ELSEVIER

Journal of Hepatology 50 (2009) 817–826

Journal of Hepatology

www.elsevier.com/locate/jhep

Review

Clinical trial design and evidence-based outcomes in the study of liver diseases☆

Jennifer M. Croswell*, Barnett S. Kramer

National Institutes of Health, Office of Disease Prevention, 6100 Executive Blvd, Suite 2B-03 Bethesda, MD 20892, USA

Current medical training often does not include the formal study of trial design, forcing clinicians to acquire this knowledge independently. This article reviews the foundational elements of clinical trial design. An overarching hierarchy of clinical evidence is introduced, and the relative strengths and limitations of the major types of study designs are discussed. A corollary to the hierarchy of evidence in trial designs is proposed for trial outcomes: the "pyramid of endpoints." This pyramid represents a spectrum of outcomes from tangible health events to intermediate markers with no direct physical impact on an individual. The potential advantages and difficulties inherent in the use of surrogate endpoints for final health outcomes are explored. Randomized controlled trials utilizing "hard" clinical endpoints are advocated as the most efficient and reliable way to directly assess the benefits and harms of a therapy; however, using a case study of treatments for hepatocellular carcinoma, we highlight the challenges that can complicate even the highest levels of evidence. All trials have a "signal-to-noise" ratio – this review emphasizes the need for careful and deliberate consideration of the potential limitations of every study, and provides basic tools to assist the practitioner in identifying common pitfalls of clinical trials.

***Keywords*: Clinical design; Evidenced-based outcomes; Surrogate endpoints**

1. Introduction

The Scottish philosopher David Hume once wrote, "A wise man proportions his belief to the evidence" [1]. Unfortunately, clinical training programs are often deficient in formal education on the fundamental principles of epidemiology and trial design [2]. Practitioners are often left to independently develop the critical skill set of analyzing and interpreting the quality of scientific evidence. In an era of increasingly demanding clinical workloads, coupled with an overwhelming amount of medical literature of heterogeneous quality, it can be challenging to sort out exactly what evidence is truly worthy of one's attention. This article will provide a framework for understanding the basic elements of clinical trial design and analysis, and will also explore specific issues related to hepatocellular carcinoma that demonstrate some of the more complex, subtle considerations accompanying clinical research designs.

2. Questions to ask about medical research: passing the "clarity test

When interpreting the results of a clinical trial, there are several basic questions to consider that have overarching implications for judging the quality and applicability of that trial. Surprisingly, published medical research articles frequently fail one or more of the criteria that make up this "clarity test" (see Table 1).

2.1. What is the exposure and what is the outcome?

It may seem obvious that one needs to have a clear definition of the intervention under study and the

Associate Editor: M. Colombo

☆ NIH study. The authors who have taken part in this study declared that they do not have anything to disclose regarding funding or conflict of interest with respect to this manuscript.

* Corresponding author. Tel.: +1 301 496 6615; fax: +1 301 480 7660.
E-mail address: croswellj@od.nih.gov (J.M. Croswell).

0168-8278 Published by Elsevier B V on behalf of the European Association for the Study of the Liver Open access under CC BY-NC-ND license.
doi:10.1016/j.jhep.2009.01.005

Table 1
The "clarity test" for medical research.

1.	What is the exposure and what is the outcome?
2.	How certain is it that the exposure actually causes the outcome?
3.	How big is the observed effect?
4.	How important is the outcome?
5.	To whom does the study apply?

outcome of interest in order to be able to properly interpret a trial. However, there are subtleties in the choice of both of these elements that can have a large impact on the ultimate utility of a study, but can be overlooked when reporting a trial. These factors will be discussed in detail throughout this article.

2.2. How certain is it that the exposure actually causes the outcome?

This question gets to the heart of study design, by highlighting the distinction between an *association* between an intervention and an outcome versus a *causal relationship*. It also addresses what is known as the *internal validity* of a study. Only experimental (i.e., randomized controlled) studies can directly establish a causal relationship between an intervention and an outcome. Observational studies may provide good evidence of an association between an exposure and an outcome. But if practical (and ethical), the association should then be tested in a more formal experimental setting to confirm causality.

2.3. How big is the observed effect? and, how important is the outcome?

It is critical to understand the clinical value and limitations of the chosen outcome(s). For example, is the outcome of interest an intermediate or surrogate marker for a "hard" clinical endpoint? If so, has the surrogate been validated as a reliable substitute for the final endpoint? Is the outcome a composite measure of several endpoints of varying clinical significance? These choices can obfuscate the conclusions one can or should derive from a study. Additionally, the value one should ascribe to an observed effect size is directly related to the choice of endpoint: a large, seemingly impressive change in a laboratory value may not be as ultimately meaningful as a more modest effect on a more tangible endpoint such as death. It is also important to pay attention to the manner in which the effect size is expressed – that is, in relative versus absolute terms – because use of rel-

