International Journal for Quality in Health Care, 2017, 29(4), 571–578
doi: 10.1093/intqhc/mzx076
Advance Access Publication Date: 23 June 2017
Article

Article

Linking quality indicators to clinical trials: an automated approach

ENRICO COIERA[1], MIEW KEEN CHOONG[1], GUY TSAFNAT[1], PETER HIBBERT[1,2], and WILLIAM B. RUNCIMAN[1,2,3]

[1]Centre for Health Informatics, Australian Institute of Health Innovation, Faculty of Medicine and Health Science, Macquarie University, Sydney, Australia, [2]Centre for Population Health Research, University of South Australia, Adelaide, South Australia, and [3]Australian Patient Safety Foundation, Adelaide, South Australia

Address reprint requests to: Enrico Coiera, Centre for Health Informatics, Australian Institute of Health Innovation, Faculty of Medicine and Health Science, Macquarie University, Sydney, Australia. Tel: +61-29-850-2403; E-mail: enrico.coiera@mq.edu.au

Editorial Decision 2 June 2017; Accepted 15 June 2017

Abstract

Objective: Quality improvement of health care requires robust measurable indicators to track performance. However identifying which indicators are supported by strong clinical evidence, typically from clinical trials, is often laborious. This study tests a novel method for automatically linking indicators to clinical trial registrations.

Design: A set of 522 quality of care indicators for 22 common conditions drawn from the CareTrack study were automatically mapped to outcome measures reported in 13 971 trials from ClinicalTrials.gov.

Intervention: Text mining methods extracted phrases mentioning indicators and outcome phrases, and these were compared using the Levenshtein edit distance ratio to measure similarity.

Main Outcome Measure: Number of care indicators that mapped to outcome measures in clinical trials.

Results: While only 13% of the 522 CareTrack indicators were thought to have Level I or II evidence behind them, 353 (68%) could be directly linked to randomized controlled trials. Within these 522, 50 of 70 (71%) Level I and II evidence-based indicators, and 268 of 370 (72%) Level V (consensus-based) indicators could be linked to evidence. Of the indicators known to have evidence behind them, only 5.7% (4 of 70) were mentioned in the trial reports but were missed by our method.

Conclusions: We automatically linked indicators to clinical trial registrations with high precision. Whilst the majority of quality indicators studied could be directly linked to research evidence, a small portion could not and these require closer scrutiny. It is feasible to support the process of indicator development using automated methods to identify research evidence.

Key words: clinical trials, text mining, concept mapping, indicator, quality of health care

Introduction

Clinical indicators are important tools for assessing the quality of health care, and for identifying and prioritizing areas for improvement [1, 2]. To be effective, such indicators need to be robust measures of system performance that correlate with the processes of interest, and be cost-effective to measure.

A lack of uniformity in reporting the rationale for selecting indicators means that it can prove difficult to know whether a given

indicator is based on research evidence. Deductive development of indicators is the most common approach taken [3], where clinical indicators are extracted from clinical guidelines or are identified in the process of guideline development [4]. Most clinical guidelines are based on systematic reviews which are syntheses, mainly of randomized controlled trials (RCTs) [5]. Thus, the development of indicators typically requires a lengthy manual process of searching for and analysis of the research evidence underpinning guidelines.

The development of methods that assist in identifying candidate indicators from the research evidence, or that validate existing indicators against the evidence base, should help increase the efficiency of indicator development, and may also improve confidence in the quality of studies based on indicators. The development of robust and reliable ways of linking indicators to the evidence base is key to achieving this. In a previous study, focussing on paediatric asthma, we demonstrated that manual linking of indicators to clinical outcomes in trial reports found a link to research for 95% of standard indicators [6].

The emergence of new methods such as computational text mining is allowing other complex processes, such as the creation of systematic reviews of the research literature, to be automated [7]. By breaking down the steps in such a complex manual process, it is possible to identify individual steps in which computational tools can assist or replace manual work, improving either efficiency or accuracy [8]. The process of indicator development can similarly be reduced to several steps that form a developmental pipeline [3, 4]. This indicator development pipeline begins with the selection of a clinical or health service process that needs to be monitored, and continues with the identification of candidate indicators, their appraisal based upon performance criteria, and then implementation and evaluation (Fig. 1).

The aim of this study is to assess the degree to which is possible to use computational text processing methods to assist in one step in the indicator development pipeline—to automatically link candidate indicators to published clinical trial registrations, to assist with indicator appraisal and selection. Our approach is to seek links between candidate indicators and the outcome measures reported for clinical trials. The rationale for this approach is that the evidence for the effectiveness of an indicator is likely captured in clinical trials that use the indicator as an outcome measure. To test the generalizability of the approach we looked to the CareTrack study [9], which measured the quality of care provided for 22 common health conditions, using 522 different indicators.

