

How to read a published clinical trial: a practical guide for clinicians

Mohamad B. Sonbol¹, Belal M. Firwana², Talal Hilal³, Mohammad Hassan Murad⁴

¹Mayo Clinic Cancer Center, Phoenix, Arizona, ²Heartland Cancer Research, National Cancer Institute Community Oncology Research Program (NCORP), Missouri Baptist Medical Center, St. Louis, Missouri, ³University of Mississippi Medical Center, Jackson, Mississippi, ⁴Evidence-Based Practice Center, Mayo Clinic, Rochester, Minnesota, USA

Access this article online

Website: www.avicennajmed.com

DOI: 10.4103/ajm.ajm_186_19

Quick Response Code:



ABSTRACT

Over the last 5 years, there have been more than 140 new drug approvals in the field of Oncology alone, all based on newly published clinical trials. These approvals have led to an ongoing change in clinical practice, offering new therapeutic options for patients. Therefore, it is important for healthcare providers to be able to appraise a clinical trial and determine its validity, understand its results, and be able to apply such results to their patients. In this guide, we provide a simplified approach tailored to practicing clinicians and trainees. The same concepts and principles apply to other medical specialties.

Key words: Apixaban, apixaban for the prevention of venous thromboembolism, critical appraisal, evidence-based medicine, evidence-based practice, randomized-clinical trials

INTRODUCTION

The modern practice of oncology is based on clinical trials, which have been increasingly conducted and published in the last 20 years. Over the last 5 years, there have been more than 140 anticancer drug approvals in the United States.^[1] These approvals have led to an ongoing change in clinical practice, offering new therapeutic options for patients with cancer. Therefore, it is important for physicians to be able to appraise a clinical trial and determine its validity, understand its results, and be able to apply such results to their patients. In this guide, we provide a simplified approach based on the User's Guide to the Medical Literature series tailored to practicing clinicians and trainees.^[2] Although most of the included examples are from the oncology literature, the same concepts and principles would apply to other medical and surgical specialties.

Clinical case

A 56-year-old man with a history of diabetes mellitus who was recently diagnosed with metastatic non-small cell lung cancer comes to your oncology clinic for clinical care. You decide to start him on chemotherapy with carboplatin and

pemetrexed. He is otherwise healthy and asymptomatic. His body mass index (BMI) is 41. He has reasonable functional status as measured by the Eastern Cooperative Oncology Group (ECOG) performance status (PS) score value of 1. His mother recently died of a pulmonary embolism (PE) and he is asking you about prevention of PE. You calculate his risk for chemotherapy-associated thrombosis using the Khorana score^[3] and find it to be 2, which suggests an intermediate risk for venous thromboembolism. You are contemplating thromboprophylaxis and proceed to review the evidence.

What is a clinical trial?

A clinical trial is any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes. Therefore, a clinical trial can be

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

Address for correspondence: Dr. M. Hassan Murad, 200 1st St SW, Rochester, MN 55905, USA.
E-mail: murad.mohammad@mayo.edu

Cite this article as: Sonbol MB, Firwana BM, Hilal T, Murad M. How to read a published clinical trial: A practical guide for clinicians. *Avicenna J Med* 2020;10:68-75.

randomized (i.e., a randomized controlled trial [RCT]) or nonrandomized. For inference purposes, nonrandomized trials are at similar risk of bias to that of observational studies and can be appraised by focusing on cohort selection, comparability of study groups, and adjustment for confounders (discussed in the Comparability of the groups at the baseline section) (i.e., just like an observational study). For the most part, when clinicians think about trials, they are usually referring to RCTs which are the gold standard study design to ascertain the effect of therapy. The RCT design creates groups of patients that are similar in all known and unknown prognostic factors (i.e., confounders) except the intervention. RCTs can randomize the patients to groups and follow them prospectively (parallel RCTs) or can switch patients at random to different treatment regimens during the course of the trial (crossover RCTs). This guide will focus on parallel-design RCTs because they are more common and are critical for the practice of internal medicine and oncology. We will also focus on an example of a superiority trial for simplicity (i.e., a trial that aims to evaluate if the experimental treatment is better than the standard treatment or placebo), although many of the constructs and principles discussed here apply to noninferiority trials (i.e., a trial that aims to evaluate whether an experimental treatment is not importantly worse than a standard treatment).