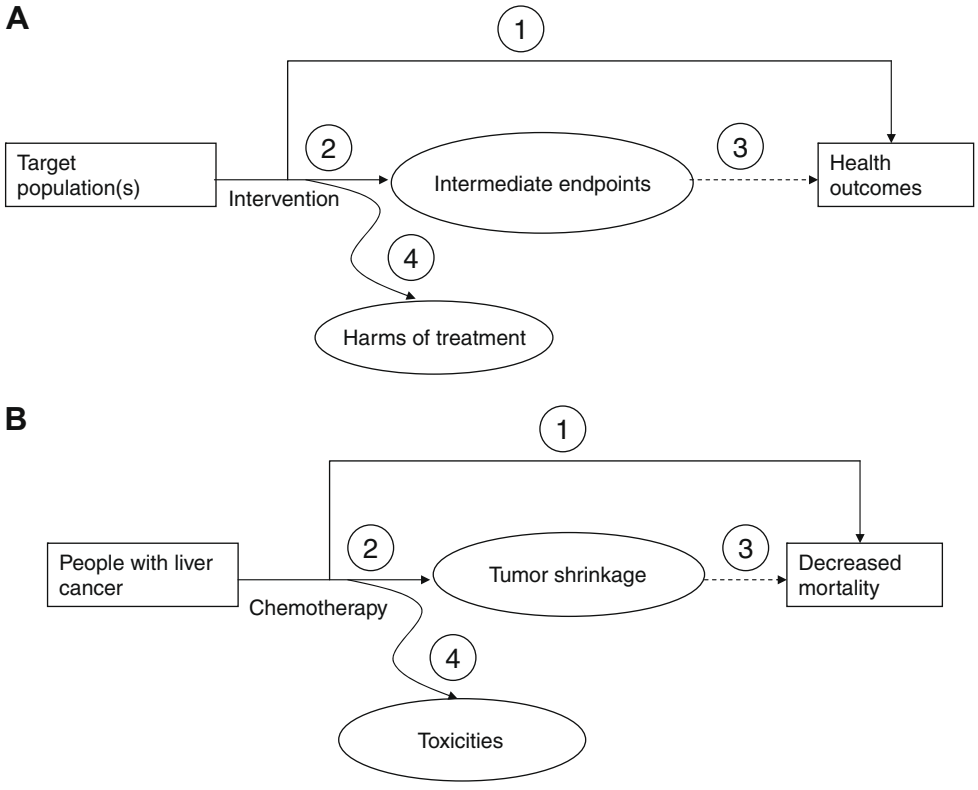

Fig. 1. Panel (A) demonstrates a generic analytic framework for medical interventions. An analytic framework makes explicit the population, therapy, and intermediate and final outcomes under investigation. It demonstrates the direct and indirect steps by which the overall balance of benefits and harms for a given treatment may be evaluated. Panel (B) provides an example of how an analytic framework might be utilized for a specific medical issue: in this case, the use of chemotherapy for people with hepatocellular carcinoma. Adapted from Ref. [3]. Used with permission.

ative rates may give a falsely inflated impression of the true impact of an intervention. Finally, statistical significance should not be confused with clinical importance. Especially in large trials, statistically significant differences may be clinically trivial.

2.4. To whom does it apply?

This question speaks to the generalizability of the findings, or the external validity of the study. As will be discussed later, this is particularly key in clinical studies of hepatocellular carcinoma, as patients with coexisting cirrhosis or liver failure (e.g., Childs-Pugh class B and C) can suffer complications from these underlying conditions that interfere with accurate evaluation of an intervention's true efficacy. The balance of benefits and harms for a given intervention may reverse when moving from a population of Childs-Pugh class A patients to B or C patients.

3. Avoiding mental shortcuts: the analytic framework

A more formalized method by which one can consider the salient elements of a clinical trial is termed an *analytic framework*, or *causal pathway*. An analytic framework is a diagram that presents an explicit representation of the chain of logic necessary to demonstrate the efficacy of an intervention on a given outcome [3]. It graphically lays out each step used to evaluate the overall balance of risks and benefits associated with a given intervention. As can be seen in Fig. 1, the analytic framework forces one to specify the study's target population, key intervention, intermediate and final health outcomes of interest, and any potential harms (e.g., toxicities, adverse events) associated with the use of the intervention. Most importantly, the framework highlights the distinction between direct and indirect proof of efficacy.

If a high-quality randomized, controlled study directly demonstrates the impact of the intervention on changes in final health outcomes (represented by path 1), this is sufficient evidence to establish a causal connection. However, frequently, this information is lacking. In this case, the analytic framework provides a visual representation of the indirect linkages between intervention and desired health outcome (paths 2 and 3). We can determine where a given study fits along the indirect causal pathway, and, thus, have a clear depiction of which critical step(s) in the pathway have not yet been addressed. Finally, path 4 represents the harms of the intervention that need to be weighed against any benefits it has been shown to provide. Thus, the framework draws attention to the areas where mental shortcuts would be apt to occur; it illuminates any remaining gaps in the evidence chain that would preclude drawing a reliable conclusion about the intervention.