Methods

Linking indicators to outcome measures in the published evidence requires a mapping to be developed. We developed a simple text-processing pipeline that takes a given indicator and attempts to map it to a collection of clinical trial reports. To evaluate the accuracy of this computational method we undertook an evaluation study that consisted of four steps:

1. Creation of a test set of candidate indicators.
2. Creation of a list of candidate outcome measures from a test set of RCTs.
3. Automatically mapping indicator and outcome measures in these two sets.
4. An analysis of the outcome of mapping clinical indicators to outcome measures using the text processing method.

Creation of a list of candidate indicator measures

To create a candidate list of indicator measures, we looked to the CareTrack Australia study [9]. An expert-driven process identified 522 clinical quality indicators across the 22 conditions.

Various levels of evidence were provided for each CareTrack indicator. The majority of indicators (71%) were Level V (consensus-based). The rest were either associated with Levels I and II

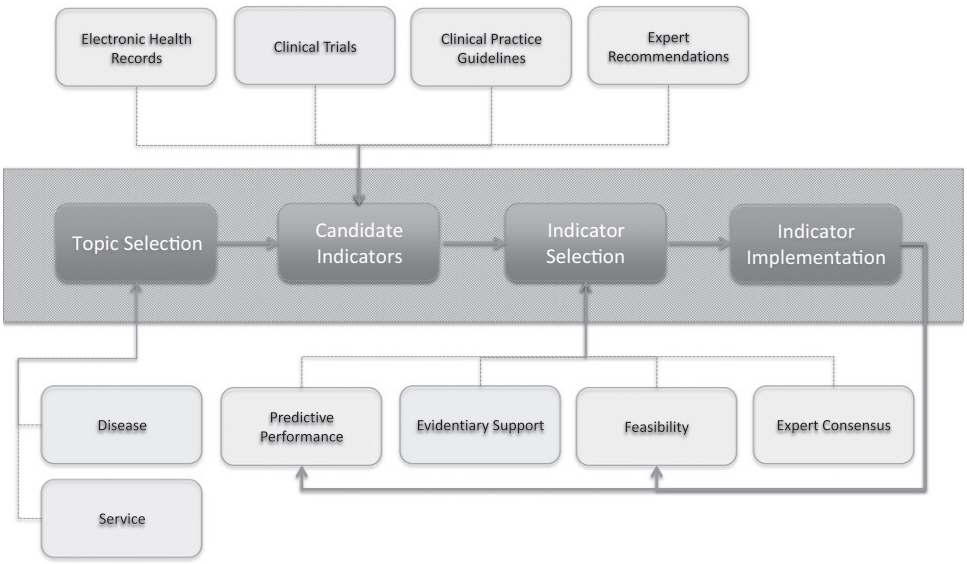

Figure 1 The indicator development pipeline. For a disease or health service process that requires monitoring, appropriate measurements or indicators are required. Such indicators can be identified by statistical analyses of electronic health records, reviews of outcome measures used in clinical trials or clinical practice guidelines, or in the absence of strong evidence, from expert recommendations. The selection of indicators from amongst these candidates is aided by evidence of the indicator's predictive performance as a measure of the process in question—taken from research or record analysis, along with technical and economic evidence about the feasibility of using the indicator in practice, and any necessary expert views. Once implemented, additional data can be gathered to update assessments of an indicator's performance and real-world feasibility.

(Level I—Systematic Reviews and Level II—Randomized Controlled Trials) evidence (13%) or Levels III and IV evidence [5] (16%).

We automatically extracted the CareTrack indicators from the list published in Appendix 1 of the main study [9]. This produced a list of noun phrases (called 'indicator phrases' hereafter). At least one indicator phrase was found for each of the 522 indicators. The method used purpose built rules that located sections of sentences in a text that were likely to contain mention of an indicator or outcome measure. Next the method filtered obviously incorrect terms, and normalized the remaining indicator phrases to a standard common format (see Appendix 1, Section 1).

Creation of a list of candidate outcome measures from RCT

To create a list of candidate outcome measures from research trials, we used ClinicalTrials.gov, a web-based clinical trial registry. The United States Food and Drug Administration Amendments Act (FDAAA 801) requires clinical studies of FDA regulated products to be prospectively registered in CliniclTrials.gov. The International Committee of Medical Journal Editors also requires prospective registration of clinical trials as a prerequisite for publication [10]. Registration includes recording the clinical trial, its methods and measured outcomes before trial commencement and reporting the results after the trial is concluded [11]. ClinicalTrials.gov specifically provides links between its registry entries and published results in research articles using a unique identifier (the NCT Number) for each study [12, 13]. To date, there are over 34 000 trials registered on ClinicalTrials.gov. Whilst ClinicalTrials.gov is not a complete list of all trials or their published results, it is large enough to act as a test resource for the present study.

