DOI: 10.1111/j.1475-6773.2006.00517.x

Evaluating Medical Effectiveness for the California Health Benefits Review Program

Harold S. Luft, Karen M. Rappaport, Edward H. Yelin, and Wade M. Aubry

An important aspect of the mandate assessments requested by the California legislature is a review of the scientific and medical literature on the medical effectiveness of the proposed health insurance benefit mandate. Although such a review bears many similarities to effectiveness reviews that might be undertaken for publication as research studies, several important differences arise from the requirements of the California legislation.

Our reviews are intended to assist the legislators in deciding whether to support a specific mandate to modify health insurance benefits in a particular way. Thus, our assessments focus on how the scientific literature bears on the proposed mandate, which may involve a complicated chain of potential effects leading from altered coverage to ultimate impact on health. Evidence may be available for only some of the links in the chain. Furthermore, not all the evidence may be directly applicable to the diverse population of California or the subpopulation affected by the mandate.

The mandate reviews, including the medical effectiveness analyses, may be used in a potentially contentious decision making setting. The legislative calendar requires that they need to be timely, yet they must be as valid, credible, and based on the best information available as possible. The focus on applicability also implies the need for informed, technical decisions concerning the relevance of the articles for the report, and these decisions need to be made as transparent as possible. These goals and constraints yield an approach that differs somewhat from an investigator-initiated review of the literature.

Key Words. California Health Benefits Review Program, medical effectiveness, efficacy, mandate review, randomized controlled trials, meta-analyses, systematic reviews, observational studies, case-control studies

INTRODUCTION

Under the provisions of Assembly Bill (AB) 1996 (California Health and Safety Code Section 127660 et seq.), the State Legislature may ask the University of California to assess legislation proposing mandated health care benefits to be provided by health care service plans and health insurers, and to prepare a written analysis (of its medical, financial, and public health impacts) in accordance with specified criteria (California Health and Safety Code 2005). Furthermore, the legislation requires an assessment of the "[m]edical impacts, including, but not limited to... [t]he extent to which the benefit or service is generally recognized by the medical community as being effective in the screening diagnosis, or treatment of a condition or disease, as demonstrated by a review of the scientific- and peer-reviewed medical literature." This overall effort, known as the California Health Benefits Review Program (CHBRP), uses staff and a task force of faculty experts at various public and private universities in California to summarize the scientific evidence in an objective manner without offering recommendations, deferring such policy making decisions to the state legislature.

Drawing on their experiences during the first 2 years of CHBRP (Table 1), the authors describe in this paper the approach taken in the medical effectiveness analysis that forms one part of each proposed mandate review (California Health Benefits Review Program 2005a, b). (The other parts are utilization and cost, and coverage impacts on public health impact.) The CHBRP medical effectiveness review process, conducted by a team of physicians, health services researchers, and staff, differs from that of an investigator-initiated review because it seeks an assessment in the context of a specific proposed mandate. The review cannot narrow its focus simply because high-quality evidence is not available. Rather, it requires as broad an assessment as is needed to address the mandate, with objective and defensible decisions about the relevance and quality of the available literature. Compounding the difficulty in making such decisions is that the review must be completed in 60 days.

Address correspondence to Harold S. Luft, Ph.D., Institute for Health Policy Studies, 3333 California Street, Suite 265, San Francisco, CA. Karen M. Rappaport, M.D., Ph.D., Edward H. Yelin, Ph.D., and Wade M. Aubry, M.D., are with the Institute for Health Policy Studies, San Francisco, CA.

Table 1: California Health Benefits Review Program Analyses (2004–2005)

Analyzed Legislation	*Topic*	*Completed Analyses*
AB 438	Osteoporosis screening	2/9/04
AB 547	Ovarian cancer screening*	2/9/04
AB 1084	Access to vision providers	2/9/04
AB 1549	Childhood asthma	2/9/04
SB 101/1192†	Substance disorder treatment	2/9/04
SB 174	Hearing aids for children	2/9/04
SB 897	Maternity services	2/9/04
SB 1555	Maternity services	4/1/04
AB 2185	Asthma management	4/14/04
AB 1927	Vision services	4/16/04
SB 1158	Hearing aids	4/19/05
SB 1157	Elimination of intoxication exclusion	4/27/04
AB 8	Mastectomies and lymph node dissections	3/7/05
AB 213	Lymphedema	4/7/05
AB 228	Transplantation services: human immunodeficiency virus	4/7/05
SB 573	Elimination of intoxication exclusion	4/7/05
SB 415	Prescription drugs: Alzheimer's disease	4/16/05
SB 572	Mental health benefits	4/16/05
SB 576	Tobacco cessation services	4/16/05
SB 749	Pervasive developmental disorders/autism	4/16/05
SB 913	Medication therapies; rheumatic diseases	4/16/05

*Subsequent to a request from the California Assembly Committee on Health to analyze AB 547, the bill was amended and no longer concerns ovarian cancer screening. The version of the bill analyzed and included here was the legislation's original language.

†Subsequent to a request from the California Senate Insurance Committee to analyze SB 101, the bill was reintroduced as SB 1192 using the same language.