4. The pyramid of evidence

There is a general hierarchy of study designs in medical literature that can be represented as a pyramid (see Fig. 2). At the base of the pyramid are the lowest forms of evidence for proving the efficacy of an intervention – expert opinion and uncontrolled case reports – and at the top are the highest – double-blind, randomized, controlled studies. The conceit of a pyramid is particularly helpful because it demonstrates that although the preponderance of clinical evidence is comprised of non-experimental study designs, the results from one well-done randomized controlled trial may overturn an entire body of observational findings. A case in point was the Women's Health Initiative, a randomized trial that overturned assumptions that routine hormone treatment of postmenopausal women would protect them against heart disease – an assumption based on a large number of observational studies [4]. This is because the randomized controlled trial is the best tool available to eliminate bias; as such, it is the most sure we can ever be that any effect we observe from an intervention is real. It is important to stress that not all randomized controlled trials are performed well. The internal validity may be poor. Thus, careful review of the trial's methods, as well as application of the "clarity test," are of course required before acceptance of any study's conclusions.

5. Major types of clinical studies

At the base of the clinical trial hierarchy (following expert opinion, which is a state of mind rather than a study design) is the *case report*. A case report is an observational study that describes findings from a single

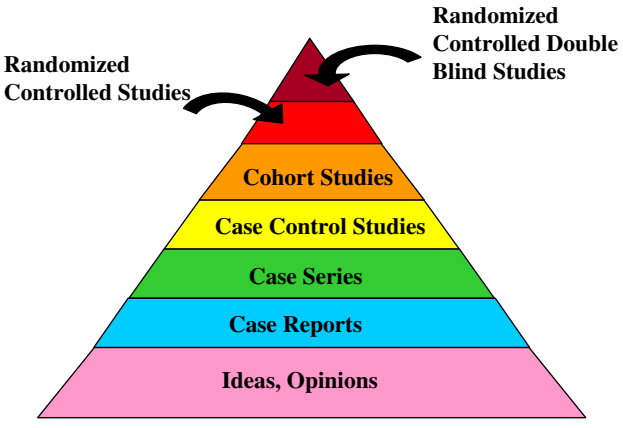

Fig. 2. The pyramid of evidence represents a general hierarchy of preferred clinical study designs. As one moves up the pyramid, the potential for bias is reduced. The pyramid graphically represents that non-experimental research designs outnumber randomized controlled trials. Adapted from SUNY Downstate's evidence-based medicine course. Used with permission.

patient with a disease or condition. There is no comparison group. Case reports often portray a "first-of-its-kind" intervention with optimistic results. They may offer important clinical insights, but they can also be misleading. A 2006 letter to the editor in *Digestive Diseases and Sciences* by Harmanci et al. provides a good example of a case report. This correspondence described the case of a 17-year-old girl with Wilson's disease taking zinc and D-penicillamine who presented with an acute hemolytic crisis and impending liver failure. Within 24 h of admission, the patient's status declined to the point where she was evaluated for emergent liver transplantation. Concurrently, she received 7 days of plasma exchange treatment, during which her hemoglobin, bilirubin, and INR normalized, her mental status stabilized, and she was discharged home. The authors concluded, "For our patient the treatment of plasma exchange has prevented liver transplantation.... Our report suggests that...early initiation of plasma exchange can be a very effective modality in reducing mortality and even preventing orthotopic liver transplantation" [5].

The next level on the pyramid of evidence is the *case series*. Uncontrolled case series are basically a group of case reports, in which the patients all receive the same intervention; as with a case report, there is no control group. The study population is often ill-defined. Most case series are not consecutive (inviting strong selection bias) or population-based (impeding generalizability). Here is an example of a case series: investigators at Penn State University reported on 5 patients with chronic hepatitis C unresponsive to other therapies treated with amatadine and ribavirin for an average of 44 months. These patients were chosen from a pool of 60 participants of a shorter-term trial of the same therapy; they were given the opportunity to continue the treatment because they had ALT normalization, the ability to tolerate the treatment without adverse reactions, and a high level of compliance. At the end of the study, all 5 had improvements in ALT, hepatitis C virus RNA, and liver histology. One patient had a flare of hepatitis after treatment discontinuation that resolved with re-initiation of the therapy. The authors felt this was "provocative data on the long-term use of ribavirin and amatadine in the HCV non-responder. The dramatic response to retreatment of our main case cannot be overlooked" [6].

How sure are we that these conclusions are valid? That is, what are the major limitations of these study designs? First, in both cases, there is no control group. We cannot compare the outcome of the girl with Wilson's disease to that of a similar patient who did not receive plasma exchange therapy, nor can we evaluate how patients with refractory chronic hepatitis C that were not treated with amatadine and ribavirin faired. The importance of such a comparison can be summarized by the following anecdote: during world war II, in the aftermath of the London blitz, rescue workers were digging in the rubble of an apartment house blown up in the air raid. They came across a dazed but unharmed old man lying stark naked in a bathtub. He was surprised but glad to see help had arrived, and exclaimed, "The *most* extraordinary thing just happened. I had finished taking my bath. But when I pulled the drain plug, the entire house blew up!" [7]. Clearly, the man could not have known what would have happened had he not pulled the plug.

It can be easy (and tempting!) to confuse a *temporal* relationship with a *causal* one. Thus, the difficulty with inferring causality from a sequence of events lacking a control group is that there is no way to know what would have happened in the absence of the intervention, and thus, no way to be certain that the intervention was actually the factor responsible for the observed outcome. The second limitation of these study designs is that they are so prone to powerful biases – a point we shall return to shortly.

