New York Department of Health Evidence-based Review Process for Coverage Determinations

Dossier Methods Guidance

Table of Contents

Introduction

Background

Historically, Medicaid benefits and covered services have rarely been examined on a systematic basis. As an essential part of the New York Medicaid Redesign Team (MRT), New York State is committed to structuring the Medicaid benefit to ensure that all beneficiaries have access to the clinically effective, efficiently delivered services they require. To that end, the New York Department of Health (DOH) has established a systematic process for making decisions about Medicaid benefits using the best available research evidence. The Evidence-based Review Process is designed to support transparent and consistent coverage and payment decisions that align with the Centers for Medicare and Medicaid Services' Triple Aim vision for health care of achieving better health, better quality, and lower costs.

Through the Dossier Process, the Department evaluates available evidence to determine coverage of health care services, procedures and devices (hereafter referred to as services). Covered services must be FDA approved when required and supported by evidence of safety and effectiveness. Services of uncertain value (e.g., high cost with lower cost alterative; high risk; questionable efficacy) will be selected to go through the Dossier Process. Individuals or entities may submit an evidence dossier. Dossiers should be comprehensive and include the most current research available. This dossier submission process will help the State better understand the body of clinical research evidence (hereafter referred to as evidence) related to the service under review, what limitations on use may be appropriate, and whether coverage of the service represents significant value to the people of the State of New York.

This document is intended to provide supplemental information regarding the Evidence-based Review Process, evaluating evidence for methodological quality, and DOH's coverage decision criteria for services under review. Specifically, this document includes discussion on the different types of evidence available in the literature and how they relate to each other, and provides guidance for assessing evidence for methodological quality, determining the overall strength of an evidence body, and assessing the impact of a service.

Dossier Submission Process

A dossier is a collection of resources that gives detailed information on a particular topic. In the context of the DOH Evidence-based Review Process, the dossier submission process is a pathway for individuals or entities to provide evidence on a specific new or existing service. The dossier submission process allows for public involvement in the evidence appraisal for services under review, and organizes the evidence into a consistent framework for DOH to evaluate.

By adding a dossier submission process to the Medicaid benefit review process, New York joins several well-established national and international health care programs in requesting the assistance of the public to collect comprehensive evidence on a service. For instance, the Drug Effectiveness Review Project, a collaboration of public entities that produce systematic, evidence-based reviews of the comparative effectiveness and safety of drugs, has utilized a dossier process since its inception in 2003. Similarly, the Centers for Medicare and Medicaid Services and Washington State Medicaid use an evidence dossier process to accept public submissions of evidence on topics (Sullivan 2009). Internationally, the Institute for Quality and Efficiency in Health Care (IQWiG) in Germany encourages industry to submit evidence dossiers for new pharmaceuticals (IQWiG 2011). The New York DOH Evidence-based Review Process, and subsequently the dossier submission process, builds on the foundation of the New York Medicaid MRT workgroups and is a continuation of New York's commitment to using the best available evidence to make transparent and consistent coverage and payment decisions that increase health, increase quality, and lower costs.

Chapter 1. Methodology of Clinical Evidence Review

Overview

Since the passage of the Patient Protection and Affordable Care Act in 2010, there has been a concentrated national focus on the use of evidence to assist in forming coverage determinations and assuring that health care decision-making is evidence based. However, there are many definitions and understandings of what it means to use evidence in healthcare decision-making and what actually constitutes sufficient evidence to inform such decisions. Additionally, the many different types and sources of evidence further complicate the process. This chapter describes the process for conducting a high quality, transparent evidence review, incorporating evidence into policy decisions, and evaluating the methodological quality of evidence.

Preparing for an Evidence Review

Services reviewed through the Dossier Process may have broad application and use. However, the literature seldom considers services or service categories as a whole and commonly focuses on a service for a specific population(s), compares a service to other specific services, or considers a service for a specific set of outcomes. Thus, in order to frame an evidence review on a service, you must first define the scope of that service. This should be accomplished using the PICO (Population, Intervention, Comparator and Outcome) framework.

PICO

The Population, Intervention, Comparator, and Outcome framework, otherwise known as the PICO, helps to define the literature search parameters and forms the basis of establishing specific research questions on a topic. For services with wide applicability, the PICO can assist in focusing the evidence review to a manageable research topic. The specific components of a PICO include:

- Population: A description of the population of interest. This could include specific health conditions, disease stage or severity, co-morbidities, and other characteristics or demographics (e.g., adults; children with muscular dystrophy; patients with Stage III small cell lung cancer).
- Intervention: The treatment or service under consideration. This could include dose, frequency, methods for administering treatment, etc. (e.g., subcutaneous insulin infusion; computed tomography; arthroscopic surgery).

- Comparator: Any alternatives to which the service is compared. For example, these could include placebo, medications, surgery, behavior modification, usual care, or no care.
- Outcome: The specific short, intermediate, and long-term results of interest. This could include morbidity, mortality, quality of life, complications, and outcomes specific to the condition.

When selecting outcomes, it is important to consider outcomes that are important to patients, rather than focusing only on intermediate or surrogate outcomes, such as incremental point changes of test results (e.g., lipid levels). Effects on morbidity and mortality are nearly always included, but it is often also appropriate to include measures related to personal functionality and other quality of life measures, as well as outcomes specific to the health condition being considered. Some PICO statements also add specifications of time frame (e.g., follow-up for at least a year) and the particular setting of treatment (e.g., primary versus tertiary care).

