

Comparative effectiveness research designs: an analysis of terms and coverage in Medical Subject Headings (MeSH) and Emtree*†

Tanja Bekhuis, PhD, MS, MLIS; Dina Demner-Fushman, MD, PhD; Rebecca Crowley, MD, MS

See end of article for authors' affiliations.

DOI: <http://dx.doi.org/10.3163/1536-5050.101.2.004>

Objectives: We analyzed the extent to which comparative effectiveness research (CER) organizations share terms for designs, analyzed coverage of CER designs in Medical Subject Headings (MeSH) and Emtree, and explored whether scientists use CER design terms.

Methods: We developed local terminologies (LTs) and a CER design terminology by extracting terms in documents from five organizations. We defined coverage as the distribution over match type in MeSH and Emtree. We created a crosswalk by recording terms to which design terms mapped in both controlled vocabularies. We analyzed the hits for queries restricted to titles and abstracts to explore scientists' language.

Results: Pairwise LT overlap ranged from 22.64% (12/53) to 75.61% (31/41). The CER design terminology (n=78 terms) consisted of terms for primary study designs and a few terms useful for evaluating evidence, such as *opinion paper* and *systematic review*. Patterns of coverage were similar in MeSH and Emtree ($\gamma=0.581$, $P=0.002$).

Conclusions: Stakeholder terminologies vary, and terms are inconsistently covered in MeSH and Emtree. The CER design terminology and crosswalk may be useful for expert searchers. For partially mapped terms, queries could consist of free text for modifiers such as *nonrandomized* or *interrupted* added to broad or related controlled terms.


INTRODUCTION

The emergent field of comparative effectiveness research (CER)‡ is beset by differences in language among stakeholders. These include methodologists in organizations that promote CER, scientists who generate original data or synthesize secondary data, panels of experts who rely on extant research to design guidelines for best practice, and policymakers who identify and prioritize future research needs. For health sciences librarians who regularly support this panoply of stakeholders, it is necessary to know about differences in order to interpret service requests. For example, the following terms are used inconsistently: CER, evidence-based medicine (EBM), and health technology assessment (HTA); randomization and random sampling; efficacy and effectiveness.

Recently, the *MLA News* published two accessible reports to introduce librarians to CER in which the

* This study was partially supported by grants awarded to Tanja Bekhuis from the National Institutes of Health, National Library of Medicine, grant no. 1K99LM010943-01A1 and grant no. 4R00LM-010943-02.

† A presentation describing this study was given to the American Medical Informatics Association (AMIA) 2012 Annual Symposium, November 3–7, Chicago, IL.

 Supplemental Table 5 is available with the online version of this journal.

‡ The Agency for Healthcare Research and Quality (AHRQ) defines comparative effectiveness research as: "Comparative effectiveness research is designed to inform health-care decisions by providing evidence on the effectiveness, benefits, and harms of different treatment options. The evidence is generated from research studies that compare drugs, medical devices, tests, surgeries, or ways to deliver health care" [1].

Highlights

- Local terminologies for study designs vary across organizations that promote comparative effectiveness research (CER).
- Coverage of design terms common to CER is similar in Medical Subject Headings (MeSH) and Emtree. Most negated or detailed terms partially map or fail to map to controlled terms.
- A crosswalk (online only) between MeSH and Emtree provides suggestions for alternative terms and query expansions.

Implications

- Librarians and trials search coordinators who support various CER stakeholder communities can consult the crosswalk for ideas when developing the design block for their search filters.
- Because scientists appear to split up concepts in detailed design phrases used for classification by methodologists, and modifiers and basic design names can occur in different places in their titles and abstracts, librarians may need to write queries that use free text for pieces of a design phrase along with controlled terms, if they exist.
- The crosswalk developed in this study could also be useful for expert searchers when designing filters requiring terms for designs in domains other than CER.

authors compare CER and EBM [2, 3]. A more thorough essay comparing CER, EBM, and HTA along several dimensions appears in *The Milbank Quarterly* [4], with some discussion of semantic differences between North America and Europe. The authors of another paper discussing infrastructure needs and capacity for conducting CER report that while capacity is adequate, the "majority of researchers are trained in either observational study methods or randomized trials, but rarely both" [5]. Thus, a lack of awareness of major approaches to research likely exacerbates the confusion in language. Note that in this paper, we use the term *language* to mean natural as opposed to formal language, with a focus on the use of phrases to communicate concepts for study designs. Jurafsky and Martin's text explains the distinction [6]. Several authors provide background papers on the structure of scientific language, sub-languages, and epistemological differences among disciplines [7–9].

An important aspect of CER is the focus on the generalizability of findings to diverse populations of real interest. Broadly, CER is concerned with answering questions regarding effectiveness rather than efficacy of interventions, which has implications for the usefulness of various study designs. Nonrandomized (NR) or observational studies, rather than randomized controlled trials (RCTs), may better answer effectiveness questions, even though well-known threats to validity exist for the former [10]. For example, consider that a well-conducted RCT ensures the statistical equivalence of groups via randomization (random assignment of treatments to experimental units or vice versa) prior to treatment and that finding a treatment effect is therefore likely to be reproducible under the same experimental conditions. However, the design of an RCT promotes internal validity at the expense of external validity (generalizability) when the investigators cannot randomly sample "units," such as patients. In contrast, researchers who conduct an NR study might randomly sample participants from populations of interest. *Random sampling*, if done well, as opposed to *random assignment* ensures that study groups will resemble the populations of interest. This is a major reason for recognizing the value of evidence derived from NR studies. In the best of worlds, a CER question would be answered by both RCTs and NR studies. This is why systematic reviewers who synthesize biomedical evidence look for both kinds of studies.

