Scott Mankowitz

2.2 Evidence-Based Patient Care

CHAPTER OUTLINE

2.2.1 EVIDENCE SOURCES

Many sources of evidence exist between the anecdotal experience of expert clinicians and quality studies that are done at individual institutions. This chapter deals primarily with published sources of evidence, such as journal articles and books. When faced with a journal article or textbook chapter, the informaticist must be able to critically analyze the presented information to assess its value.[1]

Journals typically publish articles according to certain criteria. While every journal has its own submission rules, these are the most common types of articles. For more information, see the “instructions for authors” section of your favorite journal.

1. **Original Research**—In this type of article, the author has generated a hypothesis and tested it using standard statistical methods. There are different types of research studies, which will be reviewed below.
2. **Application** (or **method**) **papers**—The author describes a particular application that he has developed or a procedure for accomplishing some task. Often, the author will attempt to prove that the new method is faster, cheaper, safer, more efficient or generally better than the current methods.
3. **Review papers**—The author collects original research papers from other authors and journals and summarizes them, highlighting the important trends and differences.
4. **Case series**—These are very small studies, usually less than ten patients, and the findings do not have enough statistical weight to be considered original research.
5. **Case report**—The author describes a novel or unusual patient presentation. Since case reports are anecdotal, they have primarily educational value and carry no statistical weight.
6. **Perspectives**—The author (usually an industry leader) gives an opinion on current trends.

[1] Unlike most other chapters, you will note that this one and the next one, Sect. 2.2.2 Evidence Grading, are full of references. Most are available for free on PubMed Central.

7. **Editorials**—One of the journal's editors gives an opinion on current events, trends or other published papers. While journals routinely solicit submissions for other categories, only editors can write editorials.

There are three questions to ask about published material:

1. The first question to ask is: *is the study meaningful?* Does this study address a real clinical question that is relevant for my patients? Does this study present any new information, or does it simply review other available information? Does it agree or disagree with other published studies on the same question?
2. The next question involves how the study was carried out. *Was the study methodology appropriate for the question at hand?* How well did the study minimize or eliminate sources of bias? For that matter, was the study actually done according to the published protocol, or did the authors change the protocol halfway through because they weren't getting the results they wanted?
3. The final question involves the conclusions. *Were the data analyzed correctly?* Do the data justify the conclusions? Statistics are useful to determine if an observation is due to a real phenomenon or could be found by chance alone. Were statistical tests used? If so, were they applied correctly?

One of the problems with this type of repeated questioning is the eventual realization that no study is perfect and that every study has a flaw if you look hard enough.[2]

Lest we become too jaded, let's talk about how to design good studies.

1. Is the study meaningful?

The meaningfulness of a study can only be interpreted in the context of where it will be used. Suppose a researcher discovers that the average centipede has 15 pairs of legs.[3] While an evolutionary biologist may find this data fascinating, a practicing physician would not. *Conclusion: the study is not meaningful for my patients.*

Another example: A computer scientist is working on user-interface research and determines that when a computer becomes unresponsive for more than 2 s, the user begins to lose focus.[4] Again, to the practicing physician, this information has little use… until he realizes that most of the nurses are ignoring the clinical decision support warnings on his EHR because the screens take too long to load. *Conclusion: the study is probably meaningful and relates indirectly to our patient care.*

Another example: When amiodarone was first introduced in 1999, it was used as an anti-arrhythmic for patients in ventricular fibrillation and pulseless ventricular tachycardia. Clinical trials showed that patients who were given amiodarone in the ambulance were much more likely to survive to hospital admission.[5] Intuitively, this sounds like a resounding success. In practice, however, most of those patients ultimately died in the intensive care unit (ICU) or subsequently in a nursing home. Critics of the study asked what could be so good about a drug if all it does is change the location of death from the Emergency Department (ED) to the ICU. *Conclusion: the study is meaningful, but may not dictate a change in practice.*

[2] Some people refer to this as *journal club nihilism* which is the belief that nothing could ever be learned from published scientific research.

[3] Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, et al. (2014) The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede *Strigamia maritima*. PLoS Biol 12(11): e1002005.

[4] Nah, F, "A Study on Tolerable Waiting Time: How Long Are Web Users Willing to Wait?" (2003). AMCIS 2003 Proceedings. Paper 285.

[5] Kudenchuk PJ, Cobb LA, Copass MK, Cummins RO, Doherty AM, Fahrenbruch CE, Hallstrom AP, Murray WA, Olsufka M, Walsh T. Amiodarone for resuscitation after out-of-hospital cardiac arrest due to ventricular fibrillation. N Engl J Med. 1999 Sep 16;341(12):871–8.

What can be done when you have two or more clinical trials that are both well designed and yet flatly contradict each other? For example, etomidate is a potent sedative used for intubation in critically ill patients. In one study, it was shown to cause adrenal suppression, resulting in increased mortality.[6] In another study, there was no significant change in mortality.[7] In addition to these two studies, there are at least a dozen additional studies showing intermediate results. How are we to interpret this conflicting information?

A **systematic review** is a thorough and comprehensive search of the medical literature. It involves several steps: (1) formulate a reasonable question; (2) develop an explicit search strategy, with inclusion and exclusion criteria; (3) Using a large database, select titles, abstracts and manuscripts based on the search strategy; (4) record the findings in a structured format.

A **meta-analysis** is a statistical approach where the data gathered from the systematic review are pooled into a single data table. Care must be taken to ensure that the populations from the various studies are **homogeneous** (i.e. similar enough to be comparable). The hope is that if you combine the data from all the smaller studies, you will come up with a more definitive answer. Sometimes, it works[8] and sometimes it just generates more controversy.[9] Searching for the best available evidence by systematic review and meta-analysis is the cornerstone of **evidence-based medicine.** The Cochrane Collaboration Controlled Trials Register is an important source of studies for meta-analysis.

When comparing data from different studies in a meta-analysis, the **mean difference** (MD) or **weighted mean difference** (WMD) can be used to combine continuous numeric measurements. In cases where the units of measurement are identical between studies (such as millimeters of mercury—mmHg—when studying blood pressure), values in one study can be directly compared with each other. The mean difference is simply the difference between the study group and the control. A mean difference of zero indicates that there is no difference between the groups.

In cases where the units of measurement are different, such as when comparing a 10-point pain scale with a 100-mm visual analog scale, the differences are normalized with respect to the pooled standard deviation of the two groups, known as the **standard mean difference** (SMD). When SMD = 0, there is no difference between the groups. Somewhat arbitrarily, SMD = 0.2 is considered weak effect; SMD = 0.5 is moderate effect; and SMD = 0.8 is strong effect.

The results of a meta-analysis are usually displayed as a forest plot (see Sect. 2.2.3—Clinical Guidelines).

2. *Was the study methodology appropriate?*

There are many different ways to study nature. The prospective, double-blind, randomized controlled trial (RCT) is considered the most reliable methodology. In this kind of trial, when patients volunteer for the study, they are **randomized** into either the study arm or the placebo arm. Neither the investigator nor the subject know which arm has been selected (i.e. **double-blind**). After all the data is gathered, the randomization is revealed in order to perform statistical analysis. For example, an investigator wants to know if a certain drug cures a disease. A hundred people with the disease sign up for the study. Half of the patients are given an injection of the drug, and the other half are given an injection of an inert ingredient such as normal saline (the **placebo**). The syringes have been prepared earlier by the

[6] Komatsu R, You J, Mascha EJ, Sessler DI, Kasuya Y, Turan A. Anesthetic induction with etomidate, rather than propofol, is associated with increased 30-day mortality and cardiovascular morbidity after noncardiac surgery. Anesth Analg. 2013 Dec;117(6):1329–37.