Lastly, although we are discussing an approach to appraise and apply an RCT, it is important to keep in mind that having a systematic review and meta-analysis of multiple RCTs would likely give more precise and valid estimates and would be preferred if available.^[4] Moreover, the fundamental principles of evidence-based medicine (EBM) are assumed, which include formulating an answerable question, identifying the best evidence, critically appraising the evidence, applying the evidence, and integrating clinical expertise and patient's values with the evidence.^[5] In this concise guide, we will critically appraise an RCT that aims to answer the following clinical question: What is the evidence supporting prophylactic anticoagulation in patients with cancer?

APPROACH

When reading a manuscript reporting the conduct and results of an RCT, one should ask three questions. How valid are the results (which is also expressed as to what extent does the risk of bias affect the trustworthiness of the results)? What are the results? How do I apply these results to patient care? This simplified approach is based on the User's Guide to the Medical Literature series and also adapted in oncology.^[2,6]

We have identified one RCT that addresses the clinical question of interest to our patient—"Apixaban to Prevent Venous Thromboembolism in Patients with Cancer"—the AVERT trial.^[7]

The trial evaluates the efficacy and safety of apixaban (2.5 mg twice daily) for thromboprophylaxis in ambulatory patients with cancer who were at intermediate-to-high risk for venous thromboembolism (Khorana score, ≥ 2) initiating chemotherapy.^[7]

How valid are the results (to what extent does the risk of bias affect the trustworthiness of the results)?

The validity of the RCT focuses on how well the study is conducted and addresses different types of bias.^[8] Appraising the study's internal validity can be achieved by evaluating the methods section and following a stepwise approach. Did the study: start well, run well, and finish well?^[6] [Figure 1].

START WELL

Was the allocation sequence random?

Randomization (also known as allocation sequence generation) ensures that the study participants have an equal chance of being assigned to either the intervention group or the control group, thereby decreasing the likelihood of an imbalance in baseline prognostic factors which can cause what is called selection bias. For example, if the fit and younger patients were assigned to one arm of a study, this arm will have better outcomes that are not caused by the intervention. Randomization is commonly performed using a computer-generated algorithm.

In the AVERT trial, the authors state in the methods section "eligible patients underwent randomization by means of a centralized, web-based randomization system to receive apixaban or placebo in a 1:1 ratio."^[7] The randomization in this trial is adequate.

Was the allocation sequence concealed until participants were enrolled and assigned to interventions?

When appraising the validity of a study, it is important to look at the method of randomization and whether it can prevent the predictability of the allocation (also known as concealment). Concealment means that both study participants and investigators are not aware, and cannot predict, which group the study participant (patient) will be assigned to. This is not to be confused with blinding of assigned interventions (discussed below). Allocation concealment happens prior to and at the time of randomization. Conversely, blinding occurs after

