic term recognition* (ATR), *automatic classification* (AC), and *reviewer terms* (RT), each with different strengths and weaknesses. In addition, we used hybrid combinations of two of these three technologies in counterpoint to one another (AC with ATR and AC with RT). There were two main reasons for using different technologies in sequence or combination: first, to maximise the number of eligible study records identified and second, to offer some degree of protection against risk of bias due to the problem of 'hasty generalisation' described previously (i.e. our assumption being that likely biases differed across the different technologies). An overview of each TM technology (ATR, AC, RT and hybrids) is provided in Appendix B.

In order to provide initial sets of records for analysis by TM, we coded title-abstract records of our early, initial corpora of studies against provisional eligibility criteria. We had developed detailed coding notes for use by researchers to guide screening decisions, and these were updated regularly during the study selection stage, as eligibility criteria evolved.

Two further essential preliminary stages in advance of applying TM technologies were to establish inter-rater reliability between those reviewers scheduled to undertake screening and to estimate a baseline inclusion rate (BIR) for each review, in order to establish a proxy metric for our unobserved counterfactual (conventional screening) against which progress of screening and performance of TM could be monitored and evaluated. Methods and results of both these stages are described in Appendix C, for the benefit of readers contemplating similar application of TM technologies in their scoping or systematic reviews.

3.2. Prioritising records for manual screening

We first deployed ATR (as described in Appendix B) to prioritise consecutive sets of unscreened records for manual screening. We prospectively monitored ATR performance in each consecutive set of screened records by comparing the observed inclusion rate (OIR, i.e. the observed rate at which eligible study records were identified in practice within each consecutive set of prioritised records assigned for manual screening) with the BIR (used in this instance as a proxy estimate of the rate at which we would have expected to identify eligible study records had we used the unobserved counterfactual method of conventional screening). A performance metric based on these two rates ($OIR : BIR_U$, i.e. the unadjusted ratio of the OIR to the BIR) was used to inform real-time decisions about when to switch from ATR to deploy an alternative (or hybrid) TM technology (see Appendix C for details).

Bearing in mind the extreme numbers of study records we were dealing with, our rule of thumb for judging the success of the TM was that, for those records screened using TM prioritisation, we expected the OIR to be five times larger than the BIR (i.e. $OIR : BIR_U \geq 5$) at any point during the screening process. When the ratio dropped below this threshold, we would consider seeking to boost performance by switching to a different TM technology. It should be acknowledged that this is an ambitious ratio to expect because it represents a proportionate saving in manual screening workload of over 80% when other studies have suggested 50% is attainable (Wallace *et al*., 2010a). The substantial efficiency anticipated was due to the nature of these reviews—to scope the literature—where the standard systematic review requirement to identify every relevant study was relaxed.

The rationale for sequencing and switching between technologies was threefold. First, AC is known to perform relatively poorly on a small training set (i.e. the balanced numbers of records in each class—'excludes' and 'includes'—used to train the classifier; see Appendix B for further details). Applying ATR first would therefore increase the number of eligible records identified—and thus the size of training dataset—before using the classifier. Second, we hypothesised that because the different technologies process terminology in different ways, they would prioritise different studies (i.e. identify different 'low hanging fruit'); thus, the utilisation of multiple technologies was a strategy to maximise early 'wins'. Third, use of different technologies aimed to mitigate the impact of hasty generalisation (i.e. the impact of focusing on some clusters of relevant studies to the exclusion of others). Our overall stopping rule for the title-abstract screening stage of these reviews was: 'either when 100% of those provisionally eligible records expected to be present within each full record set had been coded as provisionally eligible (see Appendix C), or on the date that time available to be allocated to title-abstract screening expires (25 November 2011)'.

We manually screened consecutive sets of records prioritised by the different TM technologies over periods of approximately 11 weeks (CA:8 September–25 November 2011) and 10 weeks (EE:15 September–25 November 2011). Two researchers completed the large majority of screening assignments (i.e. one researcher per review), supported by a third researcher. Because study eligibility criteria could evolve during the scoping process (Introduction), we reviewed provisional eligibility decisions periodically as criteria were refined and revised these if required. In practice, eligibility criteria were refined to a moderate extent in the CA review and a negligible extent in the EE review. This process invariably resulted in the exclusion of some study records previously coded as provisionally eligible (rather than vice versa) as the scope and boundaries of relevant evidence were progressively tightened.

3.3. Evaluation of TM performance

Once the end date of title-abstract screening stage was reached (25 November 2011), we recorded the overall OIR and the absolute number of records coded as provisionally eligible within each review. We then collated a series of performance metrics: the unadjusted ratio of the OIR to the BIR (OIR : BIR_U—see previous text); the corollary estimated reduction in screening workload in terms of both proportionate and absolute numbers of study records and the number of eligible records identified by TM as a proportion of the estimated number eligible records present in the full records sets after use of TM had been initiated. We also used the first of these metrics (OIR : BIR_U) to compare the relative performance of different TM technologies in prioritising each consecutive set of unscreened records for manual screening.

As stated above, we expected the OIR to be five times larger than the BIR and planned our work on this basis, based on previous experience with the technologies concerned. Given that we did not expect to be able to screen more than a fraction of the studies retrieved, the viability of the work depended on the anticipated reduction in workload. In the event, TM exceeded our expectations in this regard, so we also report the 'equivalent' reduction in workload, acknowledging that no less screening was carried out, to illustrate how many records would need to have been screened using conventional methods in order to identify the stated number of relevant studies.

4. Results

Figures 2 and 3 show the overall performance of TM in the CA (Figure 2) and EE (Figure 3) reviews. In each figure, the point of divergence between the dashed and solid lines represents the point at which use of TM technologies was initiated to prioritise records for manual screening (i.e. after coding initial corpus records, inter-rater reliability assessments and screening a random sample of records to estimate the BIR).

The dashed line represents the unobserved counterfactual (conventional screening). Its slope represents the average rate at which provisionally eligible records would have been identified had we used the conventional method (based on the BIR). The slope of the solid line represents the observed overall rate at which provisionally eligible records were identified in practice (based on the OIR). The vertical gap between the dashed and solid lines therefore represents the gain from use of TM at a given stage of title-abstract screening.

In the CA review, the overall OIR : BIR_U was 10.1 (Table 2). Interpretation of this ratio is that overall, TM performed ≈10 times better than screening a random sample (i.e. ≈10 times better than our proxy for conventional screening methods). This equates to a proportionate reduction in screening workload of 90.1% and an absolute reduction in screening workload of 430 839 records (Table 2). In other words, after the point at which use of TM was initiated to prioritise records for manual screening, we screened 430 839 fewer records than we would have needed to screen using conventional methods to identify and select the same number of provisionally eligible records. The total number of TM prioritised records that we manually screened within the time and resources available for this task was 47 591 records. In the CA review, we identified and selected a further 238 provisionally eligible records after the point at which use of TM was initiated; this equates to 84.9% of eligible