[7] Tekwani KL, Watts HF, Sweis RT, Rzechula KH, Kulstad EB. A comparison of the effects of etomidate and midazolam on hospital length of stay in patients with suspected sepsis: a prospective, randomized study. Ann Emerg Med. 2010;56:481–489.

[8] Gu WJ, Wang F, Tang L, Liu JC. Single-dose etomidate does not increase mortality in patients with sepsis: a systematic review and meta-analysis of randomized controlled trials and observational studies. Chest. 2015 Feb;147(2):335–46.

[9] Chan CM1, Mitchell AL, Shorr AF. Etomidate is associated with mortality and adrenal insufficiency in sepsis: a meta-analysis. Crit Care Med. 2012 Nov;40(11):2945–53.

pharmacy and are labeled with only a serial number. It is impossible to tell which syringe has drug and which has saline. The subjects are then observed to see if the disease improves. Since the data are gathered after the intervention takes place, it is called **prospective** (forward looking).

An RCT can be quite expensive. When an RCT is not possible, another option is the **cohort study**, which is an observational study in which a group (cohort) of people with a particular characteristic is compared to a group of people without that characteristic. Cohort studies can be prospective or retrospective. For example, an investigator wants to know if smokers are more likely to develop degenerative disc disease, so he enrolls 150 people. Using a survey, he finds that 35% of them are smokers. He divides the volunteers into the smoking group and the non-smoking group. He continues to survey them for 10 years until he finds that patients in the smoking cohort have a four-fold higher risk of disc disease.

Very similar to the cohort study is the **case-control study.** In this variation, the outcome (disease) is known, and the attributes (risk factors) are unknown. For example, a researcher identifies 100 people with disc disease (the cases) and 100 people without disc disease (the controls). He surveys those patients to find out which of them have a history of smoking. Since this kind of study is backward-looking, it is called **retrospective.**

A **cross-sectional study** examines a population at a single point in time. Either the entire population or a representative sample is surveyed. There are two main advantages to the cross-sectional study. Firstly, the study is relatively inexpensive and can be completed quickly. Secondly, since the whole population is studied, the prevalence of a disease can be computed, which is useful for population health. The disadvantages include the fact that the study is just a quick snapshot in time, if the study were repeated during a different timeframe, the results could be different. Also, since some diseases result in rapid death, they can be difficult to detect using this method.

Case Reports (or case series) are thorough descriptions of case presentations, management decisions and outcomes. They are very frequently combined with a review of the available literature. Since case reports represent the experience and biases of a single observer, they are termed **anecdotal evidence**.[10]

Case series are not without value. They are used when the disease or outcome is so rare that a larger study would be impossible. Case series can also be valuable when the measured effect is very large and not easily subject to bias. For example, a researcher determines that applying direct pressure to an arterial laceration reduces the amount of blood loss. Since the effect is so dramatic, further study would be considered unethical. Studies with dichotomous endpoints, often involving life and death are called **all or none studies.** If a case series is an all or none study, it is appropriate to rely on previously established disease patterns or to compare results with similar patients who did not receive treatment and were not involved in the study (**historical controls**). The term all-or-none does not imply a particular methodology. For example, an all-or-none RCT carries more weight than an all-or-none case series.

Ecological studies are studies that analyze geographical data to compare populations in terms of incidence or prevalence of disease. At least one variable is measured at the group (e.g. incidence) level instead of the individual level. For example, a researcher collected data from 71 different countries and plotted risk of prostate cancer against annual sugar consumption, and found a direct correlation.[11]

Outcomes research seeks to understand the end results of particular health care practices and interventions in terms of outcomes that matter to patients. Although outcomes research may include traditional clinical trials, it also includes studies where the intervention being evaluated is not a new drug or method, but the opening of a new clinic or enforcement of certain policies by regulatory bodies. For example, traditional researchers may measure visual acuity (VA) before and after a surgical procedure to determine if the procedure is effective. An outcomes researcher would study the effect of adding a new ophthalmology clinic on the health of seniors living in its service area. Further, instead of measuring VA,

[10] Verweij KE, van Buuren H. Oriental cholangiohepatitis (recurrent pyogenic cholangitis): a case series from the Netherlands and brief review of the literature. Neth J Med. 2016 Nov;74(9):401–405.

[11] Colli JL, Colli A. International comparisons of prostate cancer mortality rates with dietary practices and sunlight levels. Urologic Oncology 2006:24;184–194.

they might use a VF-14, a standardized series of subjective measures that are important to the patient (e.g. able to identify a curb, able to drive at night, etc).[12]

Once the study design is decided, the author must work diligently to avoid or minimize sources of bias:

TYPE OF BIAS	WHAT CAN BE DONE ABOUT IT?
Selection bias—certain patients are preferentially enrolled into the study and may not represent the population at large. For example, a study on STD recruits patients at a monastery	Select patients according to rigorous criteria that closely match the intended study population
Channeling bias—patients are steered into study arms by organizers based on preconceived notions. For example an investigator pushes healthier patients into the study arm and sicker patients into the placebo arm, so as to make the study seem more effective than it really is	Use strict, blinded randomization to assign patients to various study arms
Interviewer bias—the investigator asks different questions of the patients, knowing which study arm they are in. Similarly, a subject may behave differently because he knows he is being observed. This is sometimes called the **Hawthorne effect**	Use a double-blind design so the interviewer can not tell which arm the patient is in
Chronology bias—study patients are compared to historical controls, at a time when standard of care may have been different	Use a prospective design
Recall bias—subjects may not remember they symptoms or exposures correctly, especially when they believe that they have a disease. Similarly, reporting bias is when subjects selectively hide information about themselves (e.g. sexual history)	Use only objective data sources
Transfer bias—certain types of patients are more likely to be lost to follow up. For example, the research office is on the fourth floor and there is no elevator. Healthier subjects are more likely to follow up. Also known as attrition bias	Carefully plan ways to contact patients who would otherwise be lost to follow up
Performance bias—in cases where the intervention involves a procedure, it is possible that different operators will perform the procedure differently	Clearly define the method of the procedure, using as much detail as possible. Another possibility is to separately analyze procedures done by each operator
Publication bias—journals tend to report positive findings more readily than negative findings	Before collecting any data, register the study with the NIH registry of clinical trials at ClinicalTrials.gov

For a checklist on how to make a really great RCT, check out the Consolidated Standards of Reporting Trials web site at http://www.consort-statement.org/.

3. *Were the data analyzed correctly?*

Statistical methods are the hallmark of clinical evidence.

The **relative risk (RR)** describes the increased risk associated with an intervention. For example, suppose 21 people are randomized into a control and study group. The study group are given a drug and the control group are given a placebo. In the study group, 3 of 10 patients get sick, and in the control group 5 of 11 patients get sick. The relative risk expresses the risk in the study group compared to the control group. Mathematically,

$$RR = \frac{risk\ in\ study\ group}{risk\ in\ control\ group} = \frac{3/10}{5/11} = \frac{0.3}{0.45} \approx 0.66$$

[12] Boisjoly H, Gresset J, Fontaine N, et al. The VF-14 index of functional visual impairment in candidates for a corneal graft. Am J Ophthalmol. 1999 Jul;128(1):38–44.

If the RR is near 1, the risk in both groups is the same, and the intervention has no measurable effect. When the RR > 1, the study group has *greater* risk.