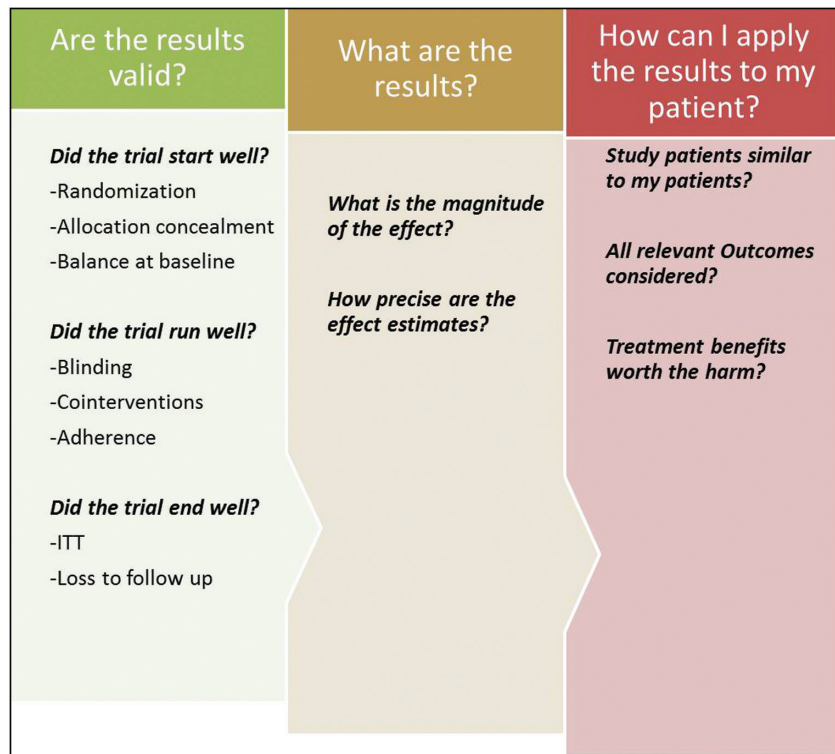


Figure 1: A framework summarizing the steps in critically appraising a randomized controlled trial. ITT = intention to treat

randomization^[9-12] [Figure 2]. Patient enrolment can be concealed but not blinded. An example of that is the biliary tract cancer (BILCAP) trial, where treatments were not masked but allocation concealment was achieved.^[13]

In the AVERT trial, the authors used a “centralized, web-based randomization” method which ensures that both participants and investigators could not foresee assignment.^[7]

Comparability of the groups at the baseline

The benefit of randomization is in minimizing the imbalance and differences in baseline characteristics and prognostic factors between the groups. These differences are sometimes referred to as “confounders.” These baseline characteristics are almost always reported in Table 1 in RCTs. When detected, it is important to evaluate the importance of the prognostic factor imbalances (confounding) by asking the following questions: (1) Does the prognostic factor affect the outcome?; (2) If yes, which group is favored?; (3) Does this change the conclusion? This accounts for known confounders, but unknown confounders can always introduce bias. Potential unknown confounder imbalance can be minimized with appropriate randomization.

In the AVERT trial, table 1 shows how the groups were comparable at the baseline, in terms of tumor type, Khorana score, PS, and others, including the use of concomitant antiplatelet medications.^[7] It is important to look at the proportions in table 1 and determine whether they are

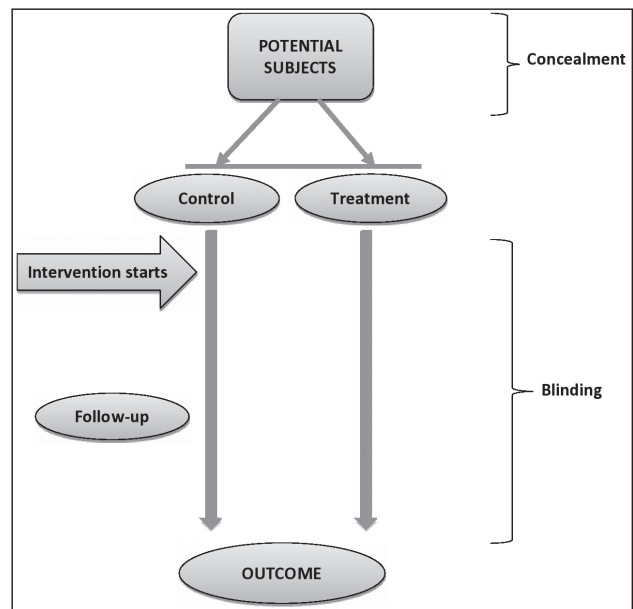


Figure 2: A flow diagram showing blinding, concealment, and randomization

clinically meaningful, and not to depend on reported P values. These P values are not meaningful (although commonly reported) because the trial is often underpowered to show significant differences in these variables.