Study Inclusion Criteria

When reviewing the evidence on a service in a systematic method, it is necessary to specify study inclusion and exclusion criteria before conducting a literature search. Developing study criteria defines the search parameters and can focus the literature search to identify relevant evidence on a service. For example, a search may focus on the comparative effectiveness of a service and thus exclude studies that solely focus on different service dosing or delivery techniques. The study selection criteria may also include:

- *Study type* – possible criteria could include limiting the evidence sources to systematic reviews with and without meta-analysis, or when systematic reviews are not available, limiting the search to randomized controlled trials and/or comparative observational studies (see further discussion below);
- *Language* – searches are often limited to English literature only; and
- *Publication date* – possible criteria could limit the search to recent publications (e.g., last 10 years), build off an existing high quality systematic review and limit search dates in order to update that systematic review search, or search from the time when a service was introduced.

Evidence Selection Criteria

There are a number of different evidence types found in the clinical literature. The New York MRT Basic Benefit Design Workgroup Principle 4 (p. 46) recognizes that the confidence the State can place in a given piece of evidence varies by its design and the quality of the study's execution. To aid in understanding these differences, the various types of evidence available are described in detail below. These descriptions are followed by a discussion of how these

categories relate to each other and their implications for use in policy decision making. While obtaining and understanding the evidence is crucial to informing policy, it does not dictate policy. Sometimes evidence is conflicting, and in other situations, it is insufficient to provide firm guidance. Judgments by policymakers are required to apply the research in a manner that appropriately considers the myriad of other factors that society considers important.

Systematic Review (with or without a meta-analysis) – Type I Evidence

Systematic reviews use specific, transparent, and reproducible methods to identify, appraise, and summarize multiple studies addressing a focused question. Results from individual studies can be summarized in a narrative, or if there is enough similarity between studies, results can be combined in a quantitative summary called a meta-analysis (Agency for Healthcare Research and Quality [AHRQ] 2011).

Generally, high quality systematic reviews include a clearly focused question, a literature search that is sufficiently rigorous to identify all relevant studies, pre-specified criteria to select studies for inclusion, appraisal of study quality, and assessments of study heterogeneity to determine if a meta-analysis would be appropriate (AHRQ 2011).

Randomized Controlled Trials – Type II Evidence

Randomized controlled trials (RCTs) randomly assign participants to two or more study groups that evaluate different interventions. The interventions are usually highly structured and delivered in tightly controlled settings. Randomized controlled trials typically use a variety of techniques such as masking (sometimes know as “blinding”) of study participants, clinicians and investigators, and standardized outcome measures to minimize the potential for bias and maximize the likelihood that study results are valid (Guyatt 2008a). Randomized controlled trials are often referred to as “experimental” because they actually set up a prospective experiment by stating a hypothesis and then conducting a unique research process to test it.

High quality RCTs clearly describe the population, setting, intervention, and comparison groups; randomly allocate patients to study groups; conceal that allocation from all people involved in the trial; have low dropout rates; and report outcomes using intention-to-treat analyses (Guyatt 2008a).

Observational Studies – Type III Evidence

Observational studies are non-experimental studies in which the exposure is not assigned by the researcher, and study groups are not randomly assigned. These types of studies typically use a variety of techniques to adjust for factors that could affect the outcome between study groups. Observational study designs include cohort, cross-sectional, case series, and case-control designs and range in methodological quality (Norris 2010). While all observational studies are non-experimental, not all observational studies are comparative and few

systematically follow groups over time. Prospective cohort studies, for example, are comparative and have the highest methodological rigor within the observational study class whereas non-comparative case series studies are subject to a much higher risk of bias.

High quality observational studies clearly describe study groups, adjust for baseline group differences and other possible confounding factors, blind outcome assessors, and have low attrition rates (Viswanathan 2011). High quality observational studies often employ prospective data collection.

Expert Panel/Professional Guidelines (Type IV Evidence) & Single Expert/Case Report (Type V Evidence)

Expert opinion consists of the opinion of individuals demonstrated to have expertise in their defined field. It can be provided by panels of experts or individuals. In addition, a person can provide an expert opinion in one field (e.g., a cardiology specialist providing expert opinion on myocardial infarction), and a lay opinion in another field (e.g., the same cardiology specialist providing an opinion about arthroscopic knee surgery). Expert opinion is seen as less dependable than other forms of evidence because experts can be influenced by a number of outside factors—including conflict of interest—and because these opinions may not have been widely tested and debated.

