Policy: **Evidence Selection and Evaluation**

Date of Implementation: **June 18, 2020**

Product: **All Products**

Related Policies:

- QM 32: Evidence Based Health Information Evaluation/Technology Assessment

American Specialty Health (ASH) lines of business include an array of services which impact the health of members in a variety of ways. These lines of business include American Specialty Health – Specialty, American Specialty Health Management, and American Specialty Health – Fitness.

- The Specialty Network benefits administration manages member benefits for access to healthcare professionals (i.e., contracted in the ASH system and out of network healthcare professional such as chiropractors, physical therapists, and others.
- The Fitness programs provide members with access to fitness facilities, on-line digital resources, and health improvement information.
- The Health Management programs provide members with digital format health information on the web and apps and, when eligible, access to coaching relative to healthy lifestyle (e.g., diet, exercise, tobacco use) chronic illness and musculoskeletal (MSK) conditions.

For each of these lines of business, there is a relevant body of scientific literature that informs ASH policies and guidelines. These bodies of literature are overlapping, but not identical. For each of these lines of business the breadth and depth of that literature varies. And for each of these lines of business there is a different set of clinical/scientific questions that must be answered.

ASH has created policies and procedures that dictate how this scientific literature is identified and evaluated. Broadly speaking, these policies and procedures conform to what has become known as "Evidence Based Medicine" (EBM). The rules for EBM do not change from one scenario to another. And this holds true within ASH. These different lines of business all respect the principles of EBM equally.

QM 33 Revision 3 - ALL
Evidence Selection and Evaluation
Revised – June 15, 2023
To QACPWG for review 05/01/2023
QACPWG reviewed 05/01/2023
To CQT for review 05/15/2023
CQT reviewed 05/15/2023
To QAC-HM for review and approval 05/23/2023
QAC-HM reviewed and approved 05/23/2023
To QIC for review and approval 06/13/2023
QIC reviewed and approved 06/13/2023
To QOC for review and approval 06/15/2023
QOC reviewed and approved 06/15/2023

What does differ is the breadth and depth of available published evidence, the quality of the published evidence, how the evidence will and can be used by different delivery systems (e.g., network providers; coaches; digital content).

Thus, a single set of principles and definitions, as defined below, are required to describe how ASH interprets the scientific literature. These principles are also used by the ASH Internal Evidence Evaluation Committee (IEEC) and the External Evidence Evaluation Committee (EEEC) to guide their work.

Sources of Evidence

Acceptable sources of evidence for Clinical/Coaching Practice Guidelines (CPGs), and Health Information Resources (HIRs), and subsequently the information contained in training manuals and consumer content may include:

- Peer-reviewed journals;
- Peer-reviewed textbooks;
- Professional clinical or health-related organizations;
- Core curriculum material from accredited educational institutions and programs;
- Expert and consensus opinion documents by professional organizations;
- Applicable State and Federal regulatory oversight agencies; and
- Applicable accreditation organizations (e.g., NCQA, URAC).

The standards for evaluation may include, but may not be limited to:

- Randomization
- Sample size
- Pre/post measurement features
- Valid/reliable measures of outcomes
- Inclusion of a control group
- Effect size
- Impartial assessment of study outcome
- Well defined inclusion/exclusion criteria
- Assessment/measurement of co-interventions
- Appropriate statistical analyses
- Appropriate accounting of all study participants: drop-outs, compliance with intervention, loss to follow-up
- Potential effects of conflicts of interest (e.g., study sponsorship or funding).

A variety of study designs may include a sufficient number of the above characteristics to be used for the evaluation of clinical effectiveness. In addition, non-experimental designs

such as prospective cohort studies with non-randomized control groups will be considered as well as variations on this basic study design. Case reports and case series studies will not be used to inform the question of treatment effectiveness; however, these may be reviewed for informational purposes in the presence of other acceptable studies.

The Clinical Evidence Hierarchy

In implementing EBM, it is necessary to understand that all evidence is not the same. Each type of evidence provides different levels of quality, insights into the proof or disproof or a hypothesis, and confidence in the outcomes and results communicated in the document. Certain types of evidence are more persuasive than others. There is general agreement over this hierarchy of evidence; the higher the hierarchy a methodology is ranked (#1 below considered the highest), the more robust and closer to objective proof of fact it is assumed to be. In descending order, the hierarchy of clinical evidence is:

1. **Systematic reviews and meta-analyses.** A systematic review is a method of identifying, reviewing, and analyzing a body of data that uses explicit methods to locate primary studies and explicit criteria to assess their quality. A meta-analysis is a statistical analysis that combines or integrates the results of several independent clinical trials considered by the analyst to be "combinable" usually to the level of re-analyzing the original data. The advantage of a meta-analysis is that it may be possible, by combining study results, to more definitively identify and quantify treatment effects. Both systematic reviews and meta-analyses are typically predicated on the existence of a sufficient body of evidence, particularly of randomized controlled trials.