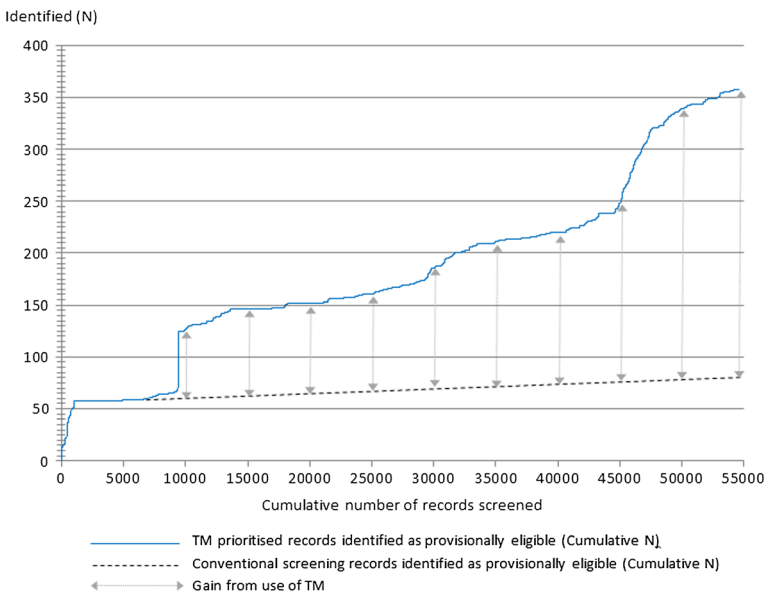

Figure 2. Overall text mining performance: choice architecture scoping review.

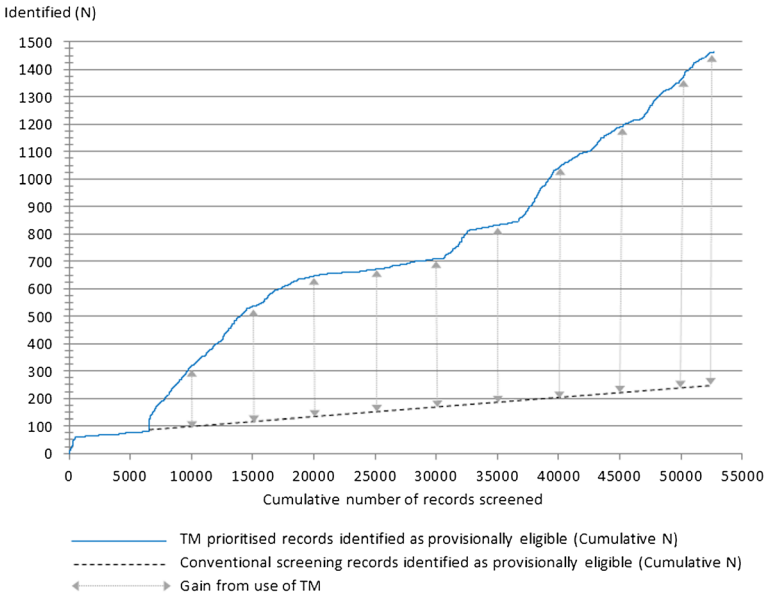

Figure 3. Overall text mining performance: economic environment scoping review.

Table 2. Summary of overall text mining performance.

Performance metric	CA	EE
Ratio of observed inclusion rate to baseline inclusion rate* (OIR : BIR_U)	10.1	8.3
Equivalent proportionate reduction in manual screening workload* (%)	90.1	88.0
Equivalent absolute reduction in manual screening workload* (N records)	430 839	376 643
Proportion of remaining provisionally eligible records identified using TM (%)	84.9	37.7
Proportion of remaining provisionally eligible records that would have been identified using conventional screening (%)	8.4	4.5

CA, choice architecture review; EE, economic environment review.
*TM versus conventional screening.

records estimated to be present in the full records set after this point. Had we used conventional methods, our estimates indicate that we would have identified <9% of these further records.

In the EE review, the overall OIR : BIR_U was 8.3 (Table 2). Overall, TM performed ≈8 times better than screening a random sample. This equates to a proportionate reduction in equivalent screening workload of 88.0% and an absolute reduction of 376 643 records (Table 2). The total number of TM prioritised records that we manually screened within the time and resources available for this task was 46 099 records. In the EE review, we identified and selected a further 1334 provisionally eligible records after the point at which use of TM was initiated. This represents 37.7% of eligible records estimated to be present in the full records set after this point. Had we used conventional methods, our estimates indicate that we would have identified <5% of these further records.

In each review, the overall OIR : BIR_U masks considerable variation between the different, sequentially applied TM technologies and hybrids in terms of their respective performance. Figures 4 and 5 show the relative performance of each TM technology in terms of its associated OIR : BIR_U in the CA (Figure 4) and EE (Figure 5) reviews.

The lower blue line in Figure 4 shows the sequence in which we used each TM technology in the CA review—from ATR, through hybrid models to using RT. It shows that while ATR prioritised screening was much more effective than conventional screening to begin with, performance quickly tailed off (ATR3–5) and thereafter, we obtained better results from hybrid models and using RT. A similar picture emerges from Figure 5, in that initial results from using ATR are good, but performance tails off before most relevant studies have been identified; a key difference in Figure 5 is that the performance of RT was clearly much poorer than in the CA review (Figure 4). In the EE review, the better results towards the end of screening were achieved using AC, at times combined with ATR.

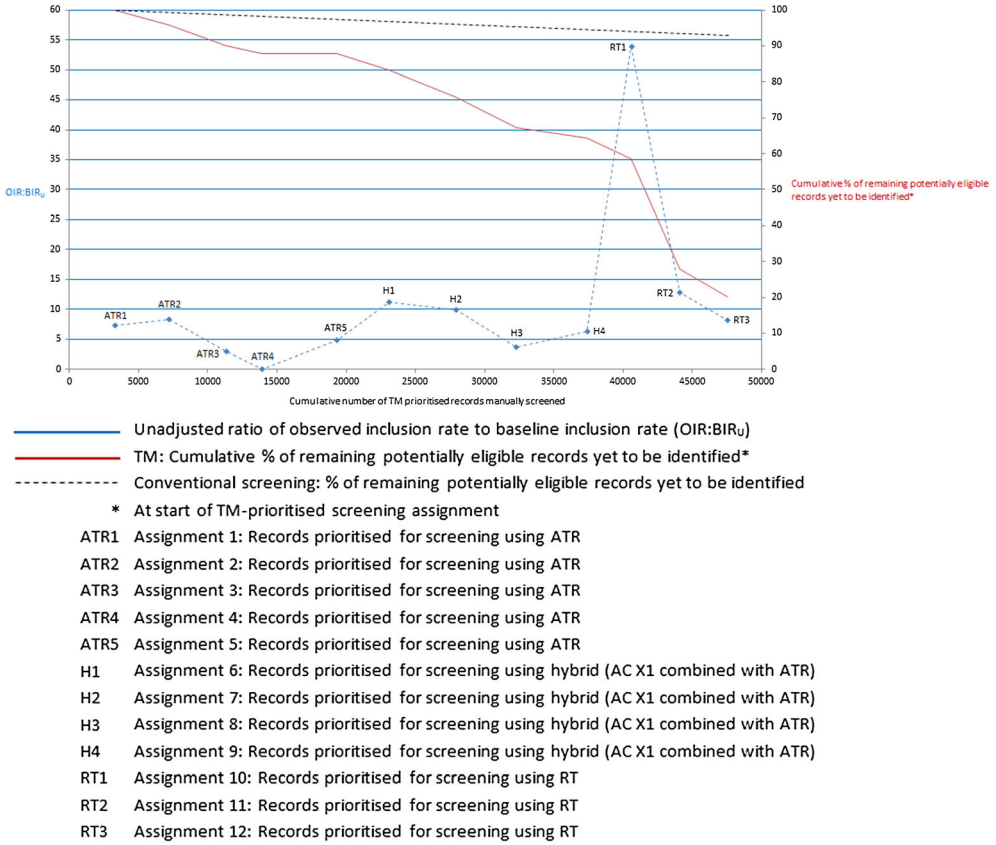

Figure 4. Performance of each text mining technology: choice architecture scoping review.

