Quantitative research approaches for public health

Contents

[1. Summary/checklist of key advice 2](#_Toc152058513)

[1.1. Descriptive research 2](#_Toc152058514)

[1.2. Causal research 2](#_Toc152058515)

[2. Aim 2](#_Toc152058516)

[3. Quantitative research questions 3](#_Toc152058517)

[4. Descriptive research questions and how to address them 4](#_Toc152058518)

[4.1. Descriptive research: key study design and data collection considerations 7](#_Toc152058519)

[4.1.1. Study design 7](#_Toc152058520)

[4.1.2. Sampling 8](#_Toc152058521)

[4.1.3. Data collection 8](#_Toc152058522)

[4.1.4. Missing data 8](#_Toc152058523)

[4.2. Descriptive research: key data analysis considerations 9](#_Toc152058524)

[4.2.1. Descriptive inference is not descriptive statistics! 9](#_Toc152058525)

[4.2.2. Descriptive associations – appropriate approaches 10](#_Toc152058526)

[4.2.3. Descriptive associations – inappropriate approaches 12](#_Toc152058527)

[4.2.4. Missing data 14](#_Toc152058528)

[4.3. The importance of descriptive research 15](#_Toc152058529)

[5. Causal research questions and how to address them 16](#_Toc152058530)

# Summary/checklist of key advice

## Descriptive research studies

1. Make use of the “[Framework for Descriptive Epidemiology](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10144679/)” when planning descriptive research and its associated reporting checklist when reporting that work, and make use of the “[Consensus-Based Checklist for Reporting of Survey Studies (CROSS)](https://link.springer.com/article/10.1007/s11606-021-06737-1#Sec11)” if reporting the results of a cross-sectional study.

2. Develop clear and fully defined descriptive research questions (use the framework) that relate to a specific target population.

3. Let the research question(s) guide your methods (e.g. if you want to measure frequency of occurrence of an event you need a longitudinal design), and choose the most robust methods you can (e.g. obtaining a representative sample, ideally via a probability sampling approach, is critical for a descriptive study).

4. Plan your analyses to address your research questions in line with the approaches recommended by the framework. Avoid thoughtless multivariable adjustment strategies involving models with lots of independent variables per outcome.

5. Include inferential measures (ideally confidence intervals) for all results that you intend to use to describe some aspect of your target population.

## Causal research studies

To follow…

# Aim

The primary aim of this document is to provide a high-level overview on how to approach planning, analysing, and reporting quantitative research in the health sciences (although the approaches arguably apply much more widely to many other fields of quantitative research). The document only considers what you might call “final stage research”, or research that aims to produce generalisable findings/conclusions, so excluding formative/development/feasibility/pilot work. A key focus is on the importance of starting by developing a clear research question or questions, from which everything else should follow. You should certainly take this approach if you are at the initial planning stages of a study where you will be able to collect data. However, it is also relevant if you have a dataset from an existing study and you’re basically trying to work out if you can use it to do something useful with, but you’re not sure what. I therefore structure the document in terms of the types of research questions we can ask using quantitative methods, with everything else following from there.

Of course, not all researchers will agree on the approaches described, but I believe they represent the consensus among prominent researchers working on methodology in epidemiology, causal inference, and applied statistics (based on my limited understanding of the relevant literature). The information and advice will be supported by key references, again from prominent researchers in epidemiology, causal inference, and statistics. In almost all cases these are provided in the form of links to open access papers for ease of viewing. I would strongly advise using these references to gain a greater and more detailed understanding of these issues as required, because it is not feasible for me to provide the huge number of finer details that relate to these issues here.

The secondary aim is to highlight *some* of the (often extremely) common practices seen in the literature that good evidence suggests are problematic (often extremely so), and to discuss what you might do instead. I cannot comprehensively cover all common problematic practices, because again this would not be feasible. However, my goal is to cover those problematic practices that I believe are both the most common. By avoiding these issues you can often make a big difference to the rigor of a study.