Medical Effectiveness Reviews in the Context of the CHBRP Rationale

State coverage mandates for screening and/or treatment vary widely. This has historically stemmed, at least in part, from differing amounts of pressure from people and organizations concerned about particular diseases as well as differences in the evidence presented for and against coverage (Holtzman 1992). The CHBRP analyses are intended to offer the legislature unbiased, evidence-based information to assist in making its decisions. The legislature is often inundated with arguments by advocacy groups or special interests that may benefit from or be threatened by the mandate. Given this potentially contentious setting, the reviews must be as complete, transparent, and evidence-based as possible.

Some argue that coverage mandates are unnecessary—if a new medical intervention is beneficial and worth more than its cost, health plans will eventually cover the service, passing the cost on through premiums. Even if the scientific evidence were clear, however, mandates might arguably be needed because: (1) medical knowledge accumulates slowly and assessing it is expensive, so insurers may lag in their assessments or not undertake them because of the public good nature of the assessment; (2) some interventions may provide health benefits to people other than those insured (externalities) and thus be undervalued in the private market; (3) it is impossible for insurers to differentially price policies at the level of detail that would allow consumers to make tradeoffs between less expensive but less effective interventions and more effective but higher cost ones; and (4) mandates might also be designed to address market failures, such as the incentive for insurers to avoid covering beneficial but expensive services needed by a small number of people in the hope they will choose to enroll in other plans. Mandates might also be intended to eliminate or create bargaining advantages for certain groups of providers, drugs, or devices such that insurers have to offer them even if other comparable alternatives are available.

In addition to the politically sensitive issues of assessing such benefits and costs, the potentially relevant data themselves may not be clear-cut. Therefore, the reviews are likely to be controversial. In describing our approach to a CHBRP medical effectiveness review, this paper addresses three types of challenges/issues: (1) the types of evidence that should be examined, and in particular, the tension between efficacy versus effectiveness; (2) issues arising from the fact that some mandates focus on expanding coverage for an intervention without an immediate effect; and (3) issues arising from attempting to be responsive to legislative needs.

Effectiveness versus Efficacy in a CHBRP Mandate Review

Medical effectiveness is defined as the benefit achieved when services are rendered under ordinary circumstances by average physicians for typical patients (D'Agostino and Kwan 1995). This is in contrast to efficacy, which is defined as how well the intervention works in the research setting, or under ideal circumstances. The CHBRP analysis focuses on evidence in the peer-reviewed scientific literature of effectiveness.

The scientific literature considers double-blind, randomized controlled trials (RCT) to be the "gold standard" for clinical decision-making purposes.

The design of a RCT limits the possibility that unforeseen characteristics might influence the outcome of interest (Victora, Habicht, and Bryce 2004; MacLehose et al. 2000). When both RCTs and nonexperimental studies of the effectiveness of an intervention are available, the latter often show larger estimates of effectiveness (MacLehose et al. 2000). Similarly, clinical trials with inadequately concealed random allocation show estimates of effect that are 40 percent larger than those of trials with well-concealed random allocation (Kunz and Oxman 1998). Thus, there is a general preference for the more tightly controlled study designs.

Numerous studies have also shown that the quality of published studies varies and that one can usually reach more valid and reliable assessments of a given question by systematically reviewing all the relevant literature, grading each study for its adherence to experimental guidelines, and then summarizing the results, preferably based on the well-conducted studies, using specific statistical methods. The Cochrane Collaboration sponsors a growing library of such meta-analyses. CHBRP effectiveness reviews therefore use a hierarchy of evidence that values meta-analyses of multiple RCTs most highly (see Table 2). Systematic reviews meet many of the same criteria but typically do not have summary measures of effect, usually because the various studies do

Table 2: Preferred Hierarchy of Articles Used in the Effectiveness Review*

Study/Publication Type	*Study/Publication Relates to Efficacy*	*Study/Publication Relates to Effectiveness*
Meta-analyses[†]	+	
Systematic reviews[‡]	+ (especially when part of a meta-analysis)	
Evidence-based guidelines		+
Individual randomized clinical trials	+ (unless it is an effectiveness trial)	
Observational studies		+
Case–control studies		+
Clinical practice guidelines based on consensus or opinion, rather than on evidence		+

**Note:* Exceptions to the hierarchy may occur, depending on the methodology used in each study. Studies or reviews critically based on evidence are given more weight.

[†]Particularly those included in the Cochrane library.

[‡]Particularly those performed by authoritative organizations such as the Agency for Healthcare Research and Quality, U.S. Preventive Services Task Force, Evidence-based Practice Centers, National Institutes of Health, and Centers, for Disease Control.