5. Discussion

This study expands an emerging corpus of empirical evidence for the use of TM to support study screening and selection in reviews (Cohen *et al.*, 2012; Cohen *et al.*, 2010; Wallace *et al.*, 2010a; Cohen *et al.*, 2006) by demonstrating its feasibility, strengths and limitations in two extremely large-scale scoping reviews of public health evidence. By reducing manual screening workload to manageable levels, TM made it possible to conduct the two largest scale reviews ever attempted, in order to assemble, configure and describe large and complex evidence bases that crossed multiple disciplinary boundaries.

Our evaluation findings indicate that, although we were unable (within available time and resources) to locate and select all potentially eligible studies expected to be present within the full record set of each review, use of TM enabled us to identify up to 10 times more potentially eligible studies than we would have identified had we used conventional screening methods with the same investment of workload. The objective of locating all available studies is relaxed to some extent in scoping reviews, which prioritise conceptual breadth over depth. Although we cannot know the performance of the TM over the entire dataset (because we did not manually screen all ≈1.85 million records), the order of magnitude of workload reduction that we observed here has previously been reported elsewhere (Wallace *et al.*, 2012). We should acknowledge, however, that empirical evidence suggests the performance of classifiers (AC) increases with the size of dataset, and so for smaller reviews, performance gains from this specific TM technology may be smaller (Wallace *et al.*, 2012).

TM technologies can hunt through the text of many thousands of study records in rapid succession to search for patterns, associations and connections that might easily elude even the most conscientious researchers undertaking screening by conventional methods (McDonald and Kelly, 2012). As a result, their use in these scoping reviews also facilitated development of a refined conceptual understanding of complex and heterogeneous sets of interventions that aim to alter environments with the goal of changing health behaviours and their proposed mechanisms of action. We therefore propose that the methods used in these scoping reviews are transferable for use in other scoping or systematic reviews that incorporate a conceptual development or explanatory dimension.

By any measure, the two examples evaluated in this study were 'extreme reviews' (Shemilt *et al.*, 2012). Practical challenges involved in assembling, de-duplicating and applying TM techniques to ≈1.85 million study records were substantial. For example, exporting study records from their source databases to reference management software was time consuming due to upper limits on numbers of records that can be exported in a single batch. Additionally, record transfer, de-duplication and TM routines required major reprogramming

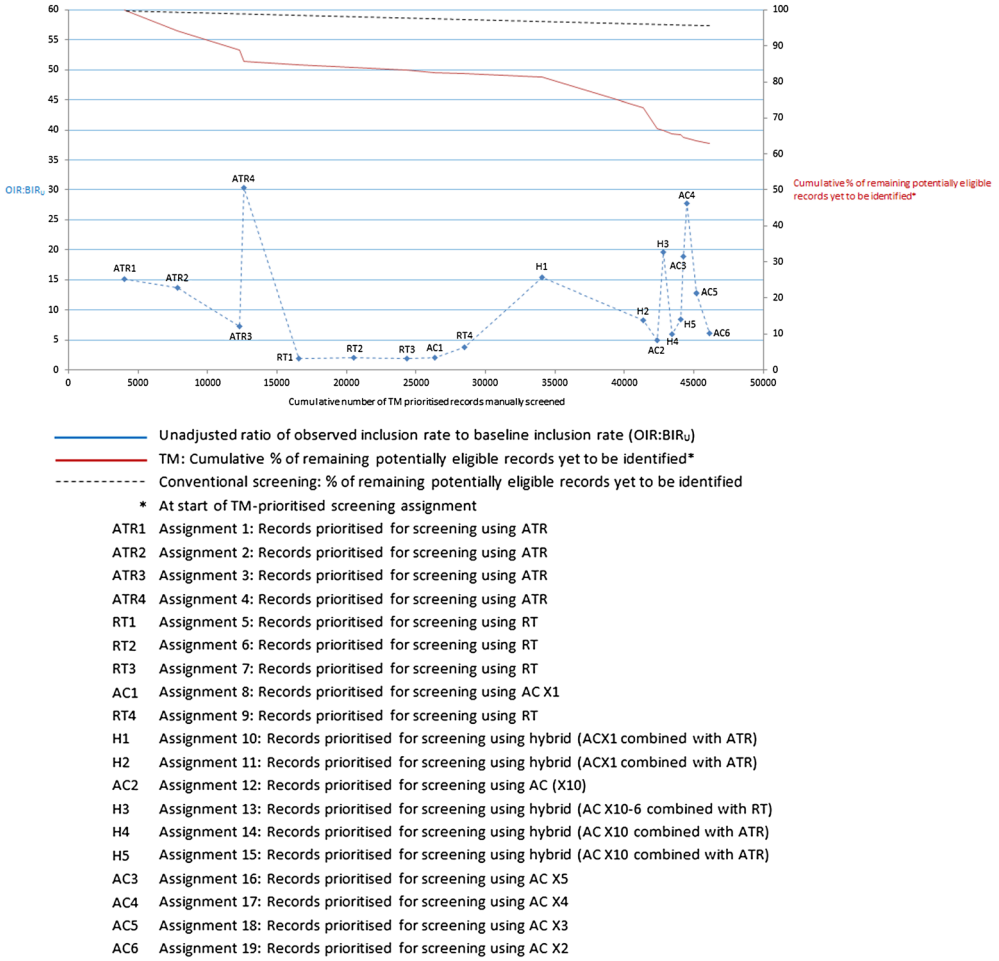

Figure 5. Performance of each text mining technology: economic environment scoping review.

of ER4 and considerable computer processing power to enable the software to handle record sets of this sheer size whilst maintaining the stability of the user interface. Combined with the human resources required to manually screen a combined total of >100 000 study records within a period of 10–11 weeks, this made the use of TM to support study screening and selection in such extremely large-scale reviews a resource intensive exercise despite the overall efficiencies gained. Use of TM technologies does not eliminate the need to screen irrelevant studies, and some relevant studies are still likely to be missed. Although the unaided time investment in manual screening of study records by researchers is likely to remain relatively consistent over time (per record), many of the practical challenges described previously and their resource implications will be ameliorated in future extremely or moderately large-scale scoping or systematic reviews that employ TM for the same purpose, because our work has now established updated software and methods that are transferable to such reviews.

