The staggering number of research journals and papers published every year within the specialty of gastroenterology, let alone within the whole spectrum of clinical and academic medicine, renders keeping up-to-date with medical literature a daunting task. GASTROENTEROLOGY itself publishes, on average, greater than 20 original research articles each month; consequently, it is hard to see how any individual could possibly read and synthesize all of the new information available to them. Throughout the course of one’s career, it is important to be able to identify important developments in one’s chosen specialty. This requires a critical appraisal of the relevant literature; in this article we aim to help you to do this in the most time-efficient manner.

Hierarchy of Studies

The appraisal of the totality of a literature is a subject which is addressed frequently via the discussion of evidence-based medicine. One of the tools from evidence-based medicine which can aid us in assessing research is a hierarchy of studies. Various versions of this have been proposed, but a fairly expansive version is suggested by the Oxford Centre for Evidence-Based Medicine,1 (Figure 1) where level 1 is considered the highest form of evidence and level 5 the lowest worthy of consideration. This list provides some help in recognizing which study types will provide the most reliable results, assuming they have been conducted well. The best evidence is provided by a well-conducted systematic review of good-quality randomized controlled trials (RCTs). A systematic review is likely to be superior to an individual RCT not primarily because it provides increased power but because one would not expect the same biases to operate in multiple trials and so the combination of studies is less likely to be biased in one direction than is any individual study. The great benefit of the RCT, in itself, is not that it is not biased but rather that it has the potential to diminish the effects of both known and unknown confounders. In this it is clearly superior to any of the observational methodologies, which at best allow for correction of suspected or recognized measurable confounders. When descending the hierarchy to the observational studies the same arguments regarding systematic reviews apply, hence systematic reviews and meta-analyses of observational studies are similarly above the individual studies from which they are constituted.

Simple Steps for Critical Appraisal of the Literature

Moving to consideration of individual studies, ideally critical appraisal of scientific literature requires knowledge of the methods, reporting, and interpretation of the techniques employed; in clinical epidemiology we consider mainly cohort studies, case control studies, RCTs, and systematic reviews. In this article we provide a brief overview of our approach to critical appraisal; more detailed discussions are available in the referenced publications2–8 and in the Journal of the American Medical Association’s “Users’ Guides to the Medical Literature” (http://jamaevidence.com/resource/520). Here we give a simple, overall approach to reading and critically appraising a research paper by suggesting you ask 5 questions of the study at hand.

(1) What is the research question? All epidemiological studies require a clear and precise question that should inform the reader of what was done and with whom.3 In epidemiology the purpose of a study can be categorized reasonably easily into those examining occurrence, diagnosis, prognosis, and treatments (Figure 1). If it is not clear from reading the introduction to the paper what the study is trying to achieve, this is often a good indication of the lack of quality of the work described thereafter.

(2) What did they do? Most epidemiological studies can be categorized broadly as a simple case series, an ecological study, case control study, cohort study, RCT, or a systematic review.3 Each of these study designs, in its own right, has a series of important methodological “building blocks” that should be reported adequately.
in the paper. The reader should attempt to understand the study design from the methods section, as the design suggested by the authors is not always accurate. The STROBE statement for the analytical observational designs (case control and cohort studies) (http://www.strobe-statement.org/), the CONSORT statement (http://www.consort-statement.org/) for randomized controlled trials and PRISMA statement (http://www.prisma-statement.org/) for systematic reviews provide excellent overviews of the component parts that should be reported. Careful consideration of the methods used in the study (with reference to these guidelines) will allow the reader the opportunity to assess the extent to which the key issues of chance, bias, confounding and reverse causality have been either avoided or dealt with in the design and analysis of the study. Having fully understood the way in which the authors carried out their study the reader can then consider our third question.

**What did they find?** Most epidemiological studies report the association between an exposure and an outcome, for example the association between celiac disease and the risk of developing non-Hodgkin’s lymphoma. Often the association is described in terms of a "relative risk", an oft misused term which may encompass many ratio measures of effect, the most common of which are rate ratios (used in cohort studies and RCTs) and odds ratios (used in case-control and cross-sectional studies). Of particular importance in the calculation of any measure of effect is the choice of reference group or baseline used for comparison. An overview of these and other epidemiological terms and their definitions can be found in the "Dictionary of Epidemiology" published by Oxford University Press. Alternatively the absolute risk of a disease, (ie, the probability of developing a disease over a specified time period) is reported for 2 groups allowing a risk difference to be calculated. The reporting of these measures has important implications for how the study will be interpreted. For example, consider the paper by Chan et al, which examined the risk of colorectal carcinoma among women with respect to their