RUN WELL

This series of questions concerns performance bias and bias due to deviations from intended intervention and includes

Table 1: Primary efficacy and safety outcomes of the AVERT trial

	Apixaban (N = 288)	Placebo (N = 275)	RR*	ARR (or RD)*	NNT*	HR (95% CI)†	P Value†
	n (%)	n (%)					
VTE	12 (4.2)	28 (10.2)	0.41	6%	16.6	0.41 (0.26–0.65)	<0.001
Major bleeding	10 (3.5)	5 (1.8)	1.94	1.7%	58.8	2.00 (1.01–3.95)	0.046

N = total number of patients, n = number of events, RR = relative risk, ARR = absolute risk reduction, RD = risk difference, NNT = number needed to treat, HR = hazard ratio, CI = confidence interval

*Value calculated

†Value exported from reported trial results

blinding, contamination, co-intervention, and compliance [Figure 1].

Were participants and investigators aware of their assigned intervention during the trial?

Blinding refers to the process by which the study participants (patients), providers (nurses and physicians), investigators, and outcome assessors are kept unaware of treatment assignment throughout the study.^[8,14] Blinding of patients and study personnel help in reducing performance bias that could occur upon the knowledge of the assignment. Performance biases arise from deviations from intended interventions. For example, if a study investigator is aware of treatment assignment, they might elect to monitor and see the patient in the novel therapy group more frequently than the control group. In addition, blinding of study participants helps in reducing the risk of the “placebo-effect” that can be detected in more subjective outcomes such as pain.^[15,16] For example, in an RCT of patients with nasopharyngeal carcinoma, acupuncture significantly lowered radiation-induced xerostomia compared to standard care group (no acupuncture).^[17] In this example, blinding of participants was not performed; however, it is hard to draw a clear conclusion from such trial when the outcomes (xerostomia and quality of life [QOL]) are subjective and could be affected by the “placebo-effect.” This has been described before where trials of acupuncture found benefit in treating pain compared to no treatment. However, this benefit was less significant when acupuncture was compared with sham control.^[18] The effect of blinding in a study should be assessed for each individual outcome; it may be less important in more objective outcomes such as overall survival (OS).

In the AVERT trial, the authors state that “The AVERT trial was a randomized, placebo controlled, double-blind clinical trial.”^[8] One can assume that “double blind” implies that patients and investigators were blinded. However, it is important to read the methods section to find out who was actually blinded.^[19]

Was there any contamination or co-intervention

The study protocol usually specifies the intended interventions in each study group. When a study participant (patient) receives a non-protocol intervention, it is usually

referred to as “co-intervention.” On the contrary, when a study participant receives the intervention that is assigned to the other study group, it is referred to as “contamination.”

In the AVERT trial, 23% and 22.6% of patients in the apixaban and placebo groups, respectively, received a concomitant antiplatelet medication (a co-intervention), which could potentially affect the primary outcomes of bleeding and clotting in such trial.^[8] However, as both groups equally received this co-intervention, this will unlikely bias the results.

Was there nonadherence to the assigned intervention regimen that could have affected participants' outcomes?

Compliance of the study participants to the intervention they are assigned to is referred to as “adherence.” It is important when appraising a trial to look at the reported adherence and whether there is a significant difference between groups. This is especially important in oncology RCTs where adverse events and safety profile of the studied medications play a major role in patients' compliance.^[20] For example, in the recently reported BILCAP trial studying the effect of adjuvant capecitabine compared to observation following surgery in patients with BILCAP, only half of the patients (55%) completed the planned eight cycles of capecitabine with third of the patients discontinuing treatment secondary to toxicity.^[13]

In the AVERT trial, the authors state that “The rate of adherence to the trial regimen was high in both groups, at 83.6% in the apixaban group and 84.1% in the placebo group.”

FINISH WELL

The method of analysis and completion of follow-up are important factors that affect trial validity.

Were all patients who entered the trial accounted for? And were they analyzed in the groups to which they were randomized? Were there any lost to follow-up?