There are two points to be raised about the metrics that we used to inform assessment of the overall performance of TM. First, it should be borne in mind that reducing the screening burden by approximately 90% may only have been possible because the usual requirement to identify every relevant study (i.e. a 100% retrieval rate) was relaxed for these scoping reviews. We may have identified the 'low hanging fruit' and considerably more workload may have been needed to identify the remaining eligible studies. This is an empirical question, for which a definitive answer would require the manual screening of all ≈1.85 million records. However, we are currently conducting follow-up simulation studies to investigate the closely related question of whether our chosen configuration of text mining technologies operated at optimal efficiency, or whether there were more optimal alternative configurations to minimise workload in these reviews. Second, we had difficulty devising an appropriate metric to capture the overall performance of the TM during the screening process. While the BIR and OIR were useful, if the OIR : BIR$_U$ ratio was 8.2 when say, 10% of eligible study reports had been identified and also when 80% of relevant study reports had been identified, these ratios are not strictly comparable because in the first case, there were 90% of eligible reports yet to find and in the second, only 20%. Thus a ratio of 8.2 in the second case actually represents better TM performance. Further methodological work is needed to develop a metric able to encapsulate the relative scarcity of remaining studies.

It is worth reflecting a little on the stark difference in comparative performance of RT between reviews (Figures 4 and 5). The exercise of using RTs to identify relevant studies was clearly successful in the CA review and unsuccessful in the EE review. This may be explained by the fact that there were only 21 such terms in the EE review, but 116 in the CA review; more terms may result in better performance. However, we went to considerable lengths in the EE review to identify more terms with limited success because we were looking for terms that *distinguished* between our 'provisionally eligible' and 'excluded' categories, and such terms were difficult to identify in the EE review. Other TM technologies performed much better than RT in the EE review, however, so the explanation is not that the application of exclusion criteria in the EE review relied on reviewer interpretation of abstracts containing similar terminology. Instead, it may be that we needed more terms to enable the RT ratio to operate effectively over a corpus of abstracts where one single term might be indicative of both relevance and irrelevance; but where terms occurring in combination with one another might distinguish better between the two. The challenge here may be in identifying those combinations of terms in a way that does not simply duplicate the way that TM technologies operate.

Work elsewhere has focused on refining TM technologies to yield the most efficient solutions in terms of performance (Wallace *et al*., 2010b; Cohen *et al*., 2006). Notwithstanding the importance of this work, our approach achieved a pragmatic balance between using technologies that offer benefits in reviews and are also sufficiently widely available–and easy to deploy–that they can be experimented with elsewhere. One of the contributions of this study has been the demonstration that relatively simple TM technologies can yield tangible benefits. Deployment of the support vector machine (AC) was a relatively challenging task technically, but one which should be within the capability of many IT departments; the use of RT and ATR should be achievable in many more situations. However, it is important to highlight that although TM technologies themselves are automated, their deployment is a semi-automated process. Careful user input is required to manage TM performance and to implement, interpret and respond to analyses of data collected from real-time performance monitoring in the pursuit of an optimal configuration (sequencing) of the different technologies. Careful user input is also needed to configure those technologies (such as RT) that rely primarily on user experience to act as a counterpoint to the automated technologies. As such (and despite the TM technologies assessed in this study being relatively easy to deploy), new users will invariably benefit from initial training and support from more experienced users in order to ensure judicious and effective use of TM technologies to support study screening and selection in reviews.

Finally, the technologies and methods utilised in this study are still in their infancy and require significant development before they can be considered to have been proven in scoping reviews and more conventional systematic reviews. The use of TM technologies to support study screening and selection does, however, offer the potential for reviews to become more efficient and thus both cheaper and more responsive to decision-making timetables—at the same time as facilitating conceptual development, broad reviews and scoping with ever-increasing numbers of potentially relevant publications.

Appendix A: MEDLINE (OVID SP) search strategies

Choice architecture review (1948 to June Week 5 2011)

1. exp diet/
2. exp diet therapy/
3. exp food/
4. exp beverages/
5. food habits/
6. food preferences/
7. fasting/
8. adolescent nutritional physiological phenomena/
9. elder nutritional physiological phenomena/
10. exp food industry/
11. exp hunger/
12. exp appetite regulation/
13. exp appetite/
14. exp digestion/
15. exp eating/
16. exp eating disorders/
17. exp child nutrition disorders/
18. exp infant nutrition disorders/
19. nutritional requirements/
20. nutritional status/
21. nutrition assessment/
22. nutrition disorders/

23. exp nutritive value/
24. (nutri$ or calori$ or diet$ or food$ or eat$ or eat or meal$ or snack$ or cook$ or restaurant$ or supermarket $ or cafe$).ti,ab.
25. (drink$ or beverage$).ti,ab.
26. exp Alcohol Drinking/
27. exp Alcohol-Related Disorders/
28. (drink$ or drunk$ or alcohol$ or beer$ or lager$ or wine$ or cider$ or alcopop$ or spirit$ or liquor$ or distilled beverage$ or whisky$ or whiskey$ or whiskies or schnapps or liqueur$ or brandy or brandies or gin$ or rum$ or tequila$ or vodka$).ti,ab.
29. exp Tobacco/
30. exp Smoking/
31. exp Smoking Cessation/
32. (cigar$ or smoking or smoke$ or tobacco$).ti,ab.
33. exp physical exertion/
34. exp human activities/
35. exp leisure activities/
36. exp locomotion/
37. exp physical education/
38. lifestyle/
39. sedentary lifestyle/
40. yoga/
41. fitness centers/
42. motor activity/
43. (physical$ adj5 (exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
44. (aerobic adj5 (exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
45. (strength$ adj5 (exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
46. (flexib$ adj5 (exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
47. (balanc$ adj5 (exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
48. (exercise$ adj5 (train$ or activit$ or fit$ or endur$)).ti,ab.
49. ((occupation$ or work$ or recreation$2 or leisure or play or household or home or domestic or commut$3 or transport$) adj5 (energ$ or exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
50. ((walk$3 or hike or hiking or climbing or run$3 or jog$3 or swim$1 or swimming or bicycl$3 or cycl$3 or bike$1 or biking or gym$ or rowing or canoe$ or kayak$ or sailing or windsurf$3 or surf$3 or diving or sport$3 or rollerblading or rollerskating or skating or skiing or yoga or pilates or calisthenics or (jump$3 adj rope$1) or (lift$3 adj weight$1) or gym$ or circuit or resistance or resilience or dance or dancing or fishing or hunting or shooting) adj5 (energ$ or exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
51. (led walk$ or health walk$).ti,ab.
52. ((leisure or fitness) adj5 (centre$ or center$ or facilit$)).ti,ab.
53. (fitness adj class$).ti,ab.
54. (fitness adj (regime$ or program$)).ti,ab.
55. cardiorespiratory fitness.ti,ab.
56. aerobic capacity.ti,ab.
57. (intensity adj2 (rest or quiet or light or moderate or vigorous)).ti,ab.
58. ((car or cars or bus or buses or train or trains or transport$) and (energ$ or activit$ or exercis$)).ti,ab.
59. (active adj (travel$4 or transport$ or commut$)).tw.
60. ((promot$ or uptak$ or encourag$ or increas$ or start$ or adher$ or sustain$ or maintain$) adj5 gym$).ti,ab.
61. ((promot$ or uptak$ or encourag$ or increas$ or start$ or adher$ or sustain$ or maintain$) adj5 physical activit$).ti,ab.
62. ((promot$ or uptak$ or encourag$ or increas$ or start$ or adher$ or sustain$ or maintain$) adj5 (circuit$ or aqua$)).ti,ab.
63. ((promot$ or uptak$ or encourag$ or increas$ or start$ or adher$ or sustain$ or maintain$) adj5 (exercis$ or exertion or keep fit or fitness class or yoga or aerobic$)).ti,ab.
64. ((decreas$ or reduc$ or discourag$) adj5 (sedentary or deskbound or inactiv$)).ti,ab.
65. (exercis$ adj aerobic$).tw.
66. (physical$ adj5 (fit$ or train$ or activ$ or endur$)).tw.
67. (exercis$ adj5 (train$ or physical$ or activ$)).tw.
68. ((lifestyle or life-style) adj5 physical$).tw.
69. ((lifestyle or life-style) adj5 activ$).tw.
70. exp behavior/
71. behavior$.ti,ab.
72. behaviour$.ti,ab.