Instead of RR, some publications refer to **risk reduction**. Using the same example, how much less risk is found in the study group? The **absolute risk reduction (ARR)** is found by subtracting the study group's risk from the that of the control group. It is usually expressed as a percentage.

$$ARR = risk\, in\, control\ group - risk\, in\, study\ group = 0.45 - 0.3 \approx 15\%$$

In order to make the risk reduction appear more dramatic, it is often expressed as the **relative risk reduction (RRR)** which is the ARR divided by the control risk.

$$RRR = \frac{risk\, in\, control\ group - risk\, in\, study\ group}{risk\, in\, control\ group} = \frac{0.45 - 0.3}{0.45} \approx 34\%$$

This tactic is most frequently used by drug manufacturers whose drugs provide modest improvement in a rare condition. For example, suppose 12 of 113 patients got sick in the study group while 15 of 106 got sick in the control group. Let's compare the RR, ARR and RRR.

$$RR = \frac{12/113}{15/106} = \frac{0.106}{0.142} \approx 0.75$$

$$ARR = 0.142 - 0.106 \approx 3.5\%$$

$$RRR = \frac{0.142 - 0.106}{0.142} \approx 25\%$$

When a clinician chooses a treatment plan, a relative risk reduction of 25% sounds much more impressive than 3.5% absolute risk reduction. It is no surprise that pharmaceutical marketing tends to report RRR over ARR.[13]

Another useful measure of effectiveness is the **number needed to treat (NNT)** which is the reciprocal of the ARR. Continuing our example above,

$$NNT = \frac{1}{0.142 - .106} \approx 28$$

Which means that you would have to treat 28 people with the study intervention to prevent one person from getting sick. In general, an NNT less than 10 is considered very useful. When the outcome is harmful, the math is the same, but the NNT is called **number needed to harm (NNH)**.

The **p-value** indicates the probability that the results that were found in this study could be a result of pure chance. The lower the p-value, the more likely the results are reliable. For most medical publications, $p < 0.05$ is considered significant. For example, suppose a study reveals that a drug is effective with $p = 0.03$. That means that if this study were repeated with a placebo, there is a 3% chance that the same or better results would have been obtained by pure chance. Said differently, there is a 97% chance that the study drug is responsible for the measured success. In our example above, $p = 0.43$, which is NOT statistically significant.

A **confidence interval** (CI) indicates the likely range for a particular measure. It is usually reported as the 95% CI, or the range in which 95% of measured values will fall. Continuing our example, the RR is 0.75, and the 95% CI ranges from 0.37 to 1.53.[14] Since the interval includes values which show harm as well as values which show benefit (i.e. it includes

[13] Othman N, Vitry A, Roughead EE. Quality of Pharmaceutical Advertisements in Medical Journals: A Systematic Review. PLoS One. 2009; 4(7): e6350.

[14] Calculation of p-values and confidence intervals is beyond the scope of the exam, but you should know what they mean and how to use them.

numbers greater and less than one), we can conclude that the result is NOT statistically significant at the p = 0.05 level.

Odds Versus Probability

Probability is the chance of having a certain outcome. Odds is the chance of having the desired outcome compared to the other outcomes. Numerically,

Probability = (chance of desired outcome)/(total number of possible outcomes)
Odds = (chance of desired outcome)/(chance of all other outcomes)

In case-control studies, a group of patients who have the risk factor are matched with control patients who do not. Since the prevalence of the risk factor is unknown, the relative risk can't be calculated. Instead, we use the **odds ratio** which compares the odds of having the finding in patients with the risk factor compared to patients without the risk factor. In general, OR > 1 means that the event is more likely to occur in the exposed group; OR < 1 means that it is less likely; and OR = 1 means that there is no difference between the two groups.

Let's try an example using the same numbers as before. A novel gene is found in 10 people (study group). These are matched with 11 people who do not have the gene (control group). In the study group, 3 people get sick and 7 do not. In the control group, 7 get sick and 4 do not.

$$OR = \frac{odds\ of\ getting\ sick\ in\ study\ group}{odds\ of\ getting\ sick\ in\ control\ group} = \frac{3/7}{7/4} = .24$$

Since the OR is less than 1, it means that getting sick is less likely in the group that has the novel gene.

For a nice primer on statistics, see Medical Statistics Made Easy by Harris and Taylor.[15] For descriptions of Odds Ratio, Relative risk, Likelihood ratios, Sensitivity, Specificity, Receiver Operating Characteristic curves, see Sect. 2.1.2.5.

2.2.2 EVIDENCE GRADING

Clinical evidence comes from many sources, with varying degree of quality. Ideally, evidence of higher quality should have a greater effect on our clinical practice. For an explanation of the different types of studies, see Sect. 2.2.1 Evidence Sources.

The highest grade of evidence is given to conclusions are unlikely to change with the introduction of new data. For example, things that are well studied by many investigators in many institutions and have all found very similar results with little variation. The lowest grade is given to recommendations or hypotheses proffered with little to no experimental evidence at all, such as the proceedings from an expert panel or basic science studies which have been extrapolated into clinical practice. Some fields of medicine are more likely to publish lower quality evidence than others.[16]

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group suggests four domains to assess the quality of studies[17]

1. **Study Design**. In general, RCTs are considered more definitive than observational studies. For example, several observational studies in the 1990's showed that hormone replacement therapy was beneficial in preventing coronary artery disease. In the early

[15] http://sumed.sun.ac.za/Portals/0/Repository/Medical-Statistics-Made-Easy.3f8ceb88-f35b-4f29-8d70-17e78ce071d8.pdf

[16] Loiselle F1, Mahabir RC, Harrop AR. Levels of evidence in plastic surgery research over 20 years. Plast Reconstr Surg. 2008 Apr;121(4):207e–11e.

[17] Grading quality of evidence and strength of recommendations. GRADE Working Group. BMJ. 2004 Jun 19; 328(7454): 1490.

2000's, there were RCTs that suggested otherwise. In this case, the RCTs were considered a higher grade of evidence than the observational studies. There are some cases where RCTs are not feasible or may not translate well to a population being studied, such as when studying rare complications of a procedure.

2. **Study Quality.** Studies are weakened by systematic flaws that introduce bias. For example, inadequate blinding, inappropriate allocation of study arms, and insufficient follow up all decrease the validity of studies
3. **Consistency.** When studies are repeated across populations, the data should essentially agree. If there is broad discrepancy between studies of the same entity, the overall confidence in the studies decreases. In some cases, however, a subgroup analysis may explain differences and bolster consistency. (**Subgroup analysis** refers to examining subsections of the study population to search for additional correlations. For example, Sacks[18] studied the effects of pravastatin on patients with previous myocardial infarction and found that pravastatin lowered the risk of recurrent coronary events by 24%. However, when he separated the population based on sex, he found that women had a 46% risk reduction, while men had only 20%)
4. **Directness.** When a study is planned, there must be a direct relationship between the population being studied and the population being treated. This may take several forms.

   (a) **Population differences.** For example, pharmaceuticals are usually tested on healthy volunteers. It is uncertain how findings of these studies may relate to patients who are older and sicker.
   (b) **Use of surrogate markers.** Studies that measure disease-oriented outcomes have less directness than those that measure patient-oriented outcomes. For example, it is known that patients with acute myocardial infarction (AMI) are at greater risk for arrhythmia. It is also known that patients with arrhythmia are at greater risk for sudden death. It seems logical, therefore that arrhythmia suppression should save lives. In other words, the arrhythmia is a **surrogate marker** for death. Using surrogate markers can lead to unexpected results. In the case of AMI, it turns out that preventing arrhythmia with anti-arrhythmics is associated with greater mortality.[19]
   (c) **No direct comparison.** When two interventions are compared to a third, but never with each other, it can lead to indirectness. For example, suppose there are studies comparing aspirin to placebo and plavix to placebo, but no trials that compared aspirin with plavix.