Unfortunately, consensus does not exist regarding how best to describe NR studies common to CER. According to the Cochrane Non-Randomised Studies Methods Group, both investigators and indexers inconsistently describe study designs [11]. Challenges arise for expert searchers, indexers, and methodologists due to the hodgepodge of terms that stakeholders use within and across disciplines. This problem is well known, and groups around the world have issued statements regarding standards for reporting studies and their designs. To improve the value of medical research, an international initiative known

as the EQUATOR Network [12] maintains a library of reporting guidelines by study type, such as STANDards for the Reporting of Diagnostic Accuracy Studies (STARD) for diagnostic accuracy studies [13], Consolidated Criteria for Reporting Qualitative Research (COREQ) [14], and STrengthening the Reporting of OBServational studies in Epidemiology (STROBE) [15]. In general, guidelines suggest that authors name their study design in the title or abstract and use a common term, but names are not standardized. Thus, inconsistent indexing and varying stakeholder language, as well as multiple reporting standards lead to serious retrieval challenges for health sciences librarians.

In this study, we investigated whether methodologists in several highly regarded CER organizations share a terminology for study designs and to what extent. By *terminology*, we mean a set of mostly phrases, which is consistent with International Organization for Standardization (ISO) 1087 "Terminology–Vocabulary Standard," described by Hammond and Cimino [16]. To compare organizational terminologies, we culled design terms and terms for related concepts from relevant documents. We then built a CER design terminology based on the documents we identified to evaluate whether and how terms for study designs used by experts correspond to terms in Medical Subject Headings (MeSH) [17] and Emtree [18], the controlled vocabularies for MEDLINE and Embase, respectively. To support librarians, we developed a crosswalk between MeSH and Emtree with suggestions for queries when design terms partially map to broad controlled terms or fail to map. We also explored whether scientists use CER design terms to describe their own studies.

METHODS

Data sources

To ensure relevancy, we elected to work with classification algorithms from respected CER organizations. Further, to ensure validity, we selected algorithms already vetted by methodologists. We therefore chose algorithms developed by organizations identified in two recent methods studies funded by the Agency for Healthcare Research and Quality (AHRQ) [19, 20]. The organizations and data sources included:

1. University of Alberta Evidence-based Practice Center (Alberta). Developing and Testing a Tool for the Classification of Study Designs in Systematic Reviews of Interventions and Exposures: Appendix G. Round Two Algorithm and Glossary [19].
2. AHRQ. Framework for Considering Study Designs for Future Research Needs: Table 1: Study design terms; Appendix A: Taxonomy for Study Designs; Table A-2: Terms Associated with Study Designs [20]. This draft report was submitted for public comment September 2011.
3. Cochrane Non-Randomised Studies Methods Group (Cochrane). Design Algorithm for Studies of Health Care Interventions; appears in Appendix H of

the Alberta report [19]. Additionally, Box 13.1.a: Some Types of NRS [non-randomized studies] Design Used for Evaluating the Effects of Interventions; Table 13.2.a: List of Study Design Features (studies with allocation to interventions at the individual level); Table 13.2.b: List of Study Design Features (studies with allocation to interventions at the group level) [11].

4. Academy of Nutrition and Dietetics, formerly known as the American Dietetic Association (ADA). Evidence Analysis Manual: Appendix 5: Algorithm for Classifying Research; Appendix 6: Glossary of Terms Related to Research Design [21].

5. Research Triangle Institute International-University of North Carolina Evidence-based Practice Center (RTI). Design Algorithm for Studies of Health Care Interventions; appears in the Alberta report, Appendix H [19].

We extracted terms from the selected resources and augmented subsequent lists with designs mentioned in corresponding glossaries, tables, and appendixes. We refer to the resultant lists of terms as local terminologies (LTs) throughout this paper.

In developing the Alberta algorithm, a steering committee with members from AHRQ and AHRQ-funded evidence-based practice centers (EPCs) identified thirty-one organizations and experts. They asked respondents to return classification tools or systems to "ensure that we have a broad spectrum" (Letter of Request to Identify Study Design Classification Tools, Appendix C; see also Appendix B for a list of contacts [19]). The identified organizations included all fifteen of the AHRQ-funded EPCs and seven other organizations, such as the Cochrane Collaboration, the Campbell Collaboration, and the National Health Service (NHS) of the United Kingdom. Additionally, nine anonymous experts were contacted. Eleven respondents returned twenty-three tools, algorithms, guidelines, or instruments for classification. Ten were selected for further analysis. Members of the steering committee independently rated the selected tools and identified the Cochrane algorithm as most suitable for further development. The ADA and RTI algorithms were rated second and third, respectively.

The Alberta algorithm is the basis for a framework currently being developed by AHRQ to promote a standard taxonomy for considering the suitability of various designs in carrying out future research. Data sources in the AHRQ report include designs that vary somewhat orthographically from the Alberta report, along with additional terms (e.g., *systematic review*, *modeling*, and *meta-analysis of individual participant data*).