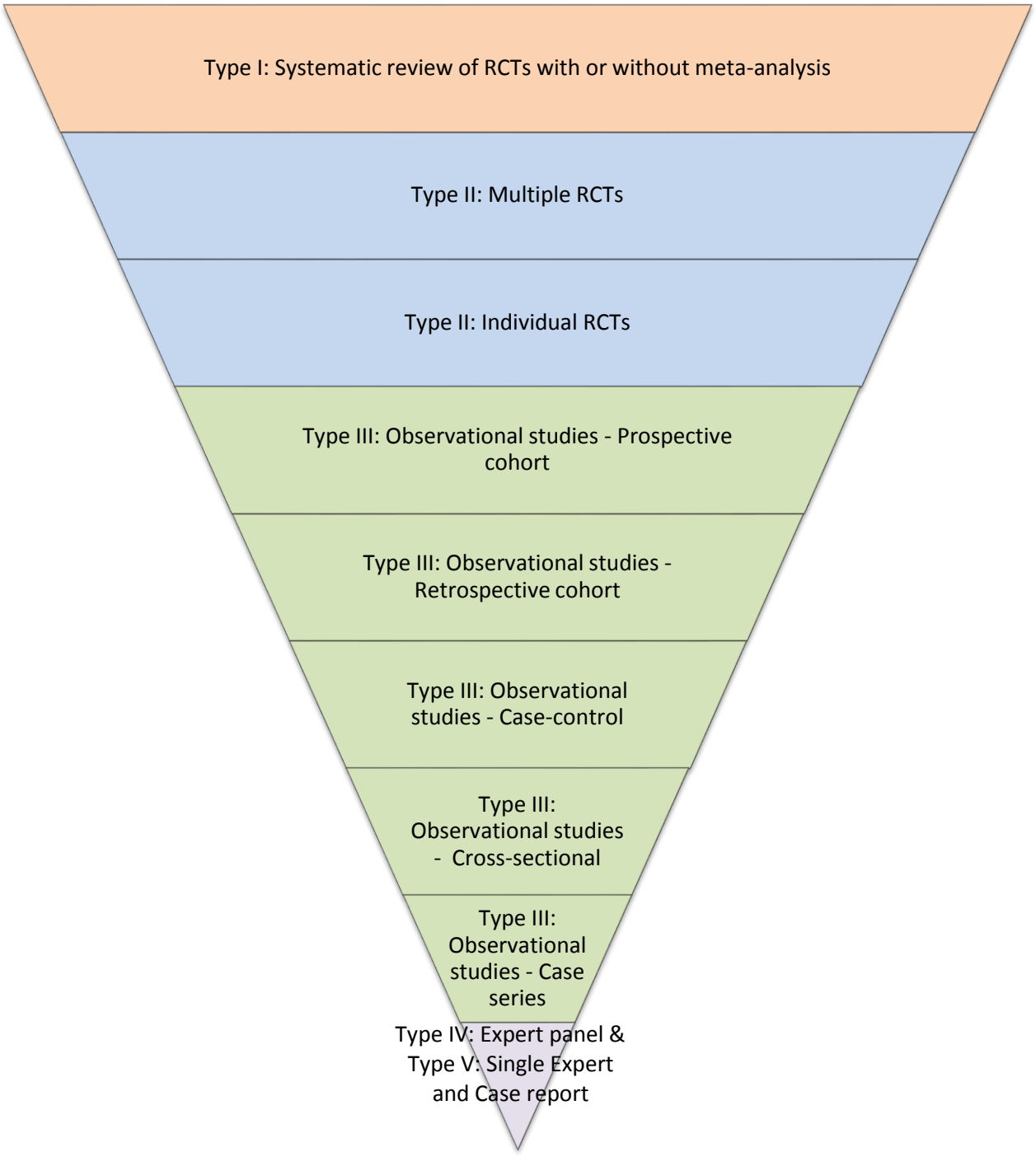
A case report is a descriptive study of a single individual. Case reports commonly include a detailed description of diagnosis, treatment, and follow-up of the patient. Case reports are not an experimental study design and it is difficult to extrapolate the findings to a broader population. While case reports are often useful in describing an unusual presentation of a disease or adverse outcome of treatment, they provide no comparative evidence.

While clinical practice guidelines may be pertinent to a service under review, DOH does not consider them an acceptable evidence source for this process. However, all good quality guidelines are supported by a systematic review of the evidence (Institute of Medicine 2011), and these systematic reviews can be included in the dossier submission.

Hierarchy of Evidence

The MRT Basic Benefit Design Workgroup Principle 4 parallels the widely accepted *hierarchy of evidence,* which provides a framework for ranking the types of evidence to be included in the review of a health care service. The hierarchy (Figure 1) is sorted based on the susceptibility of a type of evidence to bias. Bias is any factor, recognized or not, that distorts the findings of a study. In research studies, bias can influence the observations, results, and conclusions of the study and make them less accurate. Sources at the top of the hierarchy are at the lowest risk of bias (Jonas 2009) (e.g., systematic reviews of RCTs with meta-analysis), and inferences can most accurately be drawn from these sources (Guyatt 2008a).

Figure 1. Hierarchy of Evidence (adapted from Jonas 2009)



While the hierarchy of evidence provides information on the risk of bias of various types of evidence, there are gradations in quality within that hierarchy that affect the accuracy and reliability of each study. All study designs can be performed in ways that increase or decrease the risk of bias; therefore, it is essential to evaluate the methodology of a study before

assuming the results accurately reflect reality. Studies of moderate quality often lack complete information about methods, potentially obscuring important limitations. Poor quality studies have clear flaws that could introduce significant bias.

In preparing for an evidence review it is possible to set criteria for which types of studies to include in the review. For example, if the service under review is being compared to another service, it may not be fruitful to include non-comparative studies such as case series and case reports. Additionally, when conducting a review of the evidence within time and/or fiscal constraints, one option is to limit the literature search to recently published systematic reviews and meta-analyses, particularly when there is an abundance of literature.

