Environment International 123 (2019) 451–458

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Low-calorie sweeteners and health outcomes: A demonstration of rapid evidence mapping (rEM)

Juleen Lam[a,b,*], Brian E. Howard[a], Kristina Thayer[c], Ruchir R. Shah[a]

[a] Sciome LLC, RTP, NC, United States of America
[b] California State University, East Bay, Department of Health Sciences, Hayward, CA, United States of America
[c] Integrated Risk Information System (IRIS) Division, National Center for Environmental Assessment, Environmental Protection Agency, Washington, DC, United States of America

ARTICLE INFO

Handling Editor: Olga-Ioanna Kalantzi

Keywords:
Rapid evidence mapping
Evidence map
Systematic review
Evidence-based methodology
Low calorie sweeteners
Artificial sweeteners

ABSTRACT

Background: "Evidence Mapping" is an emerging tool that is increasingly being used to systematically identify, review, organize, quantify, and summarize the literature. It can be used as an effective method for identifying well-studied topic areas relevant to a broad research question along with any important literature gaps. However, because the procedure can be significantly resource-intensive, approaches that can increase the speed and reproducibility of evidence mapping are in great demand.

Methods: We propose an alternative process called "rapid Evidence Mapping" (rEM) to map the scientific evidence in a time-efficient manner, while still utilizing rigorous, transparent and explicit methodological approaches. To illustrate its application, we have conducted a proof-of-concept case study on the topic of low-calorie sweeteners (LCS) with respect to human dietary exposures and health outcomes. During this process, we developed and made publicly available our study protocol, established a PECO (Participants, Exposure, Comparator, and Outcomes) statement, searched the literature, screened titles and abstracts to identify potentially relevant studies, and applied semi-automated machine learning approaches to tag and categorize the included articles. We created various visualizations including bubble plots and frequency tables to map the evidence and research gaps according to comparison type, population baseline health status, outcome group, and study sample size. We compared our results with a traditional evidence mapping of the same topic published in 2016 (Wang et al., 2016).

Results: We conducted an rEM of LCS, for which we identified 8122 records from a PubMed search (January 1, 1946–May 1, 2014) and then utilized machine learning (SWIFT-Active Screener) to prioritize relevant records. After screening 2267 (28%) of the total set of titles and abstracts to achieve 95% estimated recall, we ultimately included 297 relevant studies. Overall, our findings corroborated those of Wang et al. (2016) and identified that most studies were acute or short-term in healthy individuals, and studied the outcomes of appetite, energy sensing and body weight. We also identified a lack of studies assessing appetite and dietary intake related outcomes in people with diabetes. The rEM approach required approximately 100 person-hours conducted over 7 calendar months.

Conclusion: Rapid Evidence Mapping is an expeditious approach based on rigorous methodology that can be used to quickly summarize the available body of evidence relevant to a research question, identify gaps in the literature to inform future research, and contextualize the design of a systematic review within the broader scientific literature, significantly reducing human effort while yielding results comparable to those from traditional methods. The potential time savings of this approach in comparison to the traditional evidence mapping process make it a potentially powerful tool for rapidly translating knowledge to inform science-based decision-making.

Abbreviations: rEM, rapid evidence mapping; LCS, low calorie sweeteners; PECO, Participants, Exposure, Comparator, and Outcomes; GRAS, generally recognized as safe; MeSH, Medical Subject Headings; SRDR, The Systematic Review Data Repository

* Corresponding author at: 2 Davis Drive, Durham, NC 27709, United States of America.
E-mail address: Juleen.Lam@sciome.com (J. Lam).

https://doi.org/10.1016/j.envint.2018.11.070
Received 25 July 2018; Received in revised form 20 November 2018; Accepted 27 November 2018
Available online 05 January 2019

1. Background

Stakeholders in the field of human health risk assessment are increasingly relying on tools and practices from the disciplines of systematic review to summarize the evidence and identify scientific consensus to support decision-making with regard to potential environmental health risks (EFSA, 2010; NRC, 2011, 2013; Rooney et al., 2014; Woodruff et al., 2011). Given the ever-accelerating pace of publications in this field, the practice of "Evidence Mapping" is now frequently being used to identify the key areas of study relevant to a given topic along with important gaps in the literature (Miake-Lye et al., 2016). This emerging tool is used to systematically, transparently, and comprehensively address broad, multi-faceted questions related to a topic of interest for which it may not be feasible to conduct a systematic review (for instance, when faced with limited resources) or when a detailed synthesis is not needed (Bragge et al., 2011; Gough et al., 2012; McKinnon et al., 2015; Miake-Lye et al., 2016; Snilstveit et al., 2013; Colquhoun et al., 2014). An evidence map may also provide an overview of a broad collection of scientific data that can be used to inform where a systematic review may be appropriate by identifying the boundaries and context of a broad topic area and providing a description of the number of studies, types of interventions, study design and study characteristics (Bragge et al., 2011). Evidence mapping aims to systematically examine the extent, range, and limitations of current scientific knowledge for a broad research question and results in the development of a catalog of the available evidence, in the form of a database, visual evidence map, or tabular count of meta-data from each study (e.g., study setting, design, interventions, populations, etc.) (James et al., 2016). In particular, the resulting evidence map characterizes in detail the quantity and nature of research in a particular area. However, constructing an evidence map can be a resource-intensive procedure, thereby limiting their utility for practical implementation.