Development of a new comparative effectiveness research (CER) design terminology

We manually extracted terms from the data sources just described. For example, if the source was a classification algorithm structured as a decision tree, we extracted the design term and any examples or

synonyms displayed at the end of each path. If the source was a glossary, we extracted each term along with any examples mentioned in its description. If the source was a table, we extracted terms in the cells or, if applicable, from the footer defining design acronyms used as column names.

We pooled terms from all five LTs, deleted duplicates, and converted to lower case. We treated as equivalent orthographic variations—such as *randomized* (US spelling) and *randomised* (British spelling); *meta-analysis*, *metaanalysis*, and *meta analysis*; and *before-after* and *before-and-after*. Similarly, we considered as equivalent singular and plural words, such as *study* and *studies*, and acronyms for corresponding terms, such as *RCT* for *randomized controlled trial* and *IPD* for *individual patient data*.

After term extraction, augmentation, and processing in the manner described, the union of terms occurring in one or more LTs defined the new CER design terminology. Additionally, the intersection of terms occurring in all five LTs defined the core set of design terms.

Match types and coverage

To evaluate coverage, we searched for CER design terms in MeSH and Emtree. We recorded the type of match per term as *exact*, *partial*, or *no match*; coverage was defined by the distribution over match type. In MeSH, if a CER term or any of its variants directly mapped to a main heading or entry term, the match was exact; if part of the term mapped to a broader or related term or to a substring in a scope note, the match was partial; otherwise, "no match" was recorded. Mapping procedures in Emtree were modified somewhat but were quite similar to those in MeSH.

Search for CER design terms in Medical Subject Headings (MeSH) and Emtree

We used the US National Library of Medicine (NLM) MeSH browser [22] to search for terms in a stepwise manner. In each step, we used a different combination of browser settings but otherwise followed the same search strategy:

- a. We selected the settings "All of the Above" (Main Headings, Qualifiers, and Supplementary Concepts) plus "Find Exact Term." If no hit was returned, we reduced the design phrase by a word or stemmed by shortening words to a base form (e.g., *nonrandomized controlled trial* became *nonrandomized controlled*, *nonrandomized*, *nonrandom*, etc.).
- b. If no hit was returned in step (a), we selected "Find Terms with ALL Fragments" and searched for strings of text as in the first step.
- c. If no hit was returned in step (b), we selected "Search as text words in Annotation & Scope Note" plus "Find Terms with ALL Fragments" and again searched for a similar sequence of text strings.

We also searched Emtree in Embase [18], a subscription database. We navigated to the "Find

Term" tab and modified terms in a stepwise manner as in MeSH, first searching for exact matches. On occasion, searching for *study* or *trial* was helpful, as it led to a variant form that we considered equivalent (e.g., *validity study* does not directly map to *validation study*, but appears under *study*).

Crosswalk between MeSH and Emtree

We created a crosswalk between MeSH and Emtree for CER design terms by recording the controlled terms to which they exactly or partially mapped. For partial and no matches, we recorded whether terms were negated or detailed, if appropriate. A negated term includes at least one word or phrase that is counter to or is in opposition to an affirmed word or phrase in another design term. For example, *interrupted time series without comparison group* is a negated term because it is counter to an *interrupted time series with comparison group*. Specifically, *without comparison group* negates *with comparison group*. Both design terms are detailed because they are multiword phrases with several modifiers, including *interrupted*, *time*, and *comparison*.

In the crosswalk, we offered suggestions regarding potential alternatives or query expansions for some design terms.

The language of scientists and CER experts

To explore whether scientists use the terms for designs and related concepts as expressed by experts in CER organizations, we used quoted strings of terms and variants and restricted our searches to titles and abstracts. Because Embase regularly adds MEDLINE records [23], we could search records from both databases via Embase, which ensured comparability of searches.

To count the number of hits per CER term by database, we compared hits from two searches. In the first, we searched de-duplicated records originating in either database using *<design term>:ab,ti*. In the second, we restricted the search to records from Embase using *<design term>:ab,ti NOT [medline]/lim AND [embase]/lim*. To find the number of hits in MEDLINE, we subtracted the count for the second search from the first. Here is a sample query:

```
'before-after study':ab,ti OR 'before-after studies':ab,ti OR
'before-after design':ab,ti OR 'before-after designs':ab,ti
OR 'before-after trial':ab,ti OR 'before-after trials':ab,ti OR
'before-and-after study':ab,ti OR 'before-and-after studies':
ab,ti OR 'before-and-after design':ab,ti OR 'before-and-after
designs':ab,ti OR 'before-and-after trial':ab,ti OR 'before-
and-after trials':ab,ti NOT [medline]/lim AND [embase]/
lim
```

Statistical analyses

We used Excel 2003 and 2010 [24, 25], as well as IBM SPSS version 20 [26], for statistical analyses of term distributions, computation of pairwise LT overlap

and overlap with the CER design terminology, evaluation of coverage, and comparison of hits for queries. By *overlap*, we mean the percentage of shared terms between LTs or between an LT and the CER design terminology.

RESULTS

Terminologies

The augmented LTs varied in length: Alberta (n=33 terms), AHRQ (n=39), Cochrane (n=32), ADA (n=36), and RTI (n=25). The CER design terminology (n=78) derived from terms that occurred in 1 or more LTs mostly consisted of terms for primary study designs and a few terms useful for evaluating evidence, such as *opinion paper* and *systematic review* (Table 1). About half the terms (47.44%, 37/78) appeared in just 1 LT. A few terms (8.97%, 7/78) were common to all LTs (Figure 1). These included *before-after study*, *case-control study*, *case series*, *cross-sectional study*, *prospective cohort study*, *retrospective cohort study*, and *randomized controlled trial*.