We used the advanced search feature of ClinicalTrials.gov to search for the 22 conditions studied in CareTrack with a publication date before 31 December 2010, to mirror the period over which the indicators were developed. Only trials that used randomized allocations, parallel/crossover/factorial intervention design (excluding single group intervention design), and that reported outcome measures, were included. A total of 13 971 trials met these inclusion criteria. We automatically extracted the noun phrases from the outcome measures recorded for the included trials with the same text processing method used to extract indicators, to create a list of 'outcome phrases' (see Appendix 1, Section 2).

Automatically mapping indicator and outcome measures between the two sets

Indicator phrases and outcome phrases were then pooled, and placed into clusters if the Levenshtein edit distance ratio between phrases was 75% or greater [14]. All the phrases in a given cluster were considered mapped to each other.

Analysis of mapping outcomes

We used both a strict and a lenient evaluation method. Under lenient evaluation, an indicator phrase was labelled 'mapped' if any of the corresponding indicator phrases could be mapped to any outcome phrases or 'miss' otherwise. Under strict evaluation, an indicator was only labelled 'mapped' if all of its corresponding indicator phrases could be linked to outcome phrases from ClinicalTrials.gov.

To establish a benchmark for the effectiveness of the phrase extraction and mapping pipeline, it was validated against a human gold standard. One hundred randomly selected clinical indicators and 100 randomly selected RCT outcome measures were manually mapped to create the gold standard. Precision, recall and F_1 scores were calculated for the performance of the automated method against this gold standard (Table 1). The precision of the mapping between indicator and outcome phrases was high at 0.88.

Results

Indicator and outcome phrase extraction

The automated method was able to extract indicator and outcome phrases for every indicator and trial (Tables 2 and 3). For the 522 indicators, an average of 2.8 phrases were extracted per indicator (1.8 unique phrases). For the 13 971 trials from ClinicalTrials.gov, an average of 2.5 outcome phrases were extracted (0.71 unique phrases) per outcome measure.

Indicator to outcome mapping

An average of 23 outcome phrases (IQR = 9) were linked to each indicator phrase. Using the strict evaluation criterion, it was possible to link all phrases associated with an indicator to one or more outcome phrases for 157/522 (30%) of indicators. Using lenient evaluation, relaxing the mapping criterion to require only one or more mappings per indicator phrase, added an additional 196 (38%) indicators, bringing the total number of indicators with a mapping to a clinical trial to 68%.

The remaining 169 (32%) indicators could not be mapped to any outcome phrase.

There were 70 CareTrack indicators known to be associated with Level I or II evidence and 71% (50/70) of these were mapped to one or more outcome phrases in clinical trials (or 37% using the strict criterion of full mapping). A further 370 indicators were understood to be consensus-based in the original CareTrack study. Amongst these, 72% (268 of 370) could be mapped to a trial (and 33% were strictly mapped). Figure 2 shows the success in mapping CareTrack indicators by level of evidence. On average, 21 studies were mapped to an indicator with Levels I and II evidence and 24 studies linked to a consensus-based indicator (Fig. 3).

Error analysis

There were 20/70 (29%) Levels I and II evidence-based indicators that could not be mapped to any outcome phrase. Manual analysis of mapping failures revealed that seven of the failed indicators appeared to be true negatives, in that manual methods were also unable to identify any clinical trial in the trial test set which contained matching outcome measures. Four indicators (5.7%) were false negatives, i.e. mappings to outcome phrases were possible but were missed by our method. For the final nine indicators, outcome phrases existed but were located in the intervention field of the clinical trial record in ClinicalTrials.gov, instead of the outcome measure field.

Table 1 Performance of the lexical pipeline for extraction of indicator and outcome phrases and for mapping each to the other, against a gold standard validation set

	Recall	Precision	F_1 score
Indicator phrases	0.93	0.51	0.66
Outcome phrases	0.98	0.64	0.77
Mapping	0.85	0.88	0.86

Table 2 Extraction and mapping results for clinical indicators and clinical trial outcome measures in 22 health conditions