Appraising Evidence

When applicable studies and references have been identified, it is then possible to appraise the literature for methodological quality, risk of bias and applicability to the PICO and research questions. Based on the hierarchy of evidence and the differences in methodological rigor for individual studies, appraising the literature can help set the stage for literature synthesis and evidence evaluation.

Quality Appraisal of Individual Studies

Once articles are selected for inclusion in the dossier, the next step is to quality appraise the methodological rigor of the selected studies. While references can appear equal based on the hierarchy of evidence, studies can differ in their level of methodological quality and risk of bias. For example, randomization may not have been done in an unbiased manner in a RCT, data analysis may use non-standard methods, or outcome measures may not be validated or meaningful. An appraisal of quality for each study included will give insight into the overall strength of the evidence for a defined outcome of a service.

The weight of evidence depends on objective indicators of validity and reliability, including the nature and source of the evidence, the empirical characteristics of the studies or trials upon which the evidence is based, and the consistency of the outcome with comparable studies. When viewed in total, this provides an estimate of the true effect, or how much the results reflect the actual benefits and harms of a service for a particular population.

There are a variety of tools that can be used to assess the quality of evidence. The tools used by New York DOH, also known as checklists, are listed in Figure 2 and are provided in the *Dossier Submission Form* document. The checklists are modeled after nationally and internationally recognized processes (adapted from the National Institute on Health and Clinical Excellence (NICE) and the Scottish Intercollegiate Guidelines Network methodologies) in order to satisfy the requirements for scientific validity in the ethical conduct of clinical research. The Quality

Appraisal Checklists provide step by step guidance on how to critically appraise the studies included in the dossier submission.

Figure 2. Quality Appraisal Tools

Study Description	Study Type	Quality Appraisal Checklists
Meta-analysis, systematic review, or technology assessment	Type I	Quality Appraisal Checklist: Systematic Reviews and Meta-analyses
Randomized controlled trial(s)	Type II	Quality Appraisal Checklist: Randomized Controlled Trials
Non-randomized studies (e.g., nonrandomized controlled, pre-post, cohort, case-control, cross-sectional, observational studies, case series, economic studies)	Type III	Quality Appraisal Checklist: Cohort Studies Quality Appraisal Checklist: Cross Over Studies Quality Appraisal Checklist: Diagnostic Test Accuracy Quality Appraisal Checklist: Case Series Quality Appraisal Checklist: Economic Evaluation
Expert panel opinion	Type IV	*n/a*
Case reports Single expert opinion	Type V	*n/a*

Guidance on filling out the Quality Appraisal Checklists is provided below. Every question should be answered with Yes, No, Unclear, or N/A as appropriate. Each Quality Appraisal Checklist has questions pertaining to the *internal validity* of the study. Internal validity refers to how well a study was conducted to minimize bias, or how likely the findings of the study are true.

There are certain elements of study design that contribute to the internal validity of a study. To assist with each Quality Appraisal Checklist, study characteristics that increase internal validity are listed below. As each study is evaluated, it is important to consider the following aspects of study design and make a judgment as to how well the current study meets each criterion:

- **All studies**
  - Study addresses an appropriate and clearly focused question
  - Study includes a clear description of the methodology used
  - Study declares conflicts of interest of authors
  - Study funding source is disclosed

- **Systematic review, meta-analysis and technology assessment**
  - Literature search is sufficiently rigorous to identify all the relevant studies
  - Study quality is assessed and taken into account
  - Criteria used to select studies for inclusion are explicit and appropriate

- **Randomized controlled trial**
  - Assignments of subjects to treatment groups are randomized
  - Adequate concealment method is used
  - Subjects and investigators are masked as to treatment allocation
  - Intervention and control groups are similar at the start of the trial
  - Intervention and control groups receive the same care apart from the intervention(s)
  - The comparison intervention is appropriate
  - Study has an appropriate length of follow-up to detect the outcome of interest
  - All relevant outcomes are measured in a standard, valid and reliable way

- **Cohort Studies**
  - The two groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation
  - Study indicates how many people asked to take part in the study actually participated for each group studied
  - Study assesses likelihood that some eligible subjects might have the outcome at the time of enrollment and incorporates it into the analysis
  - Study reports dropout rates for each arm of the study and includes comparison between full participants and those lost to follow-up, by exposure status
  - Outcomes are clearly defined
  - Assessment of outcome is made “blind” to exposure status
  - When masking of outcome assessment is not possible, discussion about how knowledge of exposure could have influence on the assessment of outcome should be present
  - Measurement of exposure is reliable
  - Evidence from other sources is used to demonstrate that the method of outcome assessment is valid and reliable
  - Exposure level or prognostic factor is assessed more than once
  - Main confounders are identified and taken into account adequately in the design and analysis
  - Confidence intervals are provided

- **Case-control studies (also case series and crossover study designs)**
  - Cases and controls are taken from comparable populations
  - Study uses same exclusion criteria for cases and controls
  - Study provides what proportion of each group (cases and controls) participated in the study
  - Study provides comparison between participants and non-participants to establish their similarities or differences
  - Cases are clearly defined and differentiated from controls
  - Study includes measures to prevent knowledge of primary exposure influencing case ascertainment
  - Exposure status is measured in a standard, valid and reliable way
  - Main potential confounders identified and taken into account in the study design and analysis
  - Confidence intervals are provided for all estimates of effect