Alberta had the most in common with the other terminologies (mean pairwise overlap=48.77%, 24 shared terms on average); RTI had the least in common (25.65%, 12 shared terms on average) (Table 2). The overlap between pairs of LTs ranged from 22.64% (12 shared terms) for AHRQ and RTI to 75.61% (31 shared terms) for Alberta and AHRQ (Table 2). The overlap of LTs with the new CER design terminology ranged from 32.05% (25/78) for RTI to 50.00% (39/78) for AHRQ (Table 3).

Coverage and characteristics of match type

Patterns of coverage in MeSH and Emtree are displayed in Table 4 and Figure 2. Coverage as defined by the distribution over match type was similar; the association was positive and statistically significant (Goodman Kruskal gamma=0.581, P=0.002). Gamma is a nonparametric measure suitable for testing the bivariate association between ordinal variables. It can be interpreted as a correlation coefficient, as it falls between -1 and +1.

Match type per term is displayed in Table 5 (online only). The terms to which CER design terms most often mapped were similar in both vocabularies. In MeSH, they were *randomized controlled trial*, *controlled clinical trial*, *longitudinal studies*, *cohort studies*, and *clinical trial*. In Emtree, they were *randomized controlled trial*, *controlled study*, *time series analysis*, and *cohort analysis*.

Frequent partial mapping indicated a broad or related MeSH or Emtree term. For example, the following CER terms mapped to the MeSH term *randomized controlled trial*: *cluster randomized controlled trial*, *cluster randomized trial*, *group randomized trial*, *open-label randomized controlled trial*, *randomized trial*, and *single-blinded randomized controlled trial*. In Emtree, the terms were the same with the exception of *open-*

Table 1
Comparative effectiveness research (CER) design terminology

Term	Frequency of appearance in local terminologies (LTs)	Term	Frequency of appearance in LTs
adaptive design	(1)	nested case-control study	(3)
analytic study	(3)	n-of-one trial	(2)
before-after study	(5)	noncomparative study	(4)
case report	(3)	nonconcurrent cohort study	(4)
case series	(5)	non-controlled trial	(2)
case study	(1)	non-experimental study	(1)
case-control study	(5)	nonrandomized comparative trial	(1)
cluster nonrandomized controlled trial	(1)	nonrandomized controlled trial	(3)
cluster quasi-randomized controlled trial	(1)	nonrandomized crossover trial	(1)
cluster randomized controlled trial	(3)	nonrandomized trial	(4)
cluster randomized trial	(2)	observational study	(4)
cohort before-and-after study	(1)	open-label randomized controlled trial	(1)
cohort study	(3)	opinion paper	(1)
community trial	(2)	parallel study	(2)
controlled before-after study	(3)	pragmatic trial	(1)
controlled cohort before-and-after study	(1)	pre-post study	(1)
controlled interrupted time series	(1)	prospective case series	(1)
controlled trial	(1)	prospective cohort study	(5)
correlational study	(1)	quasi-experimental study	(2)
crossover study	(3)	quasi-randomized controlled trial	(2)
cross-sectional study	(5)	quasi-randomized trial	(2)
data base study	(1)	randomized clinical trial	(1)
descriptive study	(4)	randomized controlled trial	(5)
diagnostic study	(1)	randomized crossover trial	(1)
double blinded randomized controlled trial	(1)	randomized trial	(4)
ecological cross-sectional study	(1)	reliability study	(1)
epidemiological study	(1)	retrospective case control study	(1)
experimental study	(3)	retrospective cohort study	(5)
factorial study	(2)	retrospective study	(2)
focus group	(1)	simulation modeling	(1)
group randomized trial	(3)	single-blinded randomized controlled trial	(1)
head-to-head study	(1)	solomon four-group study	(2)
historically controlled trial	(1)	stepped wedge study	(2)
interrupted time series study	(1)	systematic review	(2)
interrupted time series with comparison group	(2)	time series	(4)
interrupted time series without comparison group	(2)	time study	(1)
meta-analysis	(2)	trend study	(2)
meta-analysis of individual participant data	(1)	uncontrolled longitudinal study	(1)
modeling	(1)	validity study	(1)

label randomized controlled trial, which mapped to open study.

We labeled CER design terms as detailed relative to MeSH and Emtree if they consisted of more than 3 words, ignoring prepositions and hyphens (23.08%, 18/78). Almost all of the MeSH terms and entry terms, and most of the Emtree terms and synonyms to

which CER terms mapped were at most 3 words long. Examples of detailed CER terms included cluster quasi-randomized controlled trial and meta-analysis of individual participant data.

Several terms (14.10%, 11/78) involved negation, such as cluster nonrandomized controlled trial, interrupted time series without comparison group, nonrandomized crossover trial, and uncontrolled longitudinal study.

For exact matches in 1 or both controlled vocabularies (n=29), 1 term was detailed (3.45%, 1/29) and 1 negated (3.45%, 1/29): nested case-control study and non-experimental study, respectively. Emtree covered more terms exactly than MeSH (26 Emtree vs. 15 MeSH).