2. **Randomized controlled trials.** Individuals are randomly allocated to a control group and a group who receive one or more specific intervention. When assigned randomly, the groups are usually very similar on other significant variables, such as age, gender, etc. They are followed up for specific end points and evaluated using specific outcome measures. RCTs are considered the essential method for establishing clinical effectiveness.

3. **Non-randomized controlled trials. These are** similar to RCTs, with the exception that patients either self-select to a particular treatment group or are assigned (non-randomly) to a treatment group. There is a considerable reduction in methodological quality from RCT to Non-RCT. Without randomization it is much harder to determine if patients in each treatment group were comparable, especially on variables that are not measured.

4. **Prospective clinical series or pre-post study design.** A series of patients are systematically evaluated as they are being managed with a single treatment regimen. There is no randomization and no comparison group.

5. **Case reports.** A case report is a systematic analysis and published report based on a single patient or subject. A case report is best utilized as a teaching tool rather than as providing evidence of effectiveness (or ineffectiveness). Sometimes a series of case reports are collected and reported together. It must be emphasized that simply adding more case reports (even if they show consistent results) does not elevate the hierarchical status of case reports.

6. **Expert opinion.** An expert opinion is a consensus of experience and opinion from leaders or experts in a field. Although useful for certain purposes when done systematically, simply invoking expert opinion does not always carry significant methodological weight. Expert opinion with the same conclusions from different organizations and experts across disciplines adds strength to expert opinion but the conclusions must be used with caution as they remain opinion even if multiple experts opine similarly.

7. **Clinical anecdote.** Personal anecdotes and experience (either the health care practitioner's or the patient's) are meaningful only to the individual involved and by itself should not be generalized to inform clinical decisions about others.

While there are no hard and fast rules, it is generally understood that establishing the existence of clinical effectiveness requires some evidence indicating effectiveness in categories #1 and #2 above. In the absence of such evidence, it is necessary to refer to other principles of EBM to determine appropriateness.

ASH does not/will not perform its own formal systematic review of the literature, where that review already exists and is authored by institutions like the Cochrane Collaboration and the Agency for Healthcare Research and Quality (AHRQ). These organizations are the first two sources that ASH committees' access in their evidence evaluation process. As appropriate, other systematic reviews are used as are specific RCTs that represent core findings for the intervention or question at hand.

DEFINITIONS USED BY ASH EVIDENCE EVALUATION COMMITTEES

The ASH evidence evaluation committees have approved the following definitions as references to use during their review of research and evidence:

Clinical effectiveness: The capacity of a therapeutic intervention or coaching methodology to more likely than not result in or contribute to a favorable outcome for the condition and population to which it is applied.

Safety: The terms "safe" and "safety," are used only with specific reference to the absence of direct harm. Direct harm would include any injury to a patient caused by the mechanical, thermal, biological, chemical, pharmacological, electrical, electromagnetic, or psycho-dynamic properties of a diagnostic or therapeutic procedure, and as such, the procedure would be considered unsafe.

Diagnostic utility: The capacity for a diagnostic test or procedure to both:

- Provide reliable and valid information about a specific clinical condition; and
- Influence the course of management of a specific condition in a manner expected to improve prognosis and/or clinical outcomes.

Preponderance of evidence: The conclusion reached on the quality and quantity of considered sources of information regarding the safety, and clinical effectiveness or diagnostic utility. The conclusion is one of the following: Positive, Negative, or Equivocal.

Benefit versus risk profile: The relative effectiveness or utility of a therapeutic intervention, coaching methodology, or diagnostic procedure versus its potential for direct harm. The conclusion is one of the following:

- Positive (benefits outweigh risks);
- Negative (risks outweigh benefits); or
- Equivocal (available information is inconclusive).

Scientific plausibility: A belief, theory, or mechanism of health and disease that can be explained within the existing framework of scientific methods, reasoning, and available knowledge. A treatment intervention or diagnostic procedure is said to be implausible if it requires the existence of forces, mechanisms, or biological processes that are not known to exist within the current framework of scientific methods, reasoning, and available knowledge.

Substitution harm: Compromised clinical outcomes caused by:

- Utilizing a specific diagnostic or therapeutic procedure or coaching methodology whose safety, clinical effectiveness, or diagnostic utility is either unknown or is known to be unsafe, ineffective, or of no diagnostic utility, if implemented *in place of or instead of* a diagnostic or therapeutic procedure or coaching methodology known to be safe, be clinically effective, or to have diagnostic utility; or
- The utilization of a diagnostic or therapeutic procedure or coaching methodology that is substantially less effective or safe than another procedure with established safety, and clinical effectiveness or utility.