# Quantitative research questions

The primary goal of any quantitative research study is always to answer one or more research questions (I will often to a “research question” from here on, but of course you will often have more than one). However, it is not always appreciated that your methods and the interpretation of your results should all be directly guided by your research question. Therefore, your research question has a huge impact on both the direction and quality of your study, and it is critical that you develop a clear and precise research question, and then plan your study methods based on this question. Poor research questions are one of the most common weaknesses in quantitative research studies. Often this seems to occur when the research question posed is not the true question the researchers really want to answer, but they implicitly plan their study to try and address the true question, even when (as is often the case) this is not feasible. Other very common issues also appear to be having unclear or vague research questions.

A modern, widely accepted framework in epidemiology is that there are just three main distinct types of quantitative research question: 1) descriptive (including associational), 2) causal (or counterfactual), and 3) predictive. You may of course have more than one research question, and any single study may of course also address more than one of these types of quantitative research questions. See [here](https://www.hsph.harvard.edu/wp-content/uploads/sites/1268/2019/04/hernan_chance19.pdf) for an excellent and concise overview of this framework, which also briefly discuss the broad approaches we can and should take to address them. Note: be aware that while the same statistical tools, like regression modelling, may often be used to address all three types of research question, the approaches and ways in which you should use those tools differs greatly depending on the type of research question being addressed.

Below I provide a brief overview of descriptive and causal research questions, the broad approaches recommended to address them, and commonly seen problematic practices to avoid (covering these issues for each type of question separately). Note: below I talk about “descriptive research” and “causal research” when referring to the broad approaches taken to address descriptive or causal research questions in the health sciences. You will also see the term “descriptive epidemiology” and “causal epidemiology” in the literature, but I’ve stuck with slightly broader terms. I do not cover predictive questions because they are very uncommon in our field of international public health research, being primarily used in the context of predicting disease outcomes for patients based on patient characteristics (but you can read a bit about them and some examples in the reference above).

# Descriptive research questions and how to address them

Descriptive research questions are about quantifying some characteristic of a defined population and, often, also quantifying the distribution of that characteristic across subgroups in that population[[1]](#footnote-1) (i.e. describing associations[[2]](#footnote-2) between two or more characteristics in that population). I use the term “characteristic” broadly to refer to any feature/aspect of the members of a population[[3]](#footnote-3) that can be quantitatively measured (and will be stored as a variable in a dataset), such as systolic blood pressure (in mmHg), or the number of essential drug stock-outs per health facility within a given time period.

An excellent, recent framework for how to approach planning a descriptive study (or the parts of a larger study that are designed to address a descriptive research question) is described, quite concisely, in [this](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10144679/) paper, which helpfully uses an example to illustrate the issues discussed. *If you are going to plan a descriptive study, I strongly recommend reading and using this guide.* The paper also includes guidance on how to report a descriptive study, including a checklist for items to report in a results paper (in the style of CONSORT and STROBE etc).

Another useful reporting checklist exists specifically for cross-sectional survey studies (see [here](https://link.springer.com/article/10.1007/s11606-021-06737-1) but note the paper is only about the development of the checklist, which can be found in the supplementary materials). I thought they could have done a better job in places by providing more detailed, comprehensive guidance and reminders in some areas of the checklist, including on the issues that the broader, descriptive study checklist covers, but it’s still a useful additional tool to use when reporting the results specifically from a cross-sectional survey. Therefore, if you are reporting the results from a cross-sectional survey I would recommend using both checklists.

The descriptive study framework starts with, and is focused on, how to develop one or more clear descriptive research questions. In brief, the guide stresses that a well-defined descriptive research question must include and describe the following components (I have slightly rearranged the ordering).

1) A fully defined target population. This is the population that your study aims to draw quantitative conclusions about (i.e. make statistical inferences about). The target population is usually briefly stated in the description of your research question in terms of the key features of the observational units of the study, which may be individuals or entities like health facilities, but should be later fully defined in the main methods in terms of your precise study eligibility criteria (e.g. socio-demographic characteristics, health facility type etc). Your target population must also be defined in terms of geographical location (e.g. country/region etc), setting (e.g. rural/urban slum etc), and time (e.g. year). For example, community members aged >18, living in slums, in Anambra state, Nigeria, in 2023. Or public primary care and NGO primary care facilities within Dhaka, Bangladesh in 2023.