For partial matches in 1 or both controlled vocabularies (n=55), 18 terms were detailed (32.73%, 18/55) and 10 negated (18.18%, 10/55). MeSH covered more terms partially than Emtree (49 MeSH vs. 45 Emtree). Sixteen terms partially mapped to a MeSH term because of a matching substring in the scope note. For example, trend study mapped to the MeSH term sentinel surveillance because the scope note included the following excerpt: "the study of disease rates in a specific cohort, geographic area, population subgroup, etc. to estimate trends [emphasis added]."

Figure 1
Comparative effectiveness research (CER) design terms by appearance in local terminologies

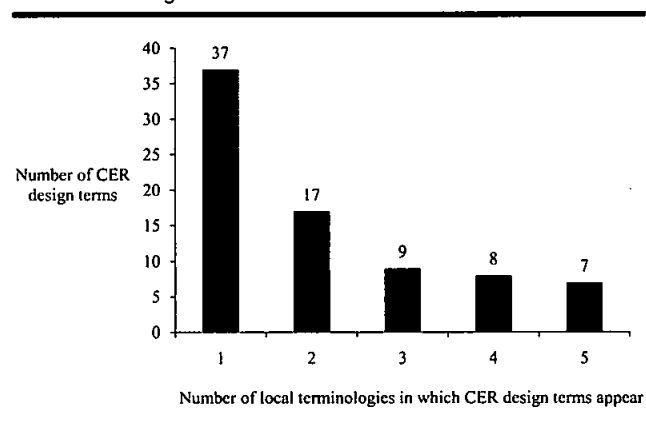


Table 2

Pairwise overlap of local terminologies: percentage overlap computed as (n terms shared/n terms in either terminology) × 100

	Agency for Healthcare Research and Quality (AHRQ)	Cochrane Non- Randomised Studies Methods Group (Cochrane)	Academy of Nutrition and Dietetics (ADA)	Research Triangle Institute International-University of North Carolina Evidence- based Practice Center (RTI)	Mean pairwise overlap
Alberta	75.61% (31/41)	57.89% (33/57)	32.69% (17/52)	28.89% (13/45)	48.77%
AHRQ		24.56% (14/57)	31.58% (18/57)	22.64% (12/53)	38.60%
Cochrane			30.77% (16/52)	23.40% (11/47)	34.16%
ADA				27.66% (13/47)	30.68%
RTI					25.65%
Mean overall pairwise overlap					35.57%

For terms not matched in 1 or both controlled vocabularies (n=15), 13.33% (2/15) were negated: *noncomparative study* and *non-experimental study*. (Note that while *non-experimental study* failed to map in MeSH, it exactly mapped to a synonym for *observational study* in Emtree.) No unmatched term was detailed. MeSH had twice as many "no matches" as Emtree (14 MeSH vs. 7 Emtree). Both controlled vocabularies failed to cover *before-after study* (including variants), which is a core term appearing in all 5 LTs. Checking whether unmapped terms appeared in any of the Unified Medical Language System (UMLS) resources [27], we found that 20% (3/15) mapped to terms in the National Cancer Institute (NCI) Thesaurus [28], including *community trial* (C1516736), *factorial study* (C2826344), and *parallel study* (C2826345).

CER design queries

The average number of hits for CER design queries restricted to titles and abstracts varied with the record source and type of match.

The median (MDN) number of MEDLINE records retrieved in Embase was 1,090 (range: 0 to 222,804); the MDN number of Embase records was 380 (range: 0 to 89,807). *Case report* yielded the most hits for both MEDLINE and Embase.

Average hits by type of match were: MeSH exact (MDN=37,750; range: 960 to 222,804), partial (MDN=735; range: 0 to 114,303), or no match (MDN=590; range: 11 to 11,915); Emtree exact (MDN=9617; range: 54 to 89,807), partial (MDN=199; range: 0 to 22,584), and no match (MDN=152; range: 9 to 432).

Based on nonparametric independent-samples median tests (MTs), the average hits varied significantly across type of match for MeSH (MT=14.022, *df*=2, *P*<0.001) and Emtree (MT=19.789, *df*=2, *P*<0.000). Pairwise differences in hits were significant for the exact versus partial category comparison (MeSH MT=14.716, *P*<0.000; Emtree MT=16.258, *P*<0.000) and the exact versus no match category (MeSH

MT=12.523, *P*<0.001; Emtree MT=8.362, *P*<0.011). Differences were statistically nonsignificant for the partial versus no match comparison in both MeSH and Emtree. *P* values were adjusted for the number of comparisons.

DISCUSSION

Terminologies

With the exception of the AHRQ and Alberta LTs, organizational terminologies varied quite a bit as measured by pairwise overlap and overlap with the new CER design terminology. The reason for this exception is that AHRQ is developing a taxonomy for study designs that builds on the Alberta classification tool. However, the overlap was not perfect because we augmented the basic set of terms that the 2 organizations share with terms from supplementary documents. Note that augmenting term lists was useful because we were not evaluating extant terminologies per se, but were interested in using documents vetted by methodologists to analyze differences in language. Thus, the mean pairwise overlap of 36% for the augmented LTs and the mean overlap of 42% with the CER design terminology substantiated what we had expected: that language varies by organization even when the domain is ostensibly the same.