- **Diagnostic studies**
  - The spectrum of patients is representative of the patients who will receive the test in practice
  - Study clearly describes selection criteria
  - The reference standard is likely to classify the condition correctly
  - The period between the reference standard and the index test is short enough to be reasonably sure that the target condition did not change between the two tests
  - The whole sample, or a random selection of the sample, was verified using a reference standard of diagnosis
  - Patients received the same reference standard regardless of the index test result
  - The reference standard was independent of the index test (i.e., the index test did not form part of the reference standard)
  - Study describes execution of index test and reference standard in enough detail to permit replication of each test
  - Study interprets the index test results and reference standard results without knowledge of the other respective test
  - Study reports uninterpretable or intermediate test results
  - Study provides explanation of participant withdrawals from study
  - The same clinical data were available when test results were interpreted as would be available when the test is used in practice

- **Economic Evaluations**

- Viewpoints of analysis are clearly stated and justified
- Study uses appropriate critical appraisal considerations for study type
- Study design is appropriate to the stated objective
- Study considers heterogeneity if multiple studies are used
- Study states time horizon of costs and benefits
- Choice of form of economic evaluation is justified in relation to the questions addressed
- Study evaluates meaningful clinical benefit including the appropriateness of outcomes used in the study as measures of effectiveness
- Study uses reasonable and valid examples when comparisons are made
- Study describes reasonable alternatives
- Study measures and values outcomes appropriately

Each Quality Appraisal Checklist assesses the overall methodological quality of an individual study through a series of questions. The Quality Appraisal Checklists are specific to study design and focus on the methodological components specific to each design. Overall study quality should be determined based on the following criteria:

- Good – All or almost all of the criteria have been fulfilled. Where criteria have not been fulfilled, it is unlikely that the conclusions of the review or study would change.
- Fair – Most of the criteria have been fulfilled. Those criteria that have not been fulfilled or not adequately described are thought unlikely to alter the conclusion significantly.
- Poor – Few or no criteria are fulfilled. Those criteria that have not been fulfilled or not adequately described are thought likely to alter the conclusion of the study significantly.

In addition to internal validity and methodological quality, it is important to consider a study's applicability to a broader population. This concept is called the *external validity* of a study. Several of the Quality Appraisal Checklists have sections below the overall quality rating that address external validity. It is important to discuss the external validity of the studies included in the dossier process and how the studies apply to the population described in the topic PICO. Discussion of the external validity is incorporated into the questions included in the *Service Rationale* section of the *Dossier Submission Form*.

Synthesizing the Literature

The synthesis of the submitted studies should be organized by outcomes and follow the format as provided in the *Dossier Submission Form*. Questionnaires and worksheets are provided to

help summarize the findings from each study and organize the study characteristics for easier interpretation. It is important to include any relevant quantification of outcomes that are included in the studies, such as the number needed to treat (NNT), percentages, and quality adjusted life years (QALYs) (see Chapter 2).

Overall Strength of a Body of Evidence

Study types and the methodological quality of individual studies are used to determine the overall strength of evidence for a body of literature. There are a variety of tools that can be used for this purpose. The DOH Dossier Process integrates the use of the Grading of Recommendation Assessment, Development and Evaluation (GRADE) process in order to satisfy the requirements for scientific validity in the ethical conduct of clinical research. The GRADE approach, a system for developing and presenting evidence summaries, was designed to create a single, universal system for evaluating evidence and continues to be developed and refined by a dedicated working group (GRADE Working Group n.d.; Guyatt 2011a). The GRADE system is internationally recognized as the leading systemic approach for synthesizing evidence and determining the strength of an evidence body.

The overall strength of the evidence should be assessed for <u>each</u> outcome[1] in the dossier submission. The GRADE system expresses the degree of confidence that future research will or will not alter the estimate of effect and considers risk of bias as well as the consistency, precision, directness and applicability of the results. It defines the overall strength of a body of evidence for an outcome in the following manner:

- **High** (Highly confident that the true effect lies close to that of the estimate of the effect)
    - Evidence typically consists of systematic reviews, meta-analyses, and randomized controlled trials <u>without</u> important limitations
- **Moderate** (Moderately confident in the estimate of of effect: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is different)
    - Evidence typically consists of systematic reviews, meta-analyses and/or randomized controlled trials with some limitations; or
    - Well-designed large observational studies with additional strengths that guard against potential bias and have large estimates of effects.
- **Low** (Limited confidence in the estimate of the effect: The true effect may be substantially different from the estimate of the effect)

[1] Recommendations and final coverage determinations are available on the DOH Evidence-based Review Process website: health.ny.gov/health_care/Medicaid/redesign/basic_benefit_ebdsp.htm

  - Evidence typically consists of systematic reviews, meta-analyses, and randomized controlled trials with a number of significant limitations; or
  - Observational studies without special strengths.

- **Very Low** (Very little confidence in the estimate of effect: The true effect is likely to be substantially different from the estimate of effect)
  - Evidence typically consists of observational studies with serious limitation and outcomes for which there is very little evidence, or studies with conflicting outcomes.

- **None** (no evidence is available).