Labeling effects: The harm that results from communicating to a patient that they have a condition or a finding that requires care/services or has an effect on their health when that condition or finding is actually incidental, not clinically significant, or not clinically valid (e.g., a diagnosis, physical or mental or functional, is defined but has no support in the evidence).

Safety of Information Provided to Participants/Members/Patients – Significant consideration is given to the process of creating consumer-directed health content to ensure the information provided does not:

- Constitute the practice of medicine, psychology/psychiatry, nursing, dietetics, or other health professions; or
- Prescribe or provide medical, pharmacy, physical therapy or any other clinical service
- Interfere with the recommendations of the participant's health care practitioner; or
- Interfere with the practitioner/patient relationship; or
- Provide information that can easily be misunderstood or misinterpreted by the coach, participant/member/patient, or health care practitioner.

Both the External Evidence Evaluation Committee (EEEC) and the Internal Evidence Evaluation Committee (IEEC) utilize a clinical decision-making algorithm along with the above considerations to determine the applicability and validity of the information being evaluated and applied in ASH program management. Both EEEC and IEEC provide clinical policy recommendations, as applicable, to the appropriate committee(s) for approval. The algorithm determines the appropriate classification (Established, Not Established, or Unacceptable Health/Safety Risk) for each technique/procedure evaluated by either the EEEC or the IEEC. The decision at each point in the algorithm is driven by the supporting information and criteria described above.

The clinical techniques/procedures are categorized as follows:

Established – The diagnostic and/or therapeutic procedure:

- Is established as safe and is established as effective; or
- Is not established as safe; however, the benefit: risk profile is favorable.

Not Established – The diagnostic and/or therapeutic procedure:

- Is established as safe, is not established as effective, and poses no substantial substitution and/or labeling risk; or
- Is not established as safe, the benefit: risk profile is not established as favorable, the benefit: risk profile is not established as unfavorable and poses no substantial substitution risk and/or labeling risk and/or the procedure is not determined to be scientifically implausible.

Health and Safety Risk – The diagnostic and/or therapeutic procedure:

- Is established as safe, is not established as effective, and poses substantial substitution and/or labeling risk; or
- Is not established as safe, the benefit: risk profile is not established as favorable, the benefit: risk profile is established as unfavorable; or
- Is not established as safe, the benefit: risk profile is not established as favorable, the benefit: risk profile is not established as unfavorable; and poses substantial substitution risk and/or labeling risk and/or the procedure is determined to be scientifically implausible.

The Bradford-Hill Criteria

A set of criteria has been established for the purpose of evaluating and analyzing all the disparate evidence that might exist on a specific clinical question. These "Bradford-Hill Criteria" (named after Sir Austin Bradford-Hill, the British epidemiologist who first identified the causal relationship between cigarette smoking and lung cancer) were first developed and used to evaluate the relationship between certain risk factors (e.g., smoking) and specific diseases (e.g., lung cancer), but these have been extended to cover most clinical science questions: safety, effectiveness, diagnostic utility, etc.

At the core of these criteria is the question of causality—is the risk factor, treatment, etc., causally related to the clinical outcome of interest? This clinical outcome might be the development of disease, the prevention of disease, or a resolution or improvement in the disease process. These criteria were explicitly established under the premise that the evidence is always flawed, incomplete, and even contradictory and that the idea of "proof" is nonsensical. These criteria assume that clinical decision-making is always done under a

cloud of uncertainty. Thus, the criteria are particularly useful when applied to practical questions (Is procedure X an appropriate modality?). Herewith, a discussion of those criteria:

1. **Strength of association.** The principle here is that large effects (either good or bad) are easily seen and that these large effects (if real) are unlikely to be produced by spurious effects. For example, the strength of the association between cigarette smoking and lung cancer is a twenty-fold increase in the likelihood of cancer among smokers. An effect of this magnitude is impossible to explain other than by a causal relationship. By contrast, many clinical studies find very small effects. Small effects (e.g., treatment A is slightly better than a placebo) are consistent both with a causal relationship or with no relationship whatever. Small and marginal effect can be explained by many means other than causality.

2. **Consistency.** The reason that small effects are suspect is that there are always many confounding variables that might have been responsible for the results. However, the set of confounding variables are not likely to be the same in different populations, different studies, etc. Thus, if different studies, with different methodologies, in different populations, produce similar results, it strongly suggests a causal relationship. If on the other hand, as is often the case, study results are inconsistent among different populations and different studies, it suggests either an absence of causality, or very small effect sizes.