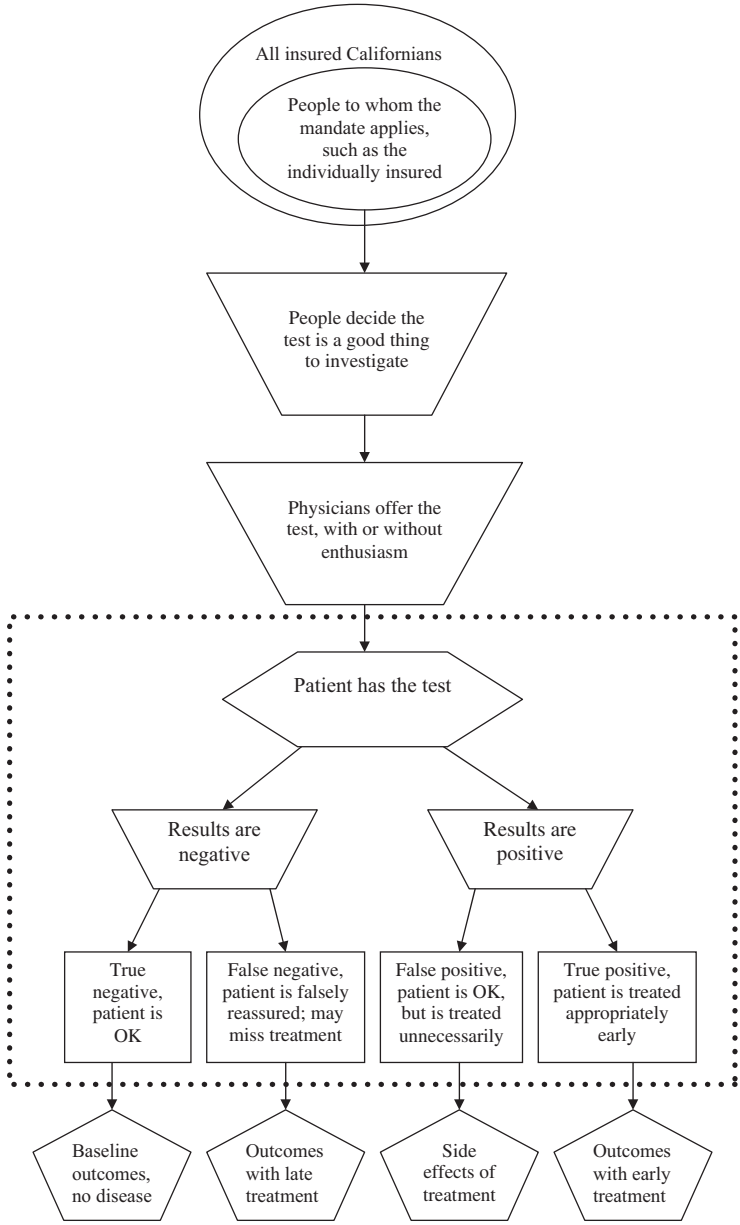
not provide comparable metrics. In theory, RCTs (as well as meta-analyses and systematic reviews) can focus on either efficacy or effectiveness. If the treatment is provided under ideal conditions, such as in a teaching hospital with rigid adherence to protocol, it would be an efficacy trial. If the intervention is administered under usual practice conditions in the community with variable implementation and adherence by clinicians and patients it would be an effectiveness trial.

The problem for CHBRP reviews arises from the fact that tight control of subjects, randomization (and thus the need for full informed consent), and blindedness of researchers and subjects to the intervention becomes increasingly difficult as one moves from efficacy to effectiveness studies. Furthermore, the costs of a trial skyrocket as the interventions become less standardized—a key aspect of effectiveness trials, the presence of confounding factors more common, and the length of time to assess outcomes greater. Even when RCTs focus on drugs or interventions in community settings, the patient population is often carefully selected for those most likely to benefit, avoiding both unnecessary risk and "statistical noise" associated with patients having potential confounding risk factors (D'Agostino and Kwan 1995; Dieppe et al. 2004). Well-done RCTs thus may provide data with a high degree of internal validity, but such studies often do not have the desired high external validity or generalizability (D'Agostino and Kwan 1995; Black 1996; Victora, Habicht, and Bryce 2004). Yet, a key intent of the mandate reviews is to address the issue of generalizability.

For example, AB 438 dealt with osteoporosis screening in healthy women between the ages of 50 and 64. The medical effectiveness team found evidence with respect to the effectiveness of screening in older or high-risk women, but very little evidence from RCTs to support screening and treatment of this younger population. More importantly, none of the evidence directly addressed whether screening actually reduced the prevalence of hip fractures or their sequelae. Instead, the evidence from some trials showed that screening could identify women with low bone density, while other studies indicated that low bone density was associated with increased risk of fracture. Yet other studies indicated that some interventions could reduce the rate at which bone mass was lost. Thus, the likely benefits of increased rates of screening depend on a long chain of causation, only parts of which might be assessed (see Figure 1).

More commonly, RCTs might not be fully applicable in the review of proposed health care mandates because new interventions are often tested only in subjects without comorbidities that may complicate the trial. For

Figure 1: Schematic of a Mandate for Covering a Test



Note: Most studies of tests focus on issues within the dotted box.

example, a RCT of a new nonsteroidal antiinflammatory agent conducted solely on younger populations does not provide us with the information we need about potential adverse drug events in elderly populations that have elevated incidences of comorbidities (D'Agostino and Kwan 1995; Dieppe et al. 2004). Unless a health care mandate is directed solely at services for a particular subgroup, such as the childhood asthma mandate (AB 1549) or the maternity services mandates (SB 897 and SB 1555), the CHBRP medical effectiveness team must consider the impact of a proposed health care mandate on all Californians regardless of age, ethnicity, or socioeconomic status.

If well-done meta-analyses are not available, the team gives preference to systematic reviews and then to evidence-based guidelines, again supplementing these as needed with RCTs published since the last review or guideline. Uncontrolled observational studies, case–controlled studies, and clinical or practice guidelines based on consensus or opinion would, ideally, carry the least weight. Because of the time constraints for the mandate reviews, the CHBRP team cannot undertake formal meta-analyses, but if no relevant meta-analyses or systematic reviews are available, less formal approaches may be used.