5. **Strength of association.** In studies where the finding of interest is very dramatic, such as all or none studies.
6. **Evidence of a dose response gradient.** Studies which show a direct correlation between increasing intensity of intervention and patient response are generally considered stronger.
7. When reasonable confounding would have pushed the effect in one direction, but the findings are the opposite, it lends credence to the results. For example, one would assume that overall care would be better in for-profit hospitals because of the greater availability of resources and intensity of services. Moreover, there would exist a selection bias which would steer wealthier, healthier patients to for-profit hospitals and sicker, poorer patients to non-profit hospitals. However, in a meta-analysis, it was found that mortality was actually higher in for-profit hospitals.[20] Since one would expect that bias would reduce the effect, the data is all the more convincing.

[18] The Effect of Pravastatin on Coronary Events after Myocardial Infarction in Patients with Average Cholesterol Levels. Sacks FM, Pfeffer MA, Moye LA, et al. N Engl J Med 1996; 335:1001–1009.

[19] Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The cardiac arrhythmia suppression trial. N Engl J Med 1991;324: 781–8

[20] Tanne, JH. Mortality higher at for-profit hospitals. BMJ. 2002 Jun 8; 324(7350): 1351.

Based on these guidelines, evidence is given a grade, as shown:

CODE	QUALITY OF EVIDENCE	DEFINITION	SOURCE
A	High	Further research is very unlikely to change our confidence in the estimate of effect	Several high-quality studies with consistent results (in special cases: one large, high-quality multi-center trial)
B	Moderate	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate	One high-quality study, or several studies with some limitations
C	Low	Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate	One or more studies with severe limitations
D	Very low	Any estimate of effect is very uncertain	Expert opinion with no direct research evidence, or one or more studies with very severe limitations

The agency for healthcare quality and research (AHRQ) supports the EPC (Evidence-based Practice Center) approach which is similar to GRADE. Like GRADE, it requires assessment of four domains: risk of bias, consistency, directness, and precision. Additional domains to be used when appropriate include dose-response association, presence of confounders that would diminish an observed effect, strength of association, and publication bias.

Precision refers to the lack of random errors in measurement, usually caused by inaccuracies in the measuring instrument. For example, a person gets on the bathroom scale and it reads 195.4. He steps off the scale and then gets back on and this time it reads 195.2. While precision is a lack of random error, **validity** refers to a lack of systematic error. **Internal validity** relates to the ability of the study's interventions to explain the findings. Poor internal validity would be caused by systematic error in the study. **External validity** refers to the ability of the study to generalize to other populations.

Another taxonomy which is commonly used is the Strength-of-Recommendation Taxonomy (SORT).

CODE	DEFINITION
A	Consistent, good-quality patient-oriented evidence
B	Inconsistent or limited-quality patient-oriented evidence
C	Consensus, disease-oriented evidence, usual practice, expert opinion, or case series for studies of diagnosis, treatment, prevention, or screening

The following list can be useful to rank the strength of evidence (in decreasing order). Assume that all systematic reviews are homogeneous.

1. Systematic review of Randomized Controlled Trials (RCT)
2. Individual RCT with narrow confidence interval
3. All or none RCT
4. Systematic review of cohort studies
5. Individual cohort study
6. RCT with quality issues (e.g. poor follow up)

7. Outcomes Research
8. Ecological studies
9. Systematic review of case-control studies
10. Individual case-control study
11. Case series
12. Lower quality cohort and case-control studies
13. Expert clinical opinion
14. Basic science studies

Specialty societies often give recommendations on how to address certain common problems. While some recommendations are strong, others are weaker. In order to define the strength of these recommendations, classification systems have been developed. (See Box 4-1).

Box 4-1: Evidence Grading and Recommendation Classification

Evidence grading should not be confused with classification of recommendation. In some cases, the evidence for a particular theory may be compelling, but it does not result in advisories for clinical practice. For example, the American Heart Association gives many recommendations according to the following scale

Classification of Recommendations

- **Class I:** Conditions for which there is evidence, general agreement, or both that a given procedure or treatment is useful and effective.
- **Class II:** Conditions for which there is conflicting evidence, a divergence of opinion, or both about the usefulness/efficacy of a procedure or treatment.
- **Class IIa:** Weight of evidence/opinion is in favor of usefulness/efficacy.
- **Class IIb:** Usefulness/efficacy is less well established by evidence/opinion.
- **Class III:** Conditions for which there is evidence, general agreement, or both that the procedure/treatment is not useful/effective and in some cases may be harmful.

The AHA employs a three-tier level of evidence scale similar to GRADE. The level of evidence is recorded separately from the classification recommendation.

Level of Evidence

- **Level of Evidence A:** Data derived from multiple randomized clinical trials
- **Level of Evidence B:** Data derived from a single randomized trial or nonrandomized studies
- **Level of Evidence C:** Consensus opinion of experts

2.2.3 CLINICAL GUIDELINES

Samuel Alfano

Clinical practice guidelines (CPGs) are statements that include recommendations intended to optimize patient care that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options.[21]

Typically, guidelines have two parts. The foundation is a systematic review of the research evidence bearing on a clinical question, focused on the strength of the evidence. The second

[21] Consensus report, Institute of Medicine. Clinical practice guidelines we can trust. March 23, 2011. http://www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust.aspx (Accessed on January 13, 2012).

part, built on this foundation, is a set of recommendations, involving both the evidence as well as the evaluator's value judgments regarding the risks and benefits of various care options, addressing how patients with that condition should be managed. Guidelines are suggestions for care, not rules. There will always be individual patients who should be managed differently. Legitimate reasons for departing from the guideline may include: biologic differences in drug metabolism; immune response; genetic endowment; the presence of comorbid conditions; availability of resources as determined by the social and economic environment; and patient preferences.

The National Guideline Clearinghouse (NGC) is a public resource maintained by the Agency for Healthcare Research and Quality (AHRQ). The NGC provides structured, standardized summaries of guidelines from most major medical organizations and clinical specialty societies. To be included in the NGC compendium, guidelines are required to meet specific criteria based on the Institute of Medicine (IOM) definition of a clinical practice guideline, such as incorporating a systematic review and an assessment of benefits and harms.

Clinical practice guidelines provide a series of steps, often in algorithm form, for providing clinical care. They may consist of any combination of text, tables, and flowcharts. The typical steps in a CPG algorithm include the following:

Action—perform a specific action.
Conditional—carry out an action (or actions) based on defined criteria.
Branch—direct flow to one or more additional steps.
Synchronization—converge paths back from branches to a common outcome/end point.

An example of a generic algorithm is shown in Fig. 4-1.

Research has shown that CPGs have the potential to reduce inappropriate practice variation, enhance translation of research into practice, and improve healthcare quality and safety. CPGs have also had an important influence on development of physician and hospital performance measures. While CPGs are useful to guide therapy, certain factors may undermine their quality and trustworthiness. These include variable quality of individual scientific studies; limitations in the systematic reviews upon which CPGs are based; lack of transparency of the development groups' methodologies (particularly with respect to evidence quality and strength of recommendation appraisals); failure to convene multi-stakeholder, multidisciplinary guideline development groups, and corresponding non-reconciliation of conflicting guidelines; unmanaged conflicts of interest (COI); and overall failure to use rigorous methodologies in CPG development.