Condition	Number of indicators	Number of extracted indicator phrases	Number of trials	Number of extracted outcome phrases	Number (%) indicators linked to outcome phrases (lenient)	Number (%) indicators linked to outcome phrases (strict)
Alcohol dependence	13	38	248	2241	12 (92.3)	7 (53.8)
Antibiotic use	5	5	11	94	5 (100)	5 (100)
Asthma	28	65	947	13 591	25 (89.3)	15 (53.6)
Atrial fibrillation	18	43	333	4287	12 (66.7)	4 (22.2)
Cerebrovascular accident	35	97	588	6908	28 (80.0)	12 (34.3)
Chronic heart failure	42	109	188	2271	37 (88.1)	8 (19.1)
Chronic obstructive pulmonary disease	39	106	677	11 112	29 (74.4)	15 (38.5)
Community acquired pneumonia	33	110	51	795	7 (21.2)	2 (6.1)
Coronary artery disease	38	127	1458	19 325	36 (94.7)	20 (52.6)
Depression	19	54	1303	14 675	14 (73.7)	9 (47.4)
Diabetes	27	68	3215	42 092	25 (92.6)	13 (48.1)
Dyspepsia	22	64	63	605	5 (22.7)	1 (4.5)
Hyperlipidaemia	15	35	491	6416	10 (66.7)	3 (20.0)
Hypertension	49	98	1546	16 944	42 (85.7)	26 (53.1)
Low back pain	10	41	233	2482	5 (50.0)	0 (0)
Obesity	9	24	1326	13 301	8 (88.9)	4 (44.4)
Osteoarthritis	21	70	631	7720	17 (81.0)	6 (28.6)
Osteoporosis	10	21	397	6214	7 (70.0)	2 (20.0)
Panic disorder	14	65	51	675	4 (28.6)	0 (0)
Preventive care	31	69	22	195	6 (19.4)	0 (0)
Surgical site infection	5	7	67	637	5 (100)	4 (80.0)
Venous thromboembolism	39	139	125	2002	14 (35.9)	1 (2.6)
Total	522	1455	13 971	174 582	353 (67.6)	157 (30.1)

Table 3 Examples extracted phrases

Measure	Indicator phrases (CareTrack)	Outcome phrases (ClinicalTrials.gov)
Forced expiratory volume in 1 s (FEV1)	Expiratory volume in 1 s (FEV1)	forced expiratory volume in one second (FEV1); Forced Expiratory Volume in 1 second (FEV1); expiratory volume in 1 second (FEV1); FEV1 (Forced Expiratory Volume in 1 Second); Forced Expiratory Volume in 1 s(FEV-1); Forced Expiratory Volume in the first second (FEV1); Forced expiratory volume in one-second (FEV1); Volume in 1 sec (FEV1); expiratory volume (FEV1); Spirometry Forced Expiratory Volume in One Second (FEV1); Forced Expiratory Volume in 1 second (FEV1): spirometry
Glycosylated hemoglobin (HbA1c)	Glycated hemoglobin (HBA1c) levels	Glycosylated hemoglobin (HbA1c) levels; glycosolated haemoglobin; Glycated hemoglobin (HbA1c); HbA1c (Glycosylated hemoglobin); HbA1C: Glycated Hemoglobin; Hemoglobin (HbA1c); Glcosylated Hemoglobin (HbA1c); HbA1c Test (Glycated hemoglobin); Glycosylated Hemoglobin A1c (HbA1c)

Discussion

Indicator development is an important, but currently difficult and manual process. The indicator development process appears well suited to automation in many, potentially all, of its stages. In this paper we have focussed on exploring the feasibility of automating just one stage in the overall indicator development pipeline—the identification of candidate indicators and the research that may support their use. Indeed, this appears to be the first study we are aware of to automatically link clinical indicators with clinical trial registrations.

Only 13% of the original CareTrack indicators were identified by their source guidelines to have Level I or II evidence behind them and 16% had Level III or IV evidence. The remaining 71% were labelled as being supported by expert consensus. Our methods

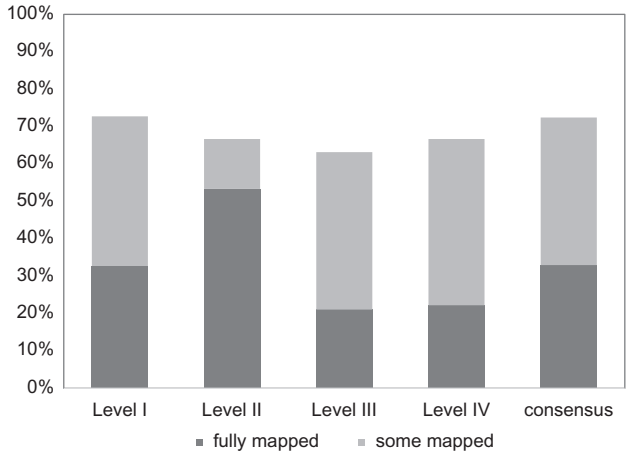

Figure 2 Percentage of indicator phrases mapped to clinical trials, by level of available evidence as understood by the indicator developers.