3. **Specificity.** The principle of specificity is one cause, one effect. That is, it is presumed to be biologically improbable that a single risk factor is a cause of multiple diseases. So, when it is reported or suspected that a particular pollutant is causing MS, breast cancer, arthritis, autism, etc., this evidence is given very little credence. This principle is also extended to the idea that a single treatment is unlikely to be effective for a broad range of conditions. One can think of exceptions to the specificity rule—smoking does cause a variety of problems and exercise is effective in treating/preventing a broad range of diseases-but in general the specificity rule is reliable.

4. **Temporality.** This criterion is applied exclusively to questions of risk factors. The problem arises in cross-sectional population studies of risk factors when it is not clear whether the risk variable was in fact present before the onset of a disease process. This issue is rarely a problem relative to evaluating treatment effectiveness.

5. **Dose-response.** The principle is that if a causal relationship exists, there is likely to be a linear relationship (within limits) between the causal agent and the effect. That is, if a little bit is good (or bad), then a little bit more is likely to be better. This is certainly

true relative to cigarette smoking—one pack a day is much worse that five cigarettes a day. It is also generally true with regard to treatment effectiveness. Assuming a treatment is effective, it should be possible to increase or decrease the treatment dose and achieve different levels of effectiveness. The absence of a dose-response finding suggests that the presumed treatment effectiveness may be spurious.

6. **Scientific plausibility.** The issue of scientific plausibility is both one of the more useful and one of the more contentious criteria. The question is whether the presumed mechanism of a treatment is or is not plausible. What is plausible to one person may not be plausible to another. But this can still be a powerful argument either for or against a particular procedure. Particularly when there is an absence of actual clinical outcomes studies it is entirely legitimate to ask: "Does this make sense? Is there a scientifically plausible explanation for the treatment's effectiveness?" If a given procedure's effectiveness, or diagnostic utility, is predicated upon mechanisms, forces, or biological processes that are not known to exist, the plausibility of that procedure is legitimately in doubt. And on the other hand, if there is a very strong (scientifically plausible) theory to explain an effect, it legitimately strengthens the case in favor of a procedure, practice, and/or lifestyle.

7. **Coherence.** This is related to the above principle of scientific plausibility but is more general. The question is whether a particular approach, theory, or treatment and the cause and effect relationship that is hypothesized is consistent with the broad base of existing knowledge on the subject. The principle is that one should always choose the explanation that requires the fewest assumptions and one should avoid theories that require re-writing the textbooks. To say that a particular theory (or a procedure associated with that theory) is coherent is to say that it is consistent with the existing set of scientific literature on the subject. An incoherent approach is one that requires that we abandon most of what we know about a subject in order to accommodate this new approach.

8. **Analogy.** The principle of analogy is a particularly useful concept when there is a complete absence of clinical literature. It permits us to observe that although we might not have any evidence of effectiveness of Procedure X on Condition B, we do have evidence of effectiveness of Procedure X on Conditions A & C, which are closely related to Condition B. The foremost example of the use of analogy would be: Procedure X = SMT; Condition B = Thoracic spine pain; Conditions A & C = Low back and cervical spine pain, respectively. There is in fact no evidence of effectiveness of Spinal Manipulative Therapy (SMT) (or any treatment, for that matter) for thoracic spine pain. So, we can observe that even though we have zero random controlled trials (RCT) evidence that SMT is effective for thoracic spine pain, we do have evidence for

the lumbar and cervical spine, and these are analogous situations. The question with any analogy is, how closely related are the analogous situations?

9. **Experimental evidence.** There is a variety of experimental evidence that can be brought to bear on a clinical question. Experiments on the mechanical, physical, thermal, electromagnetic, or physiological properties of a treatment are all relevant and may either strengthen the case or weaken the case for a particular procedure. It should always be kept in mind, however, that this type of evidence is more useful in defining how a procedure might work, rather than if it works. The latter question is best addressed directly through clinical science methods.

It is important to note that it is not presumed that each of these Bradford-Hill criteria must be met to establish a plausible cause/effect relationship. Indeed, that is very improbable. Rather, the point is to apply each of these criteria to a question and then weigh the totality of the evidence. The application of the Bradford-Hill criteria does not automatically result in a definitive conclusion, but it narrows considerably the range of possible conclusions. Probably the best instruction as to how and when to use the Bradford-Hill criteria comes from Sir Austin Bradford-Hill himself. He concluded that:

"All scientific work is incomplete – whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time."

ASH clinical/health evidence evaluation, interpretation, and application to business processes follows the principles outlined in this document. ASH committees may make changes to this policy as the knowledge and evidence management principles change within the research community. The governance and processes of policy and guideline review and adoption is future outlined in other ASH clinical policy documents.