Study types and bias—Don't judge a study by the abstract's conclusion alone

Authors   **Daniel C Norvell**
Institution   **Spectrum Research, Inc., Tacoma Washington, USA**

Why is it important to consider the study design before implementing the findings?

Not all study designs are created equal. Some designs are inherently better at minimizing bias. Bias (usually unintended) is one of the greatest threats to a study's conclusion. In this issue of EBSJ, we will discuss the strengths and weaknesses of the common study designs that you are most likely to encounter in the literature or consider for your next study.

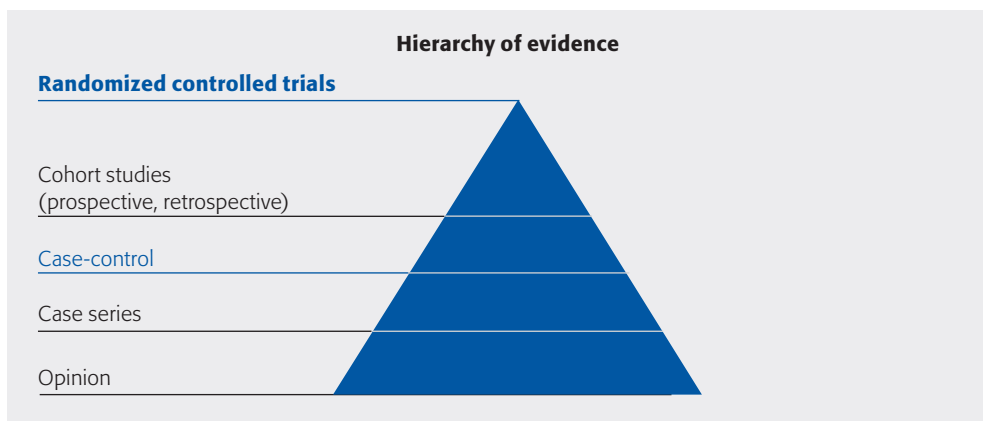### What is the primary goal of a clinical study?

The goal of most clinical studies is to evaluate a treatment method and to report the most *accurate and unbiased effect of the treatment.* One important way to help minimize bias is to select the best study design to accomplish your purpose.

There are three frequently used study designs we will discuss today: the randomized controlled trial, the cohort study, and the case series. There are costs and benefits to each that must be weighed. We will also discuss how registry studies fit into this paradigm since there is a movement toward using registries for comparative effectiveness research.

Randomized controlled trials

When comparing two treatments, the comparison groups should be comprised of participants who are similar in all respects, with the exception of the particular treatment(s) that is being studied. The best method to achieve this similarity between groups is that of random assignment.
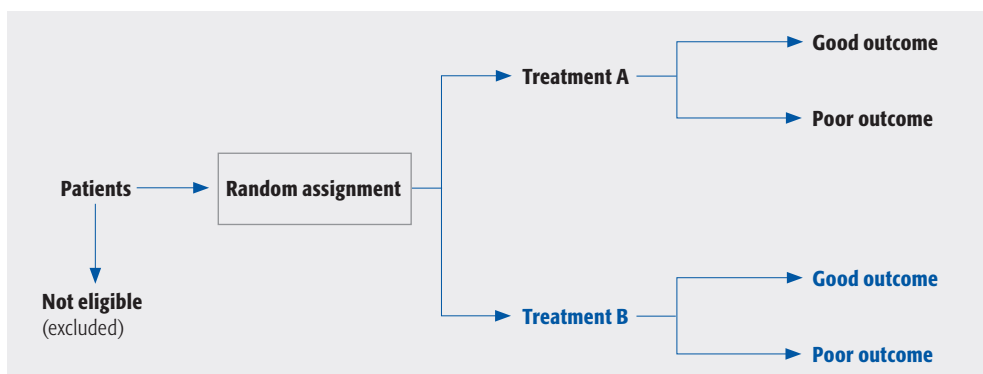
The randomized controlled trial (RCT) provides the strongest evidence for safety and effectiveness and is considered the gold standard for therapeutic studies.