While a clear hierarchy of evidence such as in Table 2 is desirable, it is often necessary to make tradeoffs between evidence drawn from studies higher on the methodological hierarchy and evidence that may be more relevant to the question at hand, but from less tightly controlled or rigorous sources. Excluding data from nonrandomized studies biases the evidence base in favor of interventions that are more easily evaluated with RCTs but may not necessarily be more effective (Des Jarlais, Lyles, and Crepaz 2004). The CHBRP team must consider evidence in the context of reasonableness and feasibility. Having chosen to be somewhat more flexible rules in an attempt to offer more useful assessments to the legislature, all decisions by the team must be clearly delineated and included in the report to avoid the appearance of arbitrariness.

Complexities of the Real World of a Mandate Review

A second level of complexity arises from the fact that the CHBRP team is assessing a proposed health insurance mandate, rather than a specific clinical intervention. This creates several problems affecting CHBRP reviews. Mandates are written to become law and as such cannot have the type of specificity one would like for a scientific study. For example, a mandate might be written

to cover all appropriate devices for the care of patients with asthma, rather than the devices made by a specific manufacturer that are the subject of published trials. Furthermore, the medical benefits, as well as the costs and savings associated with the intervention may not occur immediately, nor be clearly attributable to the mandate.

The goal of the CHBRP analysis, beginning with the medical effectiveness report, is not to merely present the results of RCTs, but also to examine the potentially far-reaching effects of adopting the intervention under consideration. Figure 1 illustrates this with a flow chart depicting steps involved with the set of mandates addressing *screening tests*, such as the osteoporosis screening or the ovarian cancer screening mandates. The effectiveness literature typically deals with questions *within* the box seen in the figure, i.e., the sensitivity and specificity of the test. Outside the box are issues related to the willingness of patients to request the test, and physicians to offer it, and the implications of treatment. While relevant to a CHBRP mandate analysis, they are not typically addressed in the empirical literature.

The passage of a mandate means only that insurance companies must cover the appropriate costs consistent with their usual policies. Passage does not mandate that physicians or patients avail themselves of services covered. Using the example of screening tests for ovarian cancer (a blood test, a sonogram or both), either a physician would first have to offer the test and the patient accept it, or the patient would have to convince the physician to order it. In the event of a positive test, a patient would then have to agree to a complete diagnostic workup that includes surgery, which has its own set of additional complications. While there will often be published studies of treatment effectiveness (separate from the screening studies), there are unlikely to be any studies of the implications for patients treated unnecessarily (the false positives), falsely reassured (false negatives), or correctly reassured (the true negatives). Although such outcomes issues are not typically the focus of a RCT, all relevant scenarios must be considered during a CHBRP medical effectiveness review.

Some mandates include a broad mix of services, such as a collection of educational programs for patients with asthma or a package of services provided as part of prenatal care. The analysis of the osteoporosis screening mandate also involved a review of exercise programs and drug treatment for the prevention of osteoporosis. The causal pathways for such interventions may involve multiple behavioral steps that are difficult to specify and measure (Victora, Habicht, and Bryce 2004). Even if there is solid evidence on various links in the chain, few studies will have examined the entire chain, much less in

a double-blinded randomized controlled manner. Studies will almost never exist that examine every combination of the components of such interventions to determine which ones are crucial.

In some instances data are absent because of ethical considerations. Senate Bill (SB) 897 would have mandated that insurance companies provide a minimum combination of maternity and neonatal services in all health plans. (The intent of the mandate was to preclude some insurers from offering plans attractive only to people not planning on becoming pregnant.) The CHBRP team found some published evidence on the effectiveness of different components of the packages, but none on the whole "package" of prenatal care because it would be unethical to deny prenatal care to pregnant women in order to test the effectiveness of specific aspects of care in a trial. Fortunately, the debate over this bill focused not on whether prenatal care was beneficial, but whether the insurance market should be segmented by such benefit exclusions.

In undertaking a review of medical effectiveness, one must first ask which outcome measures will be used. The team typically analyzes all appropriate outcome measures for which literature is available. For example, AB 1549, Childhood Asthma Management, required coverage of over-the-counter and prescription asthma medications and associated pediatric asthma outpatient self-management training and education. The medical effectiveness team examined the impact of the interventions on such outcomes as the number of days with asthma symptoms (or the number of symptom-free days), asthma symptom scores, the number of exacerbations of disease, the forced expiratory flow rate (FEV) (a measure of lung function), coping scores, knowledge scores (child and caregiver), activity restriction, emergency room utilization, use of medications, and quality of life measures, among other measures. Analysis of all known outcome variables was important, because some interventions had positive, albeit insignificant, effects on some outcomes but significant positive effects on other outcomes. If only selected outcomes were included, the medical effectiveness report could have been criticized as being biased.