The GRADE system works on a hierarchy of evidence similar to the one described above. Randomized controlled trials start as the highest strength of evidence (lowest risk of bias compared to other study designs) but may be down-graded based on methodological flaws, such as a large loss to follow-up or lack of allocation concealment (Guyatt 2011a). The GRADE rating may also be downgraded if studies of comparable type show inconsistent results, or upgraded if study results show a large magnitude of effect (Guyatt 2011a). For all study designs, the overall strength of evidence incorporates consideration of the following (Guyatt 2008b; Guyatt 2011a-2011f, 2011h):

- ***Methodological rigor of study design (e.g., randomized controlled trials vs. observational studies)*** – this captures the risk of bias inherent in a study design;
- ***Study limitation (risk of bias)*** – are assessed with study Quality Appraisal Checklists. Bias may occur from many factors related to aspects of a study's design such as lack of allocation concealment, incomplete accounting of patients and outcomes, and lack of blinding;
- ***Inconsistency of results*** – studies may show inconsistent benefit or harm of a service, and/or report conflicting results. Inconsistency of results increases the chance that future research may change the estimate of effect;
- ***Indirectness of evidence*** – this can occur in studies that do not directly measure the service, outcomes, population, or comparators of interest. Examples include studies that do not include direct head-to-head comparisons of interventions or populations of interest, or studies that use surrogate outcomes instead of patient important outcomes (e.g., cholesterol level instead of risk of myocardial infarction or death);
- ***Imprecision*** – this refers to a study's reported quantitative description of the uncertainty or imprecision in the estimate of effect (e.g., relative risk reduction). Most often studies use confidence intervals, which is a range around a study result (estimate

of effect). GRADE recommends the use of 95% or greater confidence intervals to measure degree of imprecision in a measurement. The 95% confidence interval represents the range of values in which we are 95% sure that the true effect exists. In other words, if a study was conducted 100 times, 95% of the time the true effect (estimated by the study results) would fall within the confidence interval range, and 5% of the study results would either be higher or lower than the confidence interval range. Imprecision, as represented by wide confidence intervals, may be a significant factor in studies with small sample sizes and/or infrequent outcome events;

- ***Publication biases*** – this type of bias can occur when studies are done, but their results are not published in the literature or made available by the study investigators or funding organization. GRADE suggests publication bias should be suspected in a body of literature mainly comprised of small studies that are largely commercially funded. Publication bias can lead to an overestimation of benefit; and
- ***Conflict of interest*** – researchers may receive funding from industry or other advocate sources, and/or studies may be funded by industry or advocacy organization. Conflicts of interest also generally lead to overestimation of beneficial effects.

Identification of any of these factors may lead to a *decrease* in the overall strength of the body of evidence. However, there are several factors to consider that may *increase* the overall strength of the body of evidence, such as (Guyatt 2011g):

- ***Large magnitude effect*** – GRADE suggests increasing the overall strength of the evidence body when good quality observational studies report at least a two-fold reduction or increase in risk;
- ***Plausible confounding*** – when all plausible confounders are shown to reduce a demonstrated treatment effect, this can increase the confidence in the estimate of the true effect;
- ***Dose-response gradient*** – the presence of a dose-response gradient is a strong indicator of a cause – effect relationship and may increase confidence in the estimate of effect; and
- ***Consistency of results across multiple studies*** – the more consistent results across several studies, the greater confidence in the estimate of effect.

Chapter 2. Assessing Impact

There are a number of ways to quantify the effect of a service on health outcomes. Such effects include both harms and benefits, and both must be taken into account to determine the net impact of a service. Measures of effectiveness include such calculations as odds ratio, relative risk, effect size, and number needed to treat. Diagnostic efficacy is measured using sensitivity, specificity, positive and negative predictive values and likelihood ratios. Measures of harm include such calculations as hazard ratio and number needed to harm.

Selection of which measure to use will vary depending on what outcomes are being measured. Potential measures of effect that could be included in a dossier submission are discussed below. Whenever possible, absolute measures of effect (e.g., effect size) should be used rather than relative measures of effect (e.g., odds ratio, relative risk, hazard ratio). See the *Dossier Submission Form* for example calculation of these measures.

Effectiveness Measures

Effect size: demonstrates the absolute magnitude of the difference between two groups.

$$\textbf{Effect Size} = \frac{\textbf{[Mean of experimental group] – [Mean of control group]}}{\textbf{Standard Deviation}_{\textbf{(pooled)}}}$$

$$s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

Number Needed to Treat: The number of people who need to be treated over a specific period of time to promote one additional good outcome (or prevent one additional bad outcome). It is calculated by taking the reciprocal of the *absolute difference* between *experimental groups* for a specific outcome**.**

$$\textbf{Number Needed to Treat (NNT)*} = \frac{\textbf{1}}{\textbf{Absolute Risk Reduction (ARR)}}$$

* NNTs are always rounded to the nearest whole number

$$\textbf{Absolute Risk Reduction (ARR)} = \textbf{Control Event Rate (CER) – Experimental Event Rate (EER)}$$

Odds Ratio**:** The chance of an event occurring in one group compared to the chance of it occurring in another group. The odds ratio (OR) is a measure of effect size and is commonly used to compare results in clinical trials. This is a relative measure of effect.

