MEDICAL EFFECTIVENESS ANALYSIS RESEARCH APPROACH[1]

Background

California Health Benefits Review Program (CHBRP) reports present three types of information about proposed health insurance benefit mandates or repeals: (1) the medical effectiveness of screening, diagnostic, treatment, and other health services addressed in the legislation; (2) the financial impacts of the legislation; and (3) the impact on public health. This document describes the seven steps in the process used to analyze medical effectiveness:

- Preparing to conduct the literature search;
- Conducting the literature search;
- Deciding whether to retrieve articles;
- Selecting articles for inclusion in the review;
- Reviewing the literature;
- Making a qualitative "call" on evidence of effectiveness in the literature; and
- Summarizing the quantifiable evidence for specific outcome.

Preparing to Conduct the Literature Search

A. CHBRP staff at the University of California, Office of the President (UCOP) receive a request from the California State Legislature to analyze a bill that would establish or repeal a health insurance mandate. An electronic copy of the bill is made available to all CHBRP faculty and staff.

B. CHBRP staff at UCOP work with CHBRP faculty and staff at UC campuses to determine who will work on the medical effectiveness, cost, and public health analyses.

C. CHBRP staff at UCOP complete a telephone call with the bill author's staff (and sometimes the bill sponsor) to clarify the bill author's intent. The items discussed in the telephone call are derived from a bill author questionnaire that contains standard questions as well as questions specific to the bill that have been posed by CHBRP faculty and staff. The medical effectiveness team reviews the responses to the bill author questionnaire and uses them to refine the specifications for the literature search.

[1] Prepared by Janet M. Coffman, MPP, PhD; Mi-Kyung Hong, MPH; Wade M. Aubry, MD; Chris Tonner, MPH; Patricia E. Franks; and Ed H. Yelin, PhD.

D. The medical effectiveness team, in consultation with other CHBRP faculty and staff, identifies a content expert for the bill. This person is an expert in a relevant clinical specialty who is knowledgeable about current clinical practice, as well as clinical controversies associated with the proposed mandate or repeal. The content expert is also usually familiar with clinical epidemiology, health services research, or evidence-based medicine. For some bills, two content experts may be retained to ensure that the team obtains expertise in several areas relevant to the bill. Examples include bills that would have required coverage for oral chemotherapy drugs (SB 161 and SB 961) and for diabetes-related complications (SB 1104). For both of these bills a physician and a pharmacist were retained to provide expertise on pertinent diseases and the medications used to treat them.

E. The content expert reviews the bill and assists the medical effectiveness team in clarifying the meaning of the clinical terms used in the proposed mandate or repeal. For example, in reviewing the literature pertaining to the analysis of Assembly Bill 1549 (2003), which addressed management of childhood asthma, the content expert explained what physicians mean by "treatment action plans" and the differences between types of action plans (i.e., peak flow-based vs. symptom-based).

F. The medical effectiveness team, in consultation with the content expert and the medical librarian, defines the scope of the literature search for medical effectiveness and develops a plan for analyzing the literature. The medical effectiveness team prepares a draft literature search specifications memo and circulates it to the medical librarian, the CHBRP staff lead, and the cost and public health team members working on the bill. These team members have several days to review and comment on the draft. The cost and public health sections of the memo contain "boilerplate" terms that the cost and public health team members working on the bill edit to reflect their literature search needs. The medical effectiveness and cost team members work together to revise the literature search terms in the section of the memo pertaining to literature on utilization of relevant treatments. The medical effectiveness team revises the memo to incorporate the input received and submits a final version to the medical librarian.

    1. The medical effectiveness team identifies the type of intervention(s) the bill addresses (e.g., is the intervention a screening, diagnostic, or monitoring test, a procedure, or a treatment?) and the literature needed to analyze the impact of the bill on patient outcomes and utilization of health care services.

    2. The team identifies the types of studies that contain information pertinent to the intervention(s). For example, if the mandate or repeal were about osteoporosis treatment, studies about the effectiveness of osteoporosis treatments would be included, but studies of the effects of primary prevention of osteoporosis would be excluded.

    3. The team, in consultation with the content expert, identifies the outcomes that the literature review will assess. If the language of a bill references specific outcomes, these outcomes will be included in the review. If the bill does not mention specific outcomes, the team and the content expert will identify outcomes most relevant to the proposed mandate or repeal. There is a preference for outcomes that are meaningful to consumers, including patient-reported outcomes, over physiological outcomes. Outcomes of particular interest to CHBRP include mortality, morbidity, quality of life, ability to perform everyday activities, and absences from school and work due to illness.

4. The medical effectiveness team, in consultation with the medical librarian and content expert, uses the following general inclusion/exclusion criteria:

    a. Include only studies for which an abstract has been published. The tight time frame for production of CHBRP reports (60 days from legislative request to completed report) compels the team to rely on abstracts as a screen to determine whether articles should be included in a literature review. Although some articles that do not have abstracts present research findings, most are commentaries, editorials, and letters to the editor that do not present the results of medical effectiveness studies and, thus, would not be included in CHBRP's literature reviews.

    b. Include only abstracts in English. The time frame for CHBRP reviews is too short to obtain translations of medical literature published in other languages.

    c. Limit the search to the population affected by the proposed mandate or repeal. For example, for the analysis of AB 1549 (2005), which concerned management of childhood asthma, "children" were defined as persons aged 0 to 18 years and studies in which a large proportion of the subjects were older than 18 years were excluded.

    d. Limit the search to the past 20 years. The team may shorten the time period, if there is a large body of literature on the topic and/or if the content expert has indicated that treatment has changed considerably over the past 20 years. The team may lengthen the time period it if there are few published studies.

    e. In cases in which CHBRP is asked to analyze a bill that is similar to a bill on which the program has previously issued a report, the search is limited to literature published since the previous report was issued.[2]

5. The team, in consultation with the medical librarian and the content expert, determines the databases to be searched.

    a. Peer-reviewed literature

        The following databases that index peer-reviewed literature are typically searched: The Cochrane Library, MEDLINE (PubMed), and Web of Science. EMBASE, a database that primarily contains international studies, is searched if searches of the aforementioned databases retrieve little literature. Other specialized databases of peer-reviewed literature, such as CINAHL, International Pharmaceutical Abstracts, PsycINFO, are searched if they are likely to contain articles relevant to the proposed mandate or repeal.[3]

---

[2] For example, in 2009 CHBRP was asked to analyze a bill (SB 158) that would mandate coverage for the human papillomavirus (HPV) vaccine. This bill was identical to a bill (AB 1429) CHBRP had analyzed in 2007. Because CHBRP had conducted a comprehensive search of literature published through 2006 for AB 1429, the search for SB 158 was limited to literature published from January 2007 through March 2009.

[3] Some material published in peer-reviewed journals has not been peer-reviewed. In particular, journals may publish guidelines issued by organizations whose work is of interest to their readers without peer review. For example, *Obstetrics & Gynecology* publishes guidelines issued by the American College of Obstetrics and Gynecology, and *CA: A Cancer Journal for Clinicians* publishes American Cancer Society guidelines. Some of these guidelines are based on opinion and may provide weaker evidence than peer-reviewed journal articles and some documents in the grey literature. As discussed in Section IV. C., the medical effectiveness team applies the same hierarchy of evidence to all literature regardless of whether it appears in peer-reviewed journals or the grey literature. In addition,

Cochrane reviews are authoritative, peer-reviewed systematic reviews that can be treated as a "gold standard" with regard to the rigor of the methods used to review the medical literature. Cochrane reviews are often narrow in focus and, thus, most helpful for analyses of bills that address a limited set of services. For more general bills, Cochrane reviews are used to supplement systematic reviews that address broader ranges of services, such as those conducted by the National Institute for Health and Clinical Excellence (NICE) [4] and the Agency for Healthcare Research and Quality's Evidence-based Practice Centers (AHRQ EPCs).

b. Grey literature

CHBRP also searches the grey literature, which consists of material that is not published commercially or indexed systematically in bibliographic databases. The grey literature is primarily composed of technical reports, working papers, dissertations, theses, business documents, and conference proceedings. The CHBRP medical effectiveness team draws upon grey literature from government agencies, scientific research groups, and professional societies for its reviews. Systematic reviews are among the types of grey literature most frequently analyzed for CHBRP reviews.

The medical effectiveness team has grouped the sources of grey literature into two hierarchical tiers based on the strength of the evidence.