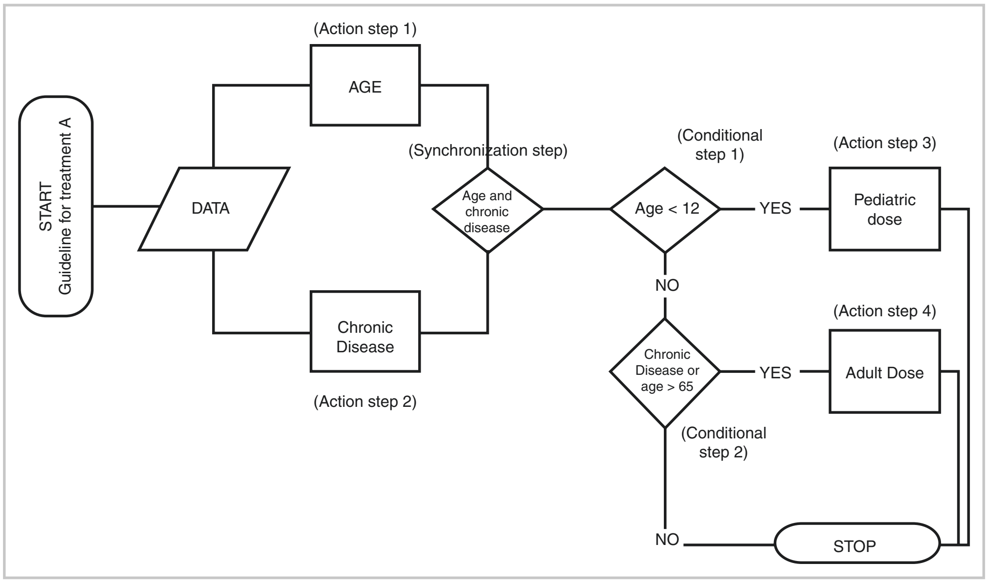


FIGURE 4-1

Example algorithm of a clinical practice guideline

When appraising a CPG, there are several components to consider. Did the developers carry out a comprehensive, reproducible literature search within the past 12 months? Are each of the recommendations rated as to the level of evidence (LOE) on which it is based? Is it linked to one or more specific citations? Is the CPG applicable only to very narrow, specific clinical settings or is it more broadly applicable (generalizable) and economically feasible?

Guidelines may have limitations in their use in clinical practice. They may not apply in complex patients with multiple comorbidities. They may be incomplete or inaccurate. They may be difficult to implement in an EHR setting and, they may have been developed by those with conflicts of interest, such as specialty societies.

The Institute of Medicine established a set of standards for guidelines to ensure validity, accuracy and reliability. They include the following metrics[22]:

- Has an explicit description of development and funding processes that is publicly accessible
- Follows a transparent process that minimizes bias, distortion, and conflicts of interest
- Is developed by a multidisciplinary panel comprising clinicians, methodological experts and representatives, including a patient or consumer, of populations expected to be affected by the guideline
- Uses rigorous systematic evidence review and considers quality, quantity, and consistency of the aggregate of available evidence
- Summarizes evidence (and evidentiary gaps) about potential benefits and harms relevant to each recommendation
- Explains the part that values, opinion, theory, and clinical experience play in deriving recommendations
- Provides a rating of the level of confidence in the evidence underpinning each recommendation and a rating of the strength of each recommendation
- Undergoes extensive external review that includes an open period for public comment
- Has a mechanism for revision when new evidence becomes available

Evidence in guidelines is often rated as to its usefulness and validity. Practice guidelines levels of evidence and grades of recommendations used by the National Guideline Clearinghouse are shown below.

Levels of Evidence:

IA	Evidence from meta-analysis of randomized controlled trials
IB	Evidence from at least one randomized controlled trial
IIA	Evidence from at least one controlled study without randomization
IIB	Evidence from at least one other type of quasi-experimental study
III	Evidence from nonexperimental descriptive studies, such as comparative studies, correlation studies, and case-control studies
IV	Evidence from expert committee reports or opinions or clinical experience of respected authorities, or both

Grades of Recommendations:

A	Directly based on Level I evidence
B	Directly based on Level II evidence or extrapolated recommendations from Level I evidence
C	Directly based on Level III evidence or extrapolated recommendations from Level I or II evidence
D	Directly based on Level IV evidence or extrapolated recommendations from Level I, II, or III evidence

The results of a meta-analysis are often reported using forest plot (see Fig. 4-2). This is a graphical comparison of the results from a number of studies. Although forest plots can take several forms, they are commonly presented with two columns. The left-hand column lists

[22] Graham R, Mancher M, Wolman DM, Greenfield S, Steinberg E, Eds. Clinical practice guidelines we can trust. Institute of Medicine. 2011. National Academies Press, Washington, DC.

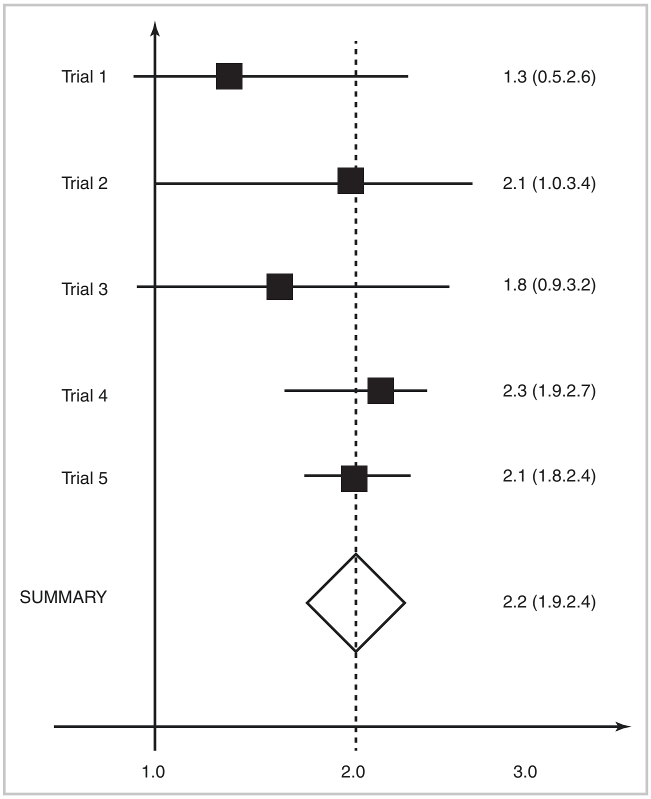


FIGURE 4-2

Forest Plot of 5 randomized studies. The x-axis shows odds ratio (OR) Trials 1,2,and 3 touch the "no effect" line (OR = 1) and thus by themselves are not statistically significant. When taken as a whole and combined with trials 4 and 5, the summary becomes significant

the names of the studies while the right-hand column is a plot of the measure of effect (*e.g.* an odds ratio) often represented by a square with confidence intervals represented by horizontal lines. The graph may be plotted on a logarithmic scale when using odds ratios or other ratio-based effect measures, so that the confidence intervals are symmetrical about the means from each study and to ensure undue emphasis is not given to odds ratios greater than 1 when compared to those less than 1. The area of each square is proportional to the study's weight in the meta-analysis. The overall measure of effect is commonly plotted as a diamond, the lateral points of which indicate confidence intervals for this estimate, with a dashed vertical line to show its position relative to other studies. A vertical line representing no effect is also plotted (e.g. at OR = 1). If the confidence intervals broach this line, it indicates a lack of statistical significance.