Such comprehensiveness, however, requires the team to provide guidance to the Legislature in comparing the various outcome variables, especially if some are favorable and others are not. This ranges from explanations of the physiologic measures, such as the FEV, to discussing whether an intervention should be "better than" or just "not worse than" the alternative. For example, AB 228, the transplantation HIV mandate, focused on coverage for transplants for patients who were HIV positive (HIV+). Advances in highly active

antiretroviral therapy (HAART) since 1996 have made transplantation a viable possibility for many HIV+ patients. For most outcome variables, including patient survival, graft survival, and measurements of viral activity, HIV+ transplant patients enjoyed outcomes comparable to that of HIV negative (HIV−) patients. In this instance, the appropriate "test" was not whether being HIV+ resulted in better transplant outcomes, but rather whether it was still associated with much worse outcomes. Among HIV+ liver transplant patients, however, those who suffered from hepatitis C tended to fare worse than other patients undergoing liver transplantation. The team felt it necessary to clarify that HIV− liver transplant patients with hepatitis C also had poorer outcomes than HIV− patients without hepatitis C. Hepatitis C thus appeared to be the biggest impediment to survival following a liver transplant, not HIV status.

Whenever possible, the medical effectiveness team looks at the language of the mandate itself as a guide for determining the outcomes of interest. However, for a multitude of reasons such as ethics, expense, or feasibility, data relevant to the outcomes of interest are not always available from RCTs or even observational studies or guidelines. A review therefore often involves analyses of less meaningful short-term end-points, such as results of bone density scans, rather than more consequential endpoints such as the number of fractures prevented.

In conducting a mandate analysis, the medical effectiveness team tabulates the various studies informing the analysis by outcome measure, listing the number of patients in each study. Ideally, one would incorporate data on the size of the trial as weights in estimating the proportionate effect attributable to the intervention. This, however, is often not possible. In the case of AB-228, Transplantation Services: HIV, the effectiveness team had to rely on case reports, case series, and observational studies, mostly from the small number of centers in the United States and Europe performing transplants on HIV+ patients. Every few years, the authors would re-publish their cases in new, peer-reviewed articles, adding new patients to their series along with up-to-date information on the survival and medical courses of earlier patients. Except for the few cases in which adequate and distinctive histories were provided, overlap with patients in earlier articles could not be determined. The team also relied on a published observational study comparing HIV+ and HIV− renal transplant patients using a national database. The patients in this database almost certainly included patients in reports from transplant centers, but this national database did not contain information on HIV status for all patients. The team decided to simply provide all information available

from all peer-reviewed articles while simultaneously cautioning readers that the true number of patients was smaller that it appeared.

The Steps Involved in a Medical Effectiveness Mandate Review

To some extent, even the limited approach described above faces challenges in the actual undertaking of a review. Not only is the 60 calendar-day timeline extremely tight, but multiple reviews by the team are usually underway simultaneously. This has led to a series of logistical and analytic adaptations.

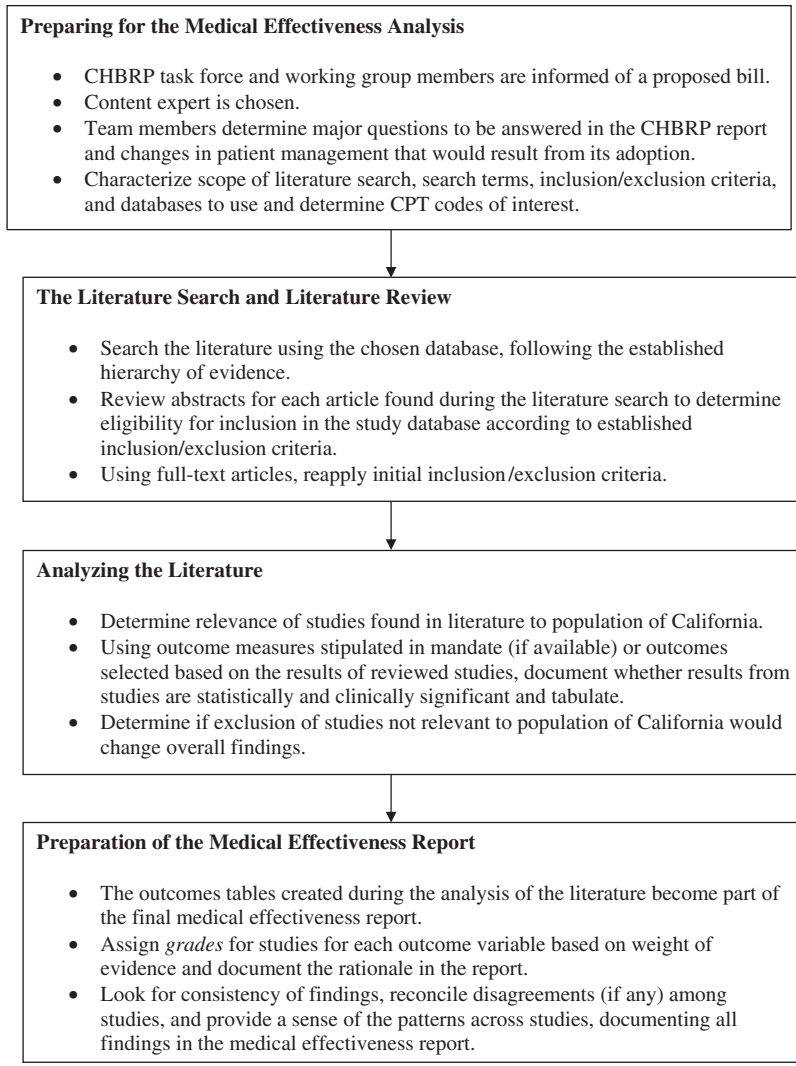
Preparing for the Medical Effectiveness Review

The CHBRP faculty and staff have developed a protocol for conducting a medical effectiveness analysis for each proposed mandate. As seen in Figure 2, the search for a content expert begins immediately because it is important to find an appropriate consultant without a real or perceived conflict of interest. Conflicts may be either financial or may reflect strong advocacy or research positions. In the case of AB 228, the transplantation-HIV mandate, the medical effectiveness team first considered a physician who, it turned out, had been instrumental in drafting the legislation.