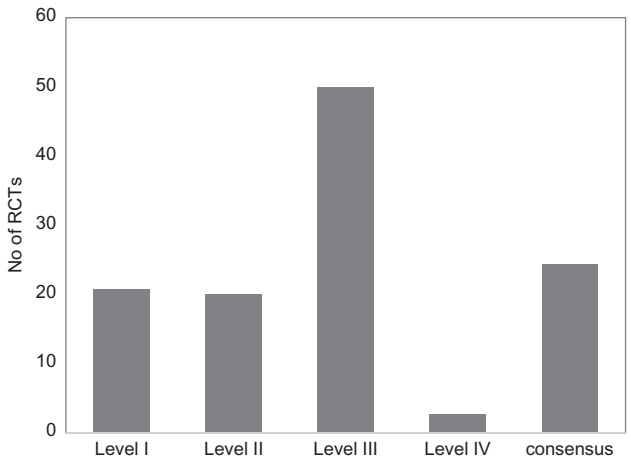

Figure 3 Average number of clinical trials found per indicator, by level of available evidence as assessed by the indicator developers.

nevertheless identified candidate clinical trials (Level II evidence) for 72% of the consensus indicators, 67% of the Level IV and 63% of Level III indicators. The CareTrack process relied on clinical guidelines to identify the level of evidence associated with a given indicator. This suggests that, despite best practice manual methods, important research evidence may be missed, either when developing guidelines, or relying on guidelines. Whilst this study did not further examine the quality of those studies, they do meet the criteria for Level I or II evidence in their clincialtrial.gov record. This is promising, as it suggests that automated methods will be able to provide a more rigorous and much more efficient approach to mapping indicators to evidence.

Our approach used simple, standard text processing methods to extract and normalize candidate phrases in documents containing mentions of indicators or clinical trial measures. Despite the simplicity of our approach in this feasibility study, initial validation of the methods against a gold standard showed the method performed with a precision of 0.88 and F_1 score of 0.86. When applied to the 70 CareTrack indicators known to have Level I or II evidence, some link to clinical trial registrations was possible for 73% of Level I and 67% of Level II indicators.

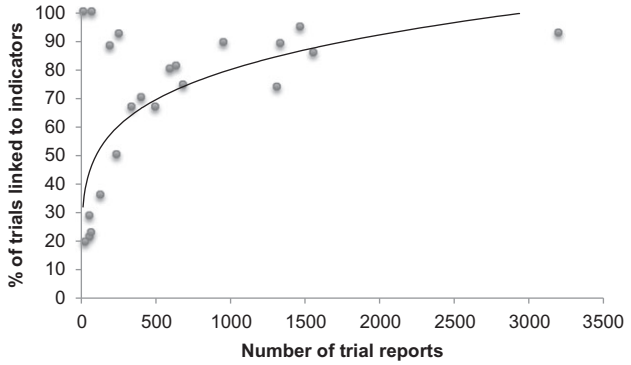

Figure 4 Scatter plot showing the association between success in finding indicators for the 22 CareTrack conditions and the sample size of RCTs available for that condition in the test corpus. A logarithmic regression function is fitted to the data.

To understand these results, it should be noted that there are two major reasons for failure to find a mapping. The first is that no mapping exists because there is no available trial registration to support the indicator. The second is that a mapping does exist but the mapping method fails to find it. Detailed analysis of the performance for Levels I and II indicators showed that for 10% no mapping was possible within the given set of trials, and for another 6% the method failed. In another 13% mapping was possible but the trial report itself was the problem, incorrectly reporting outcome measures in the wrong field.

Interestingly, the indicators that had poor mapping appeared to have a lower number of trials in the ClinicalTrials.gov test set compared to other indicators, suggesting that sample size may have contributed to poorer performance using the current methods. Figure 4 shows that there is an association between mapping success and the number of trials available to be mapped.

Performance is likely to improve with more robust methods. For example, rather than mapping raw phrases using an edit distance, the phrases could be semantically labelled using standard tools such as MMTX which comes with the UMLS metathesaurus [15].

For the overall indicator development pipeline to be supported, additional work is required at each of the stages identified in Fig. 1. For example, text-mining tools can be used to extract indicators not just from trial registrations, but also from randomized clinical trial reports, systematic reviews and clinical practice guidelines.

We also did not examine the results of the clinical trials that mapped to indicators. We did not undertake an assessment of whether the mapped trials provide evidence for the use of an indicator in a health settings or whether these indicators are economical and effective to apply. Such considerations are important further stages in the indicator development process, and different methods would be needed to support them.

