What researchers mean by...

Easy-to-understand definitions of common research terms in the health and social sciences

Institute for Work & Health

Research Excellence Advancing Employee Health

Editors: Cindy Moser, Uyen Vu
Layout: Kristina Buccat

The Institute for Work & Health is an independent, not-for-profit organization that promotes, protects and improves the safety and health of working people by conducting actionable research that is valued by employers, workers and policymakers.

The Institute for Work & health operates with the support of the Province of Ontario. The views expressed in this publication are those of the Institute and do not necessarily reflect those of the Province of Ontario

© 2017, Institute for Work & Health

This document is licensed under a Creative Commons Attribution-NonCommercialNoDerivatives 4.0 International License: http://creativecommons.org/licenses/by-nc-nd/4.0/ That means this document can be used and shared as long as IWH is credited as the source, the contents are not modified, and the contents are used for non-commercial purposes. If you wish to modify and/or use the contents for commercial purposes, please contact ip@iwh.on.ca.

Please direct questions and reprint requests to:

Institute for Work & Health
481 University Avenue, Suite 800
Toronto, Ontario M5G 2E9
416-927-2027 | info@iwh.on.ca
www.iwh.on.ca

For Sheilah Hogg-Johnson,
the Institute for Work & Health's
chief biostatistician for 25
years (1992 - 2017)

What researchers mean by...

Easy-to-understand definitions of common research terms in the health and social sciences

Note: Although some of the examples used in this booklet to help illustrate the terms come from real research, most examples are fictional. As well, many of the examples come from the world of work-place injury prevention, disability management and workers' compensation, because that is the field in which the various authors worked in during the writing of these term explanations.

Foreword

From 2005 to 2017, the Institute for Work & Health (IWH), a not-for-profit research organization based in Toronto, Canada (www.iwh.on.ca), ran a column called "What researchers mean by ...." in its quarterly newsletter *At Work*. The aim of the column was to explain research terms in plain language so that non-scientists can more easily understand the methods and findings of IWH researchers and others in the health and social sciences.

Little did the Institute know when it launched this series that it would turn out to be so popular. Some of the terms defined in WRMB (the acronym commonly used to refer to the column) have received hundreds of thousands of website visits from around the world over the years. It's heartening to see such a thirst for knowledge about research terms.

This booklet pulls together in one place most of the terms covered in the WRMB column over the past 10-plus years.

These columns were written by various people in their role as *At Work* editor, knowledge transfer & exchange professional or communications manager at the Institute. They include: Anita Dubey, Kathy Knowles-Chapeskie, Evelyne Michaels, Cindy Moser, Megan Mueller, Rhoda Reardon, Katherine Russo and Uyen Vu.

What these people all had in common was the guidance of Dr. Sheilah Hogg-Johnson, a biostatistician and senior scientist who joined the Institute in 1992 and left 25 years later upon her retirement in 2017. The column (and this booklet) would not have been possible without her expert feedback and patient explanation of research terms.

Absolute and relative risk

Absolute risk is the number of people experiencing an event in relation to the population at large. Relative risk is a comparison between two groups of people or in the same group of people over time. Knowing which type of risk is being reported is important in understanding the magnitude of the risk.

The media often mentions risk when reporting on research, but this can sometimes be misleading. For example, if a newspaper article reports on research that shows a certain gene puts people at an 800 per cent increased risk of getting a blood clot, and you have that gene, you would likely be very worried reading this news. But should you be? Understanding how risk is expressed can help determine a study's significance, or a person's chance of illness, injury or recovery. Risk can be explained in terms of absolute or relative risk. Here's a look at the difference between these terms.

Absolute risk

Let's say a study of 100 workers in factory A revealed that 20 workers experienced back pain on the job. In factory B, 30 workers in a similar workplace of 150 workers developed back pain. The absolute risk of developing back pain is simply the percentage of people affected. This is 20 per cent in both groups. In scientific terms, absolute risk is the number of people experiencing an event in relation to the population at risk.

Relative risk

Relative risk is a comparison between two groups of people, or in the same group of people over time. It can be expressed as a ratio. In the example above, the relative risk of developing back pain — comparing factory A and factory B — is 20:20 or one. That is, workers in factory A are no more (or less) likely to have back pain than workers in factory B. It's 20 per cent for both groups.

Now suppose workers in factory A were to receive exercise therapy for half an hour each day. One year later, we find that only eight of 100 workers have back pain, while the rate in factory B remains the same at 20 per cent.

The ratio now changes to 8:20. Eight is the risk per 100 workers in factory A. Twenty is the risk per 100 workers in factory B. If we divide eight by 20, this gives us 0.40, or 40 per cent. In other words, the relative risk of developing back pain in factory A is now 40 per cent of the risk in factory B.

Risk reduction

How much did the risk of back pain change due to the exercise therapy intervention? Again, this can be calculated two ways, using absolute and relative risk reduction.

Absolute risk reduction is the difference in the percentage of people who are affected. Again, recall that before the intervention, 20 per cent of workers in factory A developed back pain. Afterwards, eight per cent did. The difference is 12. Therefore, the intervention resulted in an absolute risk reduction of 12 per cent.

The relative risk reduction is the change in relative risk. Recall that before the intervention, the relative risk was one for both factory A and B. After the intervention, it dropped to 0.40. The difference is 0.60. In other words, the intervention resulted in a 60 per cent reduction in relative risk.

Which is better?

Risk expressed either way is correct. In our example, the relative risk reduction of 60 per cent appears larger than the absolute risk reduction of 12 per cent. It often helps to look at both types of risk to see how significant a change is.

For example, say the absolute risk of a work injury is two per 100 workers. Due to an intervention, it drops to one injury per 100 workers. This yields a relative risk reduction of 50 per cent. Overall, in absolute terms, this means one less injured worker per 100.

In another case, say the absolute risk of injury is 50 per 100 workers, but drops to 25 injuries per 100 workers. This will also result in relative risk reduction of 50 per cent. However, this translates to 25 fewer injured workers per 100. Even though the relative risk reduction is the same in both cases, the second intervention has a greater impact overall.

Let's go back to the example of the newspaper article on the risk of clotting due to the presence of a certain gene. The article reported on the relative risk; i.e. you are 800 per cent more likely to get a blood clot relative to those who don't have the gene. However, knowing the absolute risk is important. If the absolute risk of getting a blood clot is one in 1,000, and you are at an 800 per cent increased risk of getting a blood clot because of the presence of the gene, your risk is now eight in 1,000. So there's still a very very good chance you won't get a blood clot at all.

Bias

Bias refers to flaws in the design, conduct and analysis of research that can (usually unintentionally) creep into a study and skew the findings. If researchers can't limit the possibility of bias, they should at least report how it might impact their results.

A study often begins with a simple question. Researchers are motivated to find answers to the question and add to the overall knowledge on a topic.

However, once they publish their findings, you might hear other researchers say that they are sceptical of the results because they may be biased. What exactly are these researchers concerned about and why?

In a research sense, bias does not refer to an intentional attempt to mislead. Rather, it refers to flaws in the design, conduct and analysis that creep into the study that results in a systematic shift in the findings.

Bias can be introduced at any stage of research — from the initial stages when researchers are collecting data, to the analysis of results, to the publication of studies. Bias can also refer to things that happen before the study has started (for example, the construction of questions to include is often biased by previous research); or things that happen (or don't) when the study ends — such as publication bias (see below).

Here are some common forms of bias that occur.

Selection bias

Suppose you want to examine what young people think about their risk of getting injured at work. Ideally, you would ask this question to a random group of young workers. However, due to the difficulty in finding young workers, you select only those young workers who visit a young workers' safety website. Your selection would be biased. It is likely that this group has a better knowledge of workplace safety, or is more concerned about getting injured at work because they have visited a website with safety information. Based on the response from this group you might conclude that young people think their risk of getting injured at work is high. But because of the selection bias, this finding might be higher than the actual views of young workers in general. (See page 38 for more on selection bias)

Attrition bias

Often researchers are concerned about how conditions — such as unemployment — affect people over time. You might have a large, diverse sample of workers from the population at the start of your study. Let's say you want to see how stress levels are related to unemployment, so you survey these workers. However, over time those people who are unemployed for a long period might move, perhaps to find work elsewhere. As a result, they might not be included in a follow-up measurement a year later. Because these workers are no longer in your study, it may impact on your results. This is called attrition bias.

Measurement bias

Sometimes it's difficult for researchers to measure what they plan to. They might use a proxy or substitute for what they really want to measure. For instance, it might be difficult for researchers to go into a company and ask to measure workplace injuries, so they might use the company's lost-time work injury claims as a proxy for workplace injuries. In this situation, researchers might end up with a less accurate measure that may lead to different results.

Analysis bias

Researchers may conduct an analysis that does not consider or adjust for another potential explanation for the findings. One example would be an analysis of young workers' injury risk that does not account for how long they've worked or for the hazards in their workplace. Inexperience in general or high-hazard working conditions can also affect the risk of injury.

Publication bias

This is a type of bias in which researchers only submit studies with results that they think are likely to be published in scientific journals. It can also occur when editors of these journals accept or reject articles for publication based on the direction or strength of the findings. For instance, a study that shows an intervention works might be selected over a study that shows it has no effect.

Bias can occur in almost any study, although researchers first try and limit the possibility of bias. However, sometimes this is not possible, so the researcher's job is to better understand and report how the bias they encountered might impact on their results.

Blinding

Blinding is a practice whereby study participants are prevented from knowing certain information that may somehow influence them and, in turn, affect the study's results.

If you've done a taste test and selected 'Cola X' over 'Cola Y,' then you've already experienced what scientists call "blinding."

Blinding, in research, refers to a practice where study participants are prevented from knowing certain information that may somehow influence them—thereby tainting the results. Coke versus Pepsi taste trials are conducted in this way: Participants are, literally, blindfolded as they sample the two colas and indicate their preference.

Blinding (also called masking) is typically used in randomized controlled trials (RCTs). In RCTs, people are randomly assigned to two (or more) groups. One group receives the intervention, such as a new treatment, while the control group receives nothing, usual care or a placebo—a fake treatment, an inactive substance like sugar, distilled water or saline solution— when the treatment is a new drug. The researchers then study what happens to each group. Any differences in outcome can then be linked to the intervention, not to the participants' knowledge of whether they were receiving a new treatment or their usual care.

To ensure to the highest degree possible that the intervention is responsible for any noted differences between the two groups, people involved in gathering or analyzing the data might also be blinded to knowing who is being given the treatment and who is not. This blinding can include clinicians, data collectors, outcome assessors and data analysts. However, certain groups sometimes cannot be blinded, such as surgeons or psychologists who provide active intervention.

Why blinding is necessary

Blinding of one or more parties is done to prevent observer bias. This refers to the fact that most (if not all) researchers will have some expectations regarding the effectiveness of an intervention. Blinding of observers provides a strategy to minimize this form of bias. For example, a clinician who has established expertise in a certain procedure may believe that his or her approach is superior. If involved in a trial to explore this procedure, the clinician may tend to treat patients assigned to his or her procedure differently than patients assigned to the competing intervention.

Blinding is also done to address or control for the placebo effect, a phenomenon in which a simulated (and ineffective) treatment can sometimes improve a patient's condition, simply because the person has the expectation that it will be beneficial. Expectation is key in the placebo effect.

Landmark study: an example of blinding

In 2002, a study published in the *New England Journal of Medicine* reported on a controlled trial of arthroscopic surgery for osteoarthritis of the knee. Arthroscopic surgery is the most commonly performed type of orthopedic surgery. In this study by Moseley et al., patients with osteoarthritis—defined as a group of mechanical abnormalities involving the degradation of joints—were divided into two groups: one receiving corrective surgery (arthroscopic debridement), and the other receiving fake or sham surgery.

The patients were blinded in the sense that they did not know whether they were receiving the real or sham surgery. The results were quite surprising: Both groups of patients improved equally well regardless of whether or not they received the real surgery. This is an excellent example of the placebo effect and the need for blinding, since it implies that belief of recovery alone can have an effect, even on a mechanical knee problem.

Case control study

Case control studies start with an outcome (such as a disease) and work backwards to find exposures that may be linked to it.

Let's say your mother was recently diagnosed with breast cancer and so, too, was her best friend. The two worked together for 30 years at the town's food canning plant. You wonder if something in the workplace was the cause of their cancer.

Researchers can help find answers to this type of question using a case control study. This study design helps determine if a previous exposure is linked to a current condition, such as having a disease.

A case control study compares people who already have a condition or disease (these are the cases) with people who do not have the condition or disease but are otherwise similar (these are the controls). It then looks back to see if an exposure to something in particular (e.g. at work, in the environment, lifestyle) was more likely in the group with the condition than in the group without.

Not all studies with cases and controls are case control studies. Some studies start with a group of people with a known exposure and a comparison group (the control group) without the exposure and follow them forward to see what happens. This is the case with some cohort studies.

Case control studies are always retrospective; they always look back. The outcomes are always known—the cases do have the condition and the controls do not—and the researchers trace backwards to identify possible exposures or factors that may have contributed to the condition.

Case control study in action

Let's take our example of breast cancer and work to show how a case control study might provide some answers. The researchers begin by turning to the regional cancer treatment centre to find women within the town and the surrounding area who developed a new case of breast cancer during a six-year period and are willing to participate in the study. The researchers identify 1,000 women, the cases.

The researchers then select controls. With computer-generated phone numbers, homes are randomly called to find women in the region without breast cancer of about the same age who are willing to take part in the study. They find 1,150 women, the controls.

Both cases and controls are asked about their personal, lifestyle and reproductive pasts, including information about factors known to be associated with breast cancer (e.g. body mass index, drinking, smoking, menstrual and menopause history, use of hormone replacement therapy, birth control, family history). They are also asked about the jobs they've had over the years and for how long. The researchers take this job information to code occupation, industry and exposure, allowing them to figure out likely exposures to cancer-causing materials and endocrine disruptors (i.e. chemicals that interfere with the hormone system).