To explore coverage of designs and related concepts in MeSH and Emtree, we developed a terminology that consists of terms used by organizations dedicated to promoting CER, especially systematic reviews of medical evidence. Just seven terms were common to all five organizations, and even this core set of shared terms was inconsistently covered in the controlled vocabularies. For example, the core terms *case-control study*, *cross-sectional study*, and *randomized controlled trial* exactly mapped to controlled terms in both MeSH and Emtree; whereas, *case series* exactly mapped in Emtree and partially in MeSH; *prospective cohort study* and *retrospective cohort study* partially mapped in both; and *before-after study* failed to map in either. This

Table 3

Overlap of local terminologies with the CER design terminology: percentage overlap computed as (n terms in the local terminology/n terms in the union) × 100

Alberta	AHRQ	Cochrane	ADA	RTI	Mean overlap
42.31% (33/78)	50.00% (39/78)	41.03% (32/78)	46.15% (36/78)	32.05% (25/78)	42.31%

Table 4
Coverage of CER design terms by controlled vocabulary

Type of match	Medical Subject Headings (MeSH)		Emtree		MeSH to Emtree ratio	
No match	14	(17.9%)	7	(9.0%)	2.00	(14/7)
Partial	49	(62.8%)	45	(57.7%)	1.09	(49/45)
Exact	15	(19.2%)	26	(33.3%)	0.58	(15/26)
Total	78	(100.0%)	78	(100.0%)		

inconsistent coverage of core terms suggests that CER organizational language does not correspond well with indexing for designs.

Regarding the full set of terms for designs and related concepts in our terminology, most either partially mapped or failed to map to broad or related controlled terms. In some cases, the controlled terms were not for study designs per se, but research domains. For example, *analytic study* mapped to *analytical research* and *descriptive study* to *descriptive research* in Emtree. Interestingly, while the core term *randomized controlled trial* exactly mapped in MeSH and Emtree, the counter term *nonrandomized controlled trial* appearing in the full set did not, even though the latter is a common design. In general, negated terms rarely mapped exactly, with the exception of *non-experimental study* in Emtree.

Because CER is an emerging discipline, resources are being developed at the regional and federal level to help expert searchers. For example, the University of Pittsburgh Health Sciences Library System, a Regional Medical Library for the Middle Atlantic Region of the National Network of Libraries of Medicine, developed MedTerm Search Assist [29]. This tool promotes sharing of biomedical terms and comprehensive search strategies among librarians. One can browse for *comparative effectiveness research* to find keywords, MeSH terms, and a search filter. Currently, the fields include a few design terms, such as *cluster randomized trial* and *pragmatic clinical trial*.

At the federal level, NLM resources are available online by navigating to Comparative Effectiveness Research from Topic-Specific Queries on the PubMed home page [30]. For example, the complex query for

Observational Studies consists of several blocks: <study designs> AND <comparative terms> AND <common CER topics>. Ignoring spelling variants, most of the terms in the study design block exactly or partially match MeSH terms that we found, with the exception of *practice guidelines as topic*, *matched-pair analysis*, and *multicenter study*. However, quite a few of the relevant designs identified in this study do not appear in the PubMed query.

It is worth noting that the PubMed query for Observational Studies includes terms for retrieving registry studies [31], terms which do not appear in the documents we mined for this analytical study. However, neither the PubMed query nor our CER design terminology has a term for hospital-based case control studies, a design covered in Emtree. Both registry and hospital-based case control studies are increasingly important in CER, partly because electronic medical records facilitate data reuse within health care systems and research across institutions.

Crosswalk

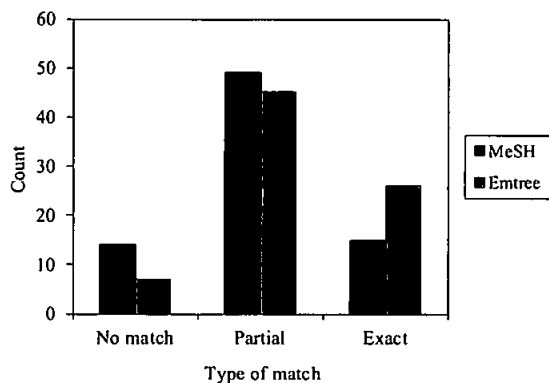
The CER design terminology and its crosswalk (Table 5, online only) may be useful for expert searchers who need to search MEDLINE and/or Embase. They could consult the crosswalk when developing queries for users who want studies in the CER domain, especially studies with designs that methodologists classify with negated or detailed phrases or terms such as *head-to-head study* and *pragmatic trial* important in CER. The latter pair failed to map in MeSH and Emtree.

Throughout Table 5, librarians will find suggestions for alternative terms or query expansions. Because this is a first effort, librarians should be alert to the potential for false positives. For example, *focus group* exactly mapped to *information processing* in Emtree because it occurs in a long list of synonyms, and *reliability study* partially mapped to *validation studies* in MeSH because *reliability* is mentioned in the scope note.

In general, MeSH terms for <design> as topic should be avoided, as this heading is usually not assigned to primary studies. However, at times it may be necessary, for example, *pre-post study* mapped to *evaluation studies as topic*.