73. environment$.ti,ab.
74. consum$.ti,ab.
75. intake$.ti,ab.
76. perform$.ti,ab.
77. exp health promotion/
78. exp primary prevention/
79. exp attention/
80. exp visual perception/
81. exp feedback, psychological/
82. exp feedback, sensory/
83. exp perception/
84. exp illusions/
85. exp psychomotor performance/
86. (change$ or alter$ or adjust$ or modif$ or adapt$ or add$ or subtract$ or restrict$ or shrink$ or shrunk or extend$ or expand$ or supplement$ or improve$ or increas$ or higher or larger or longer or remov$ or constrain$ or restrain$ or limit$ or lower$ or reduc$ or decreas$ or smaller or greater or less$ or fewer or more or choice$ or choose or chose$ or option$).ti,ab.
87. or/1-69
88. or/70-85
89. and/86-88
90. limit 89 to (english language and humans)

Economic environment review (1948 to June Week 5 2011)

1. exp diet/
2. exp diet therapy/
3. exp food/
4. exp beverages/
5. food habits/
6. food preferences/
7. fasting/
8. adolescent nutritional physiological phenomena/
9. elder nutritional physiological phenomena/
10. exp food industry/
11. exp hunger/
12. exp appetite regulation/
13. exp appetite/
14. exp digestion/
15. exp eating/
16. exp eating disorders/
17. exp child nutrition disorders/
18. exp infant nutrition disorders/
19. nutritional requirements/
20. nutritional status/
21. nutrition assessment/
22. nutrition disorders/
23. exp nutritive value/
24. (nutri$ or calori$ or diet$ or food$ or eat$ or meal$ or snack$ or cook$ or restaurant$ or supermarket$ or cafe$).ti,ab.
25. ((drink$ or beverage$) not alcohol$).ti,ab.
26. or/1-25
27. physical exertion/
28. exp human activities/
29. exp leisure activities/
30. exp locomotion/
31. exp physical education/
32. lifestyle/
33. sedentary lifestyle/
34. yoga/
35. fitness centers/
36. motor activity/
37. (physical$ adj5 (exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.

38. (aerobic adj5 (exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
39. (strength$ adj5 (exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
40. (flexib$ adj5 (exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
41. (balanc$ adj5 (exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
42. (exercise$ adj5 (train$ or activit$ or fit$ or endur$)).ti,ab.
43. ((occupation$ or work$ or recreation$2 or leisure or play or household or home or domestic or commut$3 or transport$) adj5 (energ$ or exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
44. ((walk$3 or hike or hiking or climbing or run$3 or jog$3 or swim$1 or swimming or bicycl$3 or cycl$3 or bike$1 or biking or gym$ or rowing or canoe$ or kayak$ or sailing or windsurf$3 or surf$3 or diving or sport$3 or rollerblading or rollerskating or skating or skiing or yoga or pilates or calisthenics or (jump $3 adj rope$1) or (lift$3 adj weight$1) or circuit or resistance or resilience or dance or dancing or fishing or hunting or shooting) adj5 (energ$ or exercis$ or train$ or activit$ or fit$ or endur$)).ti,ab.
45. (led walk$ or health walk$).ti,ab.
46. ((leisure or fitness) adj5 (centre$ or center$ or facilit$)).ti,ab.
47. (fitness adj class$).ti,ab.
48. (fitness adj (regime$ or program$)).ti,ab.
49. cardiorespiratory fitness.ti,ab.
50. aerobic capacity.ti,ab.
51. (intensity adj2 (rest or quiet or light or moderate or vigorous)).ti,ab.
52. ((car or cars or bus or buses or train or trains or transport$) and (energ$ or activit$ or exercis$)).ti,ab.
53. (active adj (travel$4 or transport$ or commut$)).tw.
54. ((promot$ or uptak$ or encourag$ or increas$ or start$ or adher$ or sustain$ or maintain$) adj5 gym$).ti,ab.
55. ((promot$ or uptak$ or encourag$ or increas$ or start$ or adher$ or sustain$ or maintain$) adj5 physical activit$).ti,ab.
56. ((promot$ or uptak$ or encourag$ or increas$ or start$ or adher$ or sustain$ or maintain$) adj5 (circuit$ or aqua$)).ti,ab.
57. ((promot$ or uptak$ or encourag$ or increas$ or start$ or adher$ or sustain$ or maintain$) adj5 (exercis$ or exertion or keep fit or fitness class or yoga or aerobic$)).ti,ab.
58. ((decreas$ or reduc$ or discourag$) adj5 (sedentary or deskbound or inactiv$)).ti,ab.
59. (exercis$ adj aerobic$).tw.
60. (physical$ adj5 (fit$ or train$ or activ$ or endur$)).tw.
61. (exercis$ adj5 (train$ or physical$ or activ$)).tw.
62. ((lifestyle or life-style) adj5 physical$).tw.
63. ((lifestyle or life-style) adj5 activ$).tw.
64. or/27-63
65. 26 or 64
66. (risk$ adj4 (non-communicable or non communicable or chronic)).ti,ab.
67. blood pressure/
68. hypertension/
69. blood glucose/
70. hyperglycemia/
71. cholesterol/
72. cholesterol, dietary/
73. cholesterol, hdl/
74. cholesterol, ldl/
75. cholesterol, vldl/
76. cholesterol esters/
77. hypercholesteremia/
78. exp hyperlipidemias/
79. exp body weight changes/
80. harm reduction/
81. exp overnutrition/
82. exp overweight/
83. exp obesity/
84. (overweight or over weight).ti,ab.
85. adipos$.ti,ab.
86. fat overload syndrome$.ti,ab.
87. (overeat or over eat).ti,ab.
88. weight cycling.ti,ab.
89. weight reduc$.ti,ab.
90. weight losing.ti,ab.

91. weight maint$.ti,ab.
92. weight decreas$.ti,ab.
93. weight watch$.ti,ab.
94. weight control$.ti,ab.
95. weight gain.ti,ab.
96. weight loss.ti,ab.
97. weight chang$.ti,ab.
98. (bmi or obes$ or overweight or (blood adj pressure) or hypertensi$ or (blood adj glucose) or hyperglyc?mi $ or cholester$ or hypercholester$ or hyperlipid?emia$).ti,ab.
99. or/66-98
100. (economic$ or financ$ or cost or costs or costing or pric$ or monetis$ or income$ or wage$ or salar$ or (expenditure$ not energy) or time$).ti,ab.
101. (tax$ or subsid$ or credit$ or (((cash or income) adj2 transfer) or payment) or (welfare adj benefit$) or incentiv$ or disincentiv$ or remunerat$ or retail$ or sale$ or promo$ or consumer$ or consumption$ or purchas$ or shop$ or buy$).ti,ab.
102. ((product or good or service or market) adj (innovat$ or develop$ or efficien$ or quality)).ti,ab.
103. or/100-102
104. 65 and 103
105. 99 and 103
106. 104 or 105
107. animals/
108. humans/ and animals/
109. 107 not 108
110. 106 not 109

Appendix B: Overview of TM technologies

Appendix B provides an overview of the three text mining technologies used in our two example scoping reviews and hybrids (which combine two of the three technologies).