In this paper, we describe a process we call "rapid Evidence Mapping" (rEM), which we define as a resource-efficient form of knowledge synthesis where components of the review process are simplified to produce a visual and quantitative representation of the scientific evidence from which to commission further reviews and/or primary research by identifying gaps in research. This draws from other review approaches, in particular rapid reviews and evidence mapping (Grant and Booth, 2009; Khangura et al., 2012; Ganann et al., 2010; Bragge et al., 2011; Miake-Lye et al., 2016). The process is designed to improve or enhance efficiency, while still utilizing rigorous, transparent and explicit methodological approaches. To illustrate its application, we conducted a proof-of-concept case study on the topic of low-calorie sweeteners (LCS) in human dietary exposures and health outcomes. We selected this topic because we had identified a traditional evidence map of the same topic published by Wang et al. (2016), thus enabling us to compare results as well as time/resource requirements between the two approaches. Our main goals were: 1) to assess and refine our rapid Evidence Mapping protocol, comparing the resulting outputs with those obtained using traditional Evidence Mapping methodology; 2) to explore the use of currently available semi-automated machine learning approaches to reduce the time and resource commitments required to undertake rEM assessments; and 3) to explore the feasibility of utilizing these approaches on a widespread scale by risk assessors, health assessors, and decision-makers.

2. Methods

To maximize consistency between the two approaches, we followed a similar process as outlined by Wang et al. (2016). However, our protocol additionally incorporated semi-automated machine learning methods and other potential time-saving alternative approaches at each step. We implemented a seven-step process to conduct the rEM, modifying that of Wang et al. (2016): 1) identify the scope of the evidence map; 2) develop a comprehensive search strategy; 3) establish study eligibility criteria and a systematic study selection process; 4) carry out abstract screening and selection; 5) tag/categorize studies; 6) classify study population, duration, interventions, and outcome categories; 7) create an evidence map (Fig. A.1). Our protocol is similar to the seven-step process employed by Wang et al. (2016), with a few notable modifications. First, we omitted one step ("define the roles and responsibilities of different parties: stakeholder panel and research team") because in order to replicate Wang et al.'s (2016) process we did not did not establish our own stakeholder panel but instead incorporated decisions that had been reported in Wang et al. (2016). In addition, we also replaced their "data extraction" step with "tag/categorize studies," and added one additional step to explicitly "create an evidence map," as discussed in detail below. The details of each step are provided as follows and were outlined beforehand in our publicly-available protocol (available at: https://tinyurl.com/y7gcptqg).

2.1. Identify the scope of the evidence map

To ensure that we could compare our rEM with a traditional evidence mapping, the scope of this project (study question, PECO statement, search strategy, and inclusion/exclusion criteria) was chosen to align with that defined by Wang et al. (2016). Wang et al.'s (2016) process for defining the project scope involved a stakeholder panel consisting of physicians, dietitians, policymakers, and representatives from the food industry, academia, journalism, the public, and a research funder. Their panel served as a steering committee to guide the research team through the process of evidence mapping, providing input along the way at each step. We did not establish our own steering committee, but we did incorporate the guidance of Wang et al.'s steering committee—for instance, incorporating the search strategy and eligibility criteria modified by the panel.

Aligning with Wang et al.'s (2016) approach, we targeted English-only human studies with experimental or prospective cohort study designs. We developed a PECO statement (Participants, Exposure, Comparator, and Outcomes), an aid to developing an answerable review question (Higgins and Green, 2011), reflecting Wang et al.'s (2016) inclusion criteria (Table A.1). Our PECO statement is shown below:

Participants: Humans.

Exposure: Orally administered, FDA-approved or generally recognized as safe (GRAS) LCS.

Comparator: Humans exposed to lower levels of LCS than more highly exposed humans, or humans who serve as their own control by comparing before-and-after outcomes following exposure.

Outcome: Any outcomes related to appetite, energy sensing by the brain, body weight/composition, dietary intake, or gut hormones that may influence energy homeostasis. More specific criteria defining each outcome are included in Table A.2.

2.2. Develop a comprehensive search strategy

We adopted the search strategy proposed by Wang et al. (2016) but with several modifications. Wang et al. (2016) collected key search terms from three published reviews on relevant topics and used these to develop keywords and Medical Subject Headings (MeSH) terms for the search strategy, implemented in Ovid MEDLINE. We modified this search strategy for implementation in PubMed, using the same keywords, MeSH terms, and date limit (January 1, 1946–May 1, 2014) as reported in Wang et al.'s (2016) supplemental materials. We made this decision because: 1) we did not have subscription access to Ovid MEDLINE and 2) a search in PubMed—which searches MEDLINE and several other databases—would be more comprehensive and potentially capture more relevant studies (Duffy et al., 2016). We also made additional modifications to narrow the focus for specific search terms by incorporating quotations around multiple keywords to avoid automatic

term substitution around each individual term, which would lead to a broader capture of references. For instance, searching on PubMed for the term *artificial sweetener* (as was done in the original search by Wang et al. (2016)) results in automatic substitution and search for the separate terms *artificial* and *sweetener*, which is likely not the intended outcome. Alternatively, utilizing quotation marks around "artificial sweetener" searches all fields for the combined phrase.

We implemented our modified search terms (Supplemental Materials, PubMed Search Terms) in PubMed on January 24, 2018. References were imported into EndNote where manual review of potential duplicates using EndNote's "Find Duplicates" function was used to remove duplicate citations. Similar to Wang et al. (2016), the search was also cross-referenced with published systematic reviews to check that relevant articles were included.