Methodologists used a variety of terms to classify studies involving time, including several versions of *before-after study* modified by *controlled* or *cohort*, *time series* modified by *interrupted* and *with comparison group* or *without comparison group*, *historically controlled trial*, *nonconcurrent cohort study*, *pre-post study*, several

Figure 2
Comparison of coverage in Medical Subject Headings (MeSH) and Emtree



terms modified by *prospective* or *retrospective*, and *uncontrolled longitudinal study*. All of these are not well indexed.

In sum, queries for designs with partially mapped terms could consist of free text for modifiers such as *nonrandomized* or *prospective* added to broader or related controlled terms, if they exist. Queries for designs with unmapped terms require free text by necessity.

Scientists' language

When we considered whether scientists use CER design terms, some striking discrepancies emerged. For example, scientists commonly used terms not well indexed in MeSH or Emtree, such as *before-and-after study* (1,854 total hits in Embase), *descriptive study* (16,408 hits), *diagnostic study* (4,849 hits), *prospective cohort study* (19,096 hits), and *retrospective cohort study* (14,798 hits).

On the other hand, scientists rarely used detailed terms, such as *cluster nonrandomized controlled trial*, *cohort before-and-after study*, and *interrupted time series with comparison group*. They were much more likely to describe in various parts of the titles and abstracts how their studies were carried out, effectively splitting up the concepts in detailed terms. For example, searching for "*cohort*" [tiab] AND "*before-and-after*" [tiab] in MEDLINE yielded 2,804 hits; whereas, searching for the CER design string "*cohort before-and-after study*" [tiab] yielded 0 hits (24 May 2012). As an aside, searching for just "*cohort*" [tiab] returned 190,261 hits, which was counter to Eldredge's finding that "authors rarely use the label 'cohort' when describing their methods" (p. 85) [32]. His comment together with the results of this simple MEDLINE query for *cohort* point to presumed differences in the sublanguages of librarianship and biomedicine, although this may be changing.

Limitations and future research

To improve upon the representativeness of the CER design terminology, additional documents could be mined, such as the AHRQ and Cochrane glossaries [33, 34], which are broader than the documents we used in this study. To be globally representative, documents from major international centers, such as the NHS National Institute for Health and Clinical Excellence in the United Kingdom, could be of use.

In our exploration of scientists' language, we were unable to infer detailed study designs by simple string matching of CER phrases in titles and abstracts. Thus, to improve our approach, semantic analysis [6] of texts written by scientists could be worthwhile.

The crosswalk in Table 5 (online only) could be further developed by librarians, paying attention to the potential for false positives given their users' needs and changes in indexing. An obvious extension would be to add other controlled vocabularies for databases that librarians regularly search, such as PsycINFO. Additionally, more terms such as *comment*

or *letter* for "NOTing out irrelevant content" should be added, as these can improve precision for exhaustive searches [35]. Although it was not our intention to develop a search filter, our design terminology and its crosswalk could be of use to librarians and trials search coordinators who support systematic reviewers and other comparative effectiveness researchers.

CONCLUSION

In this study, we have demonstrated that the degree to which methodologists in CER organizations share a terminology for designs varies considerably. Further, we have shown that coverage of design terms and related concepts is similar in MeSH and Emtree and that the majority of terms partially map or fail to map to controlled terms. This poses challenges for librarians who support users in various CER communities. Finally, exhaustive searches require free text for concepts appearing in detailed design phrases because scientists split up terms in their titles and abstracts.

REFERENCES

1. Agency for Healthcare Research and Quality. What is comparative effectiveness research [Internet]. The Agency [cited 1 Jun 2012]. <<http://www.effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/>>.
2. Goodrich K, Auston I, Dunn K, van Horne V, Lang L. Understanding comparative effectiveness research part 1: the controversies and the challenges. *MLA News*. 2012 Jan;52(1):1-8.
3. Auston I, Dunn K, van Horne V, Goodrich K, Lang L. Understanding comparative effectiveness research part 2: opportunities for health sciences librarians. *MLA News*. 2012 Apr;52(4):8-9.
4. Luce BR, Drummond M, Jonsson B, Neumann PJ, Schwartz JS, Siebert U, Sullivan SD. EBM, HTA, and CER: clearing the confusion. *Milbank Q*. 2010 Jun;88(2):256-76.
5. Holve E, Pittman P. A first look at the volume and cost of comparative effectiveness research in the United States. *AcademyHealth*; Jun 2009.
6. Jurafsky D, Martin JH. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. 2nd ed. Upper Saddle River, NJ: Pearson Education; 2009.
7. Harris ZS. The structure of science information. *J Biomed Inform*. 2002 Aug;35(4):215-21.
8. Grishman R, Kittredge R. *Analyzing language in restricted domains: sublanguage description and processing*. Hillsdale, NJ: Lawrence Erlbaum; 1986.
9. Carini S, Pollock BH, Lehmann HP, Bakken S, Barbour EM, Gabriel D, Hagler HK, Harper CR, Mollah SA, Nahm M, Nguyen HH, Scheuermann RH, Sim I. Development and evaluation of a study design typology for human research. *AMIA Annu Symp Proc*. 2009 Nov:81-5.
10. Cochrane Collaboration. *Methods groups* [Internet]. The Collaboration [cited 5 May 2012]. <<http://www.cochrane.org/contact/methods-groups>>.
11. Reeves B, Deeks J, Higgins J, Wells G. Chapter 13: Including non-randomized studies. In: Higgins J, Green S,