Automatic term recognition

Automatic term recognition (ATR) reorders the list of title-abstract records assigned for screening in order of priority (i.e. the most relevant, or potentially eligible, records should be located towards the top of the list). ATR operates in two phases. First, the titles and abstracts of records already identified as provisionally eligible are analysed, and a list of terms, together with a score which indicates their relative importance, is obtained. There are a range of tools which perform ATR; we used the National Centre for Text Mining's web service for TerMine (http://www.nactem.ac.uk/software/termine/), which combines linguistic and statistical analysis (Frantzi *et al*., 2000). As an example, the terms identified by TerMine for the final list of potentially eligible title-abstract records in the CA review are given in Table A2.

Second, the top 100 terms are used to search the as yet unscreened records, with search terms weighted according to the score given by TerMine. Thus, if the terms listed in Table 2 were used, the search would look something like: 'energy intake' (220.37), 'portion size' (193.46) or 'physical activity' (171.30), etc. The numbers in brackets weight the relative importance of the terms in the search; so 'food intake' (109.87) is approximately half as important as 'energy intake' (220.37). In the actual search, which runs in SQL Server 2008 (Microsoft Corporation, Redmond WA, USA), the scores are scaled to lie between 0 and 1. The search returns an ordered list of title-abstract records, with those that are 'most similar' to the records analysed by TerMine at the top of the list. Reviewers then screen the top *x* items in the list and, when more potentially eligible title-abstract records have been identified, the process is re-run, analysing the larger set of studies in TerMine. This process is repeated at regular intervals, with a larger set of potentially eligible title-abstract records analysed each time.

Automatic classification

The output from AC differs from that obtained using ATR because rather than an ordered list of title-abstract records, AC automatically classifies title-abstract records as being relevant or not relevant (Wallace *et al*., 2010b). Like ATR, it operates using existing screening decisions. However, rather than only using those title-abstract records already marked as potentially eligible, it also utilises those marked as excluded in order to build a statistical model which can then be used to predict the likely eligibility/non-eligibility of the remaining unscreened records. AC is used widely in data mining applications, and there are a wide range of algorithms, such as linear regression, Bayesian models, neural networks and support vector machines (SVM). For the purposes of this study, we used the libSVM support vector machine because this technology is known to perform well for text classification, and the libSVM implementation is actively developed and widely available (Chang and Lin, 2011); we

used the radial bias function kernel with parameters set using a 'grid' search. A challenge when using AC in a systematic review is *class imbalance*; that is, there are usually many more records marked as excluded than provisionally eligible. Because the SVM needs roughly equal numbers from each class in order to perform well, we draw a random sample of excluded records equal in number to the total number records marked as provisionally eligible at a given stage of the screening process. This is known as 'undersampling' because some relevant information is not utilised. Although this generates good results in terms of classification, multiple 'runs' of the classifier will give slightly different results due to changes in the composition of the random sample of excluded records between iterations. In order to maximise precision, we therefore ran the classifier multiple (between three and 10) times and then manually screened only those title-abstract records that were consistently classified as being potentially eligible. This component of the approach is known as 'bagging'. Empirical evidence for its superior performance over other methods when classifying the type of data found in systematic reviews has been reported by Wallace and colleagues (Wallace *et al*., 2011).

Reviewer terms

Reviewer terms (RT) operates rather differently to the previous two automated technologies because it is highly reliant on reviewer input. Operationally, it is composed of two lists of terms which are drawn up by the reviewers undertaking the screening; one list of 'relevant' terms that are subjectively judged to be indicative of a provisionally eligible record, and one list of 'irrelevant' terms that are subjectively judged to be associated with excluded records. The text contained in each title-abstract record that is yet to be screened is then analysed, and the number of 'relevant' and 'irrelevant' terms they contain is calculated. A simple ratio is then generated, dividing the number of 'relevant' terms in a given abstract by the number of 'irrelevant' terms. The natural logarithm is then taken, giving us a score where <0 indicates a balance of irrelevant terms and >0 a balance of relevant terms. Items can then be ranked according to their balance of 'relevant'/'irrelevant' terms. The purpose of this method is to act as a counterpoint to the automated technologies; whereas in ATR and AC, the results are heavily determined by those studies already identified as being relevant; RT offers another perspective on potential relevance, offering some protection against the problem of hasty generalisation (see main text article, Introduction). For further work on the use of RT, see Small and colleagues (Small *et al*., 2011).

Hybrids

As well as using each of the previous technologies alone, we also used them in combination with one another by ordering the results of the AC using ATR or RT. Although unnecessary in most reviews, these hybrid approaches were adopted because of the extremely large numbers of title-abstract records retrieved. For example, ≈10 000 records were consistently classified as being potentially eligible after running 10 iterations of the AC in the EE review. This AC-generated list of ≈10 000 records was therefore prioritised using ATR in order to identify a subset of those records most likely to be marked as provisionally eligible.

Table A2: Top TerMine terms for the included studies in the choice architecture review

Term	Score
energy intake	220.37
portion size	193.46
physical activity	171.30
stair use	135.76
food intake	109.87
energy density	84.57
food choice	64.61
large portion	48.27
nutrition information	39.75
environmental intervention	36.79
fast food	34.37
control condition	31.75
calorie information	31.00
nutrition label	28.40
vegetable consumption	24.73
healthy food	24.43
vegetable intake	24.00
physical activity level	23.61
control group	23.50
public health	22.63

References

Chang C-C, Lin C-J. 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2: 27.

Frantzi K, Ananiadou S, Mima H. 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries 3: 115–130.

Small K, Wallace BC, Brodley CE, Trikalinos TA. 2011. The constrained weight space SVM: learning with ranked features. Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA.

Wallace BC, Small K, Brodley CE, Trikalinos TA. 2010c. Active learning for biomedical citation screening. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, KDD '10, New York, NY, 173-182.

Wallace BC, Small K, Brodley CE, Trikalinos TA. 2011. Class imbalance, redux. In: IEEE 11th International Conference on Data Mining (ICDM). IEEE, 754–763.

Appendix C: Establishing inter-rater reliability and estimating baseline inclusion rate

Appendix C describes methods to establish inter-rater reliability and estimate a BIR within a text mining framework in reviews, by presenting worked examples of the methods we applied in our two scoping reviews (CA and EE), including results.

Establishing inter-rater reliability

Methods

Automatic term recognition (ATR) was used to prioritise a sample of 500 unscreened title-abstract records within each review. Each sample was assigned for manual screening by two researchers working independently, in order to establish and formally assess inter-rater reliability regarding provisional eligibility decisions. The rationale for using prioritised samples of records for this purpose was that establishment and assessment of inter-rater reliability would require screening of a mix of provisionally eligible, borderline and ineligible records (and rates of eligible or borderline records within a random sample of unscreened records drawn from record sets of this sheer size could be expected to be very low). Inter-rater reliability was assessed by calculating the joint probability of agreement and an associated Cohen's kappa coefficient (κ), to measure levels of coding agreement between two researchers classifying N records into C categories (Carletta, 1996). The formula for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Where Pr(a) is the relative observed agreement among researchers, and Pr(e) is the hypothetical probability of chance agreement.