2.3. Establish study eligibility criteria and a systematic study selection process

Study eligibility criteria mirrored those established by Wang et al. (2016) to maximize consistency. To meet the inclusion criteria, studies needed to: 1) be randomized or non-randomized, controlled, clinical trials or prospective cohort study designs; 2) investigate orally administered, FDA-approved or generally recognized as safe (GRAS) LCS; 3) report at least one health outcome within the five categories identified in the scope; 4) be English publications; and 5) use human subjects in the research. Studies were excluded if they were: 1) animal studies; 2) in vitro cell studies; 3) case-control or cross-sectional studies, reviews, interviews, bibliographies, letters, or guidelines; 4) systematic reviews and meta-analyses; 5) studies of cancer patients; 6) studies involving non-oral LCS intake. Inclusion/exclusion criteria were applied during the screening process to determine study eligibility.

2.4. Carry out abstract screening and study selection

Wang et al. (2016) performed title and abstract screening using the abstrackR citation screening tool (abstrackr.cebm.brown.edu), a free, open-source citation screening program (Rathbone et al., 2015). Although abstrackR includes machine learning features (Wallace et al., 2012), Wang et al. (2016) did not use these for creation of their evidence map (Chung, 2017). In Wang et al.'s (2016) approach, the first 1000 abstracts were screened by all four reviewers, to calibrate screening accuracy between reviewers, and the remaining abstracts were single-screened (one reviewer per abstract). Studies included at the title-and-abstract screening phase were then moved on for full-text review, with each record screened by one reviewer with a second reviewer to confirm or dispute the first reviewer's decision. Discrepancies were resolved through consensus among all five research team members.

We followed a similar but modified procedure. First, we screened the titles and abstracts using SWIFT-Active Screener (https://www.sciome.com/swift-activescreener/), a web-based, collaborative systematic review software application. SWIFT-Active Screener is designed to save screeners time and effort by using active learning and statistical modeling approaches to prioritize relevant references during the screening process and estimate the number of relevant articles in the unscreened document list. We also only screened the titles and abstracts of references and did not conduct screening of the full text, to increase the efficiency of the process by reducing the amount of time and resources required at this step.

We screened the titles and abstracts of references in duplicate for the first 500 records to calibrate the two screeners (DB and LA), and conflicts were discussed and resolved between the two reviewers with further review by two of the authors (JL and BH). When appropriate, we appended the protocol with the criteria used to resolve conflicts and the date, to document decisions that were made during the review process. One review author (BH) additionally screened a subset of the studies (50 articles) for quality assurance/quality control. Once calibration was completed, single-screening was employed for the remaining records. Screeners reviewed titles and abstracts in SWIFT-Active Screener until an estimated 95% recall was achieved—in other words, up until the machine learning algorithms predicted that we had identified 95% of all relevant (or "included") references. This approach enables the discovery of a vast majority of relevant articles after reviewing only a fraction of the total number of studies, significantly reducing the time and resources required to screen references. However, this potentially creates a trade-off between time commitment and inclusiveness, and to explore this, we conducted a sensitivity analysis to investigate the impact of screening references up to 25%, 50%, 75%, 98%, and 99% estimated recall. Our sensitivity analysis at different recall rates explores the impact on the resulting evidence map when screening fewer or more of the references to identify relevant studies. Although fewer resources are required to screen up to 25% recall (up until the machine learning algorithms predicted that we had identified 25% of all relevant (or "included") references), using this approach means that we could potentially miss relevant studies (i.e., up to 75% of studies that should be included). On the other hand, screening up to 99% recall (where virtually all relevant studies would be identified) would require more screening time and resources to achieve.

2.5. Tag/categorize studies

This step was similar to Wang et al. (2016) as "data extraction" step, where they extracted study aims, design, population, interventions, exposures, outcomes, and funding source from the full text of the included studies. The resulting evidence-map database was uploaded and made publicly available on The Systematic Review Data Repository (SRDR), an online publicly available repository of systematic review data that serves as a central archive and data extraction tool for systematic reviewers (AHRQ, 2018; Ip et al., 2012).

In contrast, to reduce the time and resource commitment for project completion, we did not manually extract data from the full text of included studies. Instead, we imported all included titles and abstracts into SWIFT-Review (Sciome Workbench of Interactive computer-Facilitated Text-mining), a freely available, interactive text mining and machine learning software application (https://www.sciome.com/swift-review/). SWIFT-Review provides tools to assist with searching, categorization, and pattern visualization in literature search results, utilizing statistical modeling and machine learning methods (Howard et al., 2016). SWIFT-Review also incorporates automatic identification of health outcomes, chemical names and synonyms, keywords, MeSH terms, etc. from included references. We utilized this functionality to automatically summarize and visualize data from the included abstracts and to identify outcome, baseline health, and comparison categories, substituting this approach for the manual extraction of data from the full text articles as performed by Wang et al. (2016). We also manually reviewed abstracts to tag study length and sample size categories, as there exist challenges with automatic parsing of this information. Through these processes, we limited our evidence map to information that could be readily identified from the titles and abstracts of the included articles.

2.6. Classify study population, duration, interventions, and outcome categories

Wang et al. (2016) classified outcomes into clinically and biologically meaningful outcome categories for their evidence-map analyses. We utilized the same final list of outcome categories (Table A.2) and employed the automated tagging/categorization and search feature functions in SWIFT-Review to tag articles with the appropriate outcome categories. Specifically, for each outcome category we used the following iterative procedure to develop and refine search terms, and to

tag articles according to the desired outcome categories:

1) The outcomes of interest identified in Table A.2 were used to define an initial set of search terms.
2) Using the title/abstract search features of SWIFT-Review, this initial set of search terms was employed to automatically assign outcome categories to the relevant abstracts.
3) The labeled studies were then manually reviewed to confirm appropriate tagging and remove incorrect tag from studies.
4) SWIFT-Review was used to compute term frequency-inverse document frequency (TF-IDF) scores for each set of correctly labeled documents (Howard et al., 2016) and to identify and rank additional potential keywords for enrichment in studies labeled with each outcome category.
5) After adding any additional search terms identified in step 4 above, steps 2–4 were repeated as necessary until additional search terms no longer provided further benefit.