Comparative Effectiveness Research (CER)

According to the Institute of Medicine, "Comparative effectiveness research is the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care."[23]

The key elements of CER are (*a*) direct comparisons of active treatments; (*b*) study patients, clinicians, and interventions that are representative of usual practice; and (*c*) a focus on helping patients, clinicians, and policy makers to make informed choices.

These elements have several implications for the principal methods of CER, which are randomized trials, observational research, and decision analysis. Because differences in treatment effects will be smaller in head-to-head comparisons of active treatments than in placebo-controlled trials, studies will require larger study populations, longer follow-up, or both. These requirements are more easily met with observational studies that use large data sets collected over the course of usual practice. The results of CER will often sacrifice internal validity (such as precision and reliability of measurements) for external validity (how well the study may be extrapolated to other populations). The best way to assure external

[23] Sox HC. Defining comparative effectiveness research: the importance of getting it right. Med Care. 2010 Jun;48(6 Suppl):S7–8.

validity is to study interventions, clinicians, and patients that are typical of community practice, which will often mean using data that were not gathered for research purposes and are therefore plagued with missing data and subject to confounding.

While there are many challenges to successful CER, especially when using community data that is not designed specifically for research purposes, statistical tools can help deal with confounding and missing data.

Appropriate use criteria (AUC, sometimes referred to as "appropriateness criteria") are a variation on clinical practice guidelines, but differ in several ways.

- Appropriateness is tailored as much as possible to the specific characteristics of individual patients.
- AUC cover a broader array of specific conditions, sometimes hundreds for a given test or treatment decision, to encompass the majority of practice situations. Appropriateness may relate to individual patient demographic characteristics, clinical history, risk scores, and/or symptoms and signs.
- AUC are based upon scientific evidence where possible but, in part as a result of the breadth of specific clinical conditions addressed, rely more on expert opinion than do rigorous clinical practice guidelines.
- Panels that formulate AUC are typically comprised of representatives from the specialties that manage patients with the clinical question under consideration.
- A common approach is to rate specific clinical scenarios (or potential indications for an intervention) on a 1–9 scale by each panelist before and after a group face-to-face discussion, and the median of the panel ratings is then used to designate each clinical scenario as appropriate, uncertain, or inappropriate.
- AUC are intended not only to help clinicians make sound clinical decisions, but also to educate patients and improve the effectiveness, efficiency, and equity of care.

An example of AUC are the Appropriate Use Criteria published by the American College of Radiology to help in ordering advanced diagnostic studies for specific conditions, such as MRI in back pain patients.

Guidance Statements

The American College of Physicians (ACP) has issued several "guidance statements" that are based on reviews of other guidelines and not a de novo systematic review of evidence. These address conditions such as breast cancer screening in women aged 40–49, prostate cancer screening, or colorectal cancer screening for which there is general agreement on all or most of the eligible trials, but difference in how the evidence from those trials is interpreted. The ACP reviews existing guidelines on the topic and appraises them using the Appraisal of Guidelines, Research, and Evaluation (AGREE II)[24] criteria. The ACP then formulates its own guidance, based on a review of the literature, recommendations in existing high-quality guidelines, and considering the needs and values of its general internist membership and their patients.

2.2.4 IMPLEMENTATION OF GUIDELINES AS CLINICAL ALGORITHMS

Samuel Alfano

Clinical practice guidelines are developed and implemented to achieve the following objectives:

- To deliver effective care based on current evidence.
- To resolve common problems in the clinical setting (e.g. poor blood pressure control).

[24] Brouwers MC, Kho ME, Browman GP, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. CMAJ 2010; 182:E839.

- To achieve excellence in care delivery by meeting or exceeding regulatory requirements.
- To introduce an innovative method of diagnosis or treatment.
- To eliminate use of interventions not recognized as best practice.

Clinical practice guidelines development is a rapidly expanding area, particularly since the introduction of electronic medical records and the use of decision support tools. Despite this, there continues to be variations in health-care practice that do not reflect best practice evidence. Moreover, there are ongoing challenges in promoting full utilization of guidelines by healthcare practitioners, particularly if they are not effectively introduced, implemented and supported. Multiple studies exist in the literature about what techniques help to influence adoption of clinical practice guidelines.

As early as 1999, Moulding[25] et al. described a multistep process for adoption of clinical changes. **Adoption** occurs through a series of communication channels over a period of time among the members of a similar group and is predictable and can be managed. It includes the following five steps.

1. **Knowledge** is the first stage of adoption. During this first stage, individuals are first exposed to an innovation, but have not yet been inspired to find out more information about it. Introduction to why the guideline was developed, what the expected outcome may be, and how the guideline will be used is presented.
2. **Persuasion** comes after introduction. The individual is interested in the innovation and actively seeks more information. It is here that a detailed understanding of how the guideline works and will be implemented is shared. Important in this stage is to encourage clinicians to become invested in the guideline and be allowed to offer critiques and suggestions for improvement.
3. **Decision** is the stage where the clinicians will take the concept of the change and weigh the advantages/disadvantages of adopting the innovation. They can then decide whether to adopt or reject the innovation. Identifying early adopters (see below) and opinion leaders among the clinical staff is important at this stage.
4. **Implementation** is where the tool is made available to clinicians and where they make a final determination to use the guideline. During this stage, the clinician decides whether or not the tool is easy to use, disrupts his workflow, and adds value to his treatment of patients.
5. **Confirmation** is where the clinician finalizes his decision to continue using the innovation. This leads to future use of the guideline as well as the clinician making it part of his routine work flow.

Further, five stages of assessment are usually needed to ensure full adoption and implementation. These have been described by Macdonald.[26]

Step 1: Assessment of Practitioners' Stage of Readiness to Change

Assessment of a practitioner's stage of readiness to change will help to ensure an appropriate mix of dissemination and implementation strategies. Different types of change strategies should be matched with each stage of readiness to change. There are commercially available evaluative tools, such as online surveys, that could be given to practitioners to provide feedback on readiness to adopt different sets of guidelines. Often, small focus groups or department meetings can be a source of discussion used to judge readiness for guideline adoption.

[25] Moulding NT, Silagy CA, Weller DP. A framework for effective management of change in clinical practice: dissemination and implementation of clinical practice guidelines, *Qual Health Care.* 1999 Sep; 8(3): 177–183.

[26] Macdonald G. Communication theory and health promotion. In: Bunton R, Macdonald G, editors. Health promotion: disciplines and diversity. London: Routledge, 1992.

Step 2: Assessment of Specific Barriers to Guideline Use
Assessment of the specific nature of competency based, social, and organizational barriers to guideline use will further ensure that appropriate strategies are selected. Assessment might be undertaken through various means, such as surveys of clinicians using questionnaires, interviews, or group consultations. The use of qualitative methods will provide a more detailed and comprehensive picture of practitioners' needs. Other relevant groups such as patients, other health professionals, and opinion leaders might also be consulted.

Step 3: Determination of Appropriate Level of Intervention
It is important to make an assessment of which level of intervention—individual/group or population—best addresses identified barriers and clinicians' stage of readiness to change before designing dissemination and implementation programs. For example, it may be unnecessary to use national opinion leaders when most practitioners are already positively disposed towards a particular guideline. It may be important, however, to work at the local level with specific groups of practitioners where there is less support.