Interpretation of the results of inter-rater reliability assessments was based on pre-set criteria as follows: Values of κ 0.40–0.59 were considered to reflect 'fair agreement'; values of κ 0.60−0.74 were considered to reflect 'good agreement' and values of κ ≥0.75 were considered to reflect 'excellent agreement' (Higgins and Deeks, 2008). Our pre-set decision rule was to proceed to the next phase of title-abstract screening once we had achieved 'good agreement' between two researchers. In the EE review, 'good agreement' was achieved in the first prioritised sample of 500 records. In the CA review, 'fair agreement' was achieved in the first prioritised sample of 500 records, so a second prioritised sample of 500 records was assigned for manual screening, and 'good agreement' was achieved in this second sample. In both reviews, the two researchers met following initial, independent screening of each consecutive set of 500 records to resolve disagreements regarding provisional eligibility by discussion of each record with reference to detailed coding notes. Once each disagreement had been resolved, the corresponding record was re-coded based on the consensus provisional eligibility decision.

Results

In the CA review, the kappa statistic (κ) associated with the joint probability of agreement between two researchers working independently to assess eligibility of the first ATR prioritised sample of 500 records was 0.55 ($N = 500$ and $C = 2$), interpreted as 'fair agreement'. In the second, ATR prioritised sample of 500 records, the corresponding κ was 0.67 ($N = 500$ and $C = 2$), interpreted as 'good agreement'. In the EE review, the kappa statistic (κ) associated with the joint probability of agreement between two researchers working independently to assess eligibility of an ATR prioritised sample of 500 records was 0.72 ($N = 500$ and $C = 2$), interpreted as 'good agreement'.

Estimating baseline inclusion rate

Methods

In the current context, a BIR can be interpreted as an estimate of the proportion of provisionally eligible records expected to be present within each full record set (i.e. within 804 919 (CA) and 1 053 908 (EE) title-abstract records) (Hoenig and Heisey, 2001; Thomas, 1997). The rationale for estimating BIRs was threefold. First, we could extrapolate from the BIR to calculate the absolute number of potentially eligible records expected to be present within the corresponding full record set, for use as a benchmark against which to monitor overall progress towards identification and selection of all potentially eligible records. Second, we could use the BIR as a baseline against which to monitor the performance of each specific TM technology, in terms of its relationship to the OIR (i.e. the rate at which eligible study records were identified in practice within each consecutive set of prioritised records assigned for manual screening). Performance monitoring was also used to guide real-time decisions about when to switch between the different TM technologies to prioritise further sets of unscreened records for manual screening. A decision to switch became more likely as the OIR within assigned, screened record sets declined towards (or below) BIRs. Third, because calculating the BIR involves screening a random sample of unscreened records (see the succeeding texts), this is analogous to conventional, quasi-random screening methods. As such, the BIR could serve as a proxy to represent our unobserved counterfactual to the use of TM (i.e. conventional screening methods), for the purposes of assessing TM performance.

Estimation of BIRs comprised two stages. In the first stage, we calculated the size of the random sample of unscreened records that would need to be screened in order to give us a reasonable degree of confidence in the estimated BIR. This calculation is analogous to a power analysis used to calculate sample sizes in the planning of statistical studies. The calculation uses the formula:

$$\mathrm{ME} = \mathrm{z}^{*}\mathrm{sqrt}\{\mathrm{pi}^{*}(1-\mathrm{pi})/(\mathrm{n}-1)\}^{*}\mathrm{sqrt}\{1-\mathrm{n}/\mathrm{N}\}$$

Where z is a critical value from the normal distribution that corresponds to a specified confidence interval (CI); pi is the sample proportion; n is the sample size; N is the population size and ME is the margin of error.

We specified z (indirectly, by specifying a CI), pi, N and ME and applied the formula to calculate n. In the current context, the sample proportion (pi) represents an a priori estimate of the upper bound of the rate of provisional eligibility records that could plausibly be expected to be present in each full record set. We judged it appropriate to specify a conservatively large value for pi (i.e. the selected value 0.01 expressed an a priori estimate that a maximum of 1% of records within each full record set would be coded as provisionally eligible at the title-abstract screening stage). Similarly, due to the sheer size of our respective full record sets (N), it was judged important to specify a conservatively small value for the margin of error (ME). This was because even relatively small errors in the estimated BIRs would equate to large discrepancies in terms of absolute numbers of records when extrapolated to the full record sets (N). For both scoping reviews, we specified the following: CI = 95%; pi = 0.01 and ME = 0.0025. In the CA review, N = 804 919, whereas in the EE review, N = 1 053 908. We used an online tool, which utilises the aforementioned formula to calculate n (Lenth, 2009). In the CA review, n = 6040 records, whereas in the EE review, n = 6051 records.

In the second stage of BIR estimation, we drew a random sample of unscreened records of size n from each review's full record set and assigned these for manual screening. The estimated BIR is equal to the rate of records coded as provisionally eligible within the screened random sample. Once estimated, we extrapolated BIRs to estimate the absolute numbers of provisionally eligible records expected to be present within each full record set (CA and EE).

Results

Baseline Inclusion Rates (BIRs) were 0.000498 (CA) and 0.003481 (EE). These rates indicated that we could expect ≈0.05% (CA) and ≈0.35% (EE) of study records within the respective full record sets to be selected as provisionally eligible at the title-abstract screening stage. Both BIRs were below the a priori estimates of 1% that we had allowed for in their calculation. We could therefore have a reasonable degree of confidence in the validity of estimated BIRs within the specified margin of error. Extrapolating BIRs to the respective full record sets generated estimated totals of 400 (CA) and 3669 (EE) study records that could expected to be selected as provisionally eligible at the title-abstract screening stage.

References

Carletta JC. 1996. Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics 22: 249–254.

Higgins JPT, Deeks JJ. 2008. Chapter 7: Selecting studies and collecting data. In Cochrane Handbook for Systematic Reviews of Interventions, Higgins JPT, Green S (eds.). John Wiley & Sons, Chichester, 151–186.

Hoenig JM, Heisey DM. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. The American Statistician 55: 1–6.

Lenth RV. 2009. Java Applets for Power and Sample Size [Computer software]. Available from: http://www.stat.uiowa.edu/~rlenth/Power. Accessed: 23 August 2011.

Thomas L. 1997. Retrospective power analysis. Conservation Biology 11: 276–280.

6. Acknowledgements

The study reported in this article was funded by the UK Department of Health Policy Research Programme (107/0001-Policy Research Unit in Behaviour and Health). The views expressed in this article are those of the authors and not necessarily those of the UK Department of Health. We are particularly grateful to Jeff Brunton, Sergio Graziosi and Irene Kwan at the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre), Department of Children and Health, Institute of Education, UK, for the technical and research support they provided to this study.

7. Conflict of Interest

EPPI REVIEWER 4, the systematic review management software used to manage the study screening and selection phases of the two systematic scoping reviews reported in this paper, is semi-commercial not-for-profit software that requires payment of subscription fees for public use. Two authors of this article (Alison O'Mara-Eves and James Thomas) are affiliated to the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre), which develops and publishes EPPI REVIEWER 4. No other conflicts of interest exist.