This process resulted in tagging relevant studies that fell into each of the five outcome group categories, predominantly utilizing automatic searching and categorizing. A similar approach was conducted for baseline health and comparison categories.

2.7. Create an evidence map

We then generated an evidence map describing study design and population characteristics, following a similar approach to Wang et al. (2016) in order to compare our results with their traditional evidence map. The original evidence map displayed the categorical features of included studies by study duration, outcome group, intervention, and baseline study population characteristics. We emulated these results using tools in SWIFT-Review to automatically generate frequency tables reporting on the categorical percentage of studies falling within the intersection of categories. We also recreated Wang et al.'s (2016) bubble plot visualizations, which are essentially a type of a weighted scatter plot. In each plot, the unit of analysis was the individual study, and because studies could report on multiple outcomes, one study could be counted multiple times. The evidence map was generated in SWIFT-Review, with modifications conducted in either R (R Core Team, 2017) or Microsoft Excel.

3. Results

Our search in PubMed spanning January 1, 1946–May 1, 2014 retrieved 8122 unique records. Of these, we screened 2267 (28%) total titles and abstracts using SWIFT-Active Screener, until 95% predicted recall was achieved, which yielded 297 potentially relevant references (Fig. A.2). As part of a sensitivity analysis, we also continued screening until 99% predicted recall was achieved, which yielded 301 potentially relevant references (Fig. A.2). In comparison, the original search by Wang et al. (2016) in PubMed resulted in 12,830 records and after screening, ultimately yielded 225 relevant references, with 115 studies overlapping from our list of included studies. Wang et al. (2016) included 70 references that we did not include, and we included 182 references that Wang et al. (2016) did not include (Supplemental Materials, Fig. B.1, Supplemental Materials, Tables B.1–B.2).

The following sections summarize descriptive analyses of the data in the LCS rapid Evidence Map that were generated to characterize the existing body of literature and to identify potential gaps for future research. We also compare these results to those reported by Wang et al. (2016).

3.1. Summarize study characteristics and design

Of the 297 included studies, the majority of studies: a) were conducted in subjects where health status was healthy, mixed or other (n = 249, 84%); b) involved interventions comparing LCS versus sugar (n = 150, 51%); and c) were acute in duration lasting < 1 day (n = 173, 58%) (Table A.4). Wang et al. (2016) reported similar proportions for health status and duration, with 83% studies with health status as healthy, mixed, or other; and 60% studies acute lasting < 1 day. For intervention comparison, Wang et al. (2016) identified a greater proportion (80%) of studies comparing LCS versus sugar interventions.

Among our 297 included studies, 128 (43%) reported energy sensing-related outcomes, 69 (23%) reported appetite-related outcomes, 60 (20%) reported glycemic-related outcomes, 54 (18%) reported dietary intake, and 23 (8%) reported body weight/composition-related outcomes (Table A.4). For comparison, Wang et al. (2016) reported 36% studies reported energy sensing-related outcomes, 40% reporting appetite-related outcomes, 37% reported glycemic-related outcomes, 30% reported dietary intake-related outcomes, and 17% reported body weight/composition. Thus, Wang et al. (2016) identified a smaller proportion of energy sensing-related outcomes (36% versus 43%), so although that outcome category was tagged for the highest proportion of our studies, this proportion was only the third highest for Wang et al. (2016), following appetite- and glycemic-related outcomes. Aside from this, the order of relative proportion of studies tagged by the remaining outcomes was identical when comparing our results with Wang et al. (2016), although the percentage of studies reporting on each outcome were not identical between the two approaches.

3.2. Summarize publication patterns

A cumulative frequency chart of the number of publications by outcome categories illustrates the publication growth of each over time (Fig. A.3). There was a consistent increasing trend in the number of publications reporting energy sensing outcomes from 1974 to 2014 and appetite from 1988 to 2014. A similar pattern was identified by Wang et al. (2016), although they identified a smaller proportion of energy-sensing outcomes overall. In comparison, publications reporting glycemic outcomes increased between 1990 and 2000 and then again from 2004 to 2014. For dietary intake, publications increased rapidly between 1988 and 1994 and then experienced consistent growth between 1996 and 2014. Publications reporting on body weight were consistent but minimal (between 1 and 3 per year) between 1976 and 2014. Cumulative numbers of publications reporting on energy sensing outcomes remained consistently higher than those reporting on appetite, glycemic, energy intake, or body weight for the entire duration (Fig. A.3). The patterns for glycemic outcomes, dietary intake, body weight, and cumulative number of publications were consistent with that reported by Wang et al. (2016).

3.3. An evidence map to identify research gaps

We generated an evidence map in the form of frequency tables and bubble plots displaying the categorical features of included studies according to study duration, outcome group, intervention, and baseline study population characteristics. We report evidence map results for the 297 studies included at 95% recall and provide evidence maps for all other recall stopping points (25%, 50%, 75%, 99%) in Supplemental Materials. The evidence map for 98% recall were identical to those for 99% recall, as no additional relevant articles were identified screening additional references from 98% to 99% recall (Fig. A.2).