<table>
<thead>
<tr>
<th>Question</th>
<th>Step 1 (Level 1*)</th>
<th>Step 2 (Level 2*)</th>
<th>Step 3 (Level 3*)</th>
<th>Step 4 (Level 4*)</th>
<th>Step 5 (Level 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>How common is the problem?</td>
<td>Local and current random sample surveys (or censuses)</td>
<td>Systematic review of surveys that allow matching to local circumstances**</td>
<td>Local non-random sample**</td>
<td>Case-series**</td>
<td>n/a</td>
</tr>
<tr>
<td>Is this diagnostic or monitoring test accurate? (Diagnosis)</td>
<td>Systematic review of cross sectional studies with consistently applied reference standard and blinding</td>
<td>Individual cross sectional studies with consistently applied reference standard and blinding</td>
<td>Non-consecutive studies, or studies without consistently applied reference standards**</td>
<td>Case-control studies, or &quot;poor or non-independent reference standard&quot;**</td>
<td>Mechanism-based reasoning</td>
</tr>
<tr>
<td>What will happen if we do not add a therapy? (Prognosis)</td>
<td>Systematic review of inception cohort studies</td>
<td>Inception cohort studies</td>
<td>Cohort study or control arm of randomized trial*</td>
<td>Case-series or case-control studies, or poor quality prognostic cohort study</td>
<td>n/a</td>
</tr>
<tr>
<td>Does this intervention help? (Treatment benefits)</td>
<td>Systematic review of randomized trials or n-of-1 trials</td>
<td>Randomized trial or observational study with dramatic effect</td>
<td>Non-randomized controlled cohort/follow-up study**</td>
<td>Case-series, case-control studies, or historically controlled studies**</td>
<td>Mechanism-based reasoning</td>
</tr>
<tr>
<td>What are the COMMON harms? (Treatment harms)</td>
<td>Systematic review of randomized trials, systematic review of nested case-control studies, n-of-1 trial with the patient you are raising the question about, or observational study with dramatic effect</td>
<td>Individual randomized trial or (exceptionally) observational study with dramatic effect</td>
<td>Non-randomized controlled cohort/follow-up study (post-marketing surveillance) provided there are sufficient numbers to rule out a common harm. (For long-term harms the duration of follow-up must be sufficient.)**</td>
<td>Case-series, case-control, or historically controlled studies**</td>
<td>Mechanism-based reasoning</td>
</tr>
<tr>
<td>What are the RARE harms? (Treatment harms)</td>
<td>Systematic review of randomized trials or n-of-1 trial</td>
<td>Randomized trial or (exceptionally) observational study with dramatic effect</td>
<td>Non-randomized controlled cohort/follow-up study**</td>
<td>Case-series, case-control, or historically controlled studies**</td>
<td>Mechanism-based reasoning</td>
</tr>
<tr>
<td>Is this (early detection) test worthwhile? (Screening)</td>
<td>Systematic review of randomized trials</td>
<td>Randomized trial</td>
<td>Non-randomized controlled cohort/follow-up study**</td>
<td>Case-series, case-control, or historically controlled studies**</td>
<td>Mechanism-based reasoning</td>
</tr>
</tbody>
</table>

* Level may be graded down on the basis of study quality, imprecision, indirectness (study PICO does not match questions PICO), because of inconsistency between studies, or because the absolute effect size is very small; level may be graded up if there is a large or very large effect size.

** As always, a systematic review is generally better than an individual study.

Figure 1. Oxford Centre for Evidence-Based Medicine Levels of Evidence Table.
use of aspirin and NSAIDs using the Nurses' Health Study which is a cohort study established in 1976 in the US of about 80,000 women. The abstract reports “Among women who regularly used aspirin the multivariate relative risk (RR) for colorectal cancer was 0.77 (95% CI, 0.67–0.88)”. This represents a 23% reduction in risk for those taking regular aspirin. However, until we read the full results section we were not aware that this 23% decrease in risk could otherwise be presented as an absolute risk of colorectal cancer of 5.56 cases per 10,000 women regularly using aspirin compared with an absolute risk of colorectal cancer of 6.36 cases per 10,000 who did not regularly use aspirin (ie, an absolute risk reduction of only about 1 case per 10,000 women). Clearly the perceived implications of these findings are highly dependent upon whether the absolute risk differences or relative risks are used. Having understood what the authors’ findings were the reader must then consider our fourth question.

(4) Is the relationship described in the paper causal? In epidemiology one of the ways in which we approach the decision about whether a relationship is causal is by considering the other options. A thorough discussion of all aspects of this approach is provided by Kenneth Rothman and Sander Greenland in their article in the American Journal of Public Health. These aspects can in general be simplified to the possibilities that the relationship is due to chance, confounding, bias, or reverse causality. In the reading of epidemiological studies we can consider the potential role of these if we know what they are, how they can occur and how they can be detected. Chance (the potential for a relationship to have occurred due to random variation) is assessed in the statistical analyses within a paper, usually using confidence intervals. A 95% confidence interval tells us that if the variation between the observed measure of effect and the true value is purely due to chance, then there is a 95% chance that the true value lies within the limits that they describe. A P value of .05 states that when considering whether a difference between observed samples is due to a true difference between populations, the observed, or a more extreme result will occur by chance on 5% of occasions if the populations do not truly differ.