By comparing the two groups, the researchers find that, taking the other risk factors into account, the women with breast cancer are more likely to have worked in certain occupations, including food canning. Although the study cannot say that your mother and her best friend's breast cancer was caused by work—case control studies cannot show causation—it does indicate that their breast cancer may be linked to their work.

Case control studies have a number of drawbacks. They cannot show causation, as mentioned; nor can they provide information on incidence (e.g. what percentage of people have a condition). As well, the information collected can be faulty or incomplete because it depends on people accurately and truthfully recalling their past.

Nonetheless, case control studies are relatively quick, inexpensive and easy. Thus, they are often used to conduct preliminary investigations of suspected risk factors. If a link is found, a more costly study that starts with a group of people and follows them forward may be justified.

Cohort study

A cohort study follows a group of people over time to understand the relationship between some attribute shared by the group of people at the beginning of the study and the eventual outcome.

Ever wonder why some injured workers return to work (RTW) after six months while others do so after a year or more? A cohort study that follows and observes a group of people who have something in common (namely, a workplace injury) could help answer this question.

A "cohort" is any group of people with a shared characteristic. For example, in a birth cohort, what's common to all individuals is their birth year.

In a cohort study, the study participants are followed over time—from weeks to years, depending on the time frame. The goal is to understand the relationship between some attribute related to the cohort at the beginning of the study and the eventual outcome.

There are five steps in a cohort study:

1. Identify the study subjects; i.e. the cohort population.
2. Obtain baseline data on the exposure; measure the exposure at the start. (The exposure may be a particular event, a permanent state or a reversible state.)
3. Select a sub-classification of the cohort—the unexposed control cohort—to be the comparison group.
4. Follow up; measure the outcomes using records, interviews or examinations. (Note: Outcomes must be defined in advance and should be specific and measurable.)
5. Do the data analysis where the outcomes are assessed and compared.

Cohort study in action

Returning to our example, a cohort study could follow a group of injured workers who were off work (and filed musculoskeletal-related claims) and observe when these workers returned to work.

Researchers in such a study could determine what's affecting the workers' RTW. At six and 12 months post-injury, the workers could be interviewed about their readiness to RTW. They may be asked if they have returned to work and, if so, if they were able to meet their job demands. They might be asked about their organization's policies and practices, and if accommodated work had been offered and accepted.

It may come to light that the workers who felt their companies were doing well in terms of policies and practices were more likely to be back at work at six months, for example, than those who didn't. If this were the case, this cohort study could likely tell us that workplace policies play an important role in RTW. Researchers could use these results to develop a tool to identify readiness for RTW and guidelines surrounding successful RTW.

Strengths of a cohort study include the fact that multiple outcomes can be observed. Weaknesses are that they can be expensive and time-consuming because they can involve large populations and long periods of time.

In terms of levels of evidence for establishing relationships between exposure and outcome, cohort studies are considered second to randomized controlled trials (RCTs) because RCTs limit the possibility for biases by randomly assigning one group of participants to an intervention/treatment and another group to non-intervention/treatment or placebo. Cohort studies are observational—meaning the researcher observes what's happening or naturally occurring, measures variables of interest and draws conclusions. RCTs, in contrast, are experimental—meaning the researcher manipulates one of the variables (assigns treatments, for example) and determines how this influences the outcome.

If cohort studies are second-best, then why use them? They may be the only way to explore certain questions. For example, it would be unethical to design an RCT deliberately exposing workers to a potentially harmful situation.

Confidence intervals

A confidence interval is the range of values above and below a finding in which the actual value is likely to fall. It represents the accuracy or precision of an estimate.

Imagine that you are trying to find out how many Canadians have taken at least two weeks of vacation in the past year. You could ask every Canadian about his or her vacation schedule to get the answer, but this would be expensive and time consuming.

To save time and money, you would probably survey a smaller group of Canadians. However, your finding may be different from the actual value if you had surveyed the whole population. That is, it would be an estimate. Each time you repeat the survey, you would likely get slightly different results.

Commonly, when researchers present this type of estimate, they will put a confidence interval (CI) around it. The CI is a range of values, above and below a finding, in which the actual value is likely to fall. The confidence interval represents the accuracy or precision of an estimate.

How confidence intervals are used

We often see CIs in newspapers when the results of polls are released. An example from the *Globe and Mail* newspaper regarding the mayoral race in Toronto read, "52 per cent [of survey respondents] said they would have voted for Mr. Miller if the election had been held last week. The margin of error is plus or minus 4.4 percentage points, 19 times out of 20."

The "margin of error" represents the confidence interval. It is the range from 47.6 to 56.4 per cent; that is, 52 per cent plus or minus 4.4 percentage points. The researchers are confident that if other surveys had been done, then 95 per cent of the time — or 19 times out of 20 — the findings would fall in this range.

The 95 per cent confidence level is used most often in research; it is a generally accepted standard. However, researchers can calculate CIs at any level of significance, such as 90 per cent or 99 per cent. The significance level simply indicates how precise they are willing to be.

Factors influencing a confidence interval

A narrow or small confidence interval indicates that if we were to ask the same question of a different sample, we are reasonably sure we would get a similar result. A wide confidence interval indicates that we are less sure and perhaps information needs to be collected from a larger number of people to increase our confidence.

Confidence intervals are influenced by the number of people that are being surveyed. Typically, larger surveys will produce estimates with smaller confidence intervals compared to smaller surveys. Other factors will include the accuracy of the measurements in a survey. If measurements are less accurate, it will likely increase confidence intervals.

Why are confidence intervals important?

Because confidence intervals represent the range of scores that are likely if we were to repeat the survey, they are important to consider when generalizing results. In the example with Mr. Miller, how confident would you be in saying that more than half of Torontonians would vote for Miller?

If you repeated the survey again, you may get a value of 47.6 per cent, which lies within your 95 per cent CI. Therefore, you may not be comfortable with such a statement. On the other hand, you would likely be more confident saying that at least 45 per cent of voters will cast their vote for Miller.

Confounding variables

A confounding variable is an unforeseen or unaccounted-for factor that may call into question the finding of a relationship between two other factors or variables. In other words, it "confounds" the relationship by being the "something else" that may explain the relationship.

Are workers who wear supportive back belts on the job less prone to back strain compared to those who don't? Before researchers design a study to answer this question, they must carefully consider all the variables that could affect their findings. If they fail to do so, the results of their study might not be valid.

Let's say a study found that, over a 12-month period, one group of lumber-yard workers who wore back belts had half the rate of back strain compared to another group of workers who didn't wear the belts. (In this case, wearing the belts is what researchers call the "independent variable," while the occurrence of back strain is the "dependent variable.")

Based on this finding, it would be tempting to recommend that all lumberyard workers protect themselves from back strain by wearing supportive belts. But are the study results valid? Was one group of workers protected by the independent variable — their use of back belts — or was something else going on?

The "something else" would be a confounding variable, defined as "an unforeseen and unaccounted-for variable that jeopardizes the reliability and validity of an experiment's outcome."

Before designing their study, the researchers should have known that the two groups of workers — who were employed in different lumberyards — didn't do the same amount of heavy lifting. One lumberyard typically used forklifts to load and deliver orders by truck, while the workers at the other location were sometimes expected to load orders into the customers' vehicles. So this variable — the amount of lifting — rather than back belt use could explain the different rates of back strain in the two groups.

Variables that might introduce errors

When researchers design a study or interpret data, they must make every effort to account for variables that might introduce errors into the results. These include participant variables like age, gender and education, situational variables — some aspect of the task or environment — or even temporary variables like hunger or fatigue that might influence what happens during the study.

It's important to understand that while many such variables exist, they are not necessarily confounding in each and every study. Also, it would be impossible for researchers to control for every possible confounding variable. In the real world, they try to control only those variables that might be relevant to the outcome.

One way researchers try to avoid confounding variables is to use a randomized experiment design. With randomization, all the background characteristics should be similar in the groups being studied, which minimizes the influence of confounding factors.

In the back belt study, they might have observed or surveyed the workers at both lumberyards to determine how much lifting they actually did and then designed the study comparing the effects of back belt use in two more similar groups of workers. Researchers can also use a number of analytic and statistical strategies such as stratified analysis and multivariate analysis to control for certain variables and thus protect the validity of their findings.

Cross-sectional vs. longitudinal studies

Cross-sectional studies make comparisons at a single point in time, whereas longitudinal studies make comparisons over time. The research question will determine which approach is best.

Study design depends greatly on the nature of the research question. In other words, knowing what kind of information the study should collect is a first step in determining how the study will be carried out (also known as the methodology).

Let's say we want to investigate the relationship between daily walking and cholesterol levels in the body. One of the first things we'd have to determine is the type of study that will tell us the most about that relationship. Do we want to compare cholesterol levels among different populations of walkers and non-walkers at the same point in time? Or, do we want to measure cholesterol levels in a single population of daily walkers over an extended period of time?

The first approach is typical of a cross-sectional study. The second requires a longitudinal study. To make our choice, we need to know more about the benefits and purpose of each study type.

Cross-sectional study

Both the cross-sectional and the longitudinal studies are observational studies. This means that researchers record information about their subjects without manipulating the study environment. In our study, we would simply measure the cholesterol levels of daily walkers and non-walkers along with any other characteristics that might be of interest to us. We would not influence non-walkers to take up that activity, or advise daily walkers to modify their behaviour. In short, we'd try not to interfere.

The defining feature of a cross-sectional study is that it can compare different population groups at a single point in time. Think of it in terms of taking a snapshot. Findings are drawn from whatever fits into the frame.

To return to our example, we might choose to measure cholesterol levels in daily walkers across two age groups, over 40 and under 40, and compare these to cholesterol levels among non-walkers in the same age groups. We might even create subgroups for gender. However, we would not consider past or future cholesterol levels, for these would fall outside the frame. We would look only at cholesterol levels at one point in time.

The benefit of a cross-sectional study design is that it allows researchers to compare many different variables at the same time. We could, for example, look at age, gender, income and educational level in relation to walking and cholesterol levels, with little or no additional cost.

However, cross-sectional studies may not provide definite information about cause-and-effect relationships. This is because such studies offer a snapshot of a single moment in time; they do not consider what happens before or after the snapshot is taken. Therefore, we can't know for sure if our daily walkers had low cholesterol levels before taking up their exercise regimes, or if the behaviour of daily walking helped to reduce cholesterol levels that previously were high.

Longitudinal study

A longitudinal study, like a cross-sectional one, is observational. So, once again, researchers do not interfere with their subjects. However, in a longitudinal study, researchers conduct several observations of the same subjects over a period of time, sometimes lasting many years.

The benefit of a longitudinal study is that researchers are able to detect developments or changes in the characteristics of the target population at both the group and the individual level. The key here is that longitudinal studies extend beyond a single moment in time. As a result, they can establish sequences of events.

To return to our example, we might choose to look at the change in cholesterol levels among women over 40 who walk daily for a period of 20 years. The longitudinal study design would account for cholesterol levels at the onset of a walking regime and as the walking behaviour continued over time. Therefore, a longitudinal study is more likely to suggest cause-and-effect relationships than a cross-sectional study by virtue of its scope.

In general, the research should drive the design. But sometimes, the progression of the research helps determine which design is most appropriate. Cross-sectional studies can be done more quickly than longitudinal studies. That's why researchers might start with a cross-sectional study to first establish whether there are links or associations between certain variables. Then they would set up a longitudinal study to study cause and effect.

Difference in differences

A method of analysis called "difference in differences" helps identify the effect of an intervention when intervention and control groups have meaningful differences.

Experimental studies are typically designed so that researchers can learn about the impact of an intervention (a drug, a therapy or a program). They do this by looking for different outcomes between the group that received the intervention (the intervention group) and the group that did not (the control group).

But what if the people in both groups start out with important differences to begin with? That's when researchers use a method of analysis called difference in differences to identify the effect of the intervention.

In controlled settings such as a randomized controlled trial, study participants are randomly placed in either the intervention group or the control group. That step helps make sure that the groups start out relatively the same so that changes in the intervention group can be more easily attributed to the intervention. In natural experiments (or observational studies), researchers don't have this ability to randomly assign participants.

That's because, in natural experiments, the interventions happen naturally, as the name would suggest. For example, a study of a school board policy that requires all school students to be vaccinated, of a province's policy to cut a cheque to everyone who lives in it so no one lives below a certain level of income, or of a town council decision to make helmets mandatory for all cyclists would all be natural experiments.

When such policies or programs are offered in one school board, one province or one town but not others, they offer researchers a valuable opportunity to study the impact of the intervention. But in natural experiments such as these, participants may start out with important differences; i.e. the people in the school board, province or town subject to the policy or program may already be different in some meaningful way from those with whom they are being compared. To overcome this, researchers don't compare one group's outcomes to those of the other. Instead, they look for how much each group changes over a period of time with respect to a certain outcome. Then they compare the extent of the change between the two groups.

An example of difference in differences

Let's take the helmet bylaw as an example. If you as a researcher want to look at the effect of that bylaw—introduced by Town A, let's say—you might hypothesize that it reduces head injuries. As a result, you take a close look at stats from emergency rooms to see whether head injuries from cycling accidents have gone down. For a control group, you look at similar stats in a neighbouring town of the same size—Town B—where a mandatory helmet bylaw does not exist.

But you know there may be prior differences between Town A and Town B. They may differ in road and traffic conditions or in how willingly people wear helmets when cycling, whether required by law or not. As a result, you don't simply look at the two towns' post-intervention stats—the number of head injuries one year after the bylaw took effect, for example—and draw a conclusion based on those two numbers. Rather, you also look at head injury stats prior to the bylaw in both towns. If head injury stats in Town A go down by 25 per cent but only by 15 per cent in Town B, you attribute that 10-per-cent difference to the effect of the bylaw.

This approach has some limitations. One is the possibility that you might be seeing regression to the mean. That would be the case if pre-bylaw injury stats in Town A were extreme or exceptional to begin with. If so, there's a strong statistical likelihood that the extreme injury rates seen at that point in time would naturally decline towards a lower average.