If you don’t clearly define your target population you may inadvertently develop methods that do not allow you to draw quantitative conclusions about the population you are interested in, and when interpreting and reporting your results you and your readers may not be able to clearly judge which population your results apply directly to. The target population should generally be the population that you sample from, although you may also argue that your results apply more widely – a concept known as transportability or applicability (see [this](https://ebm.bmj.com/content/23/1/17) paper for a concise overview of this issue and related issues of external validity, generalisability and transportability, and [this](https://bmjmedicine.bmj.com/content/2/1/e000399) paper for an interesting, if less commonly encountered, view about representativeness and generalisability). Poorly defined or even completely undefined target populations are a common weakness in descriptive studies, and not something you can easily fix after you have collected your data.

2) The characteristic (i.e. outcome) being described, the aspect of the statistical distribution of that characteristic in the target population that you seek to describe, and the summary statistic you will use (e.g. the mean systolic blood pressure [mmHg], the incidence rate of essential drug stock-outs per month in health facilities etc).

3) “Any auxiliary variables and their roles as stratification factors (to characterize the outcome distribution) or nuisance variables (to be standardized over)”. This is a direct quote from the framework. I think these are more complicated considerations than the other two, so I strongly recommend reading the section in the framework [paper](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10144679/) on “The Role of Covariates”, and also in [this](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9383568/) paper the sections on “Stratification and adjustment in descriptive epidemiology: understanding who, where, and when” and “The line between descriptive and causal epidemiology” for more guidance on this issue.[[4]](#footnote-4) However, stratification and standardisation/“nuisance variables” can mean quite different things in terms of analyses, and they are not clear about all the ways you might involve these issues. Therefore, I’m not 100% certain on what they mean for all situations here.

However, in the context of a descriptive study I think the main consideration in terms of possible “stratification” is two-fold. First, deciding whether you will describe the characteristic of interest in relation to subgroups (i.e. the stratifying variables) in the target population. For example, you may want to describe the mean systolic blood pressure (mmHg) in key subgroups within the target population based on specific socio-demographic characteristics, such as sex or age-groups. In addition to describing the level (e.g. mean) or frequency (e.g. rate/percentage) of the characteristic within specific subgroups, you may also certainly want to describe the actual differences in the level/frequency of the characteristic between those subgroups, or more generally the associations between the characteristic and one or more other characteristics, which may naturally form subgroups or naturally be on some kind of numerical scale. For example, you may want to describe the association between the number of monthly patients and the incidence rate of essential drug stock-outs per month in health facilities in continuous terms, so as a slope or trend (e.g. as a regression slope coefficient).

Second, when describing associations you may want to look at how those associations themselves differ between strata, or more generally how they vary in terms of other characteristics. For example, how the association (difference) between sex and mean systolic blood pressure (mmHg) varies between individuals aged over 40 compared to those aged 40 or less. Or how the association (slope) between the number of monthly patients and the incidence rate of essential drug stock-outs per month in health facilities varies depending on the number of staff per health facility.

Standardisation is a reweighting procedure that is most commonly used to adjust measures like mortality rates in one population for the distribution of some key characteristic that is thought to strongly determine the outcome, like age, in a different target population, so the mortality rates can, in theory, be compared on a more like-for-like basis. See the recommended reading above and any general epidemiology textbook for more details.

## Descriptive research: key study design and data collection considerations

Once you have your descriptive research question/questions defined you can plan your methods to collect your data (assuming you are not using existing data). Again, this guide is not intended to go into details on these methods. Additional guidance is in the framework, but further details on robust methods for descriptive studies would have to come from relevant, specific literature and textbooks. However, below I highlight some key considerations.

### Study design

For a descriptive study a single-point/period cross-sectional design is the most commonly used approach. However, it is important to be aware that, if this is a “true” single-point/period cross-sectional design, then you can only describe characteristics at that point/period in time. For binary/count characteristics this means you can only describe the prevalence proportion or prevalence count of those characteristics. With a repeated cross-sectional design, like the DHS survey in many countries, you can describe changes in characteristics through time, but typically you would again only be able to describe changes in prevalence measures for binary/count characteristics.

If you want to describe the frequency of occurrence of some event over some period of time (i.e. the incidence proportion or incidence rate of an outcome) you need to use a longitudinal (also called panel) type design. Note: you may be able to collect longitudinal data retrospectively though from a single point/period though (e.g. via recall). This would then not be a true cross-sectional study though, because while you would be collecting data from a single point/period your data would be longitudinal, and technically your study design would therefore be a longitudinal design, and you could then compute measures of incidence for binary/count characteristics.