The principle of intention to treat (ITT) analysis indicates that participants should be analyzed based on the

intervention group to which they were assigned, regardless of their adherence to the intervention or lost to follow-up (participant cannot be located).^[21]

This is in contrast to the per-protocol analysis, which only analyzes the individuals who adhered to the intervention. ITT analysis maintains the benefit of randomization in minimizing any prognostic differences between groups. In contrast, the problem with the per-protocol analysis is that prognostic factors might influence whether individuals receive their allocated intervention. In RCTs assessing a superiority outcome, ITT is suggested for the most part. Some trials report both ITT and per-protocol analysis; for example, the previously mentioned BILCAP trial reported the OS results using both ITT and per-protocol analyses, with significant improvement in outcome seen with per-protocol analysis, but not with ITT analysis, reducing the trustworthiness or believability of the results.^[13]

In some trials, instead of reporting ITT, a modified intention to treat (mITT) is reported. The definition of such an analysis is variable between trials and mostly generates post-randomization exclusions that potentially bias results making interpretation of such analyses challenging.^[22]

In the AVERT trial, the primary analysis was performed in the “modified intention-to-treat” population, which included all patients who undergone randomization and received at least one dose of apixaban or placebo on or before day 180.^[7] Although ITT is the preferred approach, in this study the mITT is likely adequate and would not be expected to greatly alter the observed effect size compared to ITT. This modification—analyzing patients who received at least one dose of the study drug—is commonly seen in studies assessing differences in adverse drug events between treatment groups because it could be considered inappropriate to attribute an adverse drug event to a medication never received by the patient.^[23] Although a threshold of >20% patients lost to follow-up is sometimes used to assess whether the number of patients of lost to follow-up is not acceptable, these arbitrary cutoffs can sometimes be misleading. It is important to compare the proportion lost to follow-up to the event rate in the trial. It is also important to conduct what is called a worst-case scenario in which we assume that patients lost to follow-up had bad outcome. If this new analysis shows results that are different from the original analysis, validity is then reduced.

Was outcome assessment blinded?

As described above, in addition to blinding patients and investigators, it is important to have blinding of outcome

assessors. Indeed, the effect of blinding in a study should be assessed for each individual outcome as it is probably less important in objective outcomes as OS (death or alive) compared to progression-free survival.^[24]

In the AVERT trial, outcomes were assessed by blinded investigators “All trial outcomes were adjudicated by an independent adjudication committee whose members were unaware of the treatment assignment.”

WHAT ARE THE RESULTS?

Once trial validity is established (i.e., risk of bias is low or unlikely to impact the conclusions), results need to be interpreted by asking about the magnitude of the effect and its precision.

What is the magnitude of the treatment effect?

There are several commonly used methods that are referred to as “measures of association” to assess the magnitude of treatment effect in clinical trials, including but not limited to relative risk (RR), odds ratio (OR), risk difference (RD), and hazard ratio (HR).

Relative risk and relative risk reduction

RR is the risk of disease or outcome in the treatment or exposed arm compared (relative) to the risk of the outcome in the control arm, hence the name RR.

On the contrary, relative risk reduction (RRR) is an estimate of the percentage of baseline risk (the control arm risk) that is reduced by receiving the experimental therapy, which is calculated as subtracting RR from 1 ($1 - RR$). For example, looking at the outcome table for the AVERT trial [Table 1], the risk of venous thromboembolism (VTE) in apixaban group is $12/288 = 4.2\%$ (also known as experimental event rate or EER) and the risk of VTE in placebo group is $28/275 = 10.2\%$ (also known as control event rate or CER). Compared to patients in the placebo group, patients assigned to the apixaban group have almost half of the risk (41%) of the patients in the placebo group $4.2/10.2 = 41\%$. This is also known as RR. In other words, this means that apixaban decreased the RR by $1 - 0.41$ (41%) = 59%. This is known as RRR.

One example of using RR in cancer clinical trials is when assessing response rates in the experimental and control arms. For example, in the Keynote-189 trial, comparing pembrolizumab plus chemotherapy versus chemotherapy alone in metastatic non-small-cell lung cancer,^[25] objective response rates were 47.6% versus 18.9%, with an RR of 2.5, meaning that the experimental regimen results in 2.5 times better responses compared to the control arm.