One study design that attempts to redress the lack of a comparator group in observational studies is the *case series with historical controls*. This is a study in which past experience is used as a comparison group for the cases. In this situation, both the cases and the historical controls have the same disease, but the primary intervention (and often many other things) has changed over time. For example, an early investigation into the utility of hemodialysis for encephalopathy caused by fulminant hepatic failure compared 39 patients in liver failure treated by dialysis between 1974 and 1977 with a group of 117 similar patients treated at the same institution five years earlier. Dialysis was not available at the hospital during this previous time period, so the historical controls received conservative intensive care without extracorporeal assistance. The authors found that 17 of 39 patients (43.6%) fully regained consciousness after hemodialysis versus 26 of 117 (22.2%) persons receiving previous conservative intensive care. The authors cautiously concluded that these were "hopeful" results [8].

The authors were restrained in their enthusiasm regarding the results. This is because while there was a group that did not receive the intervention to compare results to, the primary difficulty with employing a control population from the past is that it is impossible to be certain that the group is sufficiently similar to the current intervention population to make a valid comparison. Many variables besides the discovery of the intervention might also have changed over time in the population: for example, other ancillary therapies in use to treat the condition, the underlying health status of the patients, and so on. Therefore, even though there was an observed difference between the cases and the historical controls, it remains unclear whether that change was truly a result of the intervention, or actually

the result of other changes that had taken place in the intervening years.

One can at least partially avoid this limitation by utilizing a control population from the same time period and same population: this is the *case-control* study. For example, in 1981, MacMahon et al. performed a case-control study investigating potential exposures leading to pancreatic cancer [9]. The cases were 369 patients with histologically-confirmed pancreatic cancer; the controls were 644 concurrently hospitalized patients being seen by the same attending physicians caring for the cases. A startling association between coffee consumption and incidence of pancreatic cancer was found, with the relative risk 2.7 times that of controls for persons that drank 3 or more cups per day (95% CI, 1.6–4.7). Given the prevalence of coffee consumption in the general population, and the high mortality rates associated with pancreatic cancer, these findings generated intense medical and media interest at that time – as well as beverage-switching behavior. However, twenty-five years later, there are no warning labels on cans of Folgers identifying coffee as a carcinogenic substance. Furthermore, when the authors attempted to replicate the findings of this study, they failed to observe any association between coffee consumption and pancreatic cancer [10]. What could explain this?

To their credit, MacMahon et al attempted to answer this question.[11] They carefully scrutinized the characteristics of both the case and the control populations, and found the answer. In a case-control study, as in all controlled studies, one attempts to compare two populations that are as similar as possible on all known variables except the one of interest: whether they received the exposure. In this case, the investigators attempted to match the two study populations by utilizing controls that were also sick enough to be in the hospital (but not because of pancreatic cancer) and were roughly identical in age, sex distribution, smoking habits, and so on. However, the controls had been seen by the same attending physicians as the cases with pancreatic cancer – that is, gastroenterologists. It became apparent that a number of the controls had mainly been hospitalized for conditions such as gastric ulcer and esophagitis. One of the first methods of symptom management these individuals had been taught by the physicians was to avoid beverages containing caffeine – especially coffee.

Thus, the real explanation for the observed association between pancreatic cancer and coffee consumption was not a causal relationship between the exposure and the outcome, but a third, unrecognized factor that travelled with the populations under investigation. This created a spurious relationship between the exposure and disease. This is the definition of a *confounding variable*. Confounding is a concern in *any* observational study, because it is most likely to occur when someone's choice – be it a participant or an investigator – determines who is exposed versus unexposed, rather than chance. In this example, "hard-wired" bias was introduced by the use of a control population that avoided the exposure of interest. There are often fundamental differences between populations that are chosen/choose to receive a medical intervention versus those that do not, and these innate differences can silently drive the results of a study and fool the investigator into believing a casual relationship exists where it does not.

As an additional example, let us return to the case series of ribavirin and amatadine therapy for refractory chronic hepatitis C and reevaluate the criteria by which the study population was chosen. This trial was a continuation of a shorter study of the same intervention; the long-term participants were a small subset of the original trial population, chosen on the basis on their favorable response to the treatment regimen, their ability to tolerate the drugs without toxic effects, and their high compliance with their clinical care regimens [6]. It is therefore not altogether surprising that this highly restricted group continued to do better than expected of the general chronic hepatitis C population, many of whom may have been too ill to adhere to advised treatments or may not have responded to the therapy. These patients may also have not started out with normal ALT levels and might not have been as tolerant of the drugs' associated toxicities; they may have been generally less healthy overall. While the results might indeed appear "provocative" and "dramatic," it is impossible to be certain if they are, in fact, a product of the therapy alone, or, perhaps more likely, a product of the favorable clinical profile of the participants. The situation also invites selective reporting and publication bias. Had the five patients not done so well, it is unlikely the report would have seen the light of day.

This issue of confounding explains why even the *cohort study*, although considered the highest level of observational evidence, still falls short of the "gold standard" of the randomized controlled trial. A high-quality cohort study is typically a large, population-based, prospective investigation that compares a group with a specific exposure to one not exposed. It differs from the randomized controlled trial in that it is not a chance or random event whether a given participant receives the intervention, but the choice of either the investigator or the participant him/herself. Therefore, the potential for inequalities between the exposed and control populations has not been excluded, and as such, it cannot be fully established whether the observed relationship between the exposure and the outcome is a causal or spurious association. As a gastroenterologist once dryly observed, non-experimental studies are "considered guilty of bias and erroneous results until proven innocent [12];" when designing and interpreting studies, one should aim for the highest possible level of evidence upon which to base clinical decisions.