### Sampling

The goal of descriptive research is typically to describe characteristics as they exist in a clearly defined target population. Therefore, for descriptive studies it is critically important that your sampling methods maximise the chance that you will get a sample that is representative of that target population, otherwise your results will be biased and won’t accurately generalise to your target population. This ideally means using a probability sampling approach, because all common inferential statistical methods assume that your sample was collected via a probability sampling method. If you use a non-probability (also called purposive) sampling approach then there is much more risk that your sample will not be representative of your target population. Consequently, if you cannot use a probability sampling approach you must try and develop the most robust purposive approach possible.

See [this](https://ebm.bmj.com/content/23/1/17) paper for a concise overview of the concepts of external validity and generalisability (plus transportability/applicability), which are directly impacted by sampling methods, and [this](https://bmjmedicine.bmj.com/content/2/1/e000399) paper for a slightly more wide-ranging view on representativeness and generalisability.

### Data collection

Develop data collection tools that will allow you to accurately and robustly measure the characteristics you want to describe. Be aware, be realistic, and be honest about whether the outcomes you are collecting are really measuring what you ideally want to measure, or if they are actually proxy measures because the real characteristic of interest is too difficult to measure. You won’t necessarily be able to overcome such an issue, and this is often the case, but it’s better to be aware and transparent about the reality of this limitation than pretending you are measuring something you’re not. For example, many questionnaire-based tools seek to measure some characteristic, say reported antibiotic use, but it’s important to remember that they are rarely actually measuring this characteristic, but instead some proxy measure with some level of (potentially large) error.

### Missing data

When planning your study data collection methods, it is critical to consider the possible ways in which you may end up with missing data and to try to put in place measures to reduce/prevent this. Missing data are usually a key source of bias in descriptive research studies. In cross-sectional studies this most commonly and most obviously occurs due to non-response, while in longitudinal studies you also often face the problem of loss to follow-up. However, many other sources of bias can also be seen as missing data problems, like measurement error and competing events. The descriptive research framework previously referenced discusses these issues in a bit more detail.

## Descriptive research: key data analysis considerations

As with all quantitative studies you should develop a fully detailed statistical analysis plan prior to doing any data analyses. Just diving straight into an analysis almost always just leads to wasted effort, because you almost always end up changing lots of things, and poorer quality results, because you haven’t thought through all the ways to maximise the rigor of the analysis. Again, I do not go into details here but cover some of the key considerations and key problematic practices below.

### Descriptive inference is not necessarily “descriptive statistics”!

Do not confuse descriptive research or a descriptive study with what are often called “descriptive statistics”. These are also (but much more rarely) called sample statistics, although this seems a much less confusing and appropriate name for them. Similarly, do not fail to use inferential statistics when your goal is inference. Statistical terminology does not help here at all. Descriptive statistics, as the name is normally used, are statistics that provide summary measures of key features of the distribution of a variable, such as its mean or standard deviation (for numerical variables), or its frequency of occurrence (for categorical variables), *in the sample alone*. You must certainly report descriptive statistics (often in your first table) for your sample as these describe the key characteristics of the sample (e.g. the distribution of ages, the sex ratio, education levels, or the distribution of health facilities types etc). This is to allow you and your readers to understand “who” the inferential results for your outcome/outcomes may be robustly applied to, especially if you have only been able to take a purposive sample and cannot rely on sampling theory for your generalisations.

However, remember that the goal of descriptive research is typically to describe characteristics as they exist *in some target population* (or possibly even more widely), and you will almost never be able to record those characteristics in every member of that target population. Instead, you can usually only record those characteristics in a very, very small sample of that target population. And due to the random (or haphazard) nature of the sampling process every sample has sampling error, which affects the representativeness and generalisability of the results to the target population. Descriptive statistics (also called sample statistics) alone cannot tell you anything about this sampling error. Therefore, if you want to draw conclusions about characteristics in your target population you must use of inferential statistics (i.e. confidence intervals, p-values etc), which account for sampling error (in theory). This problem seems to be most common when describing means/percentages/rates etc compared to when looking at associations, probably just because software more readily presents inferential measures when doing analyses for associations (e.g. t-test/regression etc) compared to computing means/percentages/rates etc.