Step 4: Design of Dissemination and Implementation Strategies
Strategies are selected and designed based on step three. Although the above framework can distinguish in some detail which strategies are likely to have the greatest impact in each stage of change, this depends on the organization's readiness and culture. The interventions most likely to induce change are those that require the clinician's participation in the change process.[27] No single implementation strategy is effective in all circumstances for all people. It is generally understood that the simple dissemination of guidelines is likely to have no impact at all on implementation.[28]

Change will occur only if specific interventions designed to encourage it are used. All of the following can be used to support implementation and adoption of clinical guidelines:

- media marketing
- the use of opinion leaders and 'champions'
- endorsement by clinical groups
- practice visits from known experts
- education of patients
- educational materials
- seminars and conferences
- reminder systems built into a clinician's' daily work i.e. clinical decision support
- continuing quality assurance and data feedback
- local adaptation and incorporation
- local involvement in evaluation
- financial incentives

Step 5: Evaluation
Evaluating the effectiveness of the implementation strategies in changing physician behavior is a vital component of the process. Evaluation can also be used to assess how the strategies could be used in a hospital setting to implement change. The type of evaluation tool used will be dependent on the setting and type of implementation strategies used. Measuring compliance with guidelines and providing feedback to Physicians, can help improve adoption after implementation.

Organization characteristics such as size, complexity, availability of resources, culture, communication channels and decision making processes are significantly associated with guideline utilization and explain considerably more of the variance in research and guideline utilization than other factors. The environment surrounding the practice exerts a powerful influence on practitioners. This environment can encourage or discourage the process of knowledge transfer and

[27] Wise CG, Billi JE. A model for practice guideline adaptation and implementation: empowerment of the physician. Jt Comm J Qual Improv. 1995 Sep;21(9):465–76.
[28] Oxman AD, Thomson MA, Davis DA, Haynes RB. No magic bullets: a systematic review of 102 trials of interventions to improve professional practice. CMAJ. 1995 Nov 15;153(10):1423–31.

use. Structural, social and patient related factors are also important determinants of a successful implementation. Structural factors include such characteristics as decision-making structure, workload, and available resources. Social factors include such variables as the politics and personalities involved and the culture and belief systems in place. Patient related factors include patient willingness or ability to comply with evidence-based recommendations. Although some studies have found combined strategies to be more effective than single ones, the evidence in inconclusive. In addition, there seems to be no significant relationship between the number of components of multifaceted strategies and the effects measured.

Complexity is an important predictor of guideline implementation. Several systematic reviews have shown that when a guideline can be relatively easily understood and tried out, the chance is greater that the guideline will be used.[29]

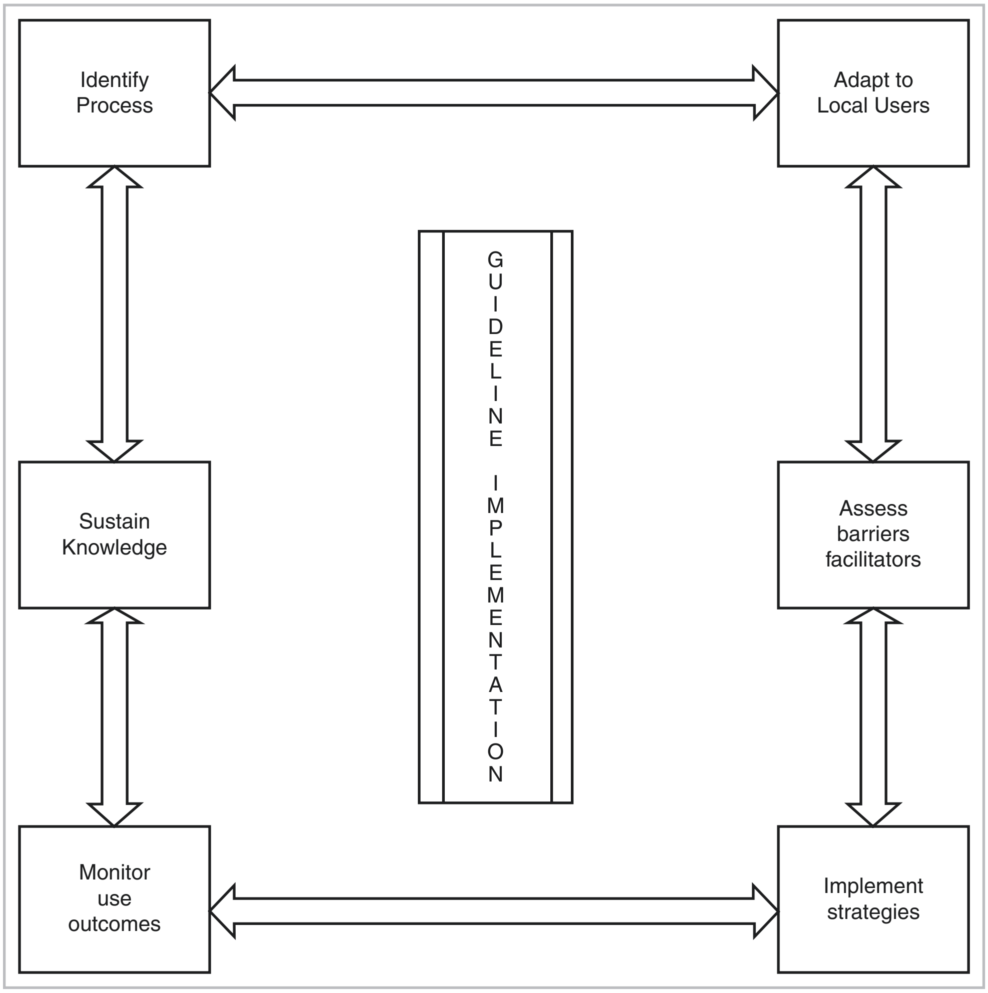
It is important therefore for guideline developers to take into account the complexity of the guidelines. Particularly for developers of multi-disciplinary guidelines directed at several target groups with varying educational levels and backgrounds (e.g. physicians, nurses, patients), it is a challenge to describe recommendations that are understandable and usable for all target groups.

Involving the professionals in the development phase enhances the chance of successful implementation. Groups that develop guidelines should be broadly composed and include all relevant health professionals. In addition, involvement of the target group may imply that the guideline is first being tested in practice before large-scale implementation takes place. Implementation may be hampered by the simple fact that professionals are often unaware that the guidelines exist, or are not familiar with their content. Clearly, it is not sufficient to merely disseminate a guideline. Targeted implementation interventions—in which professionals themselves are preferably directly and actively involved—should take place to create awareness. Characteristics of patients, too, appear to exert influence: for instance, comorbidity in a patient appears to reduce the chance that guidelines are followed. Professionals assume that guidelines are based on a general clinical picture and are insufficiently tailored to the often complex care needs of patients with comorbidity. In addition, environmental characteristics influence the implementation of guidelines. For example, support by peers in following the guidelines, and sufficient staff and clinical time appear to be important for guideline implementation.

A widely used Toolkit was conceptualized using the **Knowledge-to-Action Framework** as adapted for implementation of **Best Practice Guidelines**. The model depicts the following six essential components necessary for successful implementation.

1. **Identify** the problem: Identify, review, and select key knowledge. Identify gaps in knowledge and practice patterns using quality improvement tools and data from the institution where guidelines are to be used.
2. **Adapt** knowledge, tools, and resources to local context: Identify key stakeholders and their vested interests. Adapt the implementation process to the stakeholders identified and to the type of organization where they care for patients.
3. **Assess** barriers and facilitators to knowledge use: Identify barriers and develop strategies to overcome these. At the same time we often forget to identify and make maximal use of opinion leaders, early adopters, and other facilitators who can enhance adoption rates and implementation.
4. **Select**, tailor and implement interventions: Based on the above factors, lay out a detailed plan to support implementation and adoption. Project planning tools and dedicated project managers are valuable resources in this planning and execution.
5. **Monitor** knowledge use: Identify KPIs (Key Performance Indicators) early. Design methods to collect data on guideline use and patient outcomes. Positive early data can help win over late majority and Laggards in many cases.
6. **Sustain** knowledge use: Provide regular communication on the implementation and use of guidelines. As new information becomes available, update the guidelines and disseminate the changes. This ensures an iterative process that ensures continuous quality improvement.