Odds ratio

OR is another relative association measure that is similar to RR. However, it is a ratio of odds, not risks. Odds are events/nonevents, whereas risk is events/total exposed sample. When the event rate is low (<10%), OR and RR become very similar.^[26]

Risk difference

Although relative measures (RR and RRR, OR and HR) are very helpful to depict the direction of the association, they do not give the full picture, especially when interpreting data or discussing with patients. Therefore, reporting absolute measures is as important, namely the RD, which is the proportion of the event in the experimental arm subtracted from the proportion of the event in the control arm. In other words, it is the proportion of patients who are spared the undesired outcome having received the experimental rather than the control treatment. RD of 0 means the events occurred equally in both groups. RD is sometimes called absolute risk reduction (ARR) or absolute risk increase (ARI) based on the direction of the effect. When interpreting RCTs, it is important to look at both ARR and RRR, as looking at relative measures can be deceiving and tends to overestimate results. In a hypothetical example, an RR of 50% could represent an ARR of 30% (if the absolute risk improved from 60% to 30%), or that same RR of 50% could represent an ARR of 2% (if the absolute risk improved from 4% to 2%).

In our example [Table 1], in the AVERT trial: baseline risk of VTE in the placebo group is 10.2% and is decreased to 4.2% in the apixaban group. Therefore, giving apixaban decreased the risk of VTE by $10.2 - 4.2 = 6\%$, which is the RD.

Number needed to treat/harm

Another important measure of association is the number needed to treat (NNT). This reflects the number of patients who need to be treated in order to prevent one event (in this case, VTE). $NNT = 1/ARR$ (when ARR is in percentage, this would be $NNT = 100/ARR$). In the AVERT trial, the $NNT = 100/6 = 16.6$ patients. In the same way, we can calculate the number needed to harm (NNH), which is the number of patients who need to be treated in order to harm one patient or cause one undesired event. The risk of bleeding in the apixaban group is 3.5% and in the placebo group it is 1.8%. The RD is 1.7% (3.5–1.8). For 100 patients treated, 1.7% get harmed. The $NNH = 100/1.7 = 58.8$ patients.

These numbers are useful when evaluating the magnitude of effect and safety of the intervention by comparing NNH and NNT. For apixaban, for each 16 patients we treat we benefit 1, and for each 58 patients we treat, 1 would be harmed. We obviously seek drugs with low NNT and high NNH.

Hazard ratio

The HR is a relative association measure used for outcomes of survival in cancer clinical trials. Although calculated differently,^[27] for practical purposes it can be interpreted as an RR averaged over the course of a trial and can be expected at any given time during the follow-up. The calculation of HR includes the element of time (i.e., how long an event took to occur vs. did it occur or not). HR of 1 means no effect; HR of 2 means that the intervention doubles the risk of outcome; and HR of 0.5 means that the intervention halves the risk of outcome. HR should always be interpreted with consideration of the associated length of survival. In the trial of erlotinib plus gemcitabine compared with gemcitabine in patients with advanced pancreatic cancer,^[28] median survival time was 6.24 months in the experimental arm of combination therapy versus 5.91 months in the gemcitabine arm. Thus, although the HR of 0.82 suggests improved survival, the actual difference in survival could be trivial.

How precise is the estimate of treatment effect?

Confidence intervals (CIs) in RCTs identify a range of values within which it is probable that the true effect of treatment lies. In most trials, 95% CI is estimated to indicate that if the trial was repeated 100 times, 95% of the CI would include the true effect; the wider the CI, the less precise the estimate. For example, in the Keynote-189 trial, the HR for death was 0.49 with a reported $P < 0.001$ (which means that this effect is statistically significant because it is < 0.05 , the arbitrary cutoff for significance). This HR of 0.49 had a 95% CI of 0.38–0.64. When making a decision, one should consider precision. If our decision would be the same whether the lower or the upper boundaries were the truth, then the results are sufficiently precise. In this case, the precision is adequate.

HOW DO I APPLY THE RESULTS TO MY PATIENTS?

Applicability is a form of external validity.^[29] To assess applicability, one should ask the following questions:

Were the study patients similar to my patients?