Therefore, whenever you are aiming to describe characteristics (e.g. in terms of means/percentages etc) and/or associations between characteristics (e.g. in terms of differences/slopes) in your target population you must always compute and present inferential results. The consensus (within frequentist statistics) is to compute, report, and interpret confidence intervals as your measures of inference, as opposed to p-values, as confidence intervals allow you to judge the likely true value of a characteristic or measure of association in the target population. When describing the level of a characteristic in a target population, unless you are comparing to some specific value, which is not often the case, a p-value wouldn’t even have any obvious use. See [this](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1339793/pdf/bmjcred00225-0036.pdf) paper for a concise and accessible explanation (if you ignore the appendices) of why this broad approach makes sense. And if you are not already an expert on interpreting confidence intervals and p-values and want to improve your understanding of these measures I strongly recommend reading [this](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/) paper, which describes how to correctly interpret them in the context of the many common ways they are misinterpreted.

Therefore, it is extremely poor practice to present results that you then use to try and draw inferential conclusions about characteristics or associations in your target population without any inferential measures. However, in the literature this appears to be very common in survey studies, where you will often see tables reporting things like percentages or means of key outcomes for different subgroups (like by sex or age subgroups) without any associated confidence intervals (or indeed any other inferential measure), while those percentages/means are then clearly being interpreted as reflecting the likely percentage/mean values of those outcomes in some (often vague/undefined) target population. Unless the sample size is extremely large this is extremely misleading, as it doesn’t provide any basis to determine how precise those sample estimates are in terms of the likely values for the target population, given sampling error. Do not make this very basic mistake. It’s extremely easy to avoid.

### Descriptive associations – appropriate approaches

Remember that when addressing descriptive research questions about associations (i.e. how characteristics vary in relation to other characteristics) the goal is usually to describe those associations as they exist in the real world, and never to treat the associations as estimates of causal relationships (if you have a causal question you must use causal inference methods). In practice, this typically means estimating “unadjusted” or “crude” bivariate measures of association, because this reflects the relationship as it exists in the real world and without any attempt at making it into an estimate of a causal relationship.

In a descriptive study you can use regression type models to estimate purely descriptive associations most simply by just including one independent variable of interest at a time per model. For example, to describe how the incidence rate of essential drug stock-outs per health facility varies by facility type we could use a model with an outcome of the facility-level number of stock-outs and an independent variable for facility type. If we also wanted to describe how the incidence rate of essential drug stock-outs per health facility varies by number of patients per month we would create a new model with the same outcome but an independent variable for number of patients per month.

However, sometimes it may make sense to adjust for other independent variables. Most commonly, you may want to describe how a given association varies by one or more other characteristics. For example, say we wanted to describe how the relationship between facility type and the incidence rate of essential drug stock-outs itself varied between rural and urban areas. In a regression modelling framework we can use interactions to do this. For this example, we would include an interaction between the independent variables of facility type and facility location. When we look at how an association varies by a third variable that is binary/categorical this is typically referred as stratifying/stratification, but we can equally do the same thing with numerical variables, again by including interactions.

You may also want to adjust for other characteristics to increase the precision of your estimates for the associations of interest. There doesn’t appear to be a commonly accepted name for such variables in the context of descriptive research, but one excellent [article](https://royalsocietypublishing.org/doi/full/10.1098/rspb.2020.2815) that secondarily briefly covers broad approaches to addressing descriptive questions called them “precision covariates” (see the supplementary materials in particular). The key idea here is that these should be associated with your outcome but not (at least not strongly) associated with your independent variable of interest. You need to use subject matter knowledge and common sense, guide by a “DAG” (see the section on causal inference to understand what this is), to determine if any such variables likely exist (ideally at the planning stage so you can collect them), along with sensitivity analyses with and without such adjustments, to help you decide.

For example, if you want to describe the relationship between facility location (urban/rural) and the incidence rate of essential drug stock-outs per health facility, and you believe that drug stock-outs and the number of staff per facility are associated, but that the number of staff per facility is not associated with facility location, then it may be helpful to adjust your estimate of the association between facility location and drug stock-outs for the number of staff per facility. However, this can be a risky approach that can bias your estimate of the descriptive association of interest depending on the true relationships between the variables (see below), so if in doubt it’s probably safer to not adjust for such variables.