Another caveat to this method is that it assumes injury trends for both towns would have been the same if not for the intervention. Even if you gathered data at multiple points in time to make sure that the trends were the same leading up to the new bylaw, you have to be alert to the possibility that something else might be taking place to change that trend during the period of your study.

DOI

A DOI (or digital object identifier) is a permanent name given to studies, publications and other Internet resources to ensure a permanent link to an electronic article even when its' URL has changed.

If you read published research, you've probably noticed a vehicle for permanently housing scholarly material: the DOI or Digital Object Identifier. An alphanumeric code, it solves a lot of problems for anyone searching for documents in the vast arena of cyberspace.

A DOI is a permanent name given to documents, publications and other resources on the Internet, which is used rather than a URL (i.e. a typical web address). A URL can change over time but a DOI cannot. The International DOI Foundation, which invented and controls the system, defines a DOI as "a name, not a location, for an entity on digital networks."

Because a DOI is meant to never change, it provides a permanent link to any electronic article. Most electronically available articles have DOIs, and they can usually be found printed on the article itself. DOIs look something like this: **doi:10.1111/j.1439-0426.1997.tb00116.x**

DOIs are not dissimilar to a book's ISBN—that is, the idea of having a number associated with a document is not new. But for libraries, the change is meaningful because it makes things easier. DOIs solve a lot of problems and allow those in library sciences to locate and verify electronic documents quickly and efficiently. They allow librarians to focus and provide a unique identifier to others who would like to locate specific documents.

DOIs are particularly helpful for several reasons:

- URLs are not stable and often disappear;
- print journals often have standard bibliographic information (volume, issue, page numbers) that help track down articles, but electronic journals and documents may not;
- researchers sometimes use titles for conferences, lay articles and reports that are very similar to those used for journal articles, and it can be difficult to differentiate between them when searching by title; and
- Google-type searches can lead to hundreds of hits, making it hard to locate and verify documents.

DOI adoption has been rapid. The International DOI Foundation was established in 1998. Elsevier, the Amsterdam-based health and science, started using DOIs on all of its journal articles around 2003. By late April 2011, more than 50 million DOI names had been assigned by some 4,000 organizations.

However, unlike URLs, the DOI system is not open to everyone. Only organizations that meet the necessary contractual obligations and are willing to pay can assign DOIs.

Although some journals are not yet participating in the move to DOI, it is expected that they may do so in time, as part of the larger electronic continuum.

How to find an article using a DOI

When you see a DOI, most of the time, you can click on it to access the article, provided you have the necessary access rights. In case you see a DOI in a print document, you can do the following three steps:

1. Copy the DOI of the document you want to open.
2. Go to: www.doi.org.
3. Enter the entire DOI in the text box provided, and then click 'Go.'

Otherwise you can type the DOI into a search engine, such as Google, and the relevant study usually comes up.

Epidemiology

The cornerstone of public health, epidemiology investigates which groups in a population are affected by disease, and why.

If you've ever wondered whether vegetarians live longer than meat-eaters, or why some people suffer from chronic pain and others don't, or what the health consequences are of working nights, you're asking the same questions asked by epidemiologists—researchers who work in the field of epidemiology.

Epidemiology is considered the basic science of public health. In simple terms, it's the study of who gets sick and why. "Epidemiology" literally means "the study of what is upon the people." The word comes from the Greek epi, meaning "upon," demos, meaning "people," and logos, meaning "study.

In the early days, epidemiology concentrated on studying diseases such as cholera. Today, epidemiology is applied to all kinds of health-related conditions—diseases (e.g. influenza, cancer, depression), health problems (e.g. obesity, high blood pressure), injuries (e.g. work-related, traffic-related) and social problems (e.g. gambling, domestic violence). Its role is to describe who is affected by these conditions, why, and what can be done to treat and prevent them.

Population versus individual

A distinguishing feature of epidemiology is that it studies health-related conditions at the population level, as opposed to the individual level. A good way to understand this is to compare the differing approaches of clinicians and epidemiologists to diseases.

Doctors and other clinicians are largely concerned with the effects of disease within a single person. They work one-on-one with patients to diagnose problems and determine what can be done to make them healthier.

Epidemiologists, on the other hand, are concerned with how diseases affect society as a whole. They study groups of people to diagnose and respond to illnesses in populations: how many are affected (i.e. prevalence), who is affected and why (i.e. determinants of health), and what works and what doesn't to cure or prevent these illnesses at a societal level (e.g. treatment protocols, public health interventions).

Let's look more closely at how epidemiologists carry out their studies of disease and other conditions. To understand the "who," epidemiologists seek to describe what part of the population is affected. How does the prevalence of a disease vary by age, sex, ethnicity, income, geography, work role and so on? This analysis goes well beyond demographics. It might relate to genetic disposition, childhood exposure, living conditions and more.

Difficult to find cause

Understanding who gets sick is often the first step in learning what factors might be behind why people get sick. Sometimes, epidemiologists rely on other fields of science to get to the "why." They might learn from geneticists that certain types of people are predisposed to an illness. That might then lead them to probe more deeply about other factors that might protect certain individuals within that group from the disease.

Although epidemiologists seek to understand the why, they rarely get to say "because." Researchers must clear many hurdles before they can pronounce the cause of a health outcome. How strong is the association between event A and outcome B? Does A always occur before B? Does B always follow A? If A is altered in some way, is B altered too, and to the same degree? The more researchers can say yes to these questions, the closer they get to being able to claim A is the cause of B.

These criteria for causation should give you an idea why epidemiological studies are so difficult to carry out. They're also why epidemiologists are often so circumspect when stating the findings of their research.

Epidemiological studies are important. They form the bedrock for sound public health policies and strategies, thus protecting and improving the health of entire populations.

Generalizability

Generalizability refers to the degree to which the results of a study can be applied to a larger population, or the degree to which time- and place-specific findings, taken together, can result in a universal theory.

The goal of scientific research is to increase our understanding of the world around us. To do this, researchers study different groups of people or populations. These populations can be as small as a few individuals from one workplace or as large as thousands of people representing a cross-section of society. But how do we know if a study's results can be applied to another group or population?

To answer this question, we first need to understand the concept of generalizability. In its simplest form, generalizability can be described as making predictions based on past observations.

In other words, if something has often happened in the past, it will likely occur in the future. In studies, once researchers have collected enough data to support a hypothesis, they can develop a premise to predict the outcome in similar circumstances with a certain degree of accuracy.

Two aspects of generalizability

Generalizing to a population. Sometimes when scientists talk about generalizability, they are applying results from a study sample to the larger population from which the sample was selected. For instance, consider the question, "What percentage of the Canadian population supports the Liberal party?" In this case, it would be important for researchers to survey people who represent the population at large. Therefore they must ensure that the survey respondents include relevant groups from the larger population in the correct proportions. Examples of relevant groups could be based on race, gender or age group.

Generalizing to a theory. More broadly, the concept of generalizability deals with moving from observations to scientific theories or hypotheses. This type of generalization amounts to taking time- and place-specific observations to create a universal hypothesis or theory. For instance, in the 1940s and 1950s, British researchers Richard Doll and Bradford Hill found that 647 out of 649 lung cancer patients in London hospitals were smokers. This led to many more research studies, with increasing sample sizes, with differing groups of people, with differing amounts of smoking and so on. When the results were found to be consistent across person, time and place, the observations were generalized into a theory: "cigarette smoking causes lung cancer."

Requirements for generalizability

For generalizability we require a study sample that represents some population of interest — but we also need to understand the contexts in which the studies are done and how those might influence the results.

Suppose you read an article about a Swedish study of a new exercise program for male workers with back pain. The study was performed on male workers from fitness centres. Researchers compared two approaches. Half of the participants got a pamphlet on exercise from their therapist, and half were put on an exercise program led by a former Olympic athlete. The study findings showed that workers in the exercise group returned to work more quickly than workers who received the pamphlet.

Assuming the study was well conducted, with a strong design and rigorous reporting, we can trust the results. But to what populations could you generalize these results?

Some factors that need to be considered include: How important is it to have an Olympian delivering the exercise program? Would the exercise program work if delivered by an unknown therapist? Would the program work if delivered by the same Olympian but in a country where he or she is not well-known? Would the results apply to employees of other workplaces that differ from fitness centres? Would women respond the same way to the exercise program?

To increase our confidence in the generalizability of the study, it would have to be repeated with the same exercise program but with different providers in different settings (either worksites or countries) and yield the same results.

Grey literature

Documents and other information that haven't gone through peer review before being published are referred to as "grey literature." Magazine articles and conference proceedings, for example, fall in this category.

If you were a busy practitioner seeking information on managing back pain, where would you turn: a blog by a person describing her experiences, a fact sheet from a reputable hospital, a research study in a scientific journal or a tabloid newspaper article?

We all apply a level of trust to information based on the source and the quality we associate with it. Plus, time demands and our access to information or our ability to understand it can also influence what we choose.

Scientists generally place the most trust in information published in journals that use the peer-review process. "Peer review" means that each study submitted to a journal is sent by its editors to two or three other experts in that field. These experts, or peers, provide an anonymous critique with a view to improve the write-up of the study. If they don't think the study meets certain scientific standards, they might advise against publishing it at all. Peer review helps to maintain scientific standards.

Practitioners in workplaces may not have access to peer-reviewed journals, or the time or expertise to wade through scientific text. They're more likely to turn to other sources of information that they trust. Examples could be trade publications, government reports, survey results from a polling company or technical reports.

These documents are all considered "grey literature." The term grey literature comes from the uncertainty of the status of this information. Although there are several formal definitions, grey literature is essentially any document that hasn't gone through peer review for a publication. It can also include conference proceedings or doctoral theses.

Challenges with grey literature

When IWH reviewers conduct systematic reviews on a topic, they search for studies on that topic in peer-reviewed journals. We've found that practitioners who are consulted during reviews sometimes ask us to include the grey literature as well, to make sure that the search is as comprehensive as possible.

One concern of reviewers is the scientific quality of the studies. If an article doesn't go through peer review, it's possible for the author to make claims or interpretations that aren't supported.

Until recently, it has been more difficult to systematically search the grey literature than peer-reviewed studies. These documents often aren't indexed (or catalogued) in the major databases that are typically and systematically searched in reviews. It usually requires extra effort to find and get copies of these documents.

The format of a grey literature document can be quite diverse, unlike scientific papers that follow the structure of presenting background information, study methods, results and a discussion. So it's more difficult for reviewers to systematically extract information from grey literature as they do for peer-reviewed papers.

Benefits of grey literature

Grey literature documents can provide a richer source of detail than a scientific study. Because they aren't tied to a conventional structure, they can be longer and provide more detail. Research results can be written in a style that is more accessible and useful to a practitioner than a scientific paper.

Grey literature can also be published more quickly since it does not have to go through the potentially lengthy peer-review process. And in cases where there isn't much information on a topic in peer-reviewed research, grey literature may provide a valuable source of information.

Finally, grey literature is becoming easier to find. Increasingly, it is available on the Internet, and search engines and databases are providing ways of locating it.

Grey literature can provide a systematic review with an additional source of rich information, depending on the topic and the nature of the research. The challenges and benefits need to be weighed against each other when deciding on whether to include it in a systematic review.

Grounded theory

If you're a grounded theorist, you engage a 'zig-zag' approach to research—jumping from the field to the drawing table, then back again—in an ever-changing process of fine-tuning your findings. Grounded theory is all about having an open mind and seeing where the data take you.

Traditionally, scientists collect information to test a potential explanation or assumption. For example, let's say you are studying the role of supervisors in the return to work (RTW) of injured workers. Based on existing research, you might hypothesize that supervisors facilitate RTW in an important way, and then subsequently design a survey that asks workers about the role of supervisors to test this hypothesis.

Grounded theory, used in qualitative research, takes a different approach. First coined in the 1960s, it was an alternative to the mainstream approach in which information was collected to test a theory. Grounded theory emphasizes starting from the ground up (i.e. generating theory from data) rather than from the top down (i.e. using data to test theory). In other words, it favours an inductive approach, rather than a deductive one.

Let's return to our example. Taking the grounded theory approach, you might enter into the RTW study with similar ideas about the role of supervisor support, but you would remain open to other theories stemming from the data you collect. You might learn something wholly unexpected.

Theoretical sampling

You would start by carefully selecting the people you want to interview ("cases") and the types of workplaces you want to observe ("settings"), with the aim of getting the richest possible information. Your research plan might involve interviews or focus groups with injured workers who have and have not returned to work, in addition to supervisors and co-workers. As well, different types of workplaces, from blue- to white-collar environments, may be included in the sample. This is called theoretical sampling.

Constant comparative method

Next, you would constantly compare the information you gather with what is already known, and refine your explanations or theories as you go. This is called the constant comparative method and it is central to grounded theory. For example, you might compare supervisor/worker relationships across different jobs and types of workplaces.

Data might emerge that indicate supervisors are supportive when worker absences are brief, but not as supportive when the absences get longer. In the end, you may learn that supervisors play a relatively minor role compared to co-workers. This new knowledge would cause you to reconsider your previous understanding.

Grounded theory can take researchers in new and fruitful directions because it involves an interactive process where the overarching goal is to test and refine emerging ideas. It's easy to see how it can broaden the reach of an existing theory because it forces the researcher to change the scope of the study to incorporate new information. As such, grounded theory generates a high quality of research, revealing multi-layered interpretations of social life. A rich and detailed understanding of systems and processes is made possible.

Internal validity

Internal validity ensures a study's findings are the result of the intervention being studied and not due to chance or some other factor. In that sense, internal validity indicates how well a study was designed and carried out to prevent systematic errors or bias.

A key aspect of the quality of a study is its internal validity. Internal validity, in essence, is whether the study's findings result from the intervention being studied, and are not due to chance or some other factor. You could also say that internal validity is how well the study was set up and executed to prevent systematic errors or bias.

Let's take a fictional example to see how this plays out. Suppose researchers wanted to study the effectiveness of an ergonomics program that included staff training. The program was targeted at garment workers, who often experience wrist pain. In the study, the workers in one factory completed a test of their knowledge of postures to prevent wrist pain. Then an ergonomics program and training were introduced. Six months later, fewer workers reported pain symptoms and, when tested again, their scores were better.