6. Examining outcomes: the challenge of surrogate endpoints

Just as a hierarchy exists for study designs, there is also one for trial outcomes. Fig. 3 depicts this "pyramid of endpoints," which extends the spectrum from the most tangible and clinically important outcome (all-cause mortality), to an outcome with the least direct physical impact upon an individual (a risk factor change). For example, let us examine the disease course for hepatitis B. A change in a risk factor might be an increase in the use of a needle exchange program in illicit drug users. Better test results could include a reduction in ALT or hepatitis B virus DNA levels. The next step up could be symptomatic chronic hepatitis B. Complications of disease would include the development of decompensated liver failure. Disease-specific mortality would be represented by death from hepatitis B-induced liver failure. Finally, there is death from any cause. It is clear that as we move up through the hierarchy, the outcomes take on increasingly tangible importance for the individual patient.

One might question why a distinction is made between disease-specific and overall mortality; furthermore, one might question why overall mortality is given ascendancy over disease-specific death, given that a therapy is usually intended to only prevent the morbidity or mortality associated with a single disease process. The answer is that overall mortality is the least subject to investigator bias. Attributing a cause of death is not an entirely objective process – knowledge of the patient's various health conditions and comorbidities can strongly influence the final decision. Dead versus alive, on the other hand, tends to be a more straightforward decision.

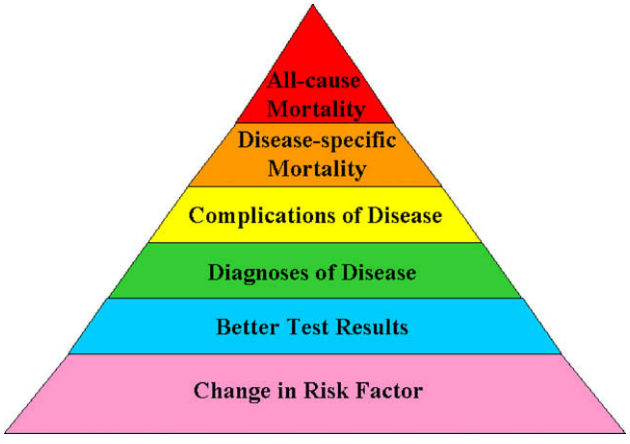

Fig. 3. The pyramid of endpoints represents a generally preferred hierarchy of study outcomes, going from those with the least clinical impact to the greatest import. It is often easier to obtain endpoints from lower down in the pyramid, as they occur with greater frequency in a population; however, they are less definitive in establishing the ultimate clinical utility of a given intervention.

Just as with the pyramid of evidence, it is also apparent that outcomes lower down in the hierarchy – those with less immediate physical import–occur with greater frequency. One of the logistical challenges of choosing "hard" clinical endpoints for trials is that they may require extensive time, money, and personnel commitments. Death may take many years to occur, and is a far less common event than changes in lab values; therefore, utilizing this endpoint can extend the length and expense of a trial, and can require a large study population to discern an effect. For this reason, *surrogate endpoints* have become increasingly popular choices in clinical trial design. A surrogate endpoint is a non-health outcome or biomarker that is used to draw a conclusion about the effect of an intervention on a health outcome observed later in time. It is used as a substitute for a clinically meaningful endpoint; i.e., one that directly reflects how a patient feels, functions or survives. The changes a therapy effects on a surrogate are presumed to fully predict the changes that therapy would have on the "hard" endpoint. Thus, although necessary, it is not sufficient for the surrogate to simply be correlated with the final outcome to be considered a valid and reliable predictor of that endpoint. A *validated* surrogate endpoint is one that yields the same inference about the effect of the treatment as the health outcome [13].

The most frequently cited potential advantage of surrogate endpoints in clinical trials has already been mentioned: they could permit shorter, more efficient trials. However, there are serious challenges accompanying their use that must be weighed and considered. As mentioned, surrogates occur with greater frequency than final health outcomes. While this is an advantage in terms of study power (as a difference in outcomes may be observed using a smaller sample size), it may also generate misleading findings – or "noise" – since there is not a 1:1 ratio between the appearance of the surrogate and progression to the actual health outcome. For example, in the case of colorectal cancer, it is accepted that some colonic adenomas are precursors to malignancy. However, the majority of adenomatous polyps will *not* progress to cancer [14]; since the overlap between the surrogate and the final outcome of interest is not perfectly aligned, interventions which are highly effective in preventing or treating all polyps could potentially give a distorted (over)estimate of the intervention's true efficacy on the "hard" health outcome. For example, an intervention could conceivably be effective only against polyps that are *not* destined to progress to cancer.