Not all published research associated with a trial is directly linked to the ClinicalTrials.gov registration, and in one study 44% of registrations with no linked publications were found to have published articles after a manual search [16]. While a lack of linkage between trial registrations and the literature did not seem to impede the identification of outcome measures that could serve as indicators, it will probably be necessary to search for these reports when assessing indicator suitability. Finally, little work has been done to utilize the data stored in electronic health records, which can also be used to identify candidate indicators based on their ability to predict specific clinical conditions or events.

Limitations

This study only examined RCTs from one source (ClinicalTrials.gov). A broader range of clinical trial repositories exists and their use would likely identify additional trials relevant to indicator development. Equally we did not search for reports associated with Levels III and IV evidence, which may also be of value during the indicator selection process.

Conclusion

We have presented a method to automatically identify clinical trial reports that may be relevant to the selection of clinical indicators. Whilst the methods used are simple, they appear to identify trials missed by the developers of clinical indicators, and so should prove to be beneficial in the indicator development process.

Funding

This research is supported by National Health and Medical Research Council Program Grant APP1054146. George Karystianis (GK) assisted in the validation of the mapping procedure.

References

1. Crampton P, Perera R, Crengle S *et al*. What makes a good performance indicator? Devising primary care performance indicators for New Zealand. *N Z Med J* 2004;**117**:1–12.
2. Mainz J. Defining and classifying clinical indicators for quality improvement. *Int J Qual Health Care* 2003;**15**:523–30.
3. Stelfox HT, Straus SE. Measuring quality of care: considering conceptual approaches to quality indicator development and evaluation. *J Clin Epidemiol* 2013;**66**:1328–37. doi:https://doi.org/10.1016/j.jclinepi.2013.05.017.
4. Kötter T, Blozik E, Scherer M. Methods for the guideline-based development of quality indicators—a systematic review. *Implement Sci* 2012;**7**:21.
5. Coleman K, Norris S, Weston A, *et al*. NHMRC additional level of evidence and grades for recommendations for developers of guidelines. In: *National Health and Medical Research Council*, ed., Pilot program 2005–2007.
6. Choong MK, Coiera E, Tsafnat G *et al*. Linking clinical quality indicators to research evidence—a case study in asthma management for children. *BMC Health Serv Res* 2017. (in press).
7. Tsafnat G, Dunn A, Glasziou P *et al*. The automation of systematic reviews. *BMJ* 2013;**346**:f139.
8. Tsafnat G, Glasziou P, Choong MK *et al*. Systematic review automation technologies. *Syst Rev* 2014;**3**:74.
9. Runciman WB, Hunt TD, Hannaford NA *et al*. CareTrack: assessing the appropriateness of health care delivery in Australia. *Med J Aust* 2012;**197**:549.
10. De Angelis C, Drazen JM, Frizelle FA *et al*. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med* 2004;**351**:1250–51.
11. Ross JS, Tse T, Zarin DA *et al*. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis 2012.
12. Zarin DA, Tse T. Sharing individual participant data (IPD) within the context of the trial reporting system (TRS). *PLoS Med* 2016;**13**:e1001946. doi:10.1371/journal.pmed.1001946.
13. Zarin DA, Tse T. Moving toward transparency of clinical trials. *Science* 2008;**319**:1340–42.
14. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*; 1966.
15. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001;17–21.
16. Huser V, Cimino JJ. Precision and negative predictive value of links between ClinicalTrials.gov and PubMed. *AMIA Annu Symp Proc* 2012;**2012**:400–08.
17. Choong MK, Tsafnat G, Hibbert P *et al*. Comparing clinical quality indicators for asthma management in children with outcome measures used in randomised controlled trials: a protocol. *BMJ Open* 2015;**5**. doi:10.1136/bmjopen-2015-008819.

Appendix 1 Lexical pipeline

Section 1: Automatic extraction of indicator phrases

Regular expressions were used to identify the measurement segment of sentences. These segments, were those that were likely to contain a mention of an indicator measure. An indicator segment in a sentence, was that portion that appeared after the regular expressions: 'received', 'were', 'had', 'have', 'are', 'was prescribed'. The CandC parser [1] then extracted noun phrases from each measurement segment, using a grammar taken from CCGbank.