Hierarchy of evidence

Randomized controlled trials

Cohort studies
(prospective, retrospective)

Case-control

Case series

Opinion

RCTs are characterized by:

- A group of patients randomly assigned to an experimental group to receive a treatment such as surgery, or to a control group (no treatment, placebo or an active alternative).
- Prospectively collected data. It is redundant to label your study a 'prospective RCT'.
- Minimizing selection bias (known and unknown). Meaning it is unlikely that there will be an appreciable imbalance in baseline factors between the groups that are also associated with the outcome. For example, smokers should be equally distributed. If they are not, the treatment with the most smokers may appear inferior when in reality it is not.
- Offering the most solid basis for an inference of cause and effect compared with the results obtained from any other study design. That is, we can assume if the results favor one treatment over another, those positive results are much more likely to be due to the treatment than if a cohort study was executed.
- A number of specific challenges when comparing surgical interventions. These include factors such as patient preferences, differential surgeon expertise, changing surgical technologies during lengthy trials and issues surrounding dealing with crossovers. These circumstances may require special methodological considerations such as sham procedures if deemed feasible and acceptable.

The RCT study design looks like this:

Good outcome

Poor outcome

Treatment A

Patients → **Random assignment**

Not eligible
(excluded)

Treatment B

Good outcome

Poor outcome

When judging an individual study's class of evidence (CoE), RCTs are given a class of I or II depending on the overall quality of the study with respect to other methodological characteristics.
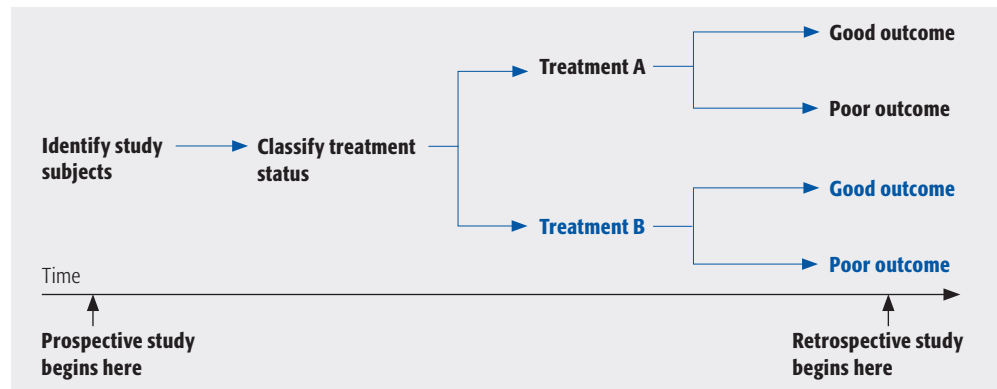
Cohort studies are characterized by:
- Comparing the outcomes of patients whose treatment differs 'naturally', ie, not as the result of random assignment. For example, comparing the outcomes of two types of spine surgeries, one done routinely by you (eg, cervical spine fusion) and one done routinely by your colleague (eg, cervical disc replacement) constitutes a cohort study. (Ideally this study is done in the same patient population. For example, your colleague works at the same institution. Comparing across institutions and across different time periods introduces additional levels of bias).
- Identifying study participants *based on treatment*, and then their outcomes are compared. In our example, the groups are formed based on the treatment they received—fusion versus disc replacement.
- The ability to establish a temporal relationship between the treatment and the outcome because the treatment precedes the outcome.
- The potential imbalance of prognostic factors (those factors that may influence outcomes apart from the treatment) between the two groups. This is one of the biggest problems with cohort studies. Some examples of factors that might have an influence on outcome that might be imbalanced between groups include age, overall health or physical condition, smoking status, and severity of degenerative changes.
- A decreased likelihood of crossover—a major problem found with RCTs. The wish of patients to have an active say in their treatment and a growing reluctance to submit to the random assignment to a specific treatment modality has increasingly hampered surgical RCTs. In some major recent spine RCTs, up to half of patients crossed over to the alternate treatment despite their consent to participate in the first place.