We categorized the 297 studies included up to 95% recall, by intervention (LCS vs Others and LCS versus Sugars) and outcome group and included results reported from Wang et al. (2016) for comparison (Table A.5). These results demonstrate a consistent pattern with greater number of studies investigating comparisons of LCS versus sugars across all outcome categories, with the single exception of glycemic outcomes, where the majority (57%) of studies were categorized as LCS versus Others instead of LCS versus Sugars (43%). This same pattern

was seen even across studies included at 25% estimated recall, and remained similar across all other percent recalls, with the percentages changing only slightly (Supplemental Materials, Table B.3). Wang et al. (2016) reported similar results, although their included studies exhibited a more consistent pattern across all five outcome categories, with a larger proportion of studies consistently favoring LCS versus Sugars intervention.

We also categorized studies by study duration (< 1 day, 1–30 days, 1–6 months, > 6 months, and unclear/not stated) and outcome group, and included results from Wang et al. (2016) for comparison (Table A.6). The category of unclear/not stated was not originally included in Wang et al.'s categorization, but we decided post-hoc to add this category because this information was often not clear or not stated in the title/abstract of references. The results illustrate that across four of the five outcome categories, there was consistently higher numbers of acute studies, with the majority of studies < 1 day in duration, followed by studies 1–30 days in duration. Very few studies were chronic (> 6 months). Again, similar results were observed across studies that were included at 25% recall, and remained similar across all other percent recalls, with the percentages changing only slightly (Supplemental Materials, Table B.4). Generally, our results were consistent with those reported in Wang et al. (2016). The exception to this was for the Body Weight/Composition outcome, where Wang et al. (2016) reported a majority of studies (56%) with duration 1–6 months, followed by 1–30 days duration (21%) and > 6 months (18%). In contrast, we found that studies were somewhat evenly split between 1 and 30 days (35%), 1–6 months (30%), and > 6 months (26%).

The third component of the evidence map is a bubble plot (Fig. A.4), where data points are grouped and plotted according to study population baseline health status (healthy, overweight, diabetes, mixed/other) and outcome category, further stratified by intervention. Each data point represents a single study, and is randomly scattered in each grid to improve visualization of the bubble (i.e., bubble position is meaningless). The size of each bubble indicates the sample size of the corresponding study (categorized as ≤10, 11–25, 26–50, 51–100, 101–200, ≥200), with larger bubbles representing larger study sample size.

Fig. A.4 shows that most of the studies utilize either generally healthy or mixed/other study populations compared to overweight and diabetic population types across all outcome categories. The empty boxes in the plot identify the areas of research where there are no existing studies. For example, there lack studies assessing appetite outcomes in people with diabetes, studies using LCS versus Sugars interventions to assess energy sensing and glycemic outcomes in overweight people, and studies comparing interventions LCS versus Others to assess energy sensing outcomes in both overweight and diabetic participants. Conversely, the bubble plot also allows one to identify the more well-studied areas of research. For example, most studies involving diabetic participants report either on glycemic or body weight/composition outcomes. This same pattern was seen even across studies included at 25% estimated recall, and remained virtually identical across all other percent recalls (Supplemental Materials, Figs. B.2–B.5).

Overall, Wang et al. (2016) reported similar findings, notably concluding that: 1) there are more studies in generally healthy populations across all outcome categories; 2) there are a lack of studies assessing appetite and dietary intake outcomes using an LCS intervention with a sugar intake comparison for people with diabetes; 3) most studies in people with diabetes reported body weight or body composition and glycemic outcomes; 4) there are a limited number of studies investigating brain energy sensing outcomes in overweight people and diabetics among trials comparing LCS to sugar.

4. Discussion and conclusions

In this work, we have reviewed the components of a traditional evidence map and described a process for a new approach, called “rapid Evidence Mapping” (rEM). rEMs are intended to serve as a resource-efficient method for knowledge synthesis that draws from methodology that has been established for systematic review, evidence mapping, and rapid reviews. The goal of an rEM is to produce visual maps of the scientific evidence in a time efficient manner, while still utilizing rigorous, transparent and explicit methodological approaches. As outlined in our publicly-available protocol (available at: https://tinyurl.com/y7gcptqg), this approach is intended to serve as a framework to guide development of other rEMs that may be completed on any research question of interest, whether narrow or broad in scope. The resulting evidence map is a tool to visually inspect, analyze, and interpret the available body of evidence relevant to the research question, while also identifying gaps in the literature to inform future research and contextualize the design of a potential future systematic review within the broader scientific literature by identifying the areas in which studies exists and where they may be lacking.

Our application to the topic of low-calorie sweeteners in human dietary exposures and the five health outcomes of “energy sensing,” “glycemic,” “appetite,” “dietary intake,” and “body weight” was intended to illustrate the potential time and resource savings of applying an rEM approach, allowing us to compare the level of effort required to that for a traditional evidence mapping on the same topic published in 2016 (Wang et al., 2016). Importantly, this design also enabled a direct comparison of the mapping results, allowing us to determine whether the rEM approach would come to the same conclusions as traditional evidence mapping.

We initiated our rEM project in January 2018, drafting a protocol outlining the study question and methodology for our approach and implementing our literature search. We began screening titles and abstracts of the resulting records in February 2018 until reaching 95% recall in April 2018, which required screening 28% of our 8122 total references. We utilized the semi-automated machine-learning functionalities in SWIFT-Review to assist with the searching, categorization, and pattern visualization of literature search results in order to automatically summarize and visualize data from included references. This greatly reduced the amount of required manual extraction of data from the articles, although some aspects were still manual, such as the manual data extraction was required for study sample size and review of automated tagging for each category. This process was completed in April 2018. The development of the evidence map was completed in May 2018. Overall, an estimated 100 person-hours and four months of calendar time was required to complete this rEM (using literature screened up to 95% recall). We submitted this manuscript for peer-review in July 2018, resulting in a total time period of 7 months from literature search to manuscript submission (Fig. A.5).