At face value, this sounds like a promising program. But in reality, something else could have caused these changes. A study with strong internal validity would be set up in a way that ruled out other explanations.

Questions to consider

The review team uses a detailed list of questions to ensure the researchers have considered these other causes and minimized bias. Here are some things the reviewers would be looking at:

- Did the researchers use a control group of workers who didn't participate in the program? A control group provides a way for researchers to see if the program led to the changes, as they can check whether any changes occurred in the control group.
- What else was happening in the workplace that might explain the results? For instance, suppose a staff ergonomist was hired after the program began. This might account for the improvements and would need to be considered.
- Was it possible that workers, over time, became more knowledgeable about preventing injuries on their own?
- Did completing the first knowledge test affect results the second time around?
- Were the workers given the same test, in the same way, both times?
- Who dropped out of the study before it ended? Maybe some workers withdrew because their pain symptoms weren't getting better. Any improvements in pain in workers remaining in the study wouldn't reflect the whole truth. The researchers need to look at the reasons that people dropped out, to see if this is an issue.
- How were workers chosen to participate in the study? The researchers need to report on how they selected the groups, and the differences between groups. If the workers who did the program volunteered, they may be more highly motivated and it would affect the findings.
- What was the average rate of reported pain before the program? Suppose the factory's management agreed to the program because in the previous year, reports of pain and work absences increased dramatically, far above the average rate each year. However, these rates may fluctuate naturally, from year to year. So the improvement may just mean the rate is coming back to the average.

Internal validity is also influenced by the way that people naturally interact. For instance, if workers in the control group found out about the program, they might try to do something similar themselves. Or, management may decide that having a control group is creating too many problems among employees, and may allow these workers to access the program or create a new one for them.

All of these scenarios show how difficult it can be to do research in workplaces. They also show how important it is to have a well-designed study when you're trying to find out if a program really works.

Overall, the higher the internal validity, the better the quality of the study. And the more sure we are that the results are due to the program, and not due to something else.

Mean, median and mode

Related to numbers-based findings, 'mean' is the average, 'median' is the number that separates the higher half from the lower half, and 'mode' is the value that occurs most often.

The game of golf can help to explain the often-misused terms of mean, median and mode. Let's say you golfed nine holes. Each number below represents the number of swings it took you to sink the ball in the hole. If you're lucky and you have some golf skills, your score is the following:

8, 4, 10, 4, 4, 5, 4, 5, 6

You go back into the clubhouse and are quite pleased with your score. You run into your friend and he says that his mean score was 6, his median was 7 and his mode was 6. So what does that mean (no pun intended)? Did you score better than your friend? Well, let's find out.

Mean

Let's define the term "mean" as it's the most common term of the three and probably the easiest to explain. Basically, the mean – which is also called the average – is the sum of all numbers divided by the number of values in the list. In your golf score, you would add up all of the numbers (which equals to 50) then divide it by 9 (the number of values) and you get 5.5.

Median

Now, let's examine median. Basically, the median is the number that separates the higher half of a sample from the lower half. To find the median, arrange the list from lowest value to highest value and pick the middle one. Using the golf scores, here is the list from lowest to highest. The bolded 5 is the median:

4, 4, 4, 4, 5, 5, 6, 8, 10

When to use mean or median

Sometimes, you need to decide if calculating the mean or median is most appropriate for what you would like determine. Hospital length of stay can be an example of data that may be skewed if the wrong term is chosen (that is, when most of the data values fall to the left or right of the mean). Most people stay in a hospital for a few days. However, some patients have hospital stays for months on end. In this example, you would likely report the median length of hospital stay, which separates the higher half from the lower half. In general, however, most people report the mean unless you have a good reason for not doing so, such as to avoid skewing the data like in the hospital example above.

Mode

While not used as frequently as mean or median, mode does have a place in certain situations. Mode is the value that occurs most frequently in a set. If you look at your golf scores, 4 is the one that's most common so, for that set, 4 is the mode. Although mode may not frequently be used in statistics, mode is more often used when describing non-numerical things. For example, if you'd like to know the most popular newborn boy name in Ontario for 2008, you may go to the Government of Ontario's website and find out that Jacob was the most popular.

You can remember mode the following way: Mode is the value that is in the set Most often.

So getting back to our golf scores example, it looks like that you likely shot a better golf score than your friend given that you had a better mean, median and mode.

Comparing the terms

Type	Description	Example	Result
Mean	Total sum divided by number of values	(8+4+10+4+4+5+4+5+6)/9	5.5
Median	Middle value that separates higher half from lower half	4, 4, 4, 4, 5, 5, 6, 8, 10	5
Mode	Most frequent number	4, 4, 4, 4, 5, 5, 6, 8, 10	4

Meta-analysis

A type of systematic review, meta-analysis integrates or adds the findings from many studies to create one large overview. By combining results, it reduces the time and energy spent looking at the difference pieces of research.

When making decisions that affect many people, policy-makers, clinicians and other decision-makers may turn to research to help inform their choices. Single studies on a topic do provide some information. However, to increase confidence in their decisions, it is better to look at all of the available research.

This is where a meta-analysis can help. A meta-analysis is a type of systematic review. In a meta-analysis, findings from many studies are integrated or "added" in a formal statistical analysis to create one large overview.

The steps of a meta-analysis are:

- define a narrow, focused question that the meta-analysis will seek to answer.
- define and follow rigorous criteria for identifying and selecting studies to include in the analysis.
- collect the data from these studies, and convert estimates or results into a common measure across studies, if possible.
- combine and analyze the data, and develop conclusions to answer the question.

In general, a meta-analysis aims to answer the questions: What is the effect of a program or treatment, based on all the relevant research to date? How large is the effect?

Meta-analysis in practice

Let's say you wanted to know if rest breaks reduced the rate of low-back pain in a particular work setting. If you gathered all the research on rest breaks and low-back pain, you might find hundreds of research articles.

You may also find studies so small that you wouldn't be confident about the findings. Various articles might seem to contradict each other, with some showing that rest breaks reduced low-back pain rates, and others finding they had no effect.

As explained earlier, in a meta-analysis these findings or outcomes would be statistically combined to provide an overall answer. But first, they need to be converted into a common measure to reach any conclusions, and this can be difficult. With low-back pain, different studies might measure back pain in workers using different scales or questionnaires. Some additional calculations would be needed to achieve a common measure.

In some cases, outcomes are routinely based on a common measure. For example, in cancer research, one widely used outcome is patients' survival rates five years after diagnosis. When many different studies use this common outcome, their results are easier to combine.

For a meta-analysis on rest breaks and back pain, the reviewer might take study findings using different low-back pain scales and calculate a standard "effect" for each study. This "effect" becomes the common measure. By statistically combining the effects from all studies, reviewers may see if there is an overall effect from rest breaks, and how large the effect is. However, the reality is that different studies on a topic may not even measure the same outcome, and there might not be a way to make all the results comparable.

Let's now compare how conclusions are expressed in meta-analysis and other systematic reviews. In the example above, a systematic review may show that six out of eight quality studies show that rest breaks reduce the rate of low-back pain. Using a meta-analysis, which integrates the effect from all the studies, you might find that the numerical size of this effect is very low.

Benefits of meta-analysis

A meta-analysis has many benefits. By combining results into one large study, it reduces the time and energy that decision-makers spend looking at research.

But the real benefit lies in the way meta-analysis can make sense of inconclusive and conflicting data from each original study. Through meta-analysis, researchers can combine smaller studies, essentially making them into one big study, which may help show an effect. Additionally, a meta-analysis can help increase the accuracy of the results. This is also because it is, in effect, increasing the size of the study.

By helping to bring into focus the sometimes blurry picture developing from the abundance of research evidence on any given topic, a meta-analysis is a very effective type of review.

Missing data

Research data may have holes for a number of reasons — from questions left blank on a survey to people dropping out of a study. Sometimes the missing information matters; sometimes it doesn't.

In a researcher's perfect world, everyone asked to participate in a study would say yes, no one would drop out along the way, and all items on a questionnaire would be answered.

Alas, researchers don't operate in a perfect world. The information they collect often has holes, and they come up against a common challenge in their pursuit of answers to research questions: missing data.

Data can be missing for a number of reasons. Study participants may not answer a certain survey question because they don't see how the question applies to them. This would be the case if a survey asked, "In what year did you get married?" and the respondent is single. Or study participants may simply refuse to answer. This sometimes occurs in response to the question, "What is your income?"

Data are also called "missing" when people decline to take part in a study or drop out along the way. Take, for example, a study looking at return to work among injured workers. Some injured workers may not take part because they don't want to "make waves," or may later withdraw from a study because their situation has changed (e.g. they feel better or worse, or have competing demands on their time).

Some studies are based not on information collected by the researcher, but on administrative information collected by a public agency. This data, too, can be incomplete. For example, studies relying on claims data from a workers' compensation board may find that a claim file is missing information — say, on marital status or employment start-date — that is relevant to a study.

Does missing data matter?

Missing data may or may not be a problem. Most important is whether or not the data are missing at random. If there is a pattern to the missing information, then drawing wrong conclusions is much more likely. For example, the results of a workers' compensation study could be skewed if those who refuse to take part largely come from a vulnerable group like recent immigrants, or if most of those who drop out do so because they have recovered. In other words, information that is consistently missing from the members of one group is a problem.

The impact of missing data also depends in part on the research question. If a study is looking at the relationship between health and socioeconomic status — as measured by income — then missing income information could be an issue. This is especially the case if people refuse to provide the information because of what their incomes are (e.g. on the high and low ends of the scale). If a study is looking at the relationship between health and marital status, then missing income information may not be important.

How much information is missing is also a factor. If two per cent is missing, then sound conclusions are likely still possible. The same can't be said if 20 per cent is missing.

What can be done about missing data?

Researchers don't necessarily call it quits when information is missing. They deal with the problem in a number of ways. Indeed, whole books have been written about missing data in the field of statistical analysis.

Researchers might simply discard any record (e.g. questionnaire or claim file) that is missing information. Or they might "fill in" the missing data using what are called "imputation," weighting or model-based procedures. These procedures are complicated. Each has its place, and none is perfect. Therefore, researchers need to be very clear in the "limitations" section of their studies about what information is missing and how that may affect results.

Multiple regression

Multiple regression is a popular technique in statistics used to measure the relationship between many variables and an outcome.

In another entry in this book, we talk about the term simple regression (see p. 39) – a statistical method used to describe the relationship between two factors. We ask you to take on the role of a researcher for a real estate agency trying to find a way to accurately price clients' homes based on house size. Using simple regression, you come up with an equation to do so. However, you don't advise the real estate agency to price clients' homes based on house size alone. You know other factors also affect selling price. This is where multiple regression comes in.

Instead of looking at a one-to-one relationship, multiple regression looks at a one-to-many relationship. It is a statistical technique that allows researchers to examine the relationship between two or more factors (called independent variables) at the same time and analyze the extent to which each predicts or explains variations in the outcome of interest (called the dependent variable). The end result is a model (which, in essence, is a mathematical formula) that can be used to explain or predict outcomes based on the presence of different factors.

Main steps in multiple regression

Multiple regression analysis is hard. It's an elaborate process, involving many steps and usually requiring sophisticated software. Let's go back to our example to take a look at some of the main steps in doing a multiple regression—most of them preparatory to ensure you are feeding the best information into the software program.

1. Determine the independent variables you want to include in your model. These variables need to make sense. Drawing on your understanding of the real estate market, you decide to include house size, neighbourhood average income, proximity to good schools, lot size, and number of bedrooms and bathrooms.
2. Collect information on each of the variables. You now randomly select, say, 100 houses that recently sold in the city. For each, you collect information on its size, neighbourhood income, proximity to good schools, lot size, number of bedrooms and bathrooms and, of course, its selling price.
3. Explore the relationship between each independent variable being considered and the dependent variable. Using the information collected, you look at the relationship between house size and house price, average neighbourhood income and price, proximity to good schools and price, and so on. You use statistical techniques to determine if a clear (i.e. statistically significant) relationship exists between the factor and house price. If yes, you are more likely to keep the factor in your model. If not, you may or may not decide to use it depending on the nature of the problem you are trying to address.
4. Explore the relationship among the independent variables. Using the same methods above, you may decide to look at how the different factors relate to each other; e.g. between house size and neighbourhood income, neighbourhood income and proximity to good schools, and so on. You may find two factors are so closely related that it would be hard to tell which is contributing to differences in house prices. This is called "multi-collinearity." Again, depending on the nature of the problem you are trying to address, you may or may not decide to keep both factors. You may also decide to look at how each factor relates to house price taking the other factors into account and, if the factor is no longer related, you may decide to remove it from your model.
5. Perform the multiple regression. For the factors you've included in your model, you enter the related information into your software program, do a lot of other statistical prep work (to take into account errors, deviations and so on), then run your program. You end up with an equation that lets you answer questions like: To what extent do each of the factors (neighbourhood income, proximity to good schools, lot size, number of bedrooms and bathrooms) account for variations in home price? What is the predicted price of a particular home knowing the value of all the variables in the model? Multiple regression lets you answer these questions and more. That's why it's a powerful tool.

Observational vs. experimental studies

Observational studies observe the effect of an intervention without trying to change who is or isn't exposed to it, while experimental studies introduce an intervention and study its effects. The type of study conducted depends on the question to be answered.

When people read about a research study, they may not pay attention to how the study was designed. But to understand the quality of the findings, it's important to know a bit about study design.

According to the widely-accepted hierarchy of evidence, the most reliable evidence comes from systematic reviews, followed by evidence from randomized controlled trials, cohort studies and then case control studies.

The latter three are research studies that fall into one of two main categories: observational studies or experimental studies.

Observational studies

Observational studies are ones where researchers observe the effect of a risk factor, diagnostic test, treatment or other intervention without trying to change who is or isn't exposed to it. Cohort studies and case control studies are two types of observational studies.