The content expert is usually a physician or other health professional practicing in a field that bears on the mandate but without known biases or the appearance of biases. For AB 213, the lymphedema mandate, for example, we identified a physical therapist trained in the specialized techniques that were addressed by the mandate. Although one of only a small number of physical therapists trained in the techniques under consideration, she was not aware of the legislation until the medical effectiveness team conferred with her about joining the team as the content expert.

The medical effectiveness review team members meet at the initiation of the literature review to characterize the scope of the search, search terms, inclusion and exclusion criteria, the databases to be used, and the relevant CPT codes (current procedural terminology codes that describe medical or psychiatric procedures performed by physicians and other health providers) needed by the cost and public health teams. They communicate closely with the other teams to reach an agreement concerning the meaning and intent of the mandate. For example, CHBRP faculty and staff concurred that the transplantation in HIV mandate should be treated as an antidiscrimination bill so the medical effectiveness team would focus on whether the outcomes of

Figure 2: Steps in a California Health Benefits Review Program (CHBRP) Medical Effectiveness Analysis



HIV+ patients undergoing transplantation were significantly worse or comparable to those of HIV− patients. If the data suggested that outcomes were similar, then excluding HIV+ patients from coverage would not be justified on the basis of medical effectiveness.

Early agreement was even more critical on the best way to analyze SB 572, a bill that would mandate that the diagnosis, treatment, and coverage of all mental health problems be on a par with those of medical illnesses. Evaluating the effectiveness of every potential intervention for each of the more than 400 distinct diagnoses included in the fourth edition of the diagnostic and statistical manual of mental disorders (DSM-IV), was impossible (American Psychiatric Association 2000). Instead, CHBRP's analysis of SB 572 was designed to provide the California State Legislature with background information on policies and legislation in California, other states, and at the federal level that affect health insurance coverage for mental health conditions, and thus focused more broadly on what is known about the effects of "mental health parity" legislation in other settings. While clearly not the "standard model" of a CHBRP effectiveness review, it was both feasible and more relevant.

The Literature Search and Review

The effectiveness team next determines the extent to which the results of the literature search (conducted by a medical librarian in an organized, pre-determined, and reproducible manner) are likely to address the question underlying the proposed mandate. If necessary, the search terms and the inclusion and exclusion criteria would be broadened. The range of specificity and clarity of the mandates varies substantially. AB 228, the transplantation-HIV mandate, for example, addressed a range of specific transplantation services (solid organ, skin, cornea, and bone marrow), each of which required searches. AB 213, the lymphedema mandate, sought to mandate the *standard of care* for lymphedema patients, but a literature search did not reveal a clearly defined *standard of care.* The medical effectiveness team pointed this out and thus reported on all types of treatments, including specialized physical therapy and pharmaceutical agents legally available in the United States.

There might not be sufficient available literature to analyze some mandates as written. AB 8, the mastectomy mandate, would require health plans to allow breast cancer patients to remain in the hospital for 48 hours following a mastectomy and 24 hours following an axillary lymph node dissection (surgical removal of lymph nodes in the armpit). Most of the recent literature concerning length of stay following surgery for breast cancer in the United States consisted of analyses of outpatient mastectomy programs, rather than the lengths specified in the bill. The CHBRP team instead used recent observational studies contrasting outpatient mastectomies involving stays of less than 24 hours with inpatient mastectomies involving hospital stays of 24 hours.

In contrast, for SB 576, which concerns health care coverage of tobacco cessation services, the targeted literature search resulted in 168 references, including nine meta-analyses. The medical effectiveness team reviewed the meta-analyses, many of which were published in the Cochrane library and were updated as recently as 2004, as well as the recommendations and conclusions of two evidence-based reviews. On the basis of meta-analyses and systemic reviews alone, the medical effectiveness team was able to review the effectiveness of counseling, brief advice, and pharmacotherapy on tobacco cessation.

At least two faculty and/or staff members of the effectiveness team review the abstract for each article found during the literature search to determine its eligibility for inclusion in the study database. The primary reason for exclusion at this stage is that the study was not conducted on a population relevant to the California population. The inclusion and exclusion criteria for articles differ for each review and become a part of the report. In general, the medical effectiveness team restricts the literature search to studies in English. For some outcome measures, such as physiological effects, results from non-United States-based populations may be relevant. For other outcomes, e.g., school absence days for children with asthma, the differences in expectations between U.S. and other settings may be so great that the reviews would be limited to U.S.-based populations. Although the medical effectiveness team strives in the analysis to adhere to the preferred hierarchy of articles as shown in Table 2, observational studies, case–control studies, and even practice guidelines based on consensus or opinion are retained in the study database pending review of the more scientifically rigorous articles.