In comparison, Wang et al. (2016) conducted their literature search in June 2014 and submitted their publication for peer-review in September 2015, for a total calendar time of 15 months (Fig. A.5). We contacted the authors to obtain an estimate of the number of person-hours to allow for a more direct comparison of the effort level; although the authors were unable to provide an exact estimate, they confirmed that the screening and data extraction for their project required 6 months calendar time and a rough time estimate of 480–960 person-hours (a team of three students and one experienced reviewer working up to 10 h/week) (Chung and Wang, 2018). In comparison, the screening and data extraction for this work required 4 months and approximately 70 person-hours. Although this indicates a potential time savings, it is challenging to directly compare between the two studies. For instance, several other confounding issues remain such as the experience of the reviewers. Wang et al. (2016) used three inexperienced student screeners which likely required training and additional time to complete the screening and extraction, whereas our two screeners were both experienced and had worked on several prior reviews. However, although it is difficult to estimate the exact time savings, these estimates provide some indication of the reduction of required resources.

Overall, a significant contributor to the potential time savings for

our rEM project likely came from the fact that we began with the already-implemented approach from Wang et al. (2016)—i.e., we adopted their existing study question and evidence map goals, and modified their existing search strategy, which was likely very time consuming, particularly with Wang et al.'s (2016) use of a large expert-based committee for these steps. However, our comparison of time estimates (Fig. A.5) is presented from implementation of the literature search to journal submission of the manuscript for each study, omitting consideration of the time to establish study questions, goals, and the literature search strategy. Therefore, this considers only the steps of the evidence mapping process in which the machine learning approach yields potential time and resource savings—i.e., through reduction of screening (using machine learning-assisted prioritization of references) and manual data extraction.

Related to the observed time savings, another important finding is the notable similarity in the overall conclusions obtained from the rapid evidence mapping at various recall stopping points. We observed remarkable consistency across the evidence map results reported at 25%, 50%, 75%, 95%, and 98/99% recall rates. Across all evidence maps, the same general patterns of endpoint, population and intervention frequencies across different tabulations appeared consistently at all recall rates, starting from 25% estimated recall (Supplemental Materials, Tables B.3 & B.4). Therefore, evaluating studies included at 25% recall would have resulted in the same conclusions regarding the current state of the science and existing research gaps. However, this would have required screening of only 495 titles and abstracts out of the 8122 total records instead of the 2267 required to achieve 95% recall (Fig. A.2). The potential time savings from this approach could be significant. Furthermore, this also makes the case for the use of rEM as a planning tool for systematic review to assist with the refinement of a focused research question where scientific evidence is known to exist. The limited time and resources required to complete an rEM, combined with the potential to further reduce these requirements by using machine learning-based prioritized screening indicate the feasibility of adopting such an approach to develop more science-based and informed systematic reviews.

For each evidence map, the majority of studies had been identified when stopping at 25% estimated recall—on average, only 12 studies were added at each step beyond screening 25% estimated recall to 99% estimated recall (i.e., 50%, 75%, 95%, 99% estimated recall) for each tabulation within evidence map. Only a small number of studies are added as one moves to higher recall percentages, and this raises the question of whether screening to a high estimated recall rate such as 99% was necessary, given that this required a screening of 4551 additional references after achieving 25% estimated recall. In particular, moving from screening to 98% estimated recall to 99% estimated recall required additional screening of 708 references (Fig. A.2), yet no additional relevant references were identified. This clearly indicates that SWIFT-Active Screener's predictions accurately and efficiently identified the vast majority of relevant articles after reviewing only a fraction of the total number of studies, and that the recall estimate in SWIFT-Active Screener tends to be conservative. The time and resource savings gained from stopping earlier in the screening process may be of useful discussion for future case studies, given the demonstration here of minimal trade-off with ultimate precision and impact on overall results.

Our overall conclusions on LCS from the rEM evidence map corroborated those reported by Wang et al. (2016) and produced similar findings regarding the areas where most studies were identified (i.e., in generally healthy populations or in diabetics reporting body weight or body composition and glycemic outcomes) and where studies were lacking (i.e., studies assessing appetite and dietary intake outcomes using an LCS intervention with a sugar intake comparison in people with diabetes or investigating brain energy sensing outcomes in overweight people and diabetics among trials comparing LCS to sugar). Although there were discrepancies in the proportion of studies reporting on certain characteristics (i.e., outcome groups) comparing our results with Wang et al. (2016), the conclusions regarding the overarching question of this work as to where current research exists and where it is lacking was remarkably consistent.

Although our overall conclusions corroborate the findings of Wang et al. (2016), we did note differences in the number of retrieved and included studies resulting from our search. For instance, Wang et al. (2016) screened 17,270 relevant citations from MEDLINE whereas our modified search retrieved 8122 relevant citations from PubMed. Although we used the same keywords, MeSH terms, and date limit reported in Wang et al. (2016), we made additional modifications to incorporate quotations around phrases to avoid automatic term substitution around each individual term, which would have instead led to a broader capture of references (see Methods). This likely led to the fewer number of studies, which in theory retrieved more relevant references applicable to the search terms.