Cohort study: For research purposes, a cohort is any group of people who are linked in some way. For instance, a birth cohort includes all people born within a given time frame. Researchers compare what happens to members of the cohort that have been exposed to a particular variable to what happens to the other members who have not been exposed.

Case control study: Here researchers identify people with an existing health problem ("cases") and a similar group without the problem ("controls") and then compare them with respect to exposure.

Experimental studies

Experimental studies are ones where researchers introduce an intervention and study the effects. Experimental studies are usually randomized, meaning the subjects are grouped by chance.

Randomized controlled trial (RCT): Eligible people are randomly assigned to two or more groups. One group receives the intervention (such as a new drug) while the control group receives nothing or an inactive placebo. The researchers then study what happens to people in each group. Any difference in outcomes can then be linked to the intervention.

Strengths and weaknesses

The strengths and weaknesses of a study design should be seen in light of the kind of question the study sets out to answer. Sometimes, observational studies are the only way researchers can explore certain questions. For example, it would be unethical to design a randomized controlled trial deliberately exposing workers to a potentially harmful situation. If a health problem is a rare condition, a case control study (which begins with the existing cases) may be the most efficient way to identify potential causes. Or, if little is known about how a problem develops over time, a cohort study may be the best design.

However, the results of observational studies are, by their nature, open to dispute. They run the risk of containing confounding biases. Example: A cohort study might find that people who meditated regularly were less prone to heart disease than those who didn't. But the link may be explained by the fact that people who meditate also exercise more and follow healthier diets. In other words, although a cohort is defined by one common characteristic or exposure, they may also share other characteristics that affect the outcome.

The RCT is still considered the "gold standard" for producing reliable evidence because little is left to chance. But there's a growing realization that such research is not perfect, and that many questions simply can't be studied using this approach. Such research is time-consuming and expensive — it may take years before results are available. Also, intervention research is often restricted by how many participants researchers can manage or how long participants can be expected to live in controlled conditions. As a result, an RCT would not be the right kind of study to pick up on outcomes that take a long time to appear or that are expected to affect a very minute number of people.

Path analysis

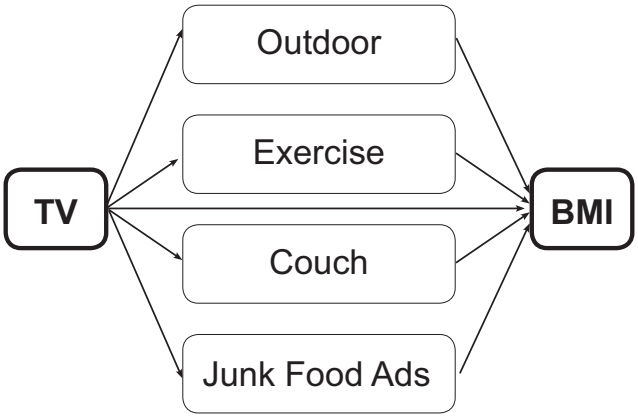
In path analysis, researchers use models to map out relationships between many variables and test them for strength.

Let's look at the link between watching TV and obesity. As a researcher, how might you learn more precisely how that link works?

You might want to find out whether watching TV affects body mass index (BMI) directly, or whether it affects something else first (e.g. less time spent on exercising, which in turn affects BMI)? Does it affect several other things first, which in turn affect BMI (e.g. less exercising and more exposure to junk food ads)? If several other factors are involved, which of them have more impact than others?

To answer these types of questions, researchers use a method called path analysis to test out the many different ways one thing can affect another. Real-world cause-and-effect relationships are complicated. Path analysis helps researchers measure which of the possible relationships matter the most, and which might turn out to be not important at all.

In a path analysis, you would take the factors (called variables) that might explain what is happening and map them out in a path model. Using our TV and obesity example, your model might look like this:



Determining what variables to include in the model is your job as a researcher. You'd have to comb through the literature to identify the variables that might play a role. For example, research showing a link between less time exercising and higher BMI would be reason to include exercise as a factor in your model.

Sometimes not much research is available to help. You might then decide to turn to focus groups to help you identify probable pathways.

If the literature on TV watching was scant, for example, you might learn from focus group participants that they hardly go outside or they sit on the couch all the time when watching TV, and that these might be the reasons higher obesity rates are seen among TV watchers.

Testing the model

Once a model is drawn up, the heavy-lifting work of testing the model begins. This is where you would examine available data to find out how well they support your model. To do that, you would run statistical analyses (usually what is known as "regression analysis"; see www.iwh.on.ca/wrmb/regression) to measure the statistical strength of each pathway.

For example, the data might show that increased TV watching has a strong association with less time exercising, and less time exercising has a strong association with higher BMI. The strength of both relationships indicates that exercise time is an important factor through which TV watching affects BMI. (Researchers sometimes use the term mediating to describe this indirect relationship, one in which a variable acts through another variable—referred to as the mediating variable—to have an impact on something else.)

The data might point to variables in the model that aren't all that important. For example, you might find a stronger relationship between TV watching and the number of junk food ads people see, but a weaker relationship between the junk food ads people see and BMI. That relationship may be so weak that you decide to drop it altogether from your model.

While statistics can help test your pathway model, they won't protect you from faulty models. For example, you might find a link between outdoor time and TV time, but neglect to consider that outdoor time might be exerting an impact on TV time instead of the other way around. In a path model, nothing indicates the direction of causality.

Similarly, if important variables are missing from the model, statistics alone might not alert you to that omission. In other words, a model might fit the data, but not necessarily fit reality.

Primary data and secondary data

Primary data and secondary data are two types of data, each with pros and cons, each requiring different kinds of skills and resources.

What does each and every research project need to get results? Data – or information – to help answer questions, understand a specific issue or test a hypothesis.

Researchers in the health and social sciences can obtain their data by getting it directly from the subjects they're interested in. This data they collect is called primary data. Another type of data that may help researchers is the data that has already been gathered by someone else. This is called secondary data.

What are the advantages of using these two types of data? Which tends to take longer to process and which is more expensive? This column will help to explain the differences between primary and secondary data.

Primary data

An advantage of using primary data is that researchers are collecting information for the specific purposes of their study. In essence, the questions the researchers ask are tailored to elicit the data that will help them with their study. Researchers collect the data themselves, using surveys, interviews and direct observations.

In the field of workplace health research, for example, direct observations may involve a researcher watching people at work. The researcher could count and code the number of times she sees practices or behaviours relevant to her interest–e.g. instances of improper lifting posture or the number of hostile or disrespectful interactions workers engage in with clients and customers over a period of time.

To take another example, let's say a research team wants to find out about workers' experiences in return to work after a work-related injury. Part of the research may involve interviewing workers by telephone about how long they were off work and about their experiences with the return-to-work process. The workers' answers–considered primary data–will provide the researchers with specific information about the return-to-work process; e.g. they may learn about the frequency of work accommodation offers, and the reasons some workers refused such offers.

Secondary data

There are several types of secondary data. They can include information from the national population census and other government information collected by Statistics Canada. One type of secondary data that's used increasingly is administrative data. This term refers to data that is collected routinely as part of the day-to-day operations of an organization, institution or agency. There are any number of examples: motor vehicle registrations, hospital intake and discharge records, workers' compensation claims records, and more.

Compared to primary data, secondary data tends to be readily available and inexpensive to obtain. In addition, administrative data tends to have large samples, because the data collection is comprehensive and routine. What's more, administrative data (and many types of secondary data) are collected over a long period. That allows researchers to detect change over time.

Going back to the return-to-work study mentioned above, the researchers could also examine secondary data in addition to the information provided by their primary data (i.e. survey results). They could look at workers' compensation lost-time claims data to determine the amount of time workers were receiving wage replacement benefits. With a combination of these two data sources, the researchers may be able to determine which factors predict a shorter work absence among injured workers. This information could then help improve return to work for other injured workers.

The type of data researchers choose can depend on many things including the research question, their budget, their skills and available resources. Based on these and other factors, they may choose to use primary data, secondary data–or both.

Primary, secondary and tertiary prevention

Primary, secondary and tertiary prevention are three terms that map out the range of interventions available to health experts.

Prevention includes a wide range of activities — known as "interventions" — aimed at reducing risks or threats to health. You may have heard researchers and health experts talk about three categories of prevention: primary, secondary and tertiary. What do they mean by these terms?

Primary prevention

Primary prevention aims to prevent disease or injury before it ever occurs. This is done by preventing exposures to hazards that cause disease or injury, altering unhealthy or unsafe behaviours that can lead to disease or injury, and increasing resistance to disease or injury should exposure occur. Examples include:

- legislation and enforcement to ban or control the use of hazardous products (e.g. asbestos) or to mandate safe and healthy practices (e.g. use of seatbelts and bike helmets)
- education about healthy and safe habits (e.g. eating well, exercising regularly, not smoking)
- immunization against infectious diseases.

Secondary prevention

Secondary prevention aims to reduce the impact of a disease or injury that has already occurred. This is done by detecting and treating disease or injury as soon as possible to halt or slow its progress, encouraging personal strategies to prevent reinjury or recurrence, and implementing programs to return people to their original health and function to prevent long-term problems. Examples include:

- regular exams and screening tests to detect disease in its earliest stages (e.g. mammograms to detect breast cancer)
- daily, low-dose aspirins and/or diet and exercise programs to prevent further heart attacks or strokes
- suitably modified work so injured or ill workers can return safely to their jobs.

Tertiary prevention

Tertiary prevention aims to soften the impact of an ongoing illness or injury that has lasting effects. This is done by helping people manage long-term, often-complex health problems and injuries (e.g. chronic diseases, permanent impairments) in order to improve as much as possible their ability to function, their quality of life and their life expectancy. Examples include:

- cardiac or stroke rehabilitation programs, chronic disease management programs (e.g. for diabetes, arthritis, depression, etc.)
- support groups that allow members to share strategies for living well
- vocational rehabilitation programs to retrain workers for new jobs when they have recovered as much as possible.

Going "upstream"

To help explain the difference, take this example. Let's say you are the mayor of a town near a swimming hole used by kids and adults alike. One summer, you learn that citizens are developing serious and persistent rashes after swimming as a result of a chemical irritant in the river. You decide to take action.

If you approach the company upstream that is discharging the chemical into the river and make it stop, you are engaging in primary prevention. You are removing the hazardous exposure and preventing rashes in the first place.

If you ask lifeguards to check swimmers as they get out of the river to look for signs of a rash that can then be treated right away, you are engaging in secondary prevention. You are not preventing rashes, but you are reducing their impact by treating them early on so swimmers can regain their health and go about their everyday lives as soon as possible.

If you set up programs and support groups that teach people how to live with their persistent rashes, you are engaging in tertiary prevention. You are not preventing rashes or dealing with them right away, but you are softening their impact by helping people live with their rashes as best as possible.

For many health problems, a combination of primary, secondary and tertiary interventions are needed to achieve a meaningful degree of prevention and protection. However, as this example shows, prevention experts say that the further "upstream" one is from a negative health outcome, the likelier it is that any intervention will be effective.

Probability

Probability provides information about the likelihood of something happening. In public health research, it looks at the likelihood of a health effect due to exposures to risk factors.

If the Weather Network informs you that the probability of precipitation is 80 per cent for the day, it might prompt you to use your umbrella.

We often use probability assessments informally in our daily lives to plan or make decisions. Formal probability theory is a fundamental tool used by researchers, health-care providers, insurance companies, stockbrokers and many others to make decisions in contexts of uncertainty.

Probability provides information about the likelihood that something will happen. Meteorologists, for instance, use weather patterns to predict the probability of rain. In epidemiology, probability theory is used to understand the relationship between exposures and the risk of health effects.

Let's start with a simple, classic example to illustrate probability: the toss of a coin. You know intuitively that there is a 50 per cent chance of getting heads, and 50 per cent chance of getting tails. If you want to actually do the math to calculate the probability of a head, here's the basic formula:

Count the number of times that the event will happen – in this case, there's just one chance of a head appearing, so it's 1. Divide this by the total number of possible outcomes. With a coin, it's either heads or tails – which is 2 outcomes. So the probability of getting heads is 1÷2, or 50 per cent.

Yet you could toss a coin 10 times and get seven heads and three tails, which is 70 per cent heads and 30 per cent tails. With this small number of repetitions, you can't determine the probability accurately. However, if you toss that coin 1,000 times or more – which a few people have done* – you will eventually begin to see that 50-50 breakdown.

This illustrates another important point about probability. It depends on the outcome or event happening over a large number of repetitions, or with a large number of people.

Use of probability in society

There are many examples of how probability is used throughout society. One common measure is the probability of developing cancer. According to the Canadian Cancer Society, 40 per cent of Canadian women and 45 per cent of men will have a diagnosis of an incident of cancer during their lifetimes. These probabilities are based on calculations from 2009 cancer statistics across the country.

While this broad information can be useful for those who plan, deliver or research health-care services, more detailed information is even more helpful. Researchers can also determine the probability of acquiring specific types of cancers at specific ages. They can also consider individual factors, which are important, too. If you have family members with breast cancer, your risk increases. If you smoke, your probability of getting lung cancer increases (smoking is estimated to account for between 88 and 90 per cent of lung cancer cases. The risk is significantly lower in never-smokers: about one per cent). These types of risk factors can be incorporated into probability calculations as well.

Another application of probability is with car insurance. Companies base your insurance premiums on your probability of having a car accident. To do this, they use information on the frequency of having a car accident by gender, age, type of car and number of kilometres driven each year to estimate an individual person's probability (or risk) of a motor vehicle accident.

Probability can fall anywhere from 0 to 1, where 1 means there's 100 per cent certainty that the event will occur. Zero means it will not.

So on a day in which the probability of precipitation was forecast at 80 per cent, but skies were sunny all day, you also have to consider that there was a 20 per cent chance that it wouldn't rain. Still, you made a wise decision to take an umbrella based on the probability you were given.

Psychometrics

Research on psychometrics examines the properties of a measure to ensure it's accurate, consistent and sensitive to change.