- eds. *Cochrane handbook for systematic reviews of interventions*. Chichester (UK): Wiley; 2008.
12. EQUATOR Network. Library for health research reporting [Internet]. The Network [rev. 13 Mar 2012; cited 5 May 2012]. <<http://www.equator-network.org/resource-centre/library-of-health-research-reporting/>>.
 13. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ*. 2003 Jan 4;326(7379):41–4. PMID: 12511463.
 14. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care*. 2007 Dec;19(6):349–57. PMID: 17872937.
 15. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening [of] the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol*. 2008 Apr;61(4):344–9. PMID: 18313558.
 16. Hammond WE, Cimino JJ. Chapter 7: Standards in biomedical informatics. In: Shortliffe EH, Cimino JJ, eds. *Biomedical informatics*. 3rd ed. New York, NY: Springer; 2006.
 17. US National Library of Medicine, National Institutes of Health. Medical subject headings [Internet]. The Library [rev. 28 Oct 2011; cited 4 Apr 2012]. <<http://www.nlm.nih.gov/mesh/>>.
 18. Elsevier. Emtree: the life science thesaurus [Internet]. [cited 4 Apr 2012]. <<http://www.embase.com/info/what-is-embase/mtree>>.
 19. Hartling L, Bond K, Harvey K, Santaguida PL, Viswanathan M, Dryden DM. Developing and testing a tool for the classification of study designs in systematic reviews of interventions and exposures: University of Alberta Evidence-based Practice Center. AHRQ methods research reports. Rockville, MD: US Agency for Healthcare Research and Quality; 2010. p. 74.
 20. Framework for considering study designs for future research needs (FRN): draft comparative effectiveness review. Rockville, MD: US Agency for Healthcare Research and Quality; 2011. p. 28.
 21. Academy of Nutrition and Dietetics. Evidence analysis manual: steps in the academy evidence analysis process [Internet]. Chicago, IL: The Academy; 2012 [cited 3 Apr 2012]. <<http://www.adaevidencelibrary.com/topic.cfm?cat=1315>>.
 22. US National Library of Medicine. MeSH browser [Internet]. Bethesda, MD: The Library [cited 27 Nov 2012]. <<http://www.nlm.nih.gov/mesh/MBrowser.html>>.
 23. Crowlesmith I. Coverage of MEDLINE in Embase [Internet]: Elsevier; May 2011 [cited 8 Apr 2012]. <http://www.pbt.up2els.com/sites/default/files/MEDLINE%20in%20Embase_Whitepaper_2011_print.12May2011.pdf>.
 24. Microsoft Office Excel 2003 [computer program]. Redmond, WA.
 25. Microsoft Office Excel 2010 [computer program]. Redmond, WA.
 26. IBM SPSS statistics, version 20 [computer program]. Armonk, NY.
 27. UMLS reference manual [Internet]., Bethesda, MD: US National Library of Medicine; 2009 [cited 8 Apr 2012]. <<http://www.ncbi.nlm.nih.gov/books/NBK9676/>>.
 28. National Cancer Institute. NCI thesaurus [Internet]. The Institute [cited 27 Nov 2012]. <<http://ncit.nci.nih.gov>>.
 29. University of Pittsburgh Health Sciences Library System. MedTerm search assist [Internet]. 2011 [cited 24 May 2012]. <<http://www.hsls.pitt.edu/terms/>>.
 30. US National Library of Medicine. NLM resources for informing comparative effectiveness [Internet]. The Library; 2009 [rev. 21 Dec 2010; cited 24 May 2012]. <<http://www.nlm.nih.gov/nichsr/cer/cerqueries.html>>.
 31. AHRQ Agency for Healthcare Research and Quality. Registries for evaluating patient outcomes: a user's guide [Internet]. 2nd ed. The Agency [rev. 3 Mar 2010; cited 27 May 2012]. <<http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=401>>.
 32. Eldredge JD. Inventory of research methods for librarianship and informatics. *J Med Lib Assoc*. 2004 Jan;92(1):83–90.
 33. Agency for Healthcare Research and Quality. Glossary of terms [Internet]. The Agency [cited 10 May 2012]. <<http://www.effectivehealthcare.ahrq.gov/index.cfm/glossary-of-terms/>>.
 34. Cochrane Collaboration. Glossary [Internet]. The Collaboration [cited 10 May 2012]. <<http://www.cochrane.org/glossary/>>.
 35. Wilczynski NL, McKibbin KA, Haynes RB. Search filter precision can be improved by NOTing out irrelevant content. *AMIA Annu Symp Proc*. 2011 Nov:1506–13.

AUTHORS' AFFILIATIONS

Tanja Bekhuis, PhD, MS, MLIS, tcb24@pitt.edu, Assistant Professor, Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, 5607 Baum Boulevard, Room 514, Pittsburgh, PA 15206; **Dina Demner-Fushman, MD, PhD**, dina.demner@nih.gov, Staff Scientist, Communications Engineering Branch (HNL32) Lister Hill National Center for Biomedical Communications, US National Library of Medicine, Building 38A, Room 10S1022, Mail Stop 3824, 8600 Rockville Pike, Bethesda, MD 20814; **Rebecca S. Crowley, MD, MS**, crowleysr@upmc.edu, Associate Professor, Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, 5607 Baum Boulevard, Room 523, Pittsburgh, PA 15206

Received June 2012; accepted September 2012

Copyright of Journal of the Medical Library Association is the property of Medical Library Association and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.