Noun phrases which were unlikely to be an indicator phrases were then filter using the following regular expressions, which were heuristically developed for an earlier study [17]: 'additionally', 'aim', 'all children', 'amount of', 'any', 'a randomized', 'at least', 'baseline', 'based', 'change(s)', 'children', 'clinical(s)', 'cm', 'combined negative outcome', 'conclusion', 'control group', 'control(s)', 'cost', 'day(s)', 'difference(s)', 'discussion', 'double-blind', 'double-dummy study', 'end of', 'evening', 'fall', 'first', 'good', 'gender', 'groups', 'high dose', '(-)hrs', 'h', 'hour(s)', 'intervention(s)', 'intervention clinicians', 'intervention components', 'it', 'l', 'lastmonths', 'low dose', 'mean', 'methods', 'min', 'minute(s)', '-monthfollow-up', 'mth(s)', 'month(s)', 'more clinicians', 'moreover', 'morning', 'nextminutes', 'no significant difference', 'number', 'objective', 'objective data', 'one', 'occurrence', 'other', 'outcomes', 'outcomes of', 'participants', 'patient(s)', 'people', 'percent', 'percentage', 'placebo', 'placebo-controlled', 'placebo-controlled study', 'placebo treatment', 'primary outcomes', 'primary efficacy variable', 'question(s)', 'randomization', 'randomization', 'randomized', 'randomized clinicians', 'readiness', 'results', 'same period', 'second', 'secondary outcomes', 'severity', 'side', 'similar changes', 'single-blind', 'study', 'standard protocols', 'study end', 'study entry', 'subjects', 'that', 'them', 'then', 'they', 'this', 'time', 'treatment', 'treatment groups', 'units', '- units', '-unitchange', 'use', 'web-based', 'wk(s)', 'week (s)', 'who', 'whole study duration', 'women', '(-)yr(s)', '-year(s)', '-yearstudy'

Section 2: Automatic extraction of outcome phrases

We again used the CandC parser to extract noun phrases from the outcome measures recorded for the included trials from ClinicalTrials.gov. The same regular expression filtering used in the extraction of indicator phrases as above were used to filter the outcome phrases.

Section 3: Natural language processing pipeline

Once noun phrases for indicators and outcomes were identified, they were transformed into a common or normalized form using the following sequence of steps:

1. Tokenization of individual noun phrases: Tokenisation is a process to break up string into words and punctuations [2].
2. Removing punctuation
3. Case normalisation: A process to convert all letters to lowercase letters.
4. Stop-word removal: A process of removing common words in English. The list of stop words is: 'a', 'about', 'above', 'after', 'again', 'against', 'all', 'am', 'an', 'and', 'any', 'are', 'as', 'at', 'be', 'because', 'been', 'before', 'being', 'below', 'between', 'both', 'but', 'by', 'can', 'did', 'do', 'does', 'doing', 'don', 'down', 'during', 'each', 'few', 'for', 'from', 'further', 'had', 'has', 'have', 'having', 'here', 'how','if', 'in', 'into', 'is', 'just', 'more', 'most', 'no', 'nor', 'not', 'now', 'of', 'off', 'on', 'once', 'only', 'or', 'other', 'out', 'over', 'own', 's', 'same', 'should', 'so', 'some', 'such', 't', 'than', 'that', 'the', 'then', 'there', 'these', 'this', 'those', 'through', 'to', 'too', 'under', 'until', 'up', 'very', 'was', 'were', 'what', 'when', 'where', 'which', 'while', 'who', 'whom', 'why', 'will', 'with'.
5. Lemmatization using WordNet lemmatizer: Lemmatization is a process of removing inflectional endings and return the base or dictionary form of a word using a vocabulary and morphological analysis of words [3].

WordNet is a semantically oriented dictionary of English, with 155 287 words and 117 659 synonym sets [2].