There is another pitfall to be cognizant of when evaluating the use of surrogate endpoints. Even validated surrogates only predict for a single health outcome, but many (if not most) interventions have more than one effect upon the human body. The art of medicine

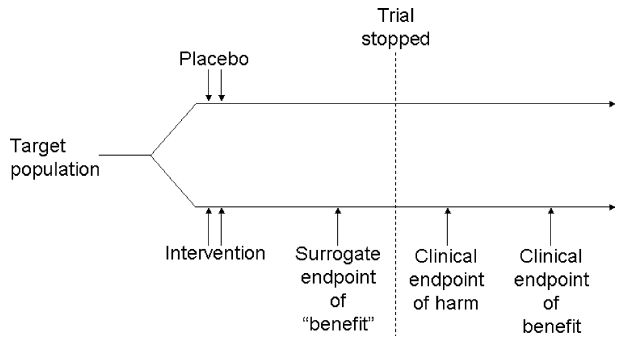

Fig. 4. Challenges to surrogate endpoint use: missed harms. This highlights one of the major limitations of the use of surrogate endpoints in clinical studies. Although the trial in this case achieved the surrogate endpoint of "benefit," and was therefore stopped, it provided no information about important harms that developed after the surrogate outcome had been obtained. Thus, use of the surrogate endpoint in this situation provided an overall incorrect picture of the relative balance of benefits and risks, as it gave no indication of the adverse events that accrued later in time. From Ref. [17]. Used with permission.

is, after all, grounded in the careful weighing of the relative risks and benefits of a therapy for an individual. This caveat is therefore most critical when considering the intervention's potential to cause adverse events. As demonstrated in Fig. 4, one of the difficulties in utilizing a surrogate endpoint is that it can only provide information up until the point that observation ceases (that is, at the point where surrogate benefit has been established and the trial is ended). However, the harms of an intervention may have a delayed onset of presentation, and if they manifest themselves after the point in time when surrogate effects have been reached, the negative effects of the treatment will not be apparent in the trial. Some adverse events are significant enough that the total risk-benefit profile of a given intervention may tip in the wrong direction; a short trial employing a surrogate endpoint may not detect this, and may therefore provide a qualitatively wrong answer about a therapy.

Although this discussion may appear to be merely an academic exercise or a theoretical postulation of harm, there are numerous examples in medical history that demonstrate the misleading potential of surrogate endpoints. A classic example is – once again – that of postmenopausal hormone therapy and heart disease. Multiple studies demonstrated that postmenopausal estrogen and progestin had favorable effects upon a woman's lipid profile: HDL levels rose, and LDL and total cholesterol levels consistently dropped. Previous experience with statin therapy had shown that modifying these lab values in a favorable direction resulted in a positive impact on cardiac event rates. As a result, the medical community presumed that since hormonal therapy achieved similar effects on these "proven" surrogates, it would be beneficial in preventing women's heart disease. However, when tested in a randomized controlled setting with cardiac events – and not merely the lipid profile – as the study's primary endpoint, postmenopausal hormone therapy did not reduce heart attacks, and in fact, showed a trend towards increasing cardiac events compared with placebo [4].

This example reveals several key limitations of surrogate endpoints. The first we have already mentioned: a surrogate outcome can only predict a single health outcome, and if harms occur after that endpoint has been reached, a study utilizing the surrogate will provide no information about those long-term adverse effects. Secondly, it illustrates that even if a surrogate has been identified as useful for one intervention (in this case, statins), it does not necessarily follow that the surrogate is valid for other therapies (e.g., hormone replacement). Although, like statins, hormone therapy did have an impact on an individual's lipid profile, it also had a secondary, competing effect: it increased hemostasis. This placed women at increased risk for thrombi, occlusive events, and resulting myocardial infarction; these events were common and severe enough to overcome any benefit gained by the changes in cholesterol levels. The surrogate endpoint only provided half of the story, with unfortunate results. While the Women's Health Initiative study is perhaps the most famous example of an exposed discordance between a surrogate and an ultimate health outcome, it is not the only one. Table 2 presents several examples of unexpected contradictions that have been discovered between proposed surrogates and their final clinical endpoints.

Table 2
Examples of discordance between surrogate endpoints and health outcomes.

Intervention	Effect on surrogate endpoint [17]	Effect on final health outcome
Postmenopausal estrogen + progestin [4]	↓ Cholesterol ↓ LDL ↑ HDL	↑ Coronary heart disease
Encainide, flecainide [18]	↓ Cardiac arrhythmias (PVCs)	↑ Sudden cardiac death
Low fat diet [19]	↓ Colonic polyps	↔ Incidence of colon cancer
Intensive blood sugar control with antiglycemic agents [20]	↓ HA1C <6.4%	↑ Death from macrovascular complications of diabetes (cardiovascular deaths)
Torcetrapib [21]	↑ HDL	↑ Heart failure, death

7. A case study: choosing endpoints for trials in hepatocellular carcinoma

We will now narrow the focus from overarching principles of clinical study design to specific challenges in liver disease: specifically, the unique issues encountered when choosing endpoints for hepatocellular carcinoma therapies. There is an extensive list of potential outcomes one might utilize in the study of new treatments for liver cancer; Table 3 describes several of them. Just as was previously suggested in the pyramid of endpoints, *overall survival* (time from random assignment until death) is the most important and objective outcome.