Although abstracts may not adequately reflect all the results in the full article, some decisions to exclude a manuscript are initially made on that basis. While abstracts may emphasize outcomes with positive rather than negative findings, we expect that few articles with empirical findings would fail to mention those findings in the abstract. We therefore feel reasonably comfortable in excluding articles whose abstracts do not indicate empirical findings. Once the full-text article is retrieved, the effectiveness review team reapplies the initial inclusion and exclusion criteria to ensure the relevance of the study to the proposed mandate. These decisions are based on whether the studies meet inclusion or exclusion criteria, without regard for the conclusions of the study.

Analyzing the Literature

The review of the articles obtained is guided by the following questions: (1) Are the results applicable to the diverse population of California? (2) Does the

intervention have a statistically significant effect? (3) Does the intervention have a clinically meaningful effect? (4) Does the article concern effectiveness as opposed to efficacy? If articles not applicable to the California population are included in meta-analyses or systematic reviews, the team attempts to determine whether their inclusion alters the overall findings of the published reviews (e.g., all the nonapplicable studies show a benefit and the evidence from the remaining studies are equivocal). As an example, if all the studies showing the value of parent training in asthma management were undertaken among highly educated, ethnically homogeneous populations in the upper Midwest and that the effectiveness was greatest during the winter, then such findings would be of limited relevance in California. Although the medical effectiveness team anticipated when CHBRP first became operational that studies would sometimes be excluded because of lack of relevancy to the population of California, no study conducted in the United States has yet to be excluded for this reason.

The full-text article is sometimes not retrieved quickly enough to meet CHBRP deadlines, forcing the team to rely on the published abstract. The abstract may omit information allowing assessment of the relevance to the particular CHBRP review or the comparability of the study participants to the population in California that would be affected by the mandate. 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. Those arriving after this date, but during the time period when a report is under review, are evaluated to see if they would alter the assessment in a substantive way, and if so, they are included.

Summarizing the Evidence and Preparation of the Medical Effectiveness Report

The effectiveness team reviews the results of meta-analyses and other studies for each outcome measure. Not all studies, however, are equally relevant. Judgment sometimes needs to be exercised to "downweight" studies because they are old relative to current medical practice, or of limited applicability to the mandate situation, or of less rigorous methodology. Such decisions are made by the group and documented with the rationale for downweighting or exclusion. Within this framework, two types of summary measures are useful. One reflects the consistency of findings across studies with respect to the measure, the other is a weighted average of the effect.

Based on the weight of the evidence available in terms of relevance, sample size, and methods used, the team assigns a "grade" for each outcome

(Table 3). This is neither a simple "vote counting" with every study counting equally, nor a simple weighted average that assumes all studies are of comparable value except for sample size. The report should present the reader with a sense of the patterns of findings and also provide a sense of the differences among studies. The effectiveness review team first looks for consistency of findings across studies. The same overall (weighted mean) effect may be generated by a situation in which all the studies indicate a benefit versus another in which some show no effect, or even harm, but one large study shows a substantial beneficial effect. The former may be more convincing, if only because it does not rely so heavily on a single study, and there is no contrary evidence to be raised by advocates. Large sample observational studies, especially if there is concern about noncomparability of groups, should not automatically overwhelm small, well-controlled studies. Contrariwise, tightly controlled studies that deal more with efficacy should not automatically overshadow observational studies addressing effectiveness. In discussing the pattern of results, the team takes into account statistical significance, sample size, and relevance, as well as the direction of the effect. A large number of statistically insignificant studies with small samples, but a totally consistent direction of effect can nonetheless be convincing.

If the conclusions of several published meta-analyses differ substantively, the review team will try to determine why. The discrepancies in conclusions might be explained by differences in the inclusion or exclusion criteria for the various meta-analyses or the RCTs comprising them, or some published meta-analyses may use less rigorous criteria. Alternatively, the screening procedure or therapy may have improved over time and be reflected in the later analyses. In such cases, the team may decide not to weight the data from some studies as much as others.

Table 3: Grading System for the Evidence for Each Outcome Measure

Favorable (statistically significant effect)	Findings are uniformly favorable, many or all are statistically significant
Pattern/trend toward favorable (but not statistically significant)	Findings are generally favorable, but there may be none that are statistically significant
Ambiguous/mixed evidence	Some significantly favorable, and some significantly unfavorable findings
Pattern toward no effect/weak evidence	Studies generally find no effect, but this may be due to a lack of statistical power
Unfavorable	Statistical evidence of no effect in literature with sufficient statistical power to make this assessment
Insufficient evidence to make a "call"	Very few relevant findings, so that it is difficult to discern a pattern

For studies with quantifiable outcomes, the team summarizes the specific outcome of interest, for example, the reduced number of emergency room visits following an asthma management program. The team begins the process by tabulating all the studies measuring the specific outcome of interest along with the reported results. They also take into account the relevance and power to detect statistical significant findings of the study. The team also considers the plausibility of the findings and the overall patterns of evidence. Such judgments and the rationale for them are recorded in the final report or its appendices.

Because samples and populations often differ across studies, calculations to determine the overall effectiveness of an intervention begin with a determination of the proportionate effect attributable to the intervention. Studies with more subjects typically have a greater effect on the statistical significance of the outcome and, therefore, are weighted more heavily in estimating the overall effect of the intervention. The studies with the highest and lowest outcome effects demarcate the range of effects observed. (Occasionally implausible extreme values may be omitted, and this is noted.)