If you've ever taken part in a questionnaire—a political poll, a customer satisfaction survey or a research study—you might not have given much thought to the types of question you were asked, how they were worded or how many there were. But researchers spend a great deal of time thinking about and creating the questions used in a study. In fact, this is an entire field of research called psychometrics.

Psychometrics is the field of study that looks at the design, delivery and interpretation of tests that measure human responses. Typically, these tests measure our knowledge or abilities (e.g. an IQ test), our personality and behaviour (e.g. whether we're more introverted or extroverted) or our attitudes and beliefs (e.g. how we feel about our level of health or the support we get in our workplace).

In health research, for example, psychometric testing is used to create measures that assess pain, fatigue, distress, anxiety, alertness, mobility, agility—the list goes on. In organizational research, psychometric testing is used to create measures that assess worker, supervisor and organizational experiences and behaviours, such as job satisfaction, perceived job characteristics (e.g. job control, work overload), organizational commitment, job stress, job roles, work-family balance/conflict, leadership styles, person-organization fit, and so on.

Psychometrics uses mathematics and statistics, as well as lots of input from individuals to whom the measure is given, to ensure a measure works the way it's intended to. It makes certain the questions asked cover a range of possible perspectives and that they get enough detail without becoming too repetitive. It ensures the questions asked give rise to results that are valid, reliable and responsive.

Validity

Psychometrics assesses a tool's validity by looking for evidence that indicates the tool measures what we think it should. For example, we might think a measure asking people about how important physical activity is to them is only valid if those individuals who say physical activity matters actually exercise more than those people who say physical activity doesn't matter. We might think it isn't valid if there are important aspects of physical activity that the questionnaire fails to include. That would be a question about content validity, just one of many different types of validity to consider.

Reliability

A tool is assessed for its reliability by determining if people give consistent answers to questions when asked those same questions under similar circumstances. For example, in developing a measure on the commuting difficulties workers face, you would run statistical analyses to find out if the questions given to the same group of workers on different occasions (but close in time) produce roughly the same results. That's an example of test-retest reliability. Some measures ask others to rate or evaluate another person's physical or psychological behaviours or health. A measure would be considered reliable if different observers score the same way. That's an example of inter-rater reliability.

Responsiveness

And then there's the question of the tool's responsiveness. Psychometrics looks at its ability to measure meaningful change. That is, if a person's situation, skills or beliefs change, is the tool sensitive enough to detect this change, and how much change has to take place before the measure will detect it? For example, if a new workplace wellness program is introduced and the program is effective, can we capture changes using a health measure? What about if the change is small—is this just random error or is it meaningful and "real"?

There's a great deal to be discussed when creating, applying and evaluating the many different measures used in research. Hopefully, this summary gives you an appreciation of the effort that researchers put into designing a questionnaire.

Qualitative research

Qualitative research aims to make sense of human experience, beliefs and actions. As such, it provides a rich source of information on social systems and processes.

It's tempting to define "qualitative research" by what it is not. It is not based on statistics or surveys or experiments; that is, it is not quantitative research.

But it's also important to understand what qualitative research is – an approach used largely in the social sciences to explore social interactions, systems and processes. It provides an in-depth understanding of the ways people come to understand, act and manage their day-to-day situations in particular settings.

To put it simply, quantitative research uses numbers to help us understand "what" is happening. Qualitative research uses words and images to help us understand more about "why" and "how."

Compare, for example, two studies that are both addressing the issue of long-term workers' compensation claims. One uses quantitative methods to find out what is driving increases in the duration of lost-time claims over the last decade. Using administrative data from a workers' compensation board, the researchers test their hypotheses that claim duration may be associated with injury severity, a changing work environment or policy changes.

The other study uses qualitative methods to explore why and how some injured workers remain on workers' compensation for long periods of time. Based on interviews with injured workers and service providers, the study finds that workers with long-term claims often try hard to return to work but encounter many roadblocks beyond their control. These may include seemingly mundane problems such as incomplete medical forms and miscommunication among the workplace parties. Taken together, such challenges prevent workers' return to work.

How qualitative research is done

Qualitative research collects information that occurs naturally; that is, it doesn't set up experiments. The main methods for collecting research include:

- conducting interviews and focus groups, during which people retell their experiences, thoughts and actions;
- observing people in their own settings;
- analyzing documents (from government reports to personal diaries); and
- analyzing conversations (as contained in documents, speeches, interviews, etc.).

With this collected information, qualitative research can be used to:

- describe the nature of what exists and how it is experienced by those in it (i.e. context); e.g. help us understand the experience of having a long-term claim;
- explain why things exist as they do; e.g. help us understand the events leading to long-term claims, the circumstances in which long-term claims occur and why they continue to occur;
- evaluate the effectiveness of interventions that aim to change what exists; e.g. help us understand the quality of any programs put in place to reduce long-term claims; and
- generate suggestions for ways to improve things, or for potential areas of new research; e.g. help us understand strategies for supporting workers on long-term claims and helping people avoid them to begin with.

Qualitative versus quantitative

Qualitative and quantitative research are often discussed as two camps, with researchers belonging to one or the other. However, this us-versus-them scenario is quickly falling by the wayside. There is a growing understanding that the two types of research share much in common.

Both strive for reliability and validity of their data, and both have developed systematic methods of doing so. As well, both aim to produce results that can be generalized and practically applied to help understand and solve problems.

In fact, the two types of research can be complementary and part of the same "toolkit" when it comes to exploring an issue, as shown in the earlier example of research into long-term claims. The choice isn't about one being more accurate, more objective or more in-depth than the other, but about what information the researchers are trying to find out.

Randomized controlled trial

One of the most powerful research tools, the randomized controlled trial is considered by some to be the "gold standard" for generating reliable evidence.

In a researcher's toolkit, the randomized controlled trial (RCT) is one of the best ways to produce valid evidence on the effectiveness of interventions, from prevention programs to treatment options. According to the established hierarchy of evidence, the most valid evidence from original research comes from RCTs, followed by cohort studies and then case control studies.

Here's how RCTs work. Study participants are deemed eligible through a recruitment process that involves specific criteria for inclusion and an informed consent process.

Those eligible are randomly assigned, in a process that's not unlike flipping a coin, into one of two groups or 'arms' of the study: (1) the intervention group, or (2) the control group. The first group receives the intervention being studied, which could be a new treatment or procedure. The second does not, and instead receives an inactive placebo, conventional treatment or nothing at all.

The cornerstone of RCTs is this: Because the allocation process is random, it minimizes the chance that people who received treatment and those who did not had different characteristics. In other words, with random allocation, any differences in outcomes between the intervention group and the control group can be attributed to the intervention, as opposed to any of the participants' attributes like age or disease.

An RCT in action

Let's say you're a scientist interested in non-medicated pain relief for fibromyalgia. Does acupuncture help? An RCT has already been conducted to answer this question.*

In the recruitment phase of this study, the research team sought to enlist female patients between the ages of 20 and 70 years diagnosed with fibromyalgia according to the 1990 American College of Rheumatology classification criteria. To be included in the study, patients needed to have reported moderate to severe pain intensity and to be using antidepressants.

In the study, 58 women with fibromyalgia were allocated randomly to receive either: (1) acupuncture with tricyclic antidepressants and exercise, or (2) tricyclic antidepressants and exercise only. Patients rated their pain on a visual rating scale, and quality of life was also evaluated using a blinded assessor (i.e. the researcher assessing the results).

At the end of 20 sessions, patients in the RCT who received acupuncture had significantly less pain than the control group. This study concluded that the addition of acupuncture to usual treatments for fibromyalgia may be beneficial for pain and quality of life for three months after the end of treatment.

This conclusion would not have been possible without the use of an RCT. Its random allocation process is one of the best ways to secure valid evidence.

*See "A randomized controlled trial of accupuncture added to usual treatment for fibromyalgia" in the July 2008 issue of *Journal of Rehabilitation Medicine* (Vol 40, No. 7, pp. 583-588)

Regression to the mean

Regression to the mean is a statistical occurrence that may result in distorted or misleading findings if not taken into account.

Suppose you're the superintendent of a school district and you want to improve the math scores of the Grade 3 students in your catchment who write compulsory province-wide exams. You hire a consulting math expert to help. The consultant starts by administering a math test to find out which students are most in need.

All 1,000 Grade 3 students in your district take the test, and the consultant chooses the 50 students with the lowest scores to receive a remedial math program. Once the program is complete, the 50 students take a second test, and their scores, on average, show a healthy improvement. On this basis, you roll out the remedial program to all Grade 3 math students in the district who are performing below par.

When the board-wide exam takes place later that year, you're disappointed. The students' scores are not much better than they were the previous year—and they certainly didn't improve to the degree you expected based upon the results of the 50 poorest performing students.

What went wrong? You might want to consider the possibility of a statistical phenomenon called regression to the mean.

Regression to the mean refers to the tendency of results that are extreme by chance on first measurement—i.e. extremely higher or lower than average—to move closer to the average when measured a second time. Results subject to regression to the mean are those that can be influenced by an element of chance. When chance or fluke gives rise to extreme scores, it's unlikely those extreme scores will be repeated on a second try.

In our school district, for example, the kids who scored the poorest on the first math test likely included some who normally know the answers but, by chance, did not that day. Perhaps they were tired, sick, distracted, etc. These kids were going to do better on the second test whether they received the remedial program or not, bringing up the average score among the 50 poorest performers.

You can see why researchers have to consider regression to the mean when they are studying the effectiveness of a program or treatment. If they don't, they may wrongly conclude that their intervention is responsible for an improvement when, in fact, regression to the mean is at play. This is especially the case when program effectiveness is based on measurements of people or organizations at the extremes—the unhealthiest, the safest, the oldest, the smartest, the poorest performing, the least educated, the largest, etc. The ones on the low extremes are all likely to do better the second time around, and those on the top are likely to do worse—even without the intervention.

Steps to account for regression to mean

Researchers can take a number of steps to account for regression to the mean and avoid making incorrect conclusions. The best way is to remove the effect of regression to the mean during the design stage by conducting a randomized controlled trial (RCT). Because an RCT randomly assigns study participants to a study group (which receives the program or treatment) or a control group (which does not), the change in the control group provides an estimate of the change caused by regression to the mean (as well as any placebo effect). Any extra improvement or decline in the study group compared to the control group (as long as it is statistically significant) can be attributed to the effect of the program or treatment.

Researchers can also take multiple baseline measurements when selecting people or organizations to be part of a study group. They can then select participants based on the average of their multiple measurements, not just on a single test.

Scientists can also identify and account for regression to the mean when analyzing their results. This involves complicated statistical calculations too difficult to describe here.

Regression toward the mean is a statistical occurrence that can get in the way and distort researchers' measurements. That's why it has to be taken into account, in the design of the study or in the analysis of findings.

Sample size and power

Sample size refers to the number of participants or observations in a study. Power refers to the probability of finding a significant relationship. Often researchers begin a study by asking what sample size is necessary to produce a desirable power.

Few of us read research reports with an eye to critiquing the methodology. The results are the main attraction, the reason for reading in the first place. But researchers spend much of their time planning how their studies will be carried out. Shouldn't we pay more attention? As any decent researcher will tell you, a study's results are only as good as its design. Sample size and power are key elements of study design.

Why is sample size important?

Sample size refers to the number of participants or observations included in a study. This number is usually represented by n. The size of a sample influences two statistical properties: 1) the precision of our estimates and 2) the power of the study to draw conclusions.

To use an example, we might choose to compare the performance of marathon runners who eat oatmeal for breakfast to the performance of those who do not. Since it would be impossible to track the dietary habits of every marathon runner in the world, we have little choice but to focus on a segment of that larger population. This might mean randomly selecting only 100 runners for our study. The sample size, or n, in this scenario is 100.

The study's findings could describe the population of all runners based on the information obtained from the sample of 100 runners. No matter how careful we are about choosing our 100 runners, there will still be some margin of error in the study results. This is because we haven't talked to everyone in our population of interest. We can't be absolutely precise about how eating oatmeal affects running performance because it would be impossible to look at every instance in which these two activities coincide. This measure of error is known as sampling error. It influences the precision of our description of the population of all runners.

Sampling error, though unavoidable, can be eased by sample size. Larger samples tend to be associated with a smaller margin of error. This makes sense. To get an accurate picture of the effects of eating oatmeal on running performance, we need plenty of examples to look at and compare. However, there is a point at which increasing sample size no longer impacts the sampling error. This phenomenon is known as the law of diminishing returns.

What about power?

Clearly, determining the right sample size is crucial for strong experimental design. But what about power?

Power refers to the probability of finding a statistically significant result (read the column on statistical significance). In our study of marathon runners, power is the probability of finding a difference in running performance that is related to eating oatmeal.

We calculate power by specifying two alternative scenarios. The first, called the null hypothesis, is one that says there's nothing going on in the population of interest. In our study of marathoners, the null hypothesis might say that eating oatmeal has no effect on performance.

The second is the alternative hypothesis. This is the often anticipated outcome of the study. In our example, it might be that eating oatmeal results in consistently better performance.

The power equation uses these two alternatives so that the study can find the answer to the research question. As researchers, we want to know if our study of marathoners can detect the difference between oatmeal having no impact on running performance (the null hypothesis) and oatmeal having a considerable impact on running performance (the alternative hypothesis).

Often researchers will begin a study by asking what sample size is necessary to produce a desirable power. This process is known as a priori power analysis. It shows nicely how sample size and power are inter-related. A larger sample size gives more power.

While the particulars of calculating sample size and power are best left to the experts, even the most mathematically-challenged of us can benefit from understanding a little bit about study design. The next time you read a research report, take a look at the methodology. You never know. It just might change the way you read the results.

Sampling

Sampling is the process of identifying the representative part of a larger whole that will allow findings from the sample to be applied to the whole. It is one of the most challenging aspects of study design.

Sampling is an act of generalization that we participate in all the time. Consider the free samples at your local grocery store.