Although the pyramid indicates that *cancer-specific survival* should be the next best choice – that is, the time from random assignment until death from liver cancer – in the case of hepatocellular carcinoma, there are several unexpected difficulties associated with this endpoint which must be considered. The endpoint is obviously of biologic importance, as the therapies under investigation are disease-specific (i.e., targeted chemotherapy); however, one must remember that in this population death often occurs from concomitant liver failure. When evaluating the cause of death for patients that are Child-Pugh class B or C, it can be challenging to sort out whether the death was specifically a result of the tumor, underlying cirrhosis, or even treatment-related toxicity [15].

Oncology is a "game of millimeters": the advantage gained in life expectancy by a new therapy compared to standard of care is often as little as weeks or months. The major issue associated with these intercurrent causes of mortality is that they can obscure real (but small) differences in treatment effect between the new intervention and the control, as they can generate imbalances of inaccurately classified deaths in either trial arm. Thus, promising new therapies might be abandoned prematurely, or ineffective therapies embraced. Even "hard" clinical outcomes near the top of the pyramid of endpoints may, in fact, be subject to biases.

Time to symptomatic progression (time from randomization to deterioration of symptoms as assessed in a standardized fashion) correlates best with the next step down on the pyramid, as it attempts to capture a tangible physical endpoint (symptoms of the disease experienced by the individual). However, once again, there are unique considerations in the case of hepatocellular carcinoma that make this choice of outcome less reliable in therapeutic trials. Time to symptomatic progression will, in theory, capture both deterioration in quality of life as well as any drug-related toxicities that are experienced. However, in Childs-Pugh B and C class patients, cirrhosis can again obscure the true differentiation of health impairments resulting from the tumor as opposed to those caused by liver failure. It has proven so difficult to accurately ascribe symptoms to one disease process versus the other that there are currently no validated tools or questionnaires to measure quality of life in hepatocellular carcinoma, and thus, no reliable and standardized way to assess this outcome in clinical trials [15]. Therefore, although quality of life is of critical importance to patients, time to symptomatic progression is not currently considered a first-line endpoint in trials of hepatocellular carcinoma therapy.

Regulatory bodies such as the Food and Drug Administration frequently accept *progression-free survival* (a composite endpoint of disease-specific death and/or evidence of radiological progression of a tumor) as an acceptable surrogate endpoint for an oncology trial when fast-tracking new drug approvals; in other solid tumors, it may represent a reasonable approach. However, hepatocellular carcinoma again presents a special challenge. The vulnerability of this population to concomitant liver failure remains an important impediment to accurate evaluation: Just as with disease-specific survival, progression-free survival is prone to inaccurate attribution of the cause of death. Secondly, the utilization of radiologic progression of a tumor as an endpoint is vulnerable to a type of time bias. This endpoint requires frequent repeated measurements equally applied in both study arms to ensure that the progression is captured at its true temporal onset, since the outcome is by definition not associated with external signs. If the interval between measurements extends for too long or is applied at different intervals, real differences between the treatment and control arms can be missed, or spurious differences introduced. Thus, as each individual element of this composite endpoint is potentially unreliable, extreme caution is advised when evaluating trials that have employed progression-free survival in the treatment of hepatocellular carcinoma.

Another indirect surrogate outcome frequently utilized in oncology trials is *disease-free survival* (a composite endpoint of time from randomization to disease-specific death and/or evidence of radiological recurrence of a tumor). Disease-free survival – as a composite of several bias-prone outcomes – suffers from the same limitations as progression-free survival.

This case study of hepatocellular carcinoma demonstrates some of the difficulties inherent not only in the use of surrogate endpoints, but even "harder" clinical outcomes like symptomatic disease and death due to a specific cause. It emphasizes the importance of going back to the "clarity test," as well as the need for a careful and deliberate consideration of potential limitations, for *every* study one evaluates.

8. Conclusion

All trials have a "signal-to-noise" ratio; the trick is to be able to correctly interpret the final balance for a given

Table 3
Potential endpoints in clinical trials in hepatocellular carcinoma.

• *Overall survival*: Time from randomization until death from any cause. Most objective of any outcome as binary choice of dead versus alive is least subject to observer bias
• *Cancer-specific survival*: Time from randomization until death from heptaocellular carcinoma. Subject to incorrect attribution of cause of death due to concomitant cirrhosis/liver failure
• *Time to symptomatic progression*: Time from randomization until occurrence of disease-related symptoms as assessed in a standardized manner. Subject to incorrect attribution of cause of symptoms due to concomitant cirrhosis/liver failure
• *Progression-free survival*: A composite endpoint of time from randomization until radiological progression of tumor or death. Subject to incorrect attribution of cause of death due to concomitant cirrhosis/liver failure as well as incorrect capture of timing of radiological progression of disease
• *Disease-free survival*: A composite endpoint of time from randomization until either radiological recurrence of tumor or death. Subject to incorrect attribution of cause of death due to concomitant cirrhosis/liver failure as well as incorrect capture of timing of radiological recurrence of disease

Adapted from Ref. [15]. Used with permission.