References

Arksey H, O'Malley L. 2005. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* **8**: 19–32.

Ananiadou S, Okazaki N, Procter R, Rea B, Sasaki Y, Thomas J. 2009. Supporting systematic reviews using text mining. *Social Science Computer Review* **27**: 509–523.

Booth A. 2008. Unpacking your literature search toolbox: on search styles and tactics. *Health Information and Libraries Journal* **25**: 313–317.

Booth A. 2003. Formulating answerable questions. In Evidence Based Practice: A Handbook for Information Professionals, Booth A, Brice A (eds.). Facet, London, 61–70.

Brunton G, Stansfield C, Thomas J. 2012. Finding relevant studies. In An Introduction to Systematic Reviews, Gough D, Oliver S, Thomas J (eds). Sage, London, 107–134.

Cabinet Office Behavioural Insights Team. 2010. Applying Behavioural Insight to Health. Cabinet Office, London.

Cohen AM, Ambert K, McDonagh M. 2012. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Medical Informatics and Decision Making* **12**: 33.

Cohen AM, Adams CE, Davis JM, Yu C, Yu PS, Meng W, Duggan L, McDonagh M, Smalheiser NR. 2010. Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. In Proceedings of the 1st ACM International Health Informatics Symposium. IHI'10, Arlington, VA, 376–380.

Cohen A, Hersh W, Peterson K, Yen P. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informantics Association.* **13**: 206–219.

Deeks JJ, Higgins JPT, Altman DG. 2008. Chapter 9: analysing data and undertaking meta-analyses. In Cochrane Handbook for Systematic Reviews of Interventions, Higgins JPT, Green S (eds.). John Wiley & Sons, Chichester, 243–296.

Dolan P, Hallsworth M, Halpern D, King D, Vlaev I. 2010. MINDSPACE: Influencing Behaviour Through Public Policy. Cabinet Office, London.

Gough D, Oliver S, Thomas J. 2012. Moving forward. In An Introduction to Systematic Reviews, Gough D, Oliver S, Thomas J (eds). Sage, London, 257–262.

Gough D, Thomas J. 2012. Commonality and diversity in reviews. In Gough D, Oliver S, Thomas J (eds). An Introduction to Systematic Reviews. Sage, London, 35–66.

Green S, Higgins JPT, Alderson P, Clarke M, Mulrow CD, Oxman AD. 2008. Chapter 1: introduction. In Cochrane Handbook for Systematic Reviews of Interventions, Higgins JPT, Green S (eds.). John Wiley & Sons, Chichester, 3–10.

Greenhalgh T, Peacock R. 2005. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ* **331**: 1064–1065.

Hammersley M. 2002. Systematic or Unsystematic, Is That The Question? Some Reflections on the Science, Art and Politics of Reviewing Research Evidence. Health Development Agency Public Health Steering Group, London.

Hollands GJ, Shemilt I, Marteau TM, Jebb SA, Kelly MP, Nakamura R, Suhrcke M, Ogilvie D. 2013. Altering Choice Architecture to Change Population Health Behaviour: a Large-Scale Conceptual and Empirical Scoping Review of Interventions Within Micro-Environments. Behaviour and Health Research Unit, Cambridge.

House of Lords Science and Technology Select Committee. 2011. Behaviour Change: 2nd Report of Session 2010–12 (HL Paper 179). The Stationery Office Limited, London.

Kaltenthaler E, Tappenden P, Paisley S, Squires H. 2011. Identifying and Reviewing Evidence to Inform the Conceptualisation and Population of Cost-Effectiveness Models. National Institute of Health and Clinical Excellence Decision Support Unit, Sheffield.

Kelly MP, Moore TA. 2012. The judgement process in evidence based medicine and health technology assessment. *Social Theory and Health* **10**: 1–19.

Lefebvre C, Manheimer E, Glanville J. 2008. Chapter 6: searching for studies. In Cochrane Handbook for Systematic Reviews of Interventions, Higgins JPT, Green S (eds.). John Wiley & Sons, Chichester, 95–150.

Marteau TM, Hollands GJ, Fletcher PC. 2012. Changing human behaviour to prevent disease: the importance of targeting automatic processes. *Science* **337**: 1492–1495.

Marteau TM, Ogilvie D, Roland M, Suhrcke M, Kelly MP. 2011. Judging nudging: can nudging improve population health? *BMJ* **342**: 263–265.

McDonald D, Kelly U. 2012. The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure. Joint Information Systems Committee, Bristol.

O'Connor D, Green S, Higgins JPT. 2008. Chapter 5: defining the review question and developing criteria for including studies. In Cochrane Handbook for Systematic Reviews of Interventions, Higgins JPT, Green S (eds). John Wiley & Sons, Chichester, 83–74.

Oliver S, Sutcliffe K.2012. Describing and analysing studies. In An Introduction to Systematic Reviews, Gough D, Oliver S, Thomas J (eds). Sage, London, 135–152.

Papaioannou D, Sutton A, Carroll C, Booth A, Wong R. 2010. Literature searching for social science systematic reviews: consideration of a range of search techniques. *Health Information and Libraries Journal* **27**: 114–122.

Shemilt I, Hollands GJ, Marteau TM, Jebb SA, Kelly MP, Nakamura R, Suhrcke M, Ogilvie D. 2013. Effects of Changes in the Economic Environment on Diet- and Physical Activity-Related Behaviours and Corollary Outcomes: a Large-Scale Scoping Review. Behaviour and Health Research Unit, Cambridge.

Shemilt I, Hollands GJ, Marteau TM, Nakamura R, Jebb SA, Kelly MP, *et al.* in press. Economic interventions for population diet and physical activity behaviour change: a systematic scoping review. *PLoS One*.

Shemilt I, Thomas J, Hollands GJ, Marteau TM, O'Mara-Eves A, Simon A, Kwan I, Ogilvie D. 2012. "Extreme reviewing": use of text-mining to reduce impractical screening workload in extremely large scoping reviews. Oral paper presentation at the 20th Cochrane Colloquium, Auckland, New Zealand.

Thaler RH, Sunstein C. 2008. Nudge: Improving Decisions About Health, Wealth, and Happiness. Yale University Press, New Haven CT.

Thomas J, McNaught J, Ananiadou S. 2011. Applications of text mining within systematic reviews. *Research Synthesis Methods* **2**: 1–14.

Valaitis R, Martin-Misener R, Wong ST, MacDonald M, Meagher-Stewart D, Austin P, Kaczorowski J, O-Mara L, Savage R, Strengthening Primary Health Care through Public Health and Primary Care Collaboration Team. 2012. Methods, strategies and technologies used to conduct a scoping literature review of collaboration between primary care and public health. *Primary Health Care Research and Development* **1**: 219–236.

Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. 2010a. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics* **11**: 55.

Wallace BC, Small K, Brodley CE, Trikalinos TA. 2010b. Active learning for biomedical citation screening. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, KDD '10, New York, NY, 173–182.

Wallace BC, Small K, Brodley CE, Lau J, Schmid CH, Bertram L, Lill CM, Cohen JT, Trikalinos TA. 2012. Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. *Genetics in Medicine* **14**: 663–669.