6. Acronym expansion from dictionary as below:

aaa: abdominal aortic aneurysm
abpa: aspergillis
acl: anterior cruciate ligament
add: attention deficit disorder
adhd: attention deficit–hyperactivity disorder
af: atrial fibrillation
age: a/biotic related
all: acute lymphocytic leukaemia
ami: acute myocardial infarction
aml: acute myelocytic leukaemia
armd: age related macular degeneration
asd: atrio-septal defect
axr: abdominal x-ray
bcc: basal cell carcinoma
bpad: bipolar affective disorder
bph: benign prostatic hyperplasia
bppv: benign paroxysmal positional vertigo
btb: break through bleeding
cabg: coronary artery bypass graft
cal: chronic airways limitation
capd: continuous ambulatory peritoneal dialysis
cea: carcino embryonic antigen
cf: cystic fibrosis
cfs: chronic fatigue syndrome
ckd: chronic kidney disease
cll: chronic lymphocytic leukaemia
cma: comprehensive medical assessment
cml: chronic myelocytic leukaemia
coad: chronic obstructive airways disease
copd: chronic obstructive pulmonary disease
cp: cerebral palsy
crps: complex regional pain syndrome
csom: chronic suppurative otitis media
cva: cerebrovascular accident
cvi: cerebrovascular insufficiency
cxr: chest x-ray
dcm: dilatation cardiomyopathy
di: diabetes insipidus
dish: diffuse idiopathic skeletal hyperostosis
dka: diabetic ketoacidosis
dmmr: domicillary medication management review
dna: did not arrive
dnw: did not wait
dub: dysfunctional uterine bleeding
ear: ext
eswl: external shock wave lithotripsy
eua: examination under anaesthesia
fb: fish bone
fdiu: fetal death in utero
fess: functional endoscopic sinus surgery
fms: fibromyalgia syndrome
fnab: fine needle aspiration biopsy
fobt: faecal occult blood test
fta: failed to attend
ftt: failure to thrive
g6pdd: glucose-6-phosphate dehydrogenase deficiency
gad: generalised anxiety disorder
ggt: gamma glutamyl transpeptidase
gih: gastrointestinal haemorrhage
gor: gastro-oesophageal reflux
gord: gastro-oesophageal reflux disease
gu: gastric ulcer
hiaa: hydroxy indole acetic acid
hnpcc: hereditary non-polyposis colon cancer
hpl: human placental lactogen
hpv: human papilloma virus
hrt: hormone replacement therapy
hsil: high grade squamous intraepithelial lesion
ht: hypertension
ibc: iron binding capacity
ibs: irritable bowel syndrome
icsi: intracytoplasmic sperm injection
iddm: insulin dependent diabetes mellitus
idk: internal derangement of knee
iec: intraepidermal carcinoma
igt: inpaired glucose tolerance
igtn: ingrown toenail
ihd: ischaemic heart disease
im: infectious mononucleosis
itp: idiopathic thrombocytopaenic purpura
iucd: intrauterine contraceptive device
iud: intrauterine device
iufd: intrauterine fetal death
iugr: intrauterine growth retardation
ivf: in-vitro-fertilisation
ivp: intravenous pyelogram
jra: juvenile rheumatoid arthritis
loc: loss of consciousness
low: loss of weight
lrti: lower respiratory tract infection
lsil: low grade squamous intraepithelial lesion
luscs: lower uterine segment caesarean section
lvf: left ventricular failure
map: morning after pill
mba: motorbike accident
mca: motor car accident
mps: mobility parking
ms: multiple sclerosis
mua: manipulation under anaesthesia
mva: motor vehicle accident
nash: non alcoholic steato hepatitis
ndss: national diabetes services scheme
niddm: non insulin dependent diabetes mellitus
nstemi: non-st-elevation myocardial infarction
nsu: non specific urethritis
oa: osteoarthritis
ocd: obsessive compulsive disorder
ocp: oral contraceptive pill
odd: oppositional defiant disorder
opd: out patient dept.
pap: papanicolaou smear
pat: paroxysmal atrial tachycardia
pcos: polycystic ovarian syndrome
pda: patent ductus arteriosus
pet: pre eclamptic toxaemia
pfo: patent foramen ovale
pms: premenstrual syndrome
pmt: premenstrual tension
pnd: paroxysmal nocturnal dyspnoea
pph: postpartum haemorrhage
psvt: paroxysmal supra ventricular tachycardia
ptsd: post traumatic stress disorder
ra: rheumatoid arthritis
rast: radio-allergo-sorbent test
rcc: red cell count
rhf: right heart failure
rmmr: residential medication management review
ros: removal of sutures
rp: retinitis pigmentosa
rpoc: retained products of conception
rrv: ross river virus

rsd: reflex sympathetic dystrophy
rta: renal tubular acidosis
rti: respiratory tract infection
rvf: right ventricular failure
sbe: subacute bacterial endocarditis
scc: squamous cell carcinoma
se: sedation
siadh: syndrome inappropriate adh secretion
sle: systemic lupus erythematosus
stemi: st-elevation myocardial infarction
sti: sexually transmitted infection
tb: tuberculosis
tca: team care arrangement
tdr: treating doctors report
thr: total hip replacement
tia: transient ischaemic attack
tkr: total knee reconstruction
tvt: tensionless vaginal tape
uhcg: urine hcg
uti: urinary tract infection
vre: vancomycin resistant enterococcal infection
vsd: ventricular septal defect
vt: ventricular tachycardia
wcc: white cell count
wpw: Wolff Parkinson white syndrome

References

1. Linguistically motivated large-scale NLP with C&C and Boxer. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions; 2007*. Association for Computational Linguistics.
2. Bird S, Klein E, Loper E. *Natural Language Processing With Python*. 'O'Reilly Media, Inc.', 2009.
3. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge: Cambridge university press, 2008.