study. There are fingerposts that point towards a more robust and reliable conclusion: these are given general rankings within the pyramids of evidence and endpoints. Constructing an analytic framework can also help to highlight – and thereby avoid – heuristic overenthusiasm and blunt epistemologic hubris. Be wary of uncontrolled trials, surrogate endpoints, and studies that permit the investigator or participant (rather than chance) to choose an exposure or intervention: all of these are prone to biases that can generate a seemingly compelling, but qualitatively wrong result. Randomized controlled trials utilizing "hard" clinical endpoints are the most efficient and reliable way to directly assess the benefits and harms of a therapy; however, even these are not infallible. When interpreting a clinical study, one cannot do better than to follow the advice of physicist Richard Feynman, who told a stadium of Cal-Tech graduates to pursue:

> *A kind of scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty – a kind of leaning over backwards.... This kind of care not to fool yourself.... You should [consider] everything that you think might make [an experiment] invalid – not only what you think is right about it: other causes that could possibly explain [the] results; and things you thought of that you've eliminated.... Details that could throw doubt on your interpretation must be given, if you know them. You must do the best you can – if you know anything at all wrong, or possibly wrong – to explain it [16].*

It is through this "care not to fool ourselves" – and the careful and deliberate application of the highest levels of evidence to clinical decision-making – that real advances in medical care continue to be made.

Acknowledgements

Elements of this review were part of a presentation on evidence-based outcomes in hepatocellular carcinoma BSK presented at the 2nd International Liver Cancer Association Annual Conference in Chicago, September 2008. The views expressed in this article are those of the authors and do not necessarily represent the views of the U.S. federal government or the National Institutes of Health.

References

[1] Hume D. Of Miracles. In: An Enquiry Concerning Human understanding. Whitefish: Kessinger Publishing; 2004.
[2] Ludmerer KM. Time to heal: American medical education from the turn of the century to the era of managed care. New York: Oxford University Press; 1999.
[3] Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Methods work group, third US preventive services task force. Current methods of the US preventive services task force: a review of the process. Am J Prev Med 2001;20:21s–35s.
[4] Manson JE, Hsia J, Johnson KC, Rossouw JE, Assaf AR, Lasser NL, et al. Estrogen plus progestin and the risk of coronary heart disease. N Engl J Med 2003;349:523–534.
[5] Harmanci O, Buyukasik Y, Bayraktar Y. Successful plasma exchange treatment in hemolytic crisis of Wilson's disease preventing liver transplantation. Dig Dis Sci 2006;51:1230.
[6] Riley TR, Taheri MR. Long-term treatment with the combination of amantadine and ribavirin in hepatitis C non-responders. A case series. Dig Dis Sci 2007;52:3418–3422.
[7] Ederer F. Why do we need controls? Why do we need to randomize? Am J Opthalmol 1975;79:758.
[8] Denis J, Opolon P, Nusinovici V, Granger A, Darnis F. Treatment of encephalopathy during fulminant hepatic failure by haemodialysis with high permeability membrane. Gut 1978;19:787–793.
[9] MacMahon B, Yen S, Trichopoulos D, Warren K, Nardi G. Coffee and cancer of the pancreas. N Engl J Med 1981;304:630–633.
[10] Yen S, Hsieh CC, MacMahon B. Consumption of alcohol and tobacco and other risk factors for pancreatitis. Am J Epidemiol 1982;116:407–414.
[11] Hsieh CC, MacMahon B, Yen S, Trichopoulos D, Warren K, Nardi G. Coffee and pancreatic cancer (Chapter 2). N Engl J Med 1986;315:587–589.
[12] Ransohoff DF. Lessons from controversy: ovarian cancer screening and serum proteomics. J Natl Cancer Inst 2005;97:315–319.
[13] Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med 1996;125:605–613.
[14] Levin B, Lieberman DA, McFarland B, Smith RA, Brooks D, Andrews KS, et al. American Cancer Society Colorectal Cancer

Advisory Group; US Multi-Society Task Force; American College of Radiology Colon Cancer Committee. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps: A joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. CA Cancer J Clin 2008;58:130–160.

[15] Llovet JM, Di Bisceglie AM, Bruix J, Kramer BS, Lencioni R, Zhu AX, et al. Panel of experts in HCC-design clinical trials. Design and endpoints of clinical trials in hepatocellular carcinoma. J Natl Cancer Inst 2008;100:698–711.

[16] Feynman R. Cargo Cult Science. In: Surely You're Joking, Mr. Feynman! New York: W.W. Norton, Inc.;1985.

[17] Levin B. Potential pitfalls in the use of surrogate endpoints in colorectal adenoma chemoprevention. J Natl Cancer Inst 2003;95:697–699.

[18] The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. N Engl J Med 1989;321:406–412.

[19] Beresford SA, Johnson KC, Ritenbaugh C, Lasser NL, Snetselaar LG, Black HR, et al. Low-fat dietary pattern and risk of colorectal cancer: the women's health initiative randomized controlled dietary modification trial. JAMA 2006;295:643–654.

[20] The Action to Control Cardiovascular Risk in Diabetes (ACCORD) Study Group. Effects of intensive glucose lowering in type 2 diabetes. N Engl J Med 2008;358:2545–2559.

[21] Barter PJ, Caufield M, Eriksson M, Grundy SM, Kastelein JJ, Komajda M, et al. ILLUMINATE investigators. Effects of torcetrapib in patients at high risk for coronary events. N Engl J Med 2007;357:2109–2122.