[29] Factors influencing the implementation of clinical guidelines for health care professionals: A systematic meta-review. Anneke L Francke, Marieke C Smit, Anke JE de Veer, Patriek Mistiaen, *BMC Medical Informatics and Decision Making* 2008 8:38.



2.2.5 INFORMATION RETRIEVAL AND ANALYSIS

Information retrieval (IR) is the process of finding material of an unstructured nature that satisfies an information need from within large collections. In medicine, the US National Library of Medicine MEDLINE/PubMed is the most commonly used search engine.[30]

Suppose I want to find documents about pheochromocytoma that are stored in a larger collection of documents (called a **corpus**). If the number of available documents is small, a simple linear search can be used. A search program simply reads through each document in the corpus, and if it finds the search text, the document is a hit. If not, it continues to the next document.

As you can imagine, this solution does not scale very well. As the number of documents in the corpus becomes large, performing a linear search becomes more and more time consuming. The solution is to construct an **index** (sometimes called an **inverted index**) of all the words in the document. Since the number of words in the index is much smaller than the number of words in the corpus, search time is much less. Moreover, since the index is already sorted, it decreases search time logarithmically. To give a rough idea how much more efficient this is, consider that in 2015, MEDLINE contained approximately 14 million English language articles with abstracts.[31] Since most abstracts at that time were truncated at 250 words, a linear search would require searching approximately a billion words. Using an

[30] With approximately 40,000 searches per second, Google is the most famous IR company in the world.
[31] See https://www.nlm.nih.gov/bsd/medline_lang_distr.html

index is much more efficient. Out of those billion words, there are only about 128,000 unique words.[32] Searching a sorted index only requires search of $\log_2(128{,}000)$ words, or about 17.

When constructing an index, it is important to omit certain extremely common words which do not differentiate one document from another (such as "and", "the", "if", etc). Words that are deliberately excluded from an index are called **stopwords**. Similarly, some words are quite ambiguous on their own, but have specific meaning when paired with other words. For example, "colic" is relatively nonspecific, but "renal colic" or "biliary colic" refer to particular entities. Incorporating multi-word expressions into an index makes it much more powerful. The process of breaking up a document into individually searchable items is called **tokenization.** It is important to make sure that the tokenizer used to parse a query is the same as the one used to parse the documents.

Another benefit to constructing an index is the inclusion of word count and word location. For example, a document that contains the word pheochromocytoma many times is likely more relevant than a document that contains the term only once. Storing word location allows users to perform queries on words that are close to each other in the text. (e.g. searching for "gates" within six words of "Microsoft").

Searchers often employ **Boolean logic** to further expand or limit a query. By using conjunctions such as AND, OR and NOT, queries can be made more specific. For example, "renal disease AND NOT hypertension" or "vomiting OR diarrhea".

Optimizing boolean queries can be quite challenging. Consider the following index from a corpus of 150,000 documents:

TERM	NUMBER OF DOCUMENTS
Renal disease	2343
Hypertension	55,232
Vomiting	17,322
Diarrhea	7442
Asthma	15,229

If someone searches for "renal disease", the search is straightforward. The database simply lists all 2343 documents which contain renal disease. This is called an index-only search because no further calculations are required.

However, if someone searches for "diarrhea OR vomiting", the database must compare the 17,322 documents with vomiting and the 7442 documents with diarrhea and remove the duplicates, which adds significant complexity to the search. Even worse, if the search is for "renal disease AND NOT hypertension", the database has to find all 94,768 documents that are not in the hypertension index and only then merge that set with the results of the renal disease portion of the query.[33]

Until this point, we have described **Boolean searches**, which means that either a document meets criteria or it does not. In practice, however, we usually get a list of results which are **ranked** from most likely relevant to least likely. There are several methods of doing this. In the case of a complex query (e.g. "asthma AND renal AND failure AND creatinine"), there may be very few documents which satisfy all the Boolean requirements. Instead, documents may be ranked by the number of matches they have with the query. Even when all the terms match, the optimizer may assign higher rank to documents which include the word multiple times, or include it in the first few words of the document. Words that are relatively rare may be given more weight in ranking because they are more specific.

[32] Suomela BP, Andrade MA. Ranking the whole MEDLINE database according to a large training set using text indexing. BMC Bioinformatics. 2005; 6: 75.

[33] It is also possible to start with the 2343 documents which contain renal disease and then weed out those that do not carry the hypertension tag, however, this can also create additional complexity which is beyond the scope of this book. For much more information, see https://docs.oracle.com/database/121/TGSQL/tgsql_optcncpt.htm

In summary, users start with **information needs**, from which they compose **queries**. They search a body of knowledge known as a **corpus** for individual **documents**. These documents are **tokenized** and stored into an **index** for rapid **retrieval**.

The goal of an IR system is to find the right documents without additional clutter. There are several measures of IR success: **Precision** (positive predictive value) measures the percentage of returned documents that are relevant to the query; **recall** (sensitivity) measures the percentage of all relevant documents in the corpus that were found. **Fall-out** (false positive rate) is the proportion of irrelevant documents that are retrieved. The F_1 score (also called the F-measure or F-score) is the harmonic mean of precision and recall. An ideal F_1 score is 1.0.

For example, suppose you have a corpus of 100 documents, and 24 of them are relevant to your search. Now, suppose you run a query that returns 20 of the relevant documents as well as 6 additional (irrelevant) documents. Calculate the precision, recall, fallout and F-score.

First we create a 2 × 2 matrix

	RELEVANT	NOT RELEVANT
Retrieved	20	6
Not retrieved	4	70

Precision = 20/(20 + 6) = 77%
Recall = 20/(20 + 4) = 83%
Fall out = 6/(6 + 70) = 8%

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = 2 \cdot \frac{0.64}{1.6} = 0.8$$

2.2.5.1 Search Skills

Although IR generally refers to the search of unstructured data, most searches rely on **metadata** (i.e. data about the data) which can be used to further improve search results. Often, this will include file creation time or author or ownership. In the case of MEDLINE, metadata includes author, journal, publication date, article type and many other data points which can be individually searched in order to get more specific results. In addition, authors supply a list of keywords to assist other researchers in finding their articles. At the same time, independent indexers from the National Library of Medicine will assign headings from the **Medical Subject Heading (MeSH)** dictionary to each article. MeSH headings are hierarchical, so that related entities can be found by traversing the tree. A thesaurus of synonyms for MeSH headings is continually updated, so that a search for “asthmatic bronchitis” or “asthma” or “reactive airways” will all map to the same MeSH heading.

Assigning standardized MeSH headings also obviates the need to re-index articles when new terminology becomes available. For example, suppose a researcher describes a case of Reiter’s Syndrome. Years later, the term becomes deprecated in favor of Reactive Arthritis. As long as the MeSH mappings are kept up to date, a modern researcher will still be able to find the older article.

2.2.5.2 Critical Analysis of Biomedical Literature

The skills required to critically analyze biomedical literature are the same as those for informatics literature. See Sect. 1.1.3.2 for more information.