This question can be answered by looking at inclusion and exclusion criteria for the RCT and compare them to the characteristics of the patient of interest. RCTs with long lists of exclusion criteria (e.g., comorbid conditions) may be less applicable in real practice. Furthermore, RCTs in oncology can be regional, a few countries in the same region, or international, spanning multiple regions and countries, which makes generalizability variable depending on the regions where the RCT was conducted. For example, the oral fluoropyrimidine, S-1, was shown to improve OS as an adjuvant chemotherapy option in patients with curatively resected gastric cancer in Japan only.^[30] It has

yet to be approved in the United States due to this regional variation, which limits generalizability of drug metabolism and efficacy data to Western patients. However, we should not expect a perfect match and we should anticipate that most of the time relative treatment effects apply to patients with various characteristics.

In the AVERT trial, one of the inclusion criteria was a Khorana score of 2, an intermediate risk category associated with only 1%–2% risk of VTE.^[3] Approximately two-thirds of participants in the AVERT trial had a Khorana score of 2. Using apixaban in this group of patients may be associated with greater harm than benefit as the baseline risk of VTE is very small.

Were all clinically meaningful outcomes considered?

When a drug produces small increments in hemoglobin level, or that a chemotherapy agent causes tumors to shrink above a specific threshold (i.e., response rate), this may not provide sufficient justification for recommending these interventions to patients. These are surrogate outcomes that may or may not lead to an improvement in clinically meaningful, patient-important outcomes, such as QOL or OS.

In the AVERT trial, investigators preemptively evaluated for the presence of VTE with imaging in the absence of symptoms or signs of VTE, a practice that is not commonly performed or indicated for most VTEs. This probably led to the diagnosis of many incidental VTE, which otherwise may have not been found or caused important morbidity or mortality.

Do treatment benefits outweigh the potential risks (harm and costs)?

We evaluate the patient's baseline risk to determine whether introducing an intervention would be worthwhile. A low baseline risk usually means the RD will be low and NNT will be high. Knowing these absolute measures can assist clinicians in helping patients weigh the benefits and risks of each potential intervention. Ultimately, the values and preferences for each individual patient need to be considered before recommending one therapy over another.

CONCLUSION OF THE CLINICAL SCENARIO

After applying the framework [Figure 1], you found that this RCT (AVERT trial) was at low risk of bias. However, after you discuss the efficacy data along with the underlying risk of bleeding in this patient, the patient decided not to start the medication. A different patient with similar characteristics (Khorana score of 2) might elect to accept

such risk in return to the benefits seen. This emphasizes the importance of shared-decision making when applying evidence to individual patients.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

- Hilal T, Sonbol MB, Prasad V. Analysis of control arm quality in randomized clinical trials leading to anticancer drug approval by the US food and drug administration. *JAMA Oncol* 2019;5:887-92.
- Guyatt G, Rennie D, Meade MO, Cook DJ. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. 3rd ed. New York: McGraw-Hill.
- Khorana AA, Kuderer NM, Culakova E, Lyman GH, Francis CW. Development and validation of a predictive model for chemotherapy-associated thrombosis. *Blood* 2008;111:4902-7.
- Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *Evid Based Med* 2016;21:125-7.
- Akobeng AK. Principles of evidence based medicine. *Arch Dis Child* 2005;90:837-40.
- Prasad K. *Fundamentals of Evidence Based Medicine*. India: Springer; 2014.
- Carrier M, Abou-Nassar K, Mallick R, Tagalakis V, Shivakumar S, Schattner A, *et al.*; AVERT Investigators. Apixaban to prevent venous thromboembolism in patients with cancer. *N Engl J Med* 2019;380:711-9.
- Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0. USA: The Cochrane Collaboration; 2011.
- Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001;323:42-6.
- Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. *Ann Intern Med* 2002;136:254-9.
- Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408-12.
- Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: Chance, not choice. *Lancet* 2002;359: 515-9.
- Primrose JN, Fox RP, Palmer DH, Malik HZ, Prasad R, Mirza D, *et al.*; BILCAP study group. Capecitabine compared with observation in resected biliary tract cancer (BILCAP): A randomised, controlled, multicentre, phase 3 study. *Lancet Oncol* 2019;20:663-73.
- Karanicolas PJ, Farrokhyar F, Bhandari M. Practical tips for surgical research: Blinding: Who, what, when, why, how? *Can J Surg* 2010;53:345-8.
- Hróbjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. *Int J Epidemiol* 2014;43:1272-83.
- Nüesch E, Reichenbach S, Trelle S, Rutjes AW, Liewald K, Sterchi R, *et al.* The importance of allocation concealment and patient blinding in osteoarthritis trials: A meta-epidemiologic study. *Arthritis Rheum* 2009;61:1633-41.
- Meng Z, Garcia MK, Hu C, Chiang J, Chambers M, Rosenthal DI, *et al.* Randomized controlled trial of acupuncture for prevention of radiation-induced xerostomia among patients with nasopharyngeal carcinoma. *Cancer* 2012;118:3337-44.