When a representative from the deli offers you a square of pizza, you are being asked to draw conclusions about the taste and value of the product itself. Offering a whole pizza to every customer would be expensive, difficult to coordinate and, in all likelihood, a waste of time and effort. Chosen well, the samples will provide customers with enough information to decide whether a whole pizza is worth purchasing. The sample is a representative part, an extract from which to generalize back to the whole.

Sampling: A scientific process

In practice, identifying a representative part of a subject, event or population of interest is one of the more challenging aspects of study design. Let's say we want to use an in-hospital survey to measure patient satisfaction. How will we select a group of patients to participate in our study?

To begin, we must differentiate between the theoretical population and the accessible population. The first might include any patient who has ever stayed in a hospital overnight. The second is limited to those who stayed in hospital on a specific night. Since we cannot hope to survey every member of the theoretical population, we must identify members of the accessible population to contact. The resulting subset of individuals will be our sampling frame.

However, we have to be cautious about introducing sampling errors and non-sampling errors into the frame. Sampling errors are the differences between the sample and the population being studied. In other words, they're errors that occur because the data is from a part rather than the whole. Non-sampling errors are statistical errors caused by human error. These can include data entry errors or biased questions in a survey. In our hospital survey, those who could not or did not respond to the survey could introduce non-sampling errors.

Probability sampling

Now that we've narrowed our population of interest, we must decide how to select the sample. Probability sampling is one of two primary strategies we might consider. In probability sampling, every member of the sampling frame has the potential to be selected for the study. Selection is random, and the probability of a member being chosen can be calculated. Knowing the probability of selection allows us to generalize to the population.

Non-probability sampling

In non-probability sampling, some members will have a greater chance of being selected than others, while some will have no chance of being selected at all. The probability of a member being chosen cannot be calculated, making it hard for researchers to know how well they have represented the theoretical population. Often researchers will turn to non-probability sampling only when other data collection methods are not possible.

Convenience sampling is a type of non-probability sampling, and it illustrates both the benefits and drawbacks of this approach. In convenience sampling, the most accessible members from the sampling frame are selected. For example, we might find that certain patients completed positive satisfaction surveys one year ago. It would be convenient to survey only those patients who already had a positive hospital experience. Probably they would be more willing to complete our survey. But in choosing only these patients, we must also ask whether it's reasonable to generalize from their experiences.

While all sampling methods are subject to error, researchers must always keep their objective in view: to obtan meaningful information about the theoretical population. Fundamental to this goal is a workable sample.

Selection bias

Selection bias is a common type of error where the decision about who to include in a study can throw findings into doubt.

Most scientific studies are designed to pinpoint the effect of something—such as the effect of a condition on developing a problem (disease, injury) or the effect of an intervention (treatment, program) on overcoming a problem. Scientists usually determine effect by taking two similar groups—the only difference being the groups' exposure to that condition or intervention—and measuring the difference in outcomes experienced by them.

But what happens when the two groups selected were not similar to begin with? What if key characteristics distinguishing the two might have played a role in producing the different outcomes? That's an example of what's called selection bias.

Bias is a type of error that systematically skews results in a certain direction. Selection bias is a kind of error that occurs when the researcher decides who is going to be studied. It is usually associated with research where the selection of participants isn't random (i.e. with observational studies such as cohort, case-control and cross-sectional studies).

For example, say you want to study the effects of working nights on the incidence of a certain health problem. You collect health information on a group of 9-to-5 workers and a group of workers doing the same kind of work, but at night. You then measure the rates at which members of both groups reported the health problem. You might conclude that night work is associated with an increase in that problem.

The trouble is, the two groups you studied may have been very different to begin with. The people who worked nights may have been less skilled, with fewer employment options. Their lower socioeconomic status would also be linked with more health risks—due to less healthy diets, less time and money for leisure activities and so on. So your finding may not be related to night work at all, but a reflection of the influence of socioeconomic status.

Selection bias also occurs when people volunteer for a study. Those who choose to join (i.e. who self-select into the study) may share a characteristic that makes them different from non-participants from the get-go. Let's say you want to assess a program for improving the eating habits of shift workers. You put up flyers where many work night shifts and invite them to participate. However, those who sign up may be very different from those who don't. They may be more health conscious to begin with, which is why they are interested in a program to improve eating habits.

If this was the case, it wouldn't be fair to conclude that the program was effective because the health of those who took part in the program was better than the health of those who did not. Due to self-selection, other factors may have affected the health of your study participants more than the program.

Minimizing selection bias

Good researchers will look for ways to overcome selection bias in their observational studies. They'll try to make their study representative by including as many people as possible. They will match the people in their study and control groups as closely as possible. They will "adjust" for factors that may affect outcomes. They will talk about selection bias in their reports, and recognize the degree to which their results may apply only to certain groups or in certain circumstances.

Another way researchers try to minimize selection bias is by conducting experimental studies, in which participants are randomly assigned to the study or control groups (i.e. randomized controlled studies or RCTs). However, selection bias can still occur in RCTs. For example, it may be that the pool of people being randomly assigned to the intervention group is not very representative of the wider population. Or it could be the researcher's allocation techniques aren't so random (e.g. when clinicians, often motivated by good intentions, manipulate the allocation method to get their patients in a treatment group instead of the control group).

Often, selection bias is unavoidable. That's why it's important for researchers to examine their study design for this type of bias and find ways to adjust for it, and to acknowledge it in their study report.

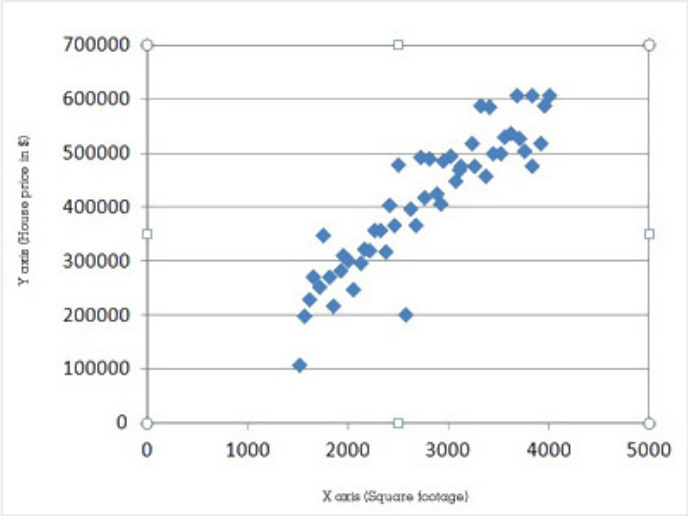
Simple regression

Simple regression helps researchers understand the relationship between two items, which can then be used to make predictions.

Suppose you are a researcher hired by a neighbourhood real estate agency, and your job is to help agents predict how much their clients' homes will sell for. One theory you keep hearing from the agents is that house prices are closely related to the size of the house. They believe they should be able to predict the price of the house based on its square footage.

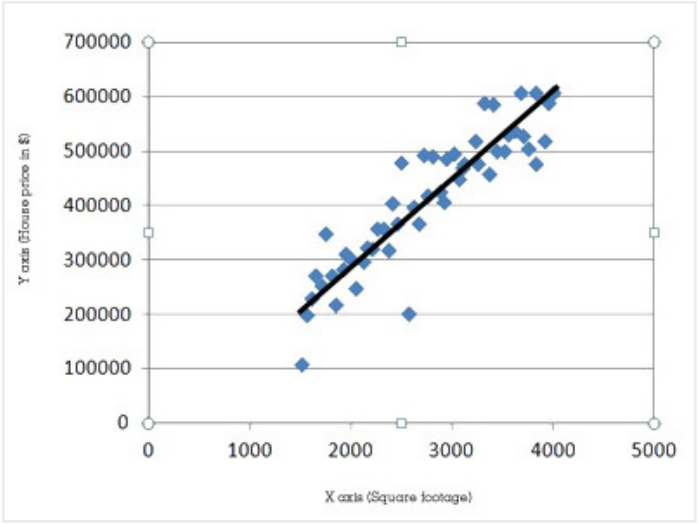
To test this theory, you would have to set up a study and use a common research technique called simple regression. This is a statistical method or tool that helps researchers understand the relationship between two items.

For your study, you first have to collect your data. You gather information on the homes that have been sold over the past year. For each house, you need to know its square footage and selling price. You then plot this information on a chart and create what is called a scatter plot (see below).



The square footage is shown along the horizontal line, which is referred to as the "X axis." The item that goes along this axis is called the independent or predictor variable because it is fixed. House price is shown on the vertical line or "Y axis." This is called the dependent or response variable because it is changeable. That is, the dependent variable (price of house) changes depending on the independent variable (size of house).

Now you conduct your simple regression. A simple regression, often calculated using a software program, creates an equation that best describes the relationship between the two things you looked at in your study or, in other words, best "fits" the dots on your scatter plot.



In this case, the simple regression shows you that the equation that best describes the relationship between house price and square footage based on the information you provided is y=150x. That is, the selling price of a house increases by $150 for every square foot increase in size. This equation is easily shown on a graph by a straight line, showing the "best fit" among all the dots on the scatter plot.This line or equation now becomes useful for predicting the selling price of a house. Knowing how big a client's house is, the real estate agent can predict how much it will sell for.

However, based on the simple regression, you wouldn't advise the real estate agent to price homes based only on their square footage. You suspect that other things besides house size might account for the price of the house and, therefore, need to be taken into consideration. That's where multiple regression comes in (see page 26).

Statistical significance

A statistically significant finding means that the differences observed in a study are likely real and not simply due to chance.

It's easy for non-scientists to misunderstand the term significant when they come across it in an article. In everyday English, the word means "important." But when researchers say the findings of a study were "statistically significant," they do not necessarily mean the findings are important.

Statistical significance refers to whether any differences observed between groups being studied are "real" or whether they are simply due to chance. These can be groups of workers who took part in a workplace health and safety intervention or groups of patients participating in a clinical trial.

Let's consider a study evaluating a new weight loss drug. Group A received the drug and lost an average of four kilograms (kg) in seven weeks. Group B didn't receive the drug but still lost an average of one kg over the same period. Did the drug produce this three-kg difference in weight loss? Or could it be that Group A lost more weight simply by chance?

Statistical testing starts off by assuming something impossible: that the two groups of people were exactly alike from the start. This means the average starting weight in each group was the same, and so were the proportions of lighter and heavier people.

Mathematical procedures are then used to examine differences in outcomes (weight loss) between the groups. The goal is to determine how likely it is that the observed difference — in this case, the three-kg difference in average weight loss — might have occurred by chance alone.

The "p" value

Now here's where it gets complicated. Scientists use the term "p" to describe the probability of observing such a large difference purely by chance in two groups of exactly-the-same people. In scientific studies, this is known as the "p-value."

If it is unlikely enough that the difference in outcomes occurred by chance alone, the difference is pronounced "statistically significant."

Mathematical probabilities like p-values range from 0 (no chance) to 1 (absolute certainty). So 0.5 means a 50 per cent chance and 0.05 means a 5 per cent chance.

In most sciences, results yielding a p-value of .05 are considered on the borderline of statistical significance. If the p-value is under .01, results are considered statistically significant and if it's below .005 they are considered highly statistically significant.

But how does this help us understand the meaning of statistical significance in a particular study? Let's go back to our weight loss study. If the results yield a p-value of .05, here is what the scientists are saying: "Assuming the two groups of people being compared were exactly the same from the start, there's a very good chance — 95 per cent — that the three-kg difference in weight loss would NOT be observed if the weight loss drug had no benefit whatsoever." From this finding, scientists would infer that the weight loss drug is indeed effective.

If you notice the p-value of a finding is .01 but prefer it expressed differently, just subtract the p-value from the number 1 (1 minus .01 equals .99). Thus a p-value of .01 means there is an excellent chance — 99 per cent — that the difference in outcomes would NOT be observed it the intervention had no benefit whatsoever.

Not all statistical testing is used to determine the effectiveness of interventions. Studies that seek associations — for example, whether new employees are more vulnerable to injury than experienced workers — also rely on mathematical testing to determine if an observation meets the standard for statistical significance.

Statistically adjusted

When determining the relationship between two factors, scientists need to take into account other factors that may affect that relationship. When they do, they statistically adjust their findings to reflect the impact of these other factors.

Let's say you need surgery and are asked to choose between two hospitals in which to have it performed. You have information about post-surgery survival rates in each hospital during the past two years, and it looks like this:

	Hospital A	Hospital B
Died	63 (3%)	16 (2%)
Survived	2,037 (97%)	784 (98%)
Total	2,100 (100%)	800 (100%)

At first glance, you would likely choose Hospital B. After all, your chances of dying after surgery in Hospital B are only two per cent compared to three per cent in Hospital A.

Scientists may express this to you as an odds ratio (OR). Comparing the risk of dying post-surgery in the two hospitals (two versus three per cent), they will tell you the odds ratio is 0.66. In other words, relatively speaking, there is a 34 per cent lower risk of dying in Hospital B than in Hospital A.

What these scientists will also tell you, however, is that this is an unadjusted or crude odds ratio. No other factors are taken into account when looking at the relationship between the hospital and the likelihood of dying. However, other factors may certainly affect the outcome. How old were the patients at each hospital? Were they in good health before surgery?

These other factors are called confounding variables. They are the "something else" that could affect the relationship between two other things – in this case, the relationship between the hospital and post-surgery outcomes.

Let's look again at the two hospitals and, this time, take into account the health of the patients going into surgery: either "good" or "poor."

With this information, you would be wise to change your mind and choose Hospital A. That's because you can now see

Good health

	Hospital A	Hospital B
Died	6 (1%)	8 (1.3%)
Survived	594 (99%)	592 (98.7%)
Total	600 (100%)	600 (100%)

Poor health

	Hospital A	Hospital B
Died	57 (3.8%)	8 (4%)
Survived	1,433 (96.2%)	192 (96%)
Total	1,500 (100%)	200 (100%)

that, for patients in good health, 1.3 per cent of patients died in Hospital B compared to only one per cent in Hospital A. Interestingly, Hospital B also did worse for patients in poor health, with four per cent dying compared to 3.8 per cent in Hospital A. The confounding variable – condition of the patient – makes a big difference.