	Treatment Group	**Control Group**	**Total**
Event Occurs	a	b	a + b
Event Does Not Occur	c	d	c + d
Total	a + c	b + d	

$$\textbf{Odds Ratio (OR)} = \frac{\textbf{a x d}}{\textbf{b x c}}$$

Relative Risk: The probability that an event will occur in the treatment group compared with the control group. Relative risk can be used to compare the risk of developing an outcome (e.g., positive treatment effect or harm) in a treatment group versus a group who receives a placebo or standard of care. This is a relative measure of effect.

$$\textbf{Relative Risk (RR)} = \frac{\textbf{a / (a + b)}}{\textbf{c / (c + d)}}$$

Diagnostic Efficacy Measures

Diagnostic efficacy measures are used to predict how accurate a diagnostic test is and can be used to select an appropriate diagnostic test or series of tests. Measures of diagnostic accuracy include sensitivity, specificity, likelihood ratios, negative predictive value, and positive predictive value and can be calculated using a 2 x 2 table.

		Target Disorder		
		Present	**Absent**	**Totals**
Diagnostic Test Result	**Positive**	a	b	a + b
	Negative	c	d	c + d
	Totals	a + c	b + d	

Likelihood Ratio: A likelihood ratio (LR) provides a direct estimate of how much a test result will change the odds of having a disease. The LR for a positive result (LR+) tells you how much the odds of the disease increase when a test is positive. The LR for a negative result (LR-) tells you

how much the odds of the disease decrease when a test is negative. Likelihood ratios are used to assess the value of performing a diagnostic test.

$$\textbf{Likelihood Ratio + (LR+)} = \frac{\textbf{Sensitivity}}{\textbf{1 - Specificity}}$$

$$\textbf{Likelihood Ratio - (LR-)} = \frac{\textbf{Specificity}}{\textbf{1 - Sensitivity}}$$

For likelihood ratios, the following are general guidelines for interpreting the results:

- >10.0 Large and often conclusive increase in the likelihood of disease
- 5.0 – 10.0 Moderate increase in the likelihood of disease
- 2.0 – 5.0 Small increase in the likelihood of disease
- 1.0 – 2.0 Minimal increase in the likelihood of disease
- 1.0 No change in the likelihood of disease
- 0.5 – 1.0 Minimal decrease in the likelihood of disease
- 0.2 – 0.5 Small decrease in the likelihood of disease
- 0.1 – 0.2 Moderate decrease in the likelihood of disease
- 0 – 0.1 Large and often conclusive decrease in the likelihood of disease

Negative Predictive Value (NPV): A measure used to interpret diagnostic test results. The NPV calculates the probability that a patient with a negative test result really is free of the condition for which the test was conducted.

$$\textbf{Negative Predictive Value (NPV)} = \frac{\textbf{d}}{\textbf{c + d}}$$

Positive Predictive Value (PPV): A measure used to interpret diagnostic test results. The PPV calculates the probability that a patient with a positive test result really does have the condition for which the test was conducted.

$$\textbf{Positive Predictive Value (PVD)} = \frac{\textbf{a}}{\textbf{a + b}}$$

Sensitivity: a measure of the proportion of actual positives that are correctly identified through a diagnostic test. Tests with a high sensitivity have a low false positive rate.

$$\textbf{Sensitivity} = \frac{a}{a + c}$$

Specificity: a measure of the proportion of actual negatives that are correctly identified through a diagnostic test. Tests with a high specificity have a low false negative rate.

$$\textbf{Specificity} = \frac{d}{b + d}$$

Harm Measures

Hazard Ratio: A relative measure of how often a particular event happens in one group compared to how often it happens in another group, over time.

$$\textbf{Hazard Ratio (HR)} = \frac{\textbf{Treatment Hazard Rate}}{\textbf{Control Hazard Rate}}$$

Number Needed to Harm: The number of people who would need to be treated over a specific period of time before one bad outcome of the treatment will occur. It is also calculated by taking the reciprocal of the *absolute difference* between *experimental groups* for a specific outcome**.**

$$\textbf{Number Needed to Harm (NNH)*} = \frac{1}{\textbf{Absolute Risk Increase (ARI)}}$$

* NNHs are always rounded to the nearest whole number

$$\textbf{Absolute Risk Increase (ARI)} = \textbf{Control Event Rate (CER)} - \textbf{Experimental Event Rate (EER)}$$

Determining Net Impact

Based on the information provided in the *Net Impact Worksheet* section of the *Dossier Submission Form*, DOH will determine the net impact of the service under review by considering the relative magnitude of benefits and harms. The net impact can range from "substantial positive net impact" to "negative net impact", and the general framework for reaching these conclusions is outlined in the table below:

Figure 3. Determining the Net Impact

	Typical Magnitude of Benefit	Typical Magnitude of Harms
Substantial Positive Net Impact	Large	Small - None
	Moderate	None
Moderate Positive Net Impact	Large	Moderate
	Moderate	Small
Small Positive Net Impact	Small	None
Zero Net Impact	None	None
	Small	Small
	Moderate	Moderate
	Large	Large
Negative Net Impact	None	Small – Large
	Small	Moderate – Large
	Moderate	Large

The quantitative data provided in the *Net Impact Worksheet* will be used to inform the Department's assessment of the magnitude of benefit and harms. For services where systematic reviews and/or meta-analyses are not available, the net impact will be derived from individual studies.

Chapter 3. Process for Determining Coverage

DOH will review all information and references included in a Dossier Submission as an integral component of the Evidence-based Review and coverage determination process. This will include independent quality appraisal of submitted references by Department staff and external reviewers, when deemed necessary. The Department, or external reviewers, will determine an independent overall strength of evidence for each outcome based on its review of the evidence and determine the net impact for each outcome.