18. Madsen MV, Götzsche PC, Hróbjartsson A. Acupuncture treatment for pain: Systematic review of randomised clinical trials with acupuncture, placebo acupuncture, and no acupuncture groups. *BMJ* 2009;338:a3115.
19. Devereaux PJ, Bhandari M, Montori VM, Manns BJ, Ghali WA, Guyatt GH. Double blind, you are the weakest link—good-bye! *ACPJ Club* 2002;136:A11.
20. Greer JA, Amoyal N, Nisotel L, Fishbein JN, MacDonald J, Stagl J, *et al.* A systematic review of adherence to oral antineoplastic therapies. *Oncologist* 2016;21:354-76.
21. Higgins JPT, Sterne JAC, Savović J, Page MJ, Hróbjartsson A, Boutron I, *et al.* A revised tool for assessing risk of bias in randomized trials. *Cochrane Methods. Cochrane Database Syst Rev* 2016;10:29-31.
22. Abraha I, Montedori A. Modified intention to treat reporting in randomised controlled trials: Systematic review. *BMJ* 2010;340:c2697.
23. Beckett RD, Loeser KC, Bowman KR, Towne TG. Intention-to-treat and transparency of related practices in randomized, controlled trials of anti-infectives. *BMC Med Res Methodol* 2016;16:106.
24. Page MJ, Higgins JP, Clayton G, Sterne JA, Hróbjartsson A, Savović J. Empirical evidence of study design biases in randomized trials: Systematic review of meta-epidemiological studies. *PLoS One* 2016;11:e0159267.
25. Gandhi L, Rodríguez-Abreu D, Gadgeel S, Esteban E, Felip E, De Angelis F, *et al.*; KEYNOTE-189 Investigators. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. *N Engl J Med* 2018;378:2078-92.
26. Murad MH, Montori VM, Walter SD, Guyatt GH. Estimating risk difference from relative association measures in meta-analysis can infrequently pose interpretational challenges. *J Clin Epidemiol* 2009;62:865-7.
27. Alahdab F, Haydour Q, Murad MH. Teaching tips: Explaining hazard ratios. In: Newsletter of the International Society for Evidence-Based Health Care. editors, Jason Busse, Paul Glasziou, Gordon Guyatt, Vol. 13; 2013.
28. Moore MJ, Goldstein D, Hamm J, Figer A, Hecht JR, Gallinger S, *et al.*; National Cancer Institute of Canada Clinical Trials Group. Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: A phase III trial of the national cancer institute of Canada clinical trials group. *J Clin Oncol* 2007;25:1960-6.
29. Murad MH, Katabi A, Benkhadra R, Montori VM. External validity, generalisability, applicability and directness: A brief primer. *BMJ Evid Based Med* 2018;23:17-9.
30. Sakuramoto S, Sasako M, Yamaguchi T, Kinoshita T, Fujii M, Nashimoto A, *et al.*; ACTS-GC Group. Adjuvant chemotherapy for gastric cancer with S-1, an oral fluoropyrimidine. *N Engl J Med* 2007;357:1810-20.