The effectiveness report provides the groundwork for the public health and cost impact analyses components of AB 1996. In some cases, the effectiveness review points to issues that should be addressed in other sections. For example, expanding coverage to a new population might generate a widespread acceptance of the intervention and, therefore, increase usage rates among people who are already covered. This scenario would increase the impact on health outcomes of the proposed mandate. On the other hand, the impact on health outcomes is likely to be small if a screening intervention would be covered by a mandate but treatments are not readily available or are not covered, or if the mandated screening intervention is already widely available and used. On the other hand, if the cost and utilization team estimates that coverage would lead to a broad expansion in the indications for use that would result in the intervention being applied to people for whom there is less evidence of a benefit, that would affect the effectiveness assessment.

CONCLUSIONS

The medical effectiveness analysis is a fundamental component of each mandate review undertaken by the CHBRP. The implications of the effectiveness assessment directly affect the public health impact estimates. Some of the effectiveness estimates are incorporated directly in the utilization analyses.

To some extent, the foundation of the effectiveness review builds on the logical steps from mandated coverage to having an impact on individuals,

specifying the scope of the procedures and interventions to be examined and the outcomes to be assessed. The scientific literature is searched for evidence, preferably well-performed meta-analyses, plus those RCTs published after the last available meta-analysis. At the same time, the team members recognize the value of nonrandomized studies and guidelines in informing public policy.

If there is little or no evidence that an intervention is effective, the arguments in favor of mandating its coverage are weaker. On the other hand, multiple well-performed meta-analyses comprised of RCTs, all suggesting that the intervention is beneficial, provide strong evidence in support of the clinical impact of the intervention. It is important, however, to distinguish the situation in which there are many large, well-powered, studies, none (or few) of which indicate the intervention is effective, from (a) the case in which few studies of effectiveness have been done, or (b) they are all of very small size. In the first instance one can say that researchers have looked, but have been unable to find an effect, in the latter two situations, one must say that the research is not available to reach a conclusion. Conveying these distinctions to legislators, rather than to researchers or reviewers, can be a challenge.

The mandate proposals that are the most difficult to assess are those in which the available evidence is not related directly to the mandate and the medical effectiveness team has to use its scientific expertise and judgment in as unbiased a manner as possible to present evidence with supporting rationale. The goal is to create, using a reasoned approach and in a brief period of time, a document with transparent methods, findings, conclusions, and rationale that can withstand critical scrutiny. This may involve occasional judgments that deviate from strict adherence to rigid protocols, but such deviations are sometimes necessary to provide legislators with useful assessments. The CHBRP goal is to provide valid and timely information to a political process. Offering precise or delayed answers to questions more narrow than the mandates we are asked to review would not achieve that goal. Whether the public will think it worth the effort to bring research-based evaluations to the political arena will have to be determined by evaluations over a period of time.

REFERENCES

American Psychiatric Association. 2000. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR®).* Arlington, VA: American Psychiatric Association.

Black, N. 1996. "Why We Need Observational Studies to Evaluate the Effectiveness of Health Care." *British Medical Journal* 312 (7040): 1215–8.

California Health Benefits Review Program. 2005a. "About CHBRP" [accessed on August 1, 2005]. Available at http://www.chbrp.org/

California Health Benefits Review Program. 2005b. "California Health Benefits Review Program Analyses" [accessed on June 20, 2005]. Available at http://www.chbrp.org/analyses.html

California Health and Safety Code. 2005. "HEALTH AND SAFETY CODE SECTION 127660-127665" [accessed on June 20, 2005]. Available at http://www.leginfo.ca.gov/cgi-bin/displaycode?section=hsc&group=127001-128000&file=127660-127665

D'Agostino, R. B., and H. Kwan. 1995. "Measuring Effectiveness." *Medical Care* 33 (4, Suppl): AS95–105.

Des Jarlais, D. C., C. Lyles, and N. Crepaz. 2004. "Improving the Reporting Quality of Nonrandomized Evaluations of Behavioral and Public Health Interventions: The TREND Statement." *American Journal of Public Health* 94 (3): 361–6.

Dieppe, P., C. Bartlett, P. Davey, L. Doyal, and S. Ebrahim. 2004. "Balancing Benefits and Harms: The Example of Non-Steroidal Anti-Inflammatory Drugs." *British Medical Journal* 329 (7456): 31–4.

Holtzman, N. A. 1992. "The Diffusion of New Genetic Tests for Predicting Disease." *FASEB Journal* 6 (10): 2806–12.

Kunz, R., and A. D. Oxman. 1998. "The Unpredictability Paradox: Review of Empirical Comparisons of Randomised and Non-Randomised Clinical Trials." *British Medical Journal* 317 (7167): 1185–90.

MacLehose, R. R., B. C. Reeves, I. M. Harvey, T. A. Sheldon, I. T. Russell, and A. M. Black. 2000. "A Systematic Review of Comparisons of Effect Sizes Derived from Randomised and Non-Randomised Studies." *Health Technology Assessment* 4 (34): 1–154.

Victora, C. G., J. P. Habicht, and J. Bryce. 2004. "Evidence-Based Public Health: Moving beyond Randomized Trials." *American Journal of Public Health* 94 (3): 400–5.