Lastly, you may also potentially want to standardise or otherwise adjust for one (or possibly a very small number of), carefully considered, additional characteristics. This is arguably a more complicated consideration with more need for careful thought and very clear justification. To quote from [this](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9383568/) paper previously referenced.

“There are important reasons descriptive statistics are sometimes more useful after adjustment for some variables (e.g., age) that are known to affect the outcome and whose distribution differs across populations (18). Such adjustment may help us hypothesize about reasons for variability within or across populations to better identify disparities in disease distribution that might otherwise be masked. For example, if we sought to describe differences in COVID mortality between doctors and nurses, we might adjust for differences in age distributions (19) to focus on occupational hazards rather than known age-related hazards. In other instances, however, age adjustment can prevent appropriate targeting or resources (20). This is because adjusting away a major difference between 2 populations, one of which is, on average older than the other, may make it seem that rates of diseases are the same, and therefore require equivalent resources, when in fact there is more disease in the older population.”

### Descriptive associations – inappropriate approaches

One extremely common but problematic approach seen in many observational studies, and seemingly almost universally done in cross-sectional survey studies, is to carry out and present one or more analyses of key outcomes via a regression-type model where a large set of independent variables are included in the model. The results (coefficients) of every independent variable included in the model are then interpreted in some way, sometimes as casual effects but more commonly as “independent effects” (or some other similar term is used like “independent relationships”). The problem is that neither interpretation necessarily makes much sense (or has much meaning) depending on the goal of the regression analysis.

If the goal of the regression analysis (i.e. the research question it is intended to address) is to estimate one or more causal relationships then you should use appropriate methods from causal inference to do this. Unfortunately, just putting a big load of independent variables into a model and treating their coefficients as all having a valid causal interpretation is deeply problematic. See the section on causal inference for more details. Briefly though, this approach has been called the [Table 2 fallacy](https://academic.oup.com/aje/article/177/4/292/147738), and will typically result in a set of estimates that are heavily biased compared to any true causal relationships, because you’re extremely unlikely to have appropriately adjusted for each causal relationship (and this ignoring lots of other considerations that should have come before this stage). They are therefore of little use from a causal inference perspective.

However, if the goal of the regression analysis is not to address a causal question then, even if the authors are not clear on this issue, it must be to address a descriptive one (assuming the goal is not predictive, which is very uncommon in our field). But as discussed, the goal of a descriptive study looking at associations is typically to describe those associations as they appear in the real world and without any causal interpretation implied. However, by adjusting for many independent variables you’re no longer describing relationships as they exist in the real world but as they exist in some very hypothetical world that will usually not have any relevance to your descriptive question or the reasons you want to address it.

For example, assume we want to inform planning to tackle essential drug stock-outs and so we want to describe how the scale of the problem of drug stock-outs varies by facility type, facility location *in some target population*. We create a regression model of the number of essential drug stock-outs per facility with independent variables for type of facility (public/NGO) and facility location (urban/rural). However, due to mutual adjustment the coefficient for facility type is then telling us what the average difference in drug stock-outs ***would be*** between public and NGO facilities ***if the distribution of facility types was******the same in both urban and rural areas***, and vice versa for the coefficient for facility location. If the distribution of facility types is very unequal between urban and rural areas then the estimated average difference in drug stock-outs between facility types won’t reflect the true difference in the target population any longer and could (depending on the underlying causal relationships between these factors) underestimate or overestimate the scale of the problem in public/NGO facilities, as it exists in the target population. Of course, if we had adjusted for many other facility characteristics this issue would likely only get worse as our model coefficients relate to an even more hypothetical world that was even less relevant to our target population.

Therefore, there is never likely to be any clear justification for blindly adjusting all independent variables for one another in an associational descriptive study, although as described above you may sometimes (correctly) need to adjust your descriptive estimates of associations for other carefully chosen variables, usually just one or very few, most often to stratify your results, or sometimes to increase precision or standardise/adjust, but this should be done using separate models for each descriptive association of interest. This approach to estimating associations in descriptive research