Wang et al. (2016) ultimately included 185 references while we included (after achieving 95% recall) 297 references, with 115 studies overlapping. Therefore, Wang et al. (2016) included 70 references that we did not include, and we included 182 references that Wang et al. (2016) did not include (Supplemental Materials, Fig. B.1, Supplemental Materials, Tables B.1–B.2). Of the 70 articles included by Wang, 44 (63%) were not retrieved by our PubMed search and 26 (37%) were retrieved, screened, and excluded—11 of these were excluded by a single reviewer and 15 were excluded by dual independent reviewers (Supplemental Materials, Table B.5). An independent review by two additional screeners (JL and BH) of the 26 references that were initially screened and excluded concluded that of these, 21 (81%) were correctly excluded (either due to unavailability of abstracts, exposures that were not low-calorie sweeteners, or outcomes that were not relevant to our study question) and 5 (19%) were incorrectly excluded and should have been included in the final set of studies. Of these 5 studies, 3 had been reviewed and excluded by a single reviewer initially whereas 2 were reviewed and excluded initially by dual reviewers. These findings suggest that although there typically exists a small chance of erroneously excluding potentially relevant studies, this is not necessarily attributable to whether references are screened singly or in duplicate. Furthermore, we note that a large portion of the discrepant studies (81%) were included by Wang et al. (2016) but did not appear to meet the inclusion criteria, indicating that there were potentially some differences arising in applying inclusion/exclusion criteria between Wang et al. (2016) and our reviewers. We were unable to evaluate the 182 references that we included but Wang et al. (2016) did not, because a list of title and abstract screening results and justification was not available in the publication and was not successfully obtained by request from the authors.

To investigate the potential impact of the missed studies, we imported and tagged the 49 references included by Wang et al. (2016) that were not included in our study (44 references that were not retrieved from our PubMed search strategy plus 5 references that were erroneously excluded at our title and abstract step) and explored the 182 references that we included but not by Wang et al. (2016). Since in general our results were comparable to Wang et al. (2016) in terms of the relative frequencies across outcome, intervention, and study duration categories, we did not expect that the missed studies included in each would vary significantly in terms of relative frequencies. The relative frequency and number of studies included in this study but not Wang et al. (2016) and vice-versa are included in Table A.5 (“Lam not Wang” and “Wang not Lam,” respectively) for outcome and intervention category. These results indicate that for Appetite, Dietary Intake, Energy Sensing, and Glycemic outcomes, there are consistently more references reporting on LCS versus Sugar interventions over LCS versus Other, even for the missed studies. An exception to this is the Body Weight outcome, where the 10 studies included by Wang et al. (2016) but not in our study were mostly LCS versus Other category (n = 8). However, because the 12 studies we included but not Wang et al. (2016) were mostly LCS versus Sugar interventions (n = 7), this

resulted in similar findings overall when comparing the two results. Similar findings are seen in Table A.5, comparing the relative frequency and number of missed studies for outcome and duration category—there is consistency across all outcomes with the exception of the Dietary Intake category, but these are balanced out by the studies that we included that were missed by Wang et al. (2016). Overall, this indicates that the missing studies in each were not impactful on the overall results.

Our rEM has several limitations, most of which are addressable given appropriate resources. First, in following Wang et al.'s (2016) approach, we only searched one database (PubMed). By limiting the scientific literature retrieved to a single source and not including more diverse sources of evidence such as the grey literature, we curtail the ability of the evidence map to speak broadly about the scientific evidence available relevant to the study question of interest. A broader search of the scientific literature could have potentially retrieved a higher number of relevant studies. We also did not work with a trained librarian to develop our search terms and only screened studies by title and abstract, and utilized single-screening for records beyond the first 500 records. Therefore, there is a possibility that we missed or excluded studies that could have been relevant to our study question. These limitations could be addressed with the time and resources to work with a librarian to develop a targeted search strategy applied to multiple databases.

Furthermore, we only extracted information based on information obtained from titles and abstracts of included references. In contrast, Wang et al. (2016) extracted information from the full text of included references. Our approach likely resulted in time savings, but simultaneously limited information available for extraction and likely reduced the accuracy of extracted information. For instance, Wang et al. (2016) were able to extract the funding source of studies (although they did not appear to incorporate this information in their evidence map), but we were not able to do so because of the unavailability of this information in the abstract. Our extraction of sample size was limited to categories (i.e., ≤10, 11–20, etc.) because of the imprecise reporting of sample size in the abstracts, and even then this information could not be extracted for all included studies. However, although Wang et al. (2016) extracted information from full text, some information such as sample size was still not reported in the full text—in both abstract-only and full-text extraction approaches, 4% of references did not report the sample size (Table A.4). Furthermore, other more nuanced information that may not be clear in the abstract (such as the intervention comparison) likely contributed to the differences in proportion of intervention types reported here compared that in Wang et al. (2016). In general, although information density is highest in abstracts, much of the information contained within each section of the full text is unique (Schuemie et al., 2004) and therefore presents a significant limitation when relying solely on the information provided in the abstract. This limitation could be addressed with additional time and resources dedicated towards screening and extracting the full text of references.

Lastly, evidence mapping does not include quality or risk of bias appraisal of the included studies. Thus, it is unknown whether the included studies are of high or poor quality; therefore, although there was a high volume of included studies, it is possible that they are of variable quality, some with limited utility for a future systematic review. However, a benefit of the rEM approach is to identify particular study areas or topics where sufficient evidence exists that can inform scoping and problem formulation for a future systematic review, in which a more formal evaluation of study quality and inferences for policy- and decision-making may be made. The rEM evidence map is also specific to the study question, PECO statement, and inclusion/exclusion criteria as developed during the scoping process and therefore are most relevant to a systematic review with a similar scope.

Evidence mapping is a rapidly evolving approach to identify, collect and evaluate the characteristics of scientific evidence. Here, we have introduced a modified protocol called “rapid Evidence Mapping (rEM),” a potentially promising approach that reduces the time, effort, and resources required to complete an evidence mapping, potentially without altering the final conclusions when compared to traditional evidence mapping. Since the main goal for evidence mapping is to quickly identify research gaps as well as opportunities for systematic review, this method, as we have shown, may have the capacity to more efficiently achieve the intended results of the more labor-intensive traditional approach. Furthermore, since this method makes substantial use of machine learning and information retrieval applications and software, continued development of such tools is likely to further enhance the capacity to perform rapid Evidence Mapping in an efficient and accurate manner.