Cohort studies may be divided into those that are prospective and retrospective.

⟹ Prospective cohort studies determine treatment at the beginning of the study with follow-up for outcome to occur in the future.

⟸ Retrospective cohort studies, on the other hand, are characterized by the treatment and outcome having already occurred at the time of study initiation.

Note that retrospective cohort studies are often assumed to have more bias since the study operations, data collected, data entry, and data quality assurance, were not planned ahead of time. Any of these areas could be compromised when relying on data that were already collected. Having said that, if the author can assure the reader that many of these areas are not compromised in their retrospective study, then the reader should give the study more credence. There will be more discussion on this when we talk about registry studies at the end of this article.

The cohort study design looks like this:

When judging an individual study's class of evidence (CoE), cohort studies are given a level of II or III depending on the overall quality of the study with respect to other methodological characteristics.

Case series are characterized by:
- Collection of multiple noteworthy clinical occurrences.
- Cases that experience a novel treatment. For example, you have developed a novel minimally invasive technique. You have performed your technique on 65 cases and now you report the outcomes from your procedure on these cases.

- Unusual cases, either those with atypical characteristics or those with unusual signs and symptoms. One example would be a group of high-performance professional athletes who had disc replacement surgery. You now have 3-year follow up in 30 of these patients and you want to report on the results.
- A lack of hypothesis or a comparison group. This is the biggest weakness of a case series. Without a contemporary comparison group, it is not possible to know with certainty what the outcome would be if the patient received a different treatment. As a result, most case series help to generate hypotheses, not answer clinical questions of efficacy or effectiveness.
- The ability to assess the safety of a new treatment where few studies have been performed evaluating it.

When judging an individual study's class of evidence (CoE), case series are given a class of IV. It is important to note one really cannot establish the efficacy of a treatment without a comparison group even if results are superior to studies in the published literature. One cannot even attempt to measure or adjust for bias in this situation; therefore, efficacy statements based on case series data should not be made or relied on for clinical implementation. On the other hand, a well planned case series may give one an overall safety profile of a specific treatment in a specific patient population.

Registry studies...

….are not a study design but rather a method of data collection. While prospective studies involve the a priori development of data collection forms with planned study operations prior to study execution, and retrospective studies rely on data that have already been collected (eg, medical records), registries may or may not possess data that were planned ahead of time. For example, some registries are a compilation of many existing databases that are merged together. On the other hand, some registries are designed similar to a clinical trial with careful planning, data collection, and quality assurance and monitoring throughout the life of the registry. Many registries fall somewhere in the middle. By definition, a study published from a registry is inherently a retrospective study. You might see it described as "prospectively collected". This is not enough to convince a thoughtful reader that high quality methods were adhered to. Cohort studies that are retrospective in nature are automatically CoE III studies (instead of II) because of the myriad of potential biases and unplanned data collection methods that are inherent in data already collected for clinical or other purposes.

The following are criteria to consider when evaluating the quality of a registry you are designing or a registry study you are evaluating. A good quality registry should have the following characteristics that are important for all studies. If all (or all but one) of these criteria are met, we would judge the study as a CoE II study even though it is retrospective in nature. Violation of two more of these data would render the study a class III or IV depending on how many are violated.

- Designed specifically for conditions evaluated
- Designed for prospective data collection
- Validation of completeness and quality of data
- Patients followed long enough for outcomes to occur
- Independent outcome assessment. Outcome assessment is independent of healthcare personnel judgment. Some examples include patient reported outcomes, death, and reoperation.
- Complete follow up of ≥ 85%
- Controlling for possible confounding. Authors must provide a description of robust baseline characteristics, and control for those that are unequally distributed between treatment groups.
- Accounting for time at risk. Equal follow-up times or for unequal follow-up times, accounting for time at risk.