While other factors may influence the coverage decision, in general, DOH non-coverage determinations will be based on the following criteria:

- Zero or negative net impact; or
- Very low strength of the body of evidence; or
- No evidence.

Coverage of the service under review will generally be granted when there is:

- A high strength of evidence; *and*
- A substantial or moderate positive net impact.

For other combinations of strength of evidence and net impact, decisions will depend on the combined results of these two components, and are at the discretion of the New York DOH.

References

Agency for Healthcare Research and Quality (AHRQ). (2011). *Methods guide for effectiveness and comparative effectiveness reviews.* AHRQ Publication No 10(11)-EHC063-EF. Rockville, MD: AHRQ. Retrieved November 10, 2011, from www.effectivehealthcare.ahrq.gov/ehc/products/60/318/MethodsGuide_Prepublication_Draft_20110824.pdf

Apraisal of Guidelines, Research and Evaluation (AGREE) Collaboration. (2009). *Appraisal of Guidelines for Research & Evaluation (AGREE) II Instrument*. Retrieved May 01, 2011 from www.agreetrust.org

GRADE Working Group. (n.d.). Frequently asked questions. Retrieved September 17, 2012, from http://www.gradeworkinggroup.org/FAQ/index.htm

Guyatt, G.H., Oxman, A.D., Aki, E.A., Kunz, R., Vist, G., Brozek, J., Norris, S., et al. (2011a). GRADE guidelines: 1. Introduction – GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology, 64*(4), 383-394.

Guyatt, G.H., Oxman, A.D., Kunz, R., Atkins, D., Brozek, J., Vist, G., et al. (2011b). GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology, 64*(4), 395-400.

Guyatt, G.H., Oxman, A.D., Kunz, R., Brozek, J., Alonso-Coello, P., Rind, D., et al. (2011c). GRADE guidelines: 6. Rating the quality of evidence – imprecision. *Journal of Clinical Epidemiology, 64*(12), 1283-1293.

Guyatt, G.H., Oxman, A.D., Kunz, R., Woodstock, J., Brozek, J., Helfand, M., et al. (2011d). GRADE guidelines: 7. Rating the quality of evidence – inconsistency. *Journal of Clinical Epidemiology, 64*(12), 1294-1302.

Guyatt, G.H., Oxman, A.D., Kunz, R., Woodstock, J., Brozek, J., Helfand, M., et al. (2011e). GRADE guidelines: 8. Rating the quality of evidence – indirectness. *Journal of Clinical Epidemiology, 64*(12), 1303-1310.

Guyatt, G.H., Oxman, A.D., Montori, V., Vist, G., Kunz, R., Bronzek, J., et al. (2011f). GRADE guidelines: 5. Rating the quality of evidence – publication bias. *Journal of Clinical Epidemiology, 64*(12), 1277-1282.

Guyatt, G.H., Oxman, A.D., Sulta, S.,. Glasziou, P., Akl, E.A., Alonso-Coello, P., et al. (2011g). GRADE guidelines: 9. Rating up the quality of evidence. *Journal of Clinical Epidemiology, 64*(12), 1311-1326.

Guyatt, G.H., Oxman, A.D., Vist, G., Kunz, R., Brozek, J., Alonso-Coello, P., et al. (2011h). GRADE guidelines: 4. Rating the quality of evidence – risk of bias. *Journal of Clinical Epidemiology, 64*(4), 407-415.

Guyatt, G.H., Oxman, A.D., Vist, G.E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., et al. (2008a). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal, 336*(7650), 924-926.

Guyatt, G.H., Oxman, A.D., Vist, G.E., Kunz, R., Falck-Ytter, Y., & Schünemann, H.J. (2008b). GRADE: What is "quality of evidence" and why is it important to clinicians? *British Medical Journal, 336*, 995-998.

Institute of Medicine. (2011). *Clinical practice guidelines we can trust*. Washington, D.C.: The National Academies Press. Retrieved December 17, from http://www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust.aspx

Institute für Qualität und Wirtschaflichkeit im Gesundheitswesen (IQWiG). (2011). *General methods*. Cologne: IQWiG. Retrieved September 10, 2012, from https://www.iqwig.de/download/General_Methods_4-0.pdf

Jonas, D., Viswanathan, M., & Crotty, K. (2009). *Selecting evidence for comparative effectiveness reviews: When to use observational studies*. Slide Presentation from the AHRQ 2009 Annual Conference. Retrieved November 10, 2011, from http://www.ahrq.gov/about/annualconf09/viswanathan2.htm

Norris, S., Atkins, D., Bruening, W., et al. (2010). *Selecting observational studies for comparing medical interventions.* Rockville, MD: Agency for Healthcare Research and Quality. Retrieved November 14, 2011, from http://www.effectivehealthcare.ahrq.gov/tasks/sites/ehc/assets/File/MethodsGuideNorris_06042010%281%29.pdf

Sullivan, S.D., Watkins, J., Sweet, B., & Ramsey, S.D. (2009). Health technology assessment in health-care decisions in the United States. *Value in Health, 12*(Suppl 2), S39-S44.

Viswanathan, M., & Berkman, N.D. (2011). Development of the RTI item bank on risk of bias and precision of observational studies. Methods Research Report. Rockville, MD: Agency for Healthcare Research and Quality. Retrieved November 14, 2011, www.effectivehealthcare.ahrq.gov/reports/final.cfm.