Declaration of interests

None.

Funding sources

The study had no funding source. Study authors were responsible for the study design; collection, analysis and interpretation of data; writing of the report; and decision to submit the article for publication.

Acknowledgements

The authors acknowledge the work of D. Bautz and L. Anderson in screening references for inclusion and R. Elmore in reviewing the draft manuscript.

We gratefully acknowledge the corresponding author of Wang et al. (2016), M. Chung, for providing additional information and data upon our request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2018.11.070.

References

AHRQ (Agency for Healthcare Research and Quality), 2018. Systematic Review Data Repository (SRDR). https://srdr.ahrq.gov/.

Bragge, P., Clavisi, O., Turner, T., Tavender, E., Collie, A., Gruen, R.L., 2011. The global evidence mapping initiative: scoping research in broad topic areas. BMC Med. Res. Methodol. 11, 92.

Chung, M., 2017. Personal Communication.

Chung, M., Wang, D., 2018. Personal Communication.

Colquhoun, H.L., Levac, D., O'Brien, K.K., Straus, S., Tricco, A.C., Perrier, L., et al., 2014. Scoping reviews: time for clarity in definition, methods, and reporting. J. Clin. Epidemiol. 67 (12), 1291–1294.

Duffy, S., de Kock, S., Misso, K., Noake, C., Ross, J., Stirk, L., 2016. Supplementary searches of PubMed to improve currency of MEDLINE and MEDLINE In-Process searches via Ovid. J. Med. Libr. Assoc. 104 (4), 309–312.

EFSA (European Food Safety Authority), 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA J. 861637. https://doi.org/10.2903/j.efsa.2010.1637.

Ganann, R., Ciliska, D., Thomas, H., 2010. Expediting systematic reviews: methods and implications of rapid reviews. Implement. Sci. 5 (1), 56.

Gough, D., Thomas, J., Oliver, S., 2012. Clarifying differences between review designs and methods. Syst. Rev. 1, 28.

Grant, M.J., Booth, A., 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Inf. Libr. J. 26 (2), 91–108.

Higgins, J.P.T., Green, S., 2011. Handbook for Systematic Reviews of Interventions Version 5.1. 0 [Updated March 2011]. The Cochrane Collaboration.

Howard, B.E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M.R., et al., 2016. SWIFT-review: a text-mining workbench for systematic review. Syst. Rev. 5, 87.

Ip, S., Hadar, N., Keefe, S., Parkin, C., Iovin, R., Balk, E.M., Lau, J., 2012. A web-based archive of systematic review data. Syst. Rev. 1, 15.

James, K.L., Randall, N.P., Haddaway, N.R., 2016. A methodology for systematic mapping in environmental sciences. Environ. Evid. 5, 1–13.

Khangura, S., Konnyu, K., Cushman, R., Grimshaw, J., Moher, D., 2012. Evidence summaries: the evolution of a rapid review approach. Syst. Rev. 1 (1), 10.

McKinnon, M.C., Cheng, S.H., Garside, R., Masuda, Y.J., Miller, D.C., 2015. Sustainability: map the evidence. Nature 528, 185–187.

Miake-Lye, I.M., Hempel, S., Shanman, R., Shekelle, P.G., 2016. What is an evidence map?

A systematic review of published evidence maps and their definitions, methods, and products. Syst. Rev. 5, 1–21.
NRC (National Research Council), 2011. Committee to review of the Environmental Protection Agency's draft IRIS assessment of formaldehyde. National Academies Press, Washington, DChttp://www.nap.edu/openbook.php?record_id=13142, Accessed date: 1 May 2018.
NRC (National Research Council), 2013. Critical aspects of EPA's IRIS assessment of inorganic arsenic: interim report. National Academies Press, Washington, DChttp://www.nap.edu/catalog.php?record_id=18594, Accessed date: 1 May 2018.
R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rathbone, J., Hoffman, T., Glasziou, P., 2015. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Syst Rev. 4, 80.
Rooney, A.A., Boyles, A.L., Wolfe, M.S., Bucher, J.R., Thayer, K.A., 2014. Systematic review and evidence integration for literature-based environmental health science assessments. Environ. Health Perspect. 122, 711–718.
Schuemie, M.J., Weeber, M., Schijvenaars, B.J.A., van Mulligen, E.M., van der Eijk, C.C., Jelier, R., Mons, B., Kors, J.A., 2004. Distribution of information in biomedical abstracts and full-text publications. Bioinformatics 20 (16), 2597–2604.
Snilstveit, B., Vojtkova, M., Bhavsar, A., Gaarder, M., World Bank, 2013. Evidence gap maps—a tool for promoting evidence-informed policy and prioritizing future research. Policy Research Working Paper; No. 6725, Washington, DChttp://documents.worldbank.org/curated/en/212651468163487838/pdf/WPS6725.pdf, Accessed date: 1 May 2018.
Wallace, B.C., Small, K., Brodley, C.E., Lau, J., Trikalinos, T.A., 2012. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In: Proceedings of the Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. ACM, pp. 819–824.
Wang, D.D., Shams-White, M., Bright, O.J., Parrott, J.S., Chung, M., 2016. Creating a literature database of low-calorie sweeteners and health studies: evidence mapping. BMC Med. Res. Methodol. 16 (1).
Woodruff, T.J., Sutton, P., Grp, N.G.W., 2011. An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. Health Aff. 30, 931–937.