Confounding occurs when an apparent association between 2 factors is influenced by their shared association with a third. A classic example is that the drinking of alcohol may be associated with lung cancer not because alcohol causes lung cancer but rather because alcohol drinkers smoke more than the average individual and smoking cigarettes causes lung cancer. To assess the possibility that confounding is of importance, one must first think of what additional factors may be present and then examine to what extent the authors have corrected for any effect in their analysis. In brief, the commonly used techniques to try to correct for confounding are to conduct stratified analyses (where all subjects in a stratum who are analyzed together are equally exposed to the potential confounder) and/or to conduct multivariate analyses.

Bias is a systematic deviation from the truth in either measurement, selection of supposedly representative samples or other factors influencing the results or inferences from a study. Bias can never be corrected for in analyses and is minimized by optimal methodology. Broadly speaking there are 2 main types of bias — selection bias, where participants within a study are systematically different from the population that they purport to represent; and information bias, where there is a systematic difference in the way in which data are recorded, dependent on participant’s exposure and/or outcome status. Particular types of bias only occur in some types of study (eg, publication bias in systematic review) and so, since bias can occur in many ways, one must always be alive to this possibility.

Reverse causality occurs when what is being regarded as an effect is in fact causing what is being regarded as an exposure. An obvious example of this would be to study the effect of wearing a Rolex watch on affluence. Wearing a Rolex may be associated with affluence, but that does not mean that by buying a Rolex one can make oneself rich. As with bias the best the reader can do is to seriously consider this possibility with regard to the relationship studied. However, once you have convinced yourself of the veracity of the report’s findings, the next and final question should be considered.

(5) Does it matter? Assuming that we believe the paper we are reading describes an association that is causal in a population, we then have to question if there is anything we can do to either enhance a beneficial association or reduce a harmful association. Essentially the whole of medical practice is based on this fundamental principle even when the evidence to support or refute an association is not as robust as perhaps we may wish it to be. Ultimately, what constitutes an association “that matters” is of course entirely at the discretion of an individual. We will all have different ideas about the relative values of life and disability, for example. One person may be willing to accept, following some intervention, a 10% increased risk of stroke for a 10% chance of increasing their life expectancy for a year, whereas another may not. Doctors’ agencies, licensing authorities, and so forth may additionally have their own guidelines as to the acceptability of risks often invoked when considering the likely benefit of introducing a specific treatment or therapeutic intervention, but decision-making at the level of the physician-patient dyad should allow for dialog and clear interpretation of all available information. When it comes to determining one’s practice the issues become even more complicated as there will be many sources of information to guide decisions and not just one paper. Here we
would refer you back to our earlier discussion of the hierarchy of evidence.

One way of assessing the value of altering a reported association is to calculate the numbers needed to treat and to harm. These numbers give simple, readily accessible, and understandable values to the potential benefit and harm of an intervention. The number needed to treat (NNT) is defined as “The number of persons needed to be treated, on average, to prevent one more event.” The number needed to harm (NNH) is defined as “The number of persons needed to be treated, on average, to produce one more adverse event.” Intuitively, low values of NNT and high values of NNH are considered better. Balancing the NNT and NNH will, to a large extent, be mediated through consideration of the relative importance of the diseases prevented or produced. Normally, the NNT and NNH are calculated from data reported from randomized clinical trials or meta-analyses of clinical trials; however, they can, with caution, be applied to the results of cohort studies, too.

Ultimately when you are reading any study you should attempt to come to your own conclusion as to whether the reported findings matter or not in the context of your own reason for reading the paper.

**Conclusion**

We hope that we will have convinced readers that attention to a few simple questions can aid in determining whether the research they read is valid and/or important. For those who wish to read more we would recommend, in addition to the resources already mentioned, the second edition of the book “GI Epidemiology” (Talley NJ, et al [Eds]. GI Epidemiology, 2nd edition. Oxford, UK: John Wiley & Sons, 2014), which contains a more extensive version of this article within a series of chapters on how to read the GI epidemiology literature.

**References**


**Reprint requests**

Address requests for reprints to: Dr. Joe West, Clinical Associate Professor and Reader in Epidemiology, Honorary Consultant Gastroenterologist, Division of Epidemiology and Public Health, Room B113 Clinical Sciences Building 2, University of Nottingham, City Hospital Campus, Hucknall Road, Nottingham, United Kingdom NG5 1PB. e-mail: joe.west@nottingham.ac.uk.

**Conflicts of interest**

The author discloses no conflicts.

**Funding**

J.W. is funded by a University of Nottingham/Nottingham University Hospitals NHS Trust Senior Clinical Research Fellowship.