(How can Hospital A do better for patients in both good and poor health, yet do worse overall? It could be that Hospital A is a teaching hospital with leading-edge surgeons, which serves seriously ill people from a wide geographic region. It attracts a much higher number of patients in poor health, who are more likely to die. As a result, Hospital A has a higher death rate overall, despite its better performance for each type of patient.)

Again, scientists may express this to you in a different way. This time, they will tell you that the odds ratio has been statistically adjusted to incorporate the effect of patient condition at the time of surgery, and is now 1.14. In other words, there is a 14 per cent higher risk of dying post- surgery in Hospital B than in Hospital A after taking the health of patients into account.

If other potentially confounding factors, such as age of patient, socioeconomic background, etc., are also taken into account, scientists will give you an odds ratio that they call fully adjusted. A fully adjusted odds ratio strips away the effects of other factors, theoretically leaving only the relationship between the two studied factors standing.

Subgroup analysis

Subgroup analysis is a tool for exploring differences in how people respond to a health intervention, but it must be used with care.

Think of a time you looked at a study and wondered if the thing being studied—a treatment, program or other intervention—was more effective for some people than others. Subgroup analysis is one way of finding out. It's a type of analysis done by breaking down study samples into subsets of participants based on a shared characteristic. The goal is to explore differences in how people respond to an intervention.

For example, let's say you want to study the effectiveness of a new drug for pain relief. You might set up a randomized controlled trial where one group gets the drug (the intervention group) and the other gets a placebo (the control group). Your goal is to find out whether those who receive the new drug report less pain compared to the control group.

However, you might also want to know if the new drug works better for certain groups of people than others. So you divide the study participants into subgroups according to factors that may be important: the type of condition causing the pain, how long the condition has been present, gender, age, etc. You may learn that the treatment works better for certain conditions and for women below a certain age—all potentially crucial information.

This might sound easy enough. But the research world struggles with subgroup analysis. That's because, when done improperly, it can lead to exaggerated or wrong findings.

How subgroup analysis can go wrong

There are two main reasons subgroups can lead to error. The sample size can be too small, and there can be too many comparisons done. When you break down your study sample into many subgroups, you may end up with too few participants in each to detect differences, or to ensure differences aren't just a matter of chance.

Take our pain relief study. Let's say there's a small but important difference in how people with neck pain respond to the treatment versus those with back pain. With enough people in the subgroups, you could find that difference, even if it's small. But if your subgroups have too few people in them, you won't have the "statistical power," as it's called, to detect the difference. As a result, you miss a difference that exists. Scientists call this a false-negative error.

Subgroup analysis can also lead you to make a false-positive error—when you see differences that aren't really there. If you slice and dice your study sample enough times, you'll eventually end up with a subgroup that responds to the pain treatment differently than the rest—such as redheads or people born in January. That would be what scientists call a spurious finding—one that doesn't make sense biologically or isn't based on sound theory.

There's also the kind of error that happens when you inappropriately define your subgroups. Take a factor such as age, for example. In your study, you might look at how the drug affects people of different ages—say, people in their 20s, 30s and 40s. But really, what's your rationale for subgroups of 10 years and not five years or 20? What if, by pure chance, the 37- and 38-year-olds respond really well to the treatment? Would you be able to resist the temptation to divvy up your sample into two-year subgroups and report on those findings? What if that meant the difference between getting your research published and not?

When subgroup analysis goes right

Despite these problems, there are certain things you can look for to tell whether a subgroup analysis has been done right:

- the subgroup analysis is a stated study objective from the start—not an afterthought;
- the researcher can explain the reason for doing the subgroup analysis (based on previous research or a sound hypothesis, for example);
- ideally, the researcher defines the subgroups upfront and states how many subgroup analyses will be done. As well, the researcher reports on all of them, not just the ones that give rise to interesting findings; and
- the study is designed so that the subgroups have large enough sample size.

Subgroup analysis is important for investigating differences in how people respond to a treatment or intervention. But when misused, it can result in misleading findings. That's why it's important to understand the risks associated with this kind of analysis and to know what to look for when you come across it.

Survival analysis

Survival analysis techniques allow researchers to study lengths of time, often to predict when a given event or end point will occur.

Survival analysis is a branch of statistics that allows researchers to study lengths of time. Historically, it was developed to study/predict time to death of patients with a disease or an illness, and it typically focused on the time between diagnosis ('start' time) and death ('end' time). As such, it is used to answer questions such as: What fraction of a population will survive past a certain time? How do particular circumstances (e.g. taking a new medication) or characteristics (e.g. age of patient) increase or decrease time to death?

However, survival analysis techniques do not always entail timelines leading to death. They can be used to study the probability of a wide range of time outcomes. For example, in the social sciences, researchers may study the "survival" of marriages, high school drop-out rates (time to drop-out), spells of unemployment and, as we will see, time to return to work following a workplace injury.

Survival times are data that measure follow-up time from a defined starting point to the occurrence of a given event or end point. However, if a study stops before all participants have reached the end point, survival analysis can accommodate this partial information; i.e. that these participants survived at least so long. For example, a researcher studying the effectiveness of a new treatment for a disease considered terminal would not want to exclude patients who survived the entire study period, because their survival reflects on the effectiveness of the treatment.

Kaplan–Meier survival curve

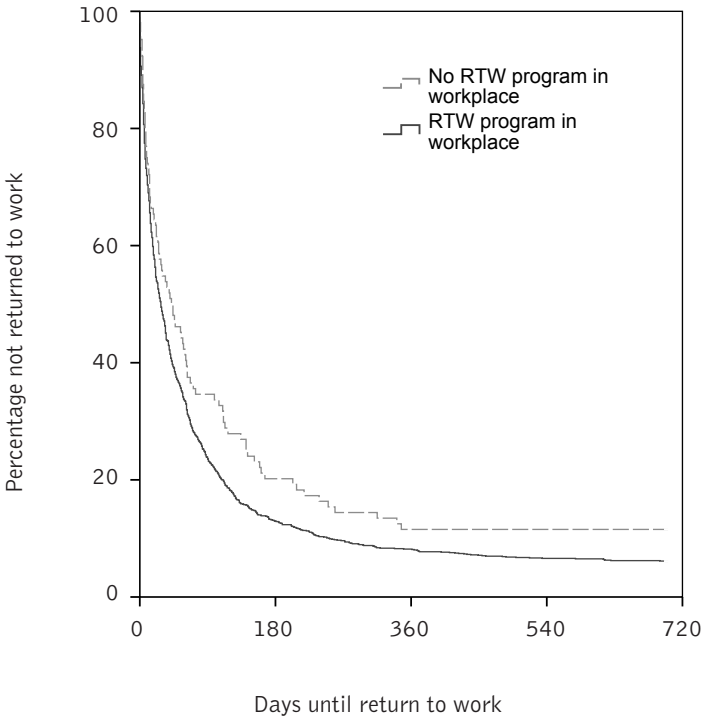
Researchers have a number of methods for analyzing data in order to show the distribution of lengths of time taken to reach a certain end point. One of the more widely used methods is the Kaplan–Meier survival curve, named after its creators Edward Kaplan and Paul Meier.

To show how this curve conveys this information, let's say you were studying return to work (RTW) among a group of injured workers with low back pain (LBP). Based on your findings, you might want to show what percentage of workers with LBP will return to work by certain points over time and how particular circumstances affect the timing of RTW.

The Kaplan-Meier curves (only available on hard copy of article) show workers who had no workplace RTW program (the lighter curve) and workers with an established RTW plan at work (the darker curve) and the number of days it took to return to work after a sick leave due to LBP. As shown in the graph, approximately 15 per cent of injured workers with an RTW program had not returned to work by 180 days, but an even greater percentage—20 per cent—had not yet returned to work by 180 days where there was no RTW program. This suggests that RTW programs are helpful in getting more injured workers back to work.

So survival analysis can do more than predict death. It can aid decision-making in a wide variety of situations, including work and health.

Kaplan-Meier curve showing workers with and without RTW programs, and the time it took for both to return to work



Systematic review

A systematic review helps users of evidence keep up to date on a body of research by synthesizing the findings of higher quality studies on a given topic.

Think of the last time you came across a research study that seemed to contradict some other study on the same question. You can probably think of a few examples, especially for health topics that are often in the news. One moment you hear that acupuncture helps relieve pain. The next, a new study says it doesn't.

If you think about how research studies are conducted, you can appreciate why discrepancies in findings arise. Different researchers studying the same question might enlist different numbers of participants. They might choose different study designs. There might be differences in how they administer the treatment or intervention or how they measure the effect of the intervention. All these things make a difference to what researchers ultimately find.

In other words, when looking for research evidence, you need to look beyond a single study and take into account the overall body of evidence. But given the amount of published research on a given topic, keeping up on the evidence can overwhelm anyone—including clinicians, researchers and policy-makers.

This is where systematic reviews come in. They help people keep up on what the overall body of research says on a topic. They're designed to take into account the reliable available evidence on a subject at a given point in time.

To do this, researchers on a systematic review team go through all the studies relevant to a topic and assess the quality of each. From the higher quality studies, they'll pull out a synthesis of the findings. Often, they'll combine the data from different studies to do what's known as meta-analysis (see www.iwh.on.ca/wrmb/meta-analysis). And as systematic reviews can only synthesize the available research at a point in time, they need to be updated regularly.

Narrative vs. systematic reviews

To better understand systematic reviews, consider traditional narrative reviews that were once more commonplace. Like systematic reviews, narrative reviews also synthesize the scientific literature on a given question. The main difference is narrative reviewers draw chiefly from their experience and expertise for their analysis. This makes narrative reviews more susceptible to bias. No clear methodology is evident to help readers understand whether reviewers have considered all the available evidence, or how and why they recommend one study over another.

Systematic reviews, in contrast, minimize this type of bias by putting methodology front and centre. Like any other scientific study, systematic reviews should be replicable. That means another research team, using the same methodology to tackle the same question, should be able to gather the same evidence and come to the same conclusion.

As such, all the steps taken in systematic reviews are clearly and transparently outlined. Right from the literature search, systematic reviews spell out what terms are used, which databases are searched, and what criteria are applied to limit the search (e.g. language of published studies). Subsequent steps are guided just as much by methodology—from deciding what studies are relevant, to assessing the studies for how rigorously they were carried out.

Another distinguishing aspect of systematic reviews is their focus. While narrative reviews might cover off a broad topic, systematic reviews centre on a single research question. This question is typically defined by applying the PICO principle; that is, the question indicates the population, intervention, comparison and outcome being considered in the review. The result might read like this statement of objective from an actual review: "A review of randomized trials of acupuncture for adults with non-specific (sub)acute or chronic low-back pain."

Systematic reviews, though relatively new, are growing more popular as people increasingly recognize the value of evidence-based practice and policy. Given the amount of new research being produced, systematic reviews have become an important tool for staying up to date.

Validity and reliability

Validity and reliability are concepts that capture the measurement properties of a survey, questionnaire or another type of measure.

Validity and reliability are important concepts in research. The everyday use of these terms provides a sense of what they mean (for example, your opinion is valid; your friends are reliable). In research, however, their use is more complex.

Suppose you hear about a new study showing depression levels among workers declined during an economic downturn. You learn that this study used a new questionnaire to ask workers about their mental health over a number of years. You decide to take a closer look at the strength of this new questionnaire. Was it valid? Was it reliable?

To assess the validity and reliability of a survey or other measure, researchers need to consider a number of things.

Ensuring the validity of measurement

At the outset, researchers need to consider the face validity of a questionnaire. That is, to a layperson, does it look like it will measure what it is intended to measure? In our example, would the people administering and taking the questionnaire think it a valid measure of depression? Do the questions and range of response options seem, on their face, appropriate for measuring depression?

Researchers also need to consider the content validity of the questionnaire; that is, will it actually measure what it is intended to measure. Researchers often rely on subject-matter experts to help determine this. In our case, the researchers could turn to experts in depression to consider their questions against the known symptoms of depression (e.g. depressed mood, sleeping problems and weight changes).

When questionnaires are measuring something abstract, researchers also need to establish its construct validity. This refers to the questionnaire's ability to measure the abstract concept adequately. In this case, the researchers could have given a questionnaire on a similar construct, such as anxiety, to see if the results were related, as one would expect. Or they could have given a questionnaire on a different construct, such as happiness, to see if the results were the opposite.

It may sometimes be appropriate for researchers to establish criterion validity; that is, the extent to which the measurement tool is able to produce accurate findings when compared to a "gold standard." In this case, the gold standard would be clinical diagnoses of depression. The researchers could see how their questionnaire results relate to actual clinical diagnoses of depression among the workers surveyed.

Ensuring the reliability of measurement

Researchers also need to consider the reliability of a questionnaire. Will they get similar results if they repeat their questionnaire soon after and conditions have not changed? In our case, if the questionnaire was administered to the same workers soon after the first one, the researchers would expect to find similar levels of depression. If the levels haven't changed, the "repeatability" of the questionnaire would be high. This is called test-retest reliability.

Another aspect of reliability concerns internal consistency among the questions. Do similar questions give rise to similar answers? In our example, if two questions are related to amount of sleep, the researchers would expect the responses to be consistent.

Researchers also look at inter-rater reliability; that is, would different individuals assessing the same thing score the questionnaire the same way. For example, if two different clinicians administer the depression questionnaire to the same patient, would the resulting scores given by the two be relatively similar?

If our depression researchers were sloppy in ensuring the validity or reliability of their questionnaire, it could affect the believability of their study's overall results. Although you can never prove reliability or validity conclusively, results will be more accurate if the measures in a study are as reliable and valid as possible.

The Institute for Work & Health (IWH) conducts and shares research that protects and improves the health of working people and is valued by policy-makers, workers and workplaces, clinicians, and health & safety professionals.

Institute for Work & Health
481 University Avenue, Suite 800
Toronto, Ontario M5G 2E9
info@iwh.on.ca
www.iwh.on.ca

Institute for Work & Health | Research Excellence Advancing Employee Health