First tier of the grey literature

The first tier of the grey literature includes systematic reviews and meta-analyses issued by authoritative organizations whose primary mission is to conduct objective analyses of the effectiveness of medical interventions that are used to develop evidence-based clinical practice guidelines. NICE and the US Preventive Services Task Force (USPSTF) are two of the most useful sources in this category, because these organizations commission systematic reviews that explicitly state their research questions, use standardized methods to assess the strength of evidence, and distill detailed findings into a small number of major conclusions. Other sources in this category include: the AHRQ EPCs, the Centers for Disease Control and Prevention Advisory Committee on Immunization Practices (CDC ACIP), the International Network of Agencies for Health Technology Assessment (INAHTA), the National Institutes of Health (NIH), the Scottish Intercollegiate Guidelines Network (SIGN), and the World Health Organization (WHO). These sources are always searched if they address the health care services for which a bill would mandate coverage (e.g., always search the USPSTF website when analyzing bills on screening tests). Systematic reviews and meta-analyses issued by these organizations are incorporated into CHBRP's literature review as described in Section IV. C. below. CHBRP relies most heavily on literature syntheses that present major findings from rigorous analyses of the evidence in a clear and concise manner.

---

the medical effectiveness team and the content expert apply their knowledge of pertinent guidelines, journals, etc., when selecting literature for inclusion in the literature reviews.

[4] NICE commissions other organizations, such as the National Collaborating Centre for Women's and Children's Health, to produce evidence-based guidelines on some topics.

Second tier of the grey literature

The second tier of grey literature consists of clinical practice guidelines issued by medical and scientific societies. They are often based on expert opinion, although some are evidence-based. The merit of these guidelines stems from the authoritative reputation of the societies. Such guidelines include those issued by AACE (American Association of Clinical Endocrinologists), AAP (American Academy of Pediatrics), AAPD (American Academy of Pediatric Dentistry), ACOG (American College of Obstetricians and Gynecologists), ADA (American Diabetes Association), APA (American Psychiatric Association), and the National Comprehensive Cancer Network (NCCN). Decisions about searches of professional society Web sites for guidelines are made on a case-by-case basis. Decisions are based on the following criteria: knowledge of the medical effectiveness team and content expert regarding guidelines issued by pertinent professional societies, the strength of evidence available from other sources, and whether the bill explicitly references a guideline or is derived from a guideline. See section I. F. 5. c. below for details.

c. Clinical practice guidelines

CHBRP has developed the following criteria to determine whether and how clinical practice guidelines should be incorporated into its medical effectiveness reviews.

Bills that reference clinical or national practice guidelines

In cases where:

- A bill mandates coverage for an intervention that is "consistent with national guidelines,"; or
- A guideline is an obvious source of bill language; or
- A guideline is specified in the bill.

In cases where:

The medical effectiveness team will select studies for inclusion per CHBRP's hierarchy of evidence (discussed in Section IV.A., below) and also will assess relevant guidelines.

Bills that DO NOT reference clinical practice guidelines

The medical effectiveness team will follow CHBRP's hierarchy of evidence, which ranks clinical practice guidelines below other sources of evidence regarding medical effectiveness. Systematic reviews and meta-analyses that are part of a guideline may be reviewed separately per the hierarchy of evidence. If a guideline appears to be evidence-based and relevant to the issue, the medical effectiveness team may reference it in the text. In a case where little or conflicting information about the issue is available, the medical effectiveness team may cite guidelines with appropriate caveats noted (i.e., strength of evidence, guideline author, etc).

For bills for which the medical effectiveness team determines that clinical practice guidelines should be reviewed, the National Guideline Clearinghouse (NGC) is always searched to identify pertinent guidelines. The medical effectiveness team uses

NGC's summaries to screen guidelines and retrieves the full text of guidelines it selects for inclusion in the literature review.

Web sites maintained by organizations that issue clinical practice guidelines are also searched, because NGC has several important limitations. NGC relies on voluntary submissions and, as a consequence, does not index all guidelines. Some of the most authoritative guidelines are not indexed by NGC. In addition, the quality of the evidence presented in guidelines indexed by NGC varies. Some guidelines are based on systematic reviews of peer-reviewed literature, whereas others are based on expert opinion. In addition, NGC's summaries of guidelines are not as authoritative or as exhaustive as the full guidelines.

G. The medical effectiveness team, content expert, and medical librarian take into account both the literal meaning and intent of the proposed mandate or repeal when developing the strategy for the literature search.

1. Some mandates and repeals address coverage for multiple types of services (e.g., medical treatment, medical supplies, physical therapy, and counseling). In such cases, the literature search will be designed to retrieve literature on all types of services to which a mandate or repeal would apply.

2. For some bills, the medical literature may be assessed in segments because it addresses a wide range of diseases and conditions. For example, if a proposed mandate or repeal addressed cancer screening, the team would need to separately analyze literature on screening of multiple types of cancer (e.g., breast, colorectal, lung, and prostate).

3. Screening, diagnostic, monitoring, and treatment interventions require different analytic approaches. For example, a treatment is typically designed to cure a disease or improve function, and designing trials to assess how well the treatment works may be relatively straightforward. On the other hand, a screening test might indicate an increased risk of a disease. This may lead to recommendations for one or more types of preventive interventions. The interventions may vary in their effectiveness, and the disease, which may or may not occur even if the result of the screening test is positive, may be treated in various ways.[5] Thus, an effectiveness assessment of an intervention will have to be built upon information available from various parts of the "evidence chain." To assess each of these links, information needs to be collected over a long period of time. Testing and treatment options continually change over time, and studies that directly address all effectiveness questions pertinent to a bill may not exist.

4. Some bills may concern the terms of coverage for different types of services rather than coverage for individual health care services per se. Examples include SB 572 (2005), which addressed parity in coverage of physical and mental health services, and SB 1198 (2008), which concerned parity in coverage for durable medical equipment. For parity bills, the medical effectiveness analysis focuses on evidence of the effects of parity, such as the effects of reduction in cost sharing on utilization of health care services and health

[5] For example, a screening test may indicate that a person has high cholesterol. Based on this result, his or her physician may recommend exercise, dietary changes, and/or medication. These preventive interventions may or may not lower the person's cholesterol or prevent him or her from developing heart disease. If he or she develops heart disease, his or her physician may recommend one of several treatments which may or may not be successful.

status, to the extent literature is available on these topics. Other bills that have addressed the terms of coverage include AB 1826 (2010), which would have prohibited "fail-first protocols" for pain medication. For this bill, the medical effectiveness team reviewed the literature on the impact of "fail-first protocols."

5. Some bills address more treatments or conditions than the medical effectiveness team can analyze within 60 days. For example, AB 219 (2013), a bill regarding coverage for oral anti-cancer medications, would have affected coverage for 54 medications that are used to treat over 50 cancers. In such cases, the medical effectiveness team assigned to a bill will work with other members of the analysis team to develop a feasible research approach. For example, for AB 219, the medical effectiveness team provided readers with general descriptive information regarding oral anti-cancer medications but did not analyze the literature on the effectiveness of any of these medications.

Conducting the Literature Search

A. The medical librarian conducts the search and contacts the medical effectiveness, cost, and public health team members working on the bill regarding questions as they arise.

B. The medical librarian provides the initial search results to the team in EndNote to the maximum extent feasible. All citations to peer-reviewed literature should be included in the EndNote file. Ideally, citations to the grey literature should be included as well, but this may not be feasible in cases in which the number of citations to the grey literature is large.

C. The medical librarian records all search terms, including Medical Subject Headings (MeSH) terms and key words.

D. The team assesses the extent to which the results of the literature search address the questions and issues underlying the proposed mandate or repeal, consulting the content expert as needed. If the initial literature search returns few results, the search criteria will be reexamined, and the medical librarian will run additional or modified searches, or the lead analyst on the medical effectiveness team will search articles from the reference lists of articles that have already been retrieved to determine if they contain any additional articles pertinent to the bill.

Deciding Whether to Retrieve Articles

A. At least two medical effectiveness team members review all abstracts returned by the search to identify articles for which the full text will be retrieved.[6] Criteria for excluding articles may include: (1) duplicate studies: (2) study subjects who are not representative of Californians who would be affected by the mandate or repeal: and (3) articles that describe interventions but do not assess their effectiveness.

B. For utilization outcomes, only studies conducted in the United States are selected. When an outcome is likely to depend on specific aspects of the US health care system, such as the effect of pediatric asthma education on emergency department visits, the results may be

[6] This approach risks excluding useful articles based on their abstracts. This risk is necessary, given the short time frame for CHBRP reports. However, abstracts often overstate, rather than understate, authors' findings.

affected by policies and norms of "usual care" that differ in other countries. However, if the outcome of interest concerns health status, international studies are included.

C. Once a full-text article is retrieved,[7] the team reapplies the initial inclusion/exclusion criteria to ensure the study is relevant to the proposed mandate or repeal.

D. There may be instances in which the full text of an article cannot be retrieved quickly enough to meet the timeline for a CHBRP review. In these instances, the team relies on the published abstract. Reliance on an abstract may omit information relevant to a CHBRP review, including some of the study's results and information about the characteristics of the study population. The team keeps a log of articles that appear relevant, but for which full text was not available in time for inclusion in the draft report circulated for review. If articles arrive after the due date for the draft report, they will be examined to determine whether they would substantively alter the team's conclusions. If the conclusions would change, the report is revised accordingly.

Selecting Studies for Inclusion in the Literature Review

A. Hierarchy of Evidence

In general, the medical effectiveness team faculty and staff adhere to the following hierarchy of evidence when determining which articles to include in a review:

1. High-quality meta-analyses[8]—particularly those included in the Cochrane Library.
2. Systematic reviews—particularly those performed by authoritative organizations, such as the AHRQ, NICE, USPSTF, and other government agencies (e.g., NIH, CDC, and the Centers for Medicare & Medicaid Services).
3. Well-designed randomized controlled trials (RCTs) and cluster RCTs.[9]

---

[7] The team retrieves full-text articles available on the Internet through the University of California libraries. If an article is not available online, but is available in hard copy at the UCSF library (or the UCSD library in cases in which a medical effectiveness analysis is completed by the UCSD team), a team member retrieves the article from the library. If an article is not available at UCSF or UCSD, the team requests the article through interlibrary loan, from the journal's website, or a commercial document delivery service.

[8] "High-quality" meta-analyses are meta-analyses that have clear objectives and hypotheses, apply appropriate inclusion/exclusion criteria, assess meaningful outcomes, and use sound methods to find, select, and evaluate studies and to generate pooled estimates of an intervention's effects. In general, results of meta-analyses of randomized controlled trials (RCTs) are likely to produce more valid estimates than meta-analyses of observational studies, because randomization of subjects reduces the risk of selection bias. In addition, meta-analyses with large numbers of observations (i.e., where the sum of observations from all studies included in a review is large) are likely to yield more valid estimates than meta-analyses with small numbers of observations because they have greater power to detect effects. *Cochrane Handbook for Systematic Reviews of Interventions 4.2.5*, Chichester, UK: John Wiley & Sons, 2005, p. 97-99; Egger M, Schneider M, Smith GD. Meta-analysis: Spurious precision? Meta-analysis of observational studies. *British Medical Journal* 1998;316:140-144. Egger M, Smith GD, Phillips AN. Meta-analysis: Principles and procedures. *British Medical Journal* 1997;315:1533-1537; Flather MD, Farkouh ME, Pogue JM, Yusuf S. Strengths and limitations of meta-analysis: Larger studies may be more reliable. *Controlled Clinical Trials*. 1997;18:568-579.

[9] "Cluster RCTs" are studies in which subjects are randomized in groups rather than as individuals. This research design is typically used in situations in which the intervention is administered to groups of subjects or in which randomization at the individual level may lead to contamination of the control group (i.e., inadvertent exposure to the intervention).

4. RCTs and cluster RCTs with major weaknesses.
5. Nonrandomized studies with comparison groups and time series analyses.
6. Case series and case reports.
7. Clinical practice guidelines and narrative reviews (i.e., "grey beard reviews").[10]

B. Implementing the Hierarchy of Evidence

1. If published meta-analyses and/or systematic reviews are available, the team generally uses them as the principal source of information for the review. The remainder of the review is then limited to individual studies published after the articles included in the meta-analyses and/or systematic reviews. For example, if a meta-analysis was published in June 2001 and included studies published up to December 1, 2000, the team would focus on individual studies published on or after December 1, 2000.
2. The team reviews published meta-analyses and/or systematic reviews for consistency. If there are several meta-analyses and/or systematic reviews that reach different conclusions, the team will consult with the content expert to identify possible explanations (e.g., the inclusion/exclusion criteria of the meta-analyses and/or systematic reviews vary, one or more meta-analyses and/or systematic reviews do not use rigorous methods). In some cases, the results of one or more meta-analyses and/or systematic reviews may be discounted. The rationale for discounting is discussed in the report.
3. If no applicable meta-analyses and/or systematic reviews are available, the medical effectiveness team proceeds down the hierarchy of evidence.
4. Where meta-analyses and/or systematic reviews are available, narrative (unsystematic) reviews are excluded from CHBRP's medical effectiveness reviews. However, when literature regarding a disease and intervention is sparse, the medical effectiveness team includes narrative reviews (e.g., AB 163 (2009) on amino-acid based elemental formula; AB 30 (2007) on inborn errors of metabolism).
5. Strict adherence to the hierarchy of evidence may not be possible or advisable in all cases. For example, if a mandate or repeal addresses coverage for a new screening test and there are meta-analyses of the sensitivity and specificity of the test, but only nonrandomized studies of the test's effects on utilization and clinical outcomes, the meta-analyses cannot fully substitute for the nonrandomized studies. The rigor of the former studies must be balanced against the relevance of the latter.[11]

---

[10] Clinical practice guidelines are ranked below other sources of evidence because strength of the evidence on which they are based varies widely. Some guidelines contain recommendations based on meta-analyses, systematic reviews, or multiple RCTs, whereas others are based solely on expert opinion. This wide variation exists across organizations that issue guidelines and among guidelines issued by individual organizations. For example, a study of guidelines issued by the American College of Cardiology and the American Heart Association found that most recommendations contained in these guidelines were based on expert opinion and only that 11% were based on evidence from meta-analyses or multiple RCTs. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC. Scientific evidence underlying the ACC/AHA clinical practice guidelines. *Journal of the American Medical Association*. 2009; 301:831-841.

[11] CHBRP's analysis of AB 259, a bill that would allow women to obtain services from a certified nurse midwife (CNM) directly without a physician's referral, illustrates the trade-off between rigor and relevance. Most RCTs on the effectiveness of midwives that have been conducted in developed countries were carried out in Australia,

C. Use of Grey Literature

1. The hierarchy of evidence is applied in a consistent fashion to both the peer-reviewed literature and the grey literature. Systematic reviews and clinical practice guidelines are the most frequently cited types of grey literature.
2. The medical librarians conduct literature searches jointly for grey literature and peer-reviewed literature, and are instructed to search for those sources of grey literature most likely to publish high-quality literature syntheses. For further discussion of literature search methods, see Section II: *Conducting the Literature Search* (pgs. 4-6).
3. Grey literature and peer-reviewed literature about the medical effectiveness of an intervention may contain varying levels of detail. For example, some organizations that develop clinical practice guidelines, such as the USPSTF, publish summaries in peer-reviewed journals and the full guidelines and associated systematic reviews as grey literature. In such cases, the grey literature version of the guideline is reviewed to obtain additional detail not found in the peer-reviewed version.

D. Selecting Studies for Inclusion in the Utilization Literature Section

1. The medical effectiveness team and the cost team will independently review the literature identified in the literature search that addresses the impact of coverage on utilization and select studies for inclusion in their respective sections. In the event that the teams use different criteria for selecting literature to include in the report write-up, a discussion of these discrepancies will be included in the cost section.

Reviewing the Literature

A. The medical effectiveness team will generally not have time to undertake as detailed a review of the methods and quality of individual studies as the authors of a meta-analysis can.

B. Once articles have been selected for inclusion in the review, the team prepares a table that records information from each article regarding the study's research design, the population studied, the location in which the study was conducted, and the intervention and comparison groups. This table appears in an appendix to the report. Table 1 presents an example of the information recorded for studies included in CHBRP's report on AB 264, a bill that would have mandated coverage for pediatric asthma self-management.

C. Some of the full-text articles retrieved may ultimately be excluded from the review if the medical effectiveness team, in consultation with the content expert, determines that the study is not relevant to the proposed mandate or repeal, is not generalizable to the population

---

Canada, New Zealand, and the United Kingdom. Midwives in these countries work within health care systems that are quite different from that of the United States. The level and type of education mandated for midwifery practice in these countries also differs from that required of CNMs in the United States. The medical effectiveness team decided that its literature review for this bill should go beyond RCTs to also include observational studies with comparison groups that were conducted in the United States (CHBRP 2009e). Although the observational studies are weaker methodologically (in particular, they may be subject to selection bias), their findings are more generalizable to the providers to which the bill would apply (i.e., CNMs) than non-U.S. studies.

addressed by the mandate or repeal, or has major methodological problems that affect the validity of its findings.

Table 1. Summary of Published Studies on the Effectiveness of Pediatric Asthma Self-Management and Training Interventions

Citation	Type of Trial*	Intervention vs. Comparison Group	Population Studied	Location
Huss et al., 2003	Level III	Education and computer-based instructional asthma game vs. education alone	Inner-city children	Baltimore, MD
Krishna et al., 2003	Level II	Internet-enabled, interactive multi-media asthma education, conventional education, and asthma action plans vs. conventional education and asthma action plans	Children who visited a pediatric pulmonary clinic	St. Louis, MO
LeBaron et al., 1985	Level II	Education vs. usual care	Children treated at private pediatric allergy practices whose families had a wide range of incomes	San Antonio, TX

Note: *Level I=Well-implemented RCTs and cluster RCTs, Level II=RCTs and cluster RCTs with major weaknesses, Level III=Nonrandomized studies that include an intervention group and one or more comparison group, time series analyses, and cross-sectional surveys, Level IV=Case series and case reports, Level V=Clinical/practice guidelines based on consensus or opinion.

D. As indicated in Section I.F., above, in the cases where (1) a bill may mandate coverage for an intervention that is "consistent with national guidelines", (2) a guideline is an obvious source of bill language, or (3) a guideline is specified in the bill, the medical effectiveness team will select studies for inclusion per CHBRP's hierarchy of evidence and also will assess relevant guidelines. In addition, the medical effectiveness team will construct a table that summarizes and rates pertinent guidelines according to CHBRP criteria.

   The rating system was developed prior to the 2011analytic season. Since that time, CHBRP has not been asked to analyze a bill that met CHBRP's criteria for including an assessment of relevant guidelines. If in the future CHBRP is asked to analyze a bill of this sort, the medical effectiveness team will proceed as described above. Based on the rating system, the medical effectiveness team may include a discussion of the consistency of the medical effectiveness review's conclusions with guidelines.

Making a Qualitative "Call" on Evidence of Effectiveness in the Literature

A. In a conference call or group meeting, the medical effectiveness team members review the results of relevant studies for each outcome and decide collectively, based on the weight of the evidence available, on the effectiveness of the intervention across three dimensions.

B. In making a "call" for each outcome measure, the team considers the number of studies as well the strength of the evidence. To grade the evidence for each outcome measured, the team uses a grading system that has the following categories:

- Research design;
- Consistency of findings; and
- Generalizability of findings to the population whose coverage would be affected by a mandate.

Each of these categories is described below along with the criteria that are used to classify studies within each category. Once studies have been classified within categories, a conclusion about the medical effectiveness of an intervention can be made. The language that is used to describe the medical effectiveness team's overall conclusion regarding the medical effectiveness of the intervention is also discussed.

1. Research Design

This category contains information about the strength of the research designs of individual studies that evaluate an intervention's effect on an outcome of interest. Studies are assigned to one of five levels adapted from ranking systems developed by the American College of Chest Physicians and the North American Spine Society.[12] ***The levels refer to the strength of the research designs of individual studies. They do not refer to the overall strength of the evidence regarding an intervention's effect on an outcome.*** Level I studies have the strongest research designs and Level V studies have the weakest research designs. The five levels are as follows:

- Level I: Well-implemented RCTs and cluster RCTs (Strong RCTs);
- Level II: RCTs and cluster RCTs with major weaknesses (Weak RCTs);
- Level III: Nonrandomized studies that include an intervention group and one or more comparison groups and time series analyses;
- Level IV: Case series and case reports; and
- Level V: Clinical practice guidelines and narrative reviews.

Level I groups RCTs and cluster RCTs because either research design may be more or less appropriate than the other depending on the intervention studied. The RCT design is more appropriate than the cluster RCT design when an intervention is delivered to individuals and is provided in such a manner that the control or comparison group is unlikely to be inadvertently exposed to the intervention. Conversely, the cluster RCT

[12] Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of Evidence and Clinical Recommendations on Use of Antithrombotic Agents (Third ACCP Consensus Conference on Antithrombotic Therapy). *Chest*. 1992;102(4):305S-311S. North American Spine Society. Levels of evidence for primary research question. www.spine.org/forms/LevelsofEvidenceFinal.pdf. Accessed on October 4, 2006.

design is more appropriate when an intervention is delivered to groups of individuals or in situations in which the control or comparison group could be contaminated.[13]

"Well-implemented RCTs and cluster RCTs" are defined as studies that have: (1) sample sizes that are sufficiently large to detect statistically significant differences between the intervention and control groups (100 or more subjects); (2) low attrition rates (less than 20%); (3) made use of intent-to-treat methods;[14] and (4) intervention and control groups that are statistically equivalent prior to the intervention, with respect to baseline measures of the outcome and important factors associated with the outcome. To be considered well-implemented, a cluster RCT must also use appropriate statistical methods to determine whether observations are clustered at the level at which randomization occurs and, if so, to adjust for clustering. Such adjustment is necessary to ensure that the statistical significance of findings is not overstated.

Level II includes RCTs and cluster RCTs that have major weaknesses, such as small sample sizes, high attrition rates without use of intent-to-treat methods, or intervention and control groups that are not statistically equivalent at baseline and, in the case of cluster RCTs, do not test for clustering of observations and adjust for clustering if it is present.

Levels III through V are used to classify studies in which subjects are not randomly assigned to either an intervention or a comparison group. Studies that do not randomize subjects are not as well designed as RCTs for assessing the efficacy of an intervention (i.e., detecting causal inference), because they do not control for selection bias.[15]

Level III encompasses time series analyses and nonrandomized studies that have intervention and comparison groups. Time series studies analyze multiple observations on subjects before and after exposure to an intervention, which enables researchers to separate the effects of interventions from other factors that influence trends in outcomes

[13] For example, the RCT design can be easily used for studies of pharmaceuticals because drugs are dispensed to individuals and because drugs and placebos can be made to appear identical. However, the RCT design is problematic for health education classes taught to children in schools, because children who receive the intervention and their teachers may interact with children in the control group and their teachers. Such interaction could involve sharing of knowledge about self-management that might lead to changes in self-care behavior among children in the control group, which would limit the study's ability to discern differences between the intervention and control groups. In such cases, a cluster RCT design under which schools rather than children are randomized would be more appropriate than an RCT design.

[14] Intent-to-treat analysis addresses the problem of attrition bias by preserving randomization. If a study has a high rate of attrition, the persons in the intervention group who receive the full treatment may be systematically different from persons who drop out of the study. For example, persons who believe the treatment is not helpful may be more likely to drop out. In such cases, analyzing data only for those persons who completed the study could lead researchers to overestimate the effectiveness of the treatment. Intent-to-treat analysis eliminates this bias because all subjects are included in the groups to which they were randomized regardless of whether they received the full treatment. Some experts in intent-to-treat analysis believe it is sufficient to analyze data only for those subjects for whom complete data are available, whereas others believe that data should be imputed for subjects for whom data are missing (Cochrane Collaboration. *Cochrane Handbook for Systematic Reviews of Interventions Version 4.2*.5. Oxford, UK: The Cochrane Collaboration, 2005).

[15] Selection bias is a formal term used to characterize situations in which the intervention and control groups are not equivalent except for exposure to the intervention due to some consistent factor that is not measured.

over time. Nonrandomized studies with comparison groups include quasi-experimental studies, cohort studies, case-control studies, and before-after studies. In cases in which most studies of an outcome are nonrandomized studies with comparison groups, the effectiveness team will parse these studies to distinguish studies with stronger and weaker research designs.

Level IV studies are those without comparison groups. This level encompasses cross-sectional studies of a single group of subjects exposed to an intervention and case reports on individual subjects exposed to an intervention.

Level V consists of clinical practice guidelines and narrative reviews.

Meta-analyses and systematic reviews are assigned to the research design level to which most of the studies reviewed correspond. For example, the meta-analyses included in the effectiveness review on Alzheimer's drugs for SB 415 (2004) would be classified as Level I, because most of the studies synthesized in these meta-analyses were well-implemented RCTs. In contrast, a systematic review of multiple types of prosthetic ankle-foot mechanisms that was examined for the report on AB 2012 (2006) would be classified as Level IV, because most studies included in that review were cross-over studies that compared the effects of two or more prosthetic ankle-foot mechanisms on a single group of subjects.

A research design level is assigned to each article included in a medical effectiveness review for a CHBRP report. The articles are aggregated by level for each outcome assessed and the aggregate results are reported in a summary table that appears in the effectiveness section of the text of the report.

The numbers of studies at each level reflect the studies included in a medical effectiveness review and not necessarily the totality of studies on the topic. For some bills, CHBRP relies primarily on meta-analyses, systematic reviews, RCTs, or cluster RCTs, and does not consider studies lower in the hierarchy.

2. Consistency

CHBRP evaluates consistency of findings across three dimensions:

- Statistical significance;
- Direction of effect; and
- Size of effect.

a. Statistical Significance

Statistical significance is an important consideration in assessing the effectiveness of an intervention. If a finding is statistically significant, one has greater confidence that it did not occur by chance. CHBRP considers a finding to be statistically significant if there is a 95% or greater probability that a difference in outcomes between the intervention and control or comparison groups did not occur by chance (i.e., if the p value is 0.05 or less). The 95% confidence interval is a conventional threshold for

determining statistical significance. Most studies report the results of formal tests of statistical significance, although some case reports and studies with very small samples do not.

Each study that assesses an outcome will be assigned to one of three categories:

- Finding was statistically significant; or
- Finding was not statistically significant; or
- Results of a test of statistical significance were not reported.

The studies are then grouped by the three categories and the numbers of studies in each category are reported in the summary table that appears in the effectiveness section of the text of the report.

In cases in which most studies of an outcome report have strong research designs and report the 95% confidence intervals around point estimates of effects, the medical effectiveness team also examines the 95% confidence intervals to determine how similar the results are across studies.

b. Direction of Effect

The direction of the relationship between an outcome and an intervention indicates whether the intervention has a favorable effect on the outcome. A favorable effect may be an increase or a decrease in an outcome depending on the nature of the outcome and the intended effect of the intervention. For example, one would expect a drug for Alzheimer's disease to improve cognitive outcomes, whereas one would expect a biological medication for rheumatic disease to reduce joint pain and swelling. In some cases, there may be no relationship between an outcome and an intervention.

For each outcome, studies that address the outcome are categorized into three groups based on the direction of the effect:

- Intervention associated with better outcomes for the intervention group; or
- Intervention had no effect or negligible effect; or
- Intervention associated with poorer outcomes for the intervention group.

The "no effect or negligible effect" category includes studies in which the intervention had no effect on the outcome and studies in which the effect was very small, regardless of whether it was statistically significant. Examples of negligible effects found in studies previously reviewed by CHBRP include a 1% difference in severity of asthma symptoms, a 2% difference in scores on an instrument measuring cognitive functioning of persons with Alzheimer's disease, and a 0.7% difference in the performance of hearing aids.

Once individual studies have been coded they are grouped by the three categories. The numbers of studies in each category (i.e., better outcomes, no or negligible effect,

and poorer outcomes) are reported in the summary table that appears in the effectiveness section of the report.

c. Size of Effect/Clinical Significance

Policymakers need to know whether an intervention's effect on an outcome is large enough to be meaningful to patients and/or their caregivers.[16] The minimum clinically meaningful effect depends on the disease or condition addressed in a bill, the outcome of interest, and the manner in which the outcome is measured. In general, the minimum clinically meaningful effect is greater for diseases and conditions for which effective treatments are widely available than for terminal or severely debilitating illnesses for which no other treatments exist. With respect to measurement, a difference of two points may be very meaningful for an outcome measured by a single question on a five-point Likert scale, but probably is not meaningful for an outcome measured by an instrument that has multiple items and a maximum score of 100 points. For all outcomes assessed, the medical effectiveness team consults the content expert to determine whether minimum clinically meaningful effects have been established through research or expert opinion.[17]

The measures used to assess clinical significance vary across outcomes depending on the availability of research on minimum meaningful differences and the measures used in studies of the intervention in question.

CHBRP cites the effects reported in studies included in its reviews. Some studies report continuous outcomes (e.g., differences in means or medians), whereas others report binary outcomes (e.g., percent changes, relative risks, odds ratios). Statistically significant point estimates are cited in the text. Both point estimates and confidence intervals are reported in the tables. Where minimum clinically meaningful effects have been established, the team will note in the text whether the effects reported by the studies included in the review meet or exceed minimum clinically meaningful effects.

The medical effectiveness team's conclusions regarding the statistical significance, direction, and size of effects are based on findings reported in studies published in peer-reviewed publications. These conclusions may be overstated in cases in which there is bias in the reporting of research findings. Forms of bias include publication bias, multiple publication bias, citation bias, and language bias. Studies have found

[16] Statistical significance and the size of an effect are related, but not synonymous. For example, the apparent effect in a diet study may be large, e.g., a 20-pound weight reduction, but measured with such imprecision due to small sample size that it could also be a weight increase. Perhaps more importantly, a very large study might show statistically significant effects that are not meaningful. For example, with a sufficient number of cases, a new diet might show convincingly that it achieves an average weight reduction of one pound—perhaps statistically significant, but not a meaningful effect.

[17] An example of a research-based approach to determining minimum meaningful effects is the American College of Rheumatology (ACR) Response Rate clinical scoring system that was used in many of the studies synthesized in CHBRP's report on SB 913, which would have mandated coverage for biological medications for rheumatic disease. Under the ACR-20 instrument used in many of these studies, a medication was determined to have a meaningful effect if patients experienced a 20% reduction in the number of tender joints, the number of swollen joints, laboratory test results, and patient and physician assessment of severity of disease.

that some journal editors are more likely to accept studies with statistically significant and favorable findings, and that some researchers are more likely to submit statistically significant findings for publication. Multiple publication bias arises when researchers publish findings for a group of patients multiple times, as was the case in the literature CHBRP analyzed on transplantation services for persons with human immunodeficiency virus. Citation bias occurs when studies with statistically significant findings are cited more frequently than studies with nonsignificant findings and, thus, more easily retrieved when searching for studies. Language bias is an especially important challenge for CHBRP, because CHBRP reviews are limited to studies published in English. Studies conducted in countries in which English is not the primary language are more likely to be published in English-language journals if their findings are statistically significant.[18]

The extent and nature of bias probably vary across topics. The problem is probably greatest where most studies are funded by industry and where most studies have weak research designs. However, except for the few topics on which empirical studies have been published, the magnitude and consequences of bias are unknown. The 60-day time frame for CHBRP reports precludes the team from undertaking its own research to determine whether unpublished studies (i.e., studies not published by commercial publishers or issued by government agencies, professional associations, or other organizations) exist and assess their impact on the team's conclusions.

The team inserts a brief paragraph in every CHBRP report that states that our conclusions are based on the best available evidence from peer-reviewed and grey literature. The paragraph also indicates that unpublished studies are not reviewed because the results of such studies, if they exist, cannot be obtained within the 60-day timeframe for CHBRP reports.

3. Generalizability

   Generalizability refers to the extent to which a study's findings can be generalized to a population of interest. For CHBRP, the population of interest is the segment of California's diverse population to which a proposed mandate or repeal would apply. Although some studies enroll persons who are very similar to the population addressed by a proposed mandate or repeal, others enroll different populations (e.g., adults vs. children) or populations with different health care needs than many persons to whom an intervention is typically provided (e.g., persons who are less severely ill or do not have co-morbidities). Findings from studies that enroll persons who are different from the population to which a mandate or repeal would apply are less useful in determining whether a mandate or repeal would benefit Californians, even if the studies are well-designed and report statistically and clinically significant findings that favor the

[18] The information presented in this paragraph was derived from the following sources: Cochrane Collaboration. *Cochrane Handbook for Systematic Reviews of Interventions Version 4.2*.5. Oxford, UK: The Cochrane Collaboration, 2005; Lee KP, Boyd EA, Holroyd-Leduc JM, Bacchetti P, Bero LA. Predictors of publication: characteristics of submitted manuscripts associated with acceptance at major biomedical journals. *Medical Journal of Australia*. 2006; 184:621-626; Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. Chichester, UK: John Wiley & Sons, LTD, 2000; Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of effect on publication bias on meta-analyses. *British Medical Journal*. 2000; 320:1574-1577.

intervention. However, concerns about generalizability must be balanced against the need to provide information about medical effectiveness to the Legislature. It is unrealistic to restrict literature reviews only to studies that enroll Californians similar to persons to whom the mandate or repeal would apply because doing so could lead to an undersampling of studies of a treatment or technology.

The medical effectiveness team addresses generalizability in two ways. First, the team selects studies for inclusion in reviews that are most likely to be generalizable to the population to which a mandate or repeal would apply. To the extent possible, the parameters for the literature search are set to retrieve studies that enroll persons similar to those to which a proposed mandate or repeal would apply. For example, the search for AB 264 (2006), a bill on pediatric asthma education, was limited to studies that enrolled children. Once the literature search is completed, the team takes generalizability into account when selecting studies for inclusion in the review. For AB 264, the team included only studies conducted in the US, because several of the most important outcomes concerned use of health care services. For AB 259 (2009), the medical effectiveness team decided that its literature review for this bill should go beyond RCTs conducted in other developed countries to also include observational studies with comparison groups that were conducted in the United States because the findings from the US studies were more likely to be generalizable to California.

Once studies are selected for inclusion in a review, the team screens them to assess the degree of generalizability to the population to whom a mandate or repeal would apply. Findings regarding the generalizaiblity of studies are summarized in the text of the report. It is unlikely that a review would include studies that are not at all generalizable to the population that would be affected by a mandate or repeal, because such studies should have been excluded from the review.

4. Conclusion

   The last step in evaluating the evidence of medical effectiveness involves making an overall conclusion regarding the strength of the evidence based on research design, consistency of findings, and generalizability of findings to the population whose coverage would be affected by the proposed mandate or repeal). The following terms are used to characterize the body of evidence regarding the medical effectiveness of the intervention on the outcome:

   - Clear and convincing evidence.
   - Preponderance of evidence.
   - Ambiguous/conflicting evidence.
   - Insufficient evidence.

Table 2. Summary of Proposed Revisions to the ME Grading System

	Clear and Convincing	Preponderance	Ambiguous/ Conflicting	Insufficient
Research Design	Multiple RCTs	2+ studies with a comparison group	Any	All uncontrolled observational studies
Consistency	Large majority have similar findings	≥60% have similar findings	<60% have similar findings	Any
Generalizability	Most are highly generalizable	Most are generalizable	Any	Any
Cumulative Impact of Evidence	Additional RCTs would not alter conclusion			
		Preponderance of Evidence Levels Strong research designs Moderate research designs Weak research designs		

a. Criteria for Grading Bodies of Evidence as "Clear and Convincing Evidence"

Bodies of evidence should be graded as "Clear and Convincing Evidence" if all of the following conditions are met:

- **Research Design:** There are multiple RCTs of the intervention (meta-analyses or systematic reviews of these RCTs are not required, although having such syntheses would strengthen the evidence regarding an intervention's effect).
- **Consistency:** The large majority of studies have similar findings with respect to statistical significance, direction of effect, and size of effect.[19]
- **Generalizability:** The studies are highly generalizable to the intervention in question and the population whose coverage would be affected by a proposed mandate.
- **Cumulative Impact of Evidence:** It is unlikely that publication of additional RCTs would change the medical effectiveness team's conclusion about the effectiveness of the intervention.

Use of the grade "Clear and Convincing Evidence" is limited to bodies of evidence that include RCTs because even the best designed nonrandomized studies cannot fully control for selection bias. The requirement for multiple RCTs recognizes that multiple studies are needed to determine whether there is a consistent pattern of findings across studies. A specific number of RCTs is not required because the strength of evidence depends on multiple factors, including the number of studies, their sample sizes, and how well they are implemented (e.g., whether randomization

[19] The conclusion could be that the intervention has a beneficial or detrimental effect on a pertinent outcome or that it does not affect the outcome.

is successful, whether attrition rates are low in both intervention and control groups). A smaller number of well-implemented RCTs with large sample sizes may yield more compelling evidence than a larger number of poorly implemented RCTs with small sample sizes.

A fixed threshold has not be established to determine whether the "large majority" of studies present consistent evidence so that the medical effectiveness team can have some discretion to distinguish among bodies of evidence that consist primarily of stronger vs. weaker RCTs.

Assessing generalizability inherently requires some level of judgment. The medical effectiveness team considers studies highly generalizable if the intervention assessed is similar to the intervention for which a bill would mandate coverage and if the population studied is similar to the population whose coverage would be affected.

Determining whether publication of additional RCTs would change the medical effectiveness team's conclusion about the effectiveness of an intervention is admittedly a judgment call. In some cases, such as bills that would mandate coverage for tobacco cessation (e.g., AB 1738), the volume of evidence from RCTs is so large and the findings are so consistent that one can easily conclude that publication of additional RCTs would not alter the conclusion regarding the intervention's effects. Other cases are not so clear cut.

b. Criteria for Grading Bodies of Evidence as "Preponderance of Evidence"

Bodies of evidence should be graded as "Preponderance of Evidence" if all of the following conditions are met:

- **Research Design:** There are at least two controlled studies of the intervention. These studies could be randomized controlled trials (RCTs) or nonrandomized/observational studies with comparison groups.
- **Consistency:** The majority of studies (> 60%) have similar findings with respect to statistical significance, direction of effect, and size of effect.[20]
- **Generalizability:** The studies are generalizable to the intervention in question and the population whose coverage would be affected by a proposed mandate.

If most studies are RCTs, the studies are highly generalizable, and it is unlikely that publication of additional RCTs would change our conclusion about the effectiveness of the intervention, the grade of "Clear and Convincing Evidence" should be assigned instead of "Preponderance of Evidence."

If all studies are uncontrolled observational studies (i.e., case series, case studies), the grade of "Insufficient Evidence" should be assigned instead of "Preponderance of Evidence."

Sub-dividing the Preponderance of Evidence Category. Bodies of evidence to which the medical effectiveness team assigns the grade "Preponderance of Evidence" are sub-divided into three categories based on the strength of the research designs of

[20] The conclusion could be that the intervention has a beneficial or detrimental effect on a pertinent outcome or that it does not affect the outcome.

the majority of the pertinent studies. The medical effectiveness team focuses on research designs because that is an aspect of the strength of evidence that can easily be determined by reading publications we include in our literature reviews. Although all studies with the same research design do not provide equally strong evidence, studies with stronger research designs generally provide stronger evidence than studies with weaker research designs. The categories and the types of research designs in each category are listed below:

- **Strong research designs**: RCTs and quasi-experimental studies (i.e., nonrandomized controlled trials for which data are collected prospectively, efforts are made to select a comparison group that is similar to the intervention group, and instrumental variables, propensity scores, or other statistical techniques are used to control for selection bias).
- **Moderate research designs**: Nonrandomized/observational studies with a concurrent comparison group (e.g., cohort studies, case-control studies) and interrupted time series studies.
- **Weak research designs**: Nonrandomized/observational studies that do not have a concurrent comparison group (e.g., studies with before-after designs, studies with historical comparison groups).

The medical effectiveness team considers both RCTs and quasi-experimental studies to have strong research designs because sophisticated, quasi-experimental studies can provide strong evidence of effectiveness. In fact, some of these studies provide stronger evidence than some RCTs, especially small RCTs in which randomization does not result in intervention and control groups that are equivalent in all measurable respects except for exposure to the intervention.

The rationale for distinguishing observational studies with a concurrent versus a historical comparison group is that having concurrent controls helps to rule out the possibility that a difference in outcomes between the intervention and comparison groups is due to a secular trend (i.e., something that changes over time that is unrelated to the intervention).

When the medical effectiveness team makes a call of "preponderance of evidence," the phrase "from studies with strong/moderate/weak research designs" should be included in CHBRP's report to indicate the strength of the research designs of the studies about which this conclusion is reached. For example, in the case of SB 189, the CHBRP report on wellness programs, instead of stating that "the preponderance of evidence from RCTs" indicates that participating in work-based wellness programs does not reduce blood pressure, blood sugar, or cholesterol, the medical effectiveness team would instead state that "the preponderance of evidence from studies with strong research designs . . ."

c. Criteria for Grading Bodies of Evidence as "Ambiguous/Conflicting"

Bodies of evidence should be graded as "Ambiguous/Conflicting Evidence" if the following criteria are met:

- **Research design:** RCTs or controlled nonrandomized/observational studies.

- **Consistency:** Less than 60% of studies have similar findings with respect to statistical significance, direction of effect, and size of effect.
- **Generalizability:** All levels.

The grade "ambiguous/conflicting" should be used regardless of whether the evidence comes from RCTs or from *controlled* nonrandomized/observational studies. In cases in which the only available evidence is from *uncontrolled* observational studies, the grade "insufficient evidence" is always used regardless of whether or not findings are similar across studies.

d. Criteria for Grading Bodies of Evidence as "Insufficient Evidence"

Bodies of evidence should be graded as "Insufficient Evidence" if the following criteria are met:

- **Research Design:** The only published studies of the intervention are *uncontrolled* observational studies (i.e., case series or case studies) or there are no published research studies of the intervention (i.e., the only evidence available is based on expert opinion or narrative reviews).
- **Consistency:** All levels.
- **Generalizability:** All levels.

The rationale for assigning the grade of insufficient evidence to bodies of evidence that include only uncontrolled observational studies is that without a comparison group, one cannot know whether outcomes that occur in an intervention group are due to the intervention versus another factor. Case studies are even less sufficient than case series because they include only one person. When only one person is studied, one cannot determine whether outcomes are similar across persons who receive the intervention.

Bodies of evidence that consist solely of narrative reviews are also classified as "insufficient evidence" because the literature searches for such reviews are not conducted systematically. There is no way for the Medical Effectiveness team to know whether the authors have synthesized all available evidence versus intentionally picking and choosing studies that support their opinions regarding the effectiveness of an intervention.

5. One way to understand these groupings is to imagine that after the assessment was completed a new well-designed RCT was published with findings contrary to those of the report. Such a single contradictory study would do little to change the overall assessment of findings labeled as "clear and convincing," but might call into question findings previously labeled as "preponderance, "and might become the basis for reevaluating findings previously labeled "ambiguous/conflicting."
6. In cases in which only a single study of an intervention's effect on an outcome has been published, the medical effectiveness team states that "findings from a single study with a strong/moderate/weak research design . . ." Studies are classified as having "strong," "moderate," and "weak" research designs in accordance with the criteria used to distinguish bodies of evidence that are graded as "preponderance of evidence.

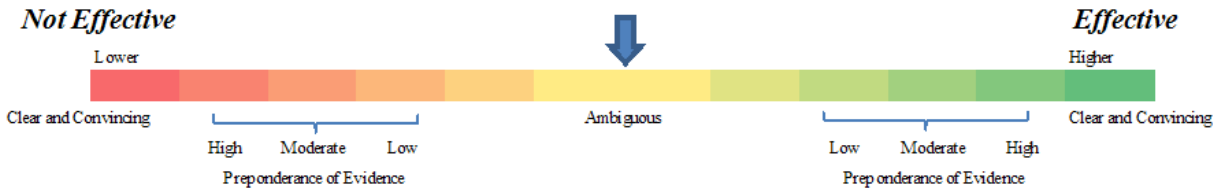
C. Table 3 provides an example of a table that appears at the end of the medical effectiveness section of the report that presents findings regarding the three dimensions assessed and the medical effectiveness team's conclusions regarding an intervention's effects on pertinent outcomes.

Table 3. Studies that Examined the Effectiveness of Different Numbers of Prenatal Visits

Outcome	Citation(s)	Research Design	Statistical Significance	Direction of Effect	Size of Effect	Conclusion
Low birth weight	Fiscella, 1995; Villar et al., 2001	1 meta-analysis and 1 systematic review of Level II studies	No statistically significant difference	No effect	No effect	The preponderance of evidence from studies with strong research designs suggests that changing the number of prenatal visits does not affect the odds of having a low–birth weight infant
Preterm birth	Fiscella, 1995; Villar et al., 2001	1 meta-analysis and 1 systematic review of Level II studies	No statistically significant difference	No effect	No effect	The preponderance of evidence from studies with strong research designs suggests that changing the number of prenatal visits does not affect the odds of giving birth preterm
Admission to neonatal intensive care unit	Fiscella, 1995; Villar et al., 2001	1 meta-analysis and 1 systematic review of Level II studies	No statistically significant difference	No effect	No effect	The preponderance of evidence from studies with strong research designs suggests that changing the number of prenatal visits does not affect the odds that a newborn will be admitted to a neonatal intensive care unit

The medical effectiveness team also includes a graphic in its sections of the executive summary and the text that presents the medical effectiveness team's overall assessment of pertinent literature. Figure 1 provides an example of this graphic.

Figure 1



As indicated above, the graphic presents the medical effectiveness team's overall conclusion. Conclusions regarding individual outcomes assessed by previous studies may differ from the overall assessment. For example, for most measures of the effectiveness of a treatment the preponderance of evidence from studies with strong research designs may favor the intervention but for a few measures there may be insufficient evidence. In such a case, the overall call would be preponderance of evidence from studies with strong research designs.

Summarizing the Quantifiable Evidence for Specific Outcomes

A. Where feasible, the medical effectiveness team also reports pooled estimates of the effects of the intervention on select medical effectiveness outcomes. These estimates may be used by the cost and public health teams to assess a proposed mandate or repeal bill's impact on utilization of health care services and its effect on public health.

B. In some cases, the medical effectiveness team reports quantitative estimates from meta-analyses or individual studies.

   1. Quantitative estimates from recent high-quality meta-analyses are used whenever possible, because the authors of meta-analyses may have greater expertise and more time to thoroughly review the pertinent literature than CHBRP's medical effectiveness team, and may use more sophisticated statistical methods to generate quantitative estimates of effects.[21] In cases in which a meta-analysis has been published, the team asks the content expert to assess whether the meta-analysis adequately addresses current practice in the prevention, diagnosis, or treatment of the disease(s) or condition(s) to which the bill would apply.

      a. Many meta-analyses (particularly those included in the Cochrane Library) report their results as standardized mean differences (SMDs), which is a unitless measure. To obtain values in meaningful units consistent with those assessed in individual studies, such as the number of physician visits, the team extracts data from the individual studies included in a meta-analysis.

[21] Findings from systematic reviews that present a qualitative assessment of the literature without an accompanying meta-analysis are excluded because they do not provide quantitative estimates of treatment effects.

2. In some cases, a single study may be much more rigorous[22] than other studies that analyze an outcome.[23] The point estimate from such a study is likely to be more accurate than a point estimate derived from pooling this study with less rigorous studies. When deciding whether to use the point estimate from a single study, the medical effectiveness team also considers whether the study enrolled persons who are representative of the population to which the proposed mandate or repeal would apply.

C. The medical effectiveness team generates its own new quantitative estimate of an intervention's effect on an outcome if the following conditions are met:

1. The outcome is relevant to consumers and policymakers. For all proposed mandates or repeals, the team determines which outcomes will be assessed in consultation with the members of the analytic team for the bill, the content expert, and State Legislature staff responsible for a bill.
2. There are no recent high-quality meta-analyses on the topic or the findings of the most recent studies differ significantly from findings of studies synthesized in meta-analyses.
3. There is not a single large, well-executed RCT that is much more rigorous than other studies that assess an outcome and that analyzes subjects who are representative of the population to which the proposed mandate or repeal would apply.
4. The studies that measure the outcome are methodologically rigorous. RCTs generally provide the best estimates of a proposed mandate or repeal's effect on an outcome, because they provide the greatest assurance that a change in the outcome is due to the intervention and not some other factor. If the majority of studies of an outcome are RCTs or cluster RCTs, the team only pools estimates from RCTs. If a majority of the relevant studies are observational studies, a biostatistician is consulted to assess the appropriateness of pooling the observational studies with one another and with RCTs that assess the outcome. Quantitative estimates are not generated if the only pertinent studies do not randomize subjects, have very small samples, and/or do not include control groups.

D. If the criteria for a quantitative estimate are met, the medical effectiveness team uses the following procedure to calculate these estimates:

---

[22] "Rigorous" can encompass a variety of characteristics of a study such as selecting a sample that is sufficiently large to provide adequate power to detect differences between the intervention and control or comparison groups, designing the sampling procedure to maximize the likelihood that the intervention and control or comparison groups are equivalent at baseline, using appropriate statistical methods to adjust for lack of equivalence, implementing procedures to prevent contamination of the intervention and control groups, and concealing allocation to the intervention and control groups to the maximum extent feasible. The assessment of "rigor" in this case is considered within the context of studies that address the questions needed for the review. Thus, a methodologically rigorous study that focused only on a narrow subset of the population to whom the mandate or repeal would be applied would not necessarily "trump" other studies.

[23] For example, CHBRP relied on a single study in its analysis of the literature on the effect of high-deductible health plans on use of preventive services. The medical effectiveness team found that the literature consisted of one, large, rigorous RCT, the RAND Health Insurance Experiment (HIE), a few small RCTs, and a number of retrospective observational studies. The RAND HIE was a highly generalizable study that enrolled children and non-elderly adults with low or moderate household incomes from six urban and rural communities across the United States into various types of health plans, including a high deductible plan.

1. In general, pool results only from studies in which similar comparisons are made. There are two major types of medical effectiveness studies: (1) studies that compare a group of subjects who receive an intervention to a group that receives either no intervention or a placebo; and (2) studies that compare groups of subjects who receive different interventions (e.g., two different drugs used to treat persons with Alzheimer's disease, chiropractic services vs. surgery for low back pain) or receive the same intervention at different intensities (e.g., different dosage, different number of visits). Estimates from studies that make these two different types of comparisons should not be combined, because combining them is likely to generate pooled results that reflect neither an intervention's effectiveness relative to no intervention nor its effectiveness relative to a different or more/less intensive intervention. The team consults with the content expert if its members have difficulty making such distinctions. The team always calculates pooled estimates for studies that compare an intervention group to a group that receives a placebo or no intervention. Studies that compare two different interventions may be pooled, if there are multiple studies that compare the same two interventions.
2. For all studies, review pre-intervention data on the outcome of interest to ascertain whether the intervention and control or comparison groups are equivalent at baseline. Estimates should be pooled only if both pre- and post-intervention data are reported and appropriate multivariate methods are used to adjust for significant baseline differences between the intervention and control groups.[24] If the intervention and control or comparison groups are not equivalent, differences in outcomes may be due to differences between the two groups prior to exposure to the intervention rather than to the intervention. Randomization does not necessarily produce equivalent intervention and control groups, particularly when the sample size is small.[25] Observational studies are even more vulnerable to selection bias, especially if researchers do not use multivariate analytic methods to adjust for baseline differences between the intervention and comparison groups.
3. If a study reports an overall "adjusted" effect of an intervention that takes into account important differences that may exist between the intervention and comparison groups, that estimate is used to calculate the pooled estimate of effects across studies.
4. If a study does not report an overall "adjusted" measure of the effect, the medical effectiveness team calculates the proportionate effect attributable to the intervention and then applies it to the overall study population (intervention plus comparison group).
    a. Raw data from the study are inserted into a spreadsheet. A sample calculation for Krishna and colleagues' study (2003) appears in Table 3 below. This study assessed the effects of an asthma education intervention on a variety of outcomes, including the number of days children with asthma were absent from school.

[24] Use of multivariate methods mitigates selection bias only if the additional variables added to an analysis are the only factors other than the intervention that are likely to affect the outcome of interest. This method does not eliminate the possibility that there may be unmeasured variables that are associated with the outcome but not correlated with any of the other variables included in the analysis. However, studies that make an effort to adjust for baseline differences are preferable to studies that ignore them.

[25] Randomization of subjects only produces equivalent groups if the trial is repeated many times or if the sample is very large. Well-executed RCTs with small samples may have non-equivalent intervention and control groups just by chance.

b. Baseline data and post-intervention data for the study appear in Table 3. In this instance, the intervention group had a somewhat higher rate of school absences (7.90) at baseline than the control group (6.40). The difference for the intervention group (-6.50) equals the post-intervention rate (1.40) minus the baseline rate (7.90).

c. Baseline data for the intervention and comparison groups (7.15) are averaged. (Implicitly, averaging assumes that the two groups are the same, as they would be if randomization were successful, and that any observed differences are due to chance variation.) If the study reports the numbers of cases in each group, they are used as weights. If not, the two groups are assumed to be of equal size.

Table 3. Calculating the Overall Effectiveness of an Intervention: Proportionate Reduction in School Absences

Trial		Intervention Group	Control Group	Average
Krishna et al., 2003	Baseline	7.90	6.40	7.15
	Post-intervention	1.40	5.40	
	Difference	-6.50	-1.00	
	% difference	-82.3%	-15.6%	
	Expected difference	-5.88	-1.12	
	Expected reduction in days absent			-4.77
	Expected days absent in the control group			6.03
	Proportionate reduction in days absent in intervention group			-79.0%

- The % difference (-82.3%) = difference (-6.50)/baseline (7.90). This is the observed percentage reduction in the intervention group.
- Expected difference (-5.88) = % reduction in the intervention group (-82.3) times the baseline average for all subjects (7.15)
- Expected reduction in days absent (-4.77) = the expected difference in the intervention group (-5.88) – the expected difference in the control group (-1.12)
- Expected days absent in the control group (6.03) = baseline average (7.15) + expected difference in the control group (-1.12).
- Proportionate reduction in days absent in intervention group (-79.0%) = expected reduction in days absent (-4.77)/expected days absent in the control group (6.03). This last calculation compares the results for the intervention and control groups. Even if the intervention group experiences a reduction in days absent, this calculation may appear to indicate an increase in the number of absences in the intervention group, if the control group experiences a greater reduction in absences than the intervention group.

d. For studies that publish only post-intervention data, the proportionate reduction = (control – intervention)/control (see Table 4).

Table 4. Calculating Proportionate Reduction in School Absences with Post-Intervention Results Only

Trial		Intervention Group	Control Group	Difference
Fireman et al., 1981	Post-intervention	0.5	4.6	-89.1%

e. Next, a weighted average calculation is made to estimate the overall proportionate reduction in days absent for the intervention groups in the studies being pooled. The results for each study are weighted by sample size so that results from studies with more subjects will be weighted more heavily. Table 5 illustrates the weighted average for the effect of asthma education on school absences.

Table 5. Calculating the Weighted Average to Find the Overall Proportionate Reduction in School Absences

Trial	Total Subjects	% Reduction	(Weighted)
Clark 2004	835	0.0%	0
Christiansen et al., 1997	42	-19.8%	-0.3
Evans et al., 1987	204	-3.8%	-0.3
Fireman et al., 1981	26	-89.1%	-1.0
Horner 2004	44	18.3%	0.3
Morgan 2004	937	-50.1%	-19.6
Perrin et al., 1992	56	-79.1%	-1.8
Persaud et al., 1996	36	-15.8%	-0.2
Rubin et al., 1986	54	-0.9	0.0
Velsor-Friedrich 2004	102	-28.0%	-1.2
Wilson et al., 1996	59	-60.0%	-5.0
Total	**2395**		**-25.7%**

5. After a new, pooled estimate of the effect of an intervention on an outcome has been completed, a sensitivity analysis is conducted to determine whether the pooled estimate is highly sensitive to the results of one or two studies. If one or two studies have samples that are much larger than those of other studies with which they are pooled, the pooled estimate will be dominated by the results of those studies. Pooled estimates may also be sensitive to studies with anomalous results, regardless of sample size, particularly if the

total number of studies pooled is small.[26] Sensitivity analyses are performed by omitting each study sequentially, repeatedly recalculating the pooled estimate, and comparing the pooled estimate obtained when all studies are included to the pooled estimate obtained when a study is omitted. If one or two studies to which a pooled estimate is highly sensitive are large, well-implemented RCTs, the medical effectiveness team may choose to rely on estimates reported in these studies rather than on the pooled estimate from the larger group of studies. If the studies in question are not large, well-implemented RCTs, the team reports the pooled estimate but also reports the results of the sensitivity analysis.

[26] For example, in the analysis of AB 264 the pooled estimate of the effect of pediatric asthma self-management education on mean hospitalizations for asthma is highly sensitive to the results of the one study of this outcome that found no association between the intervention and the outcome. All other studies found a reduction in mean hospitalizations. If the study with anomalous results were omitted from the pooled estimate, the estimated size of the effect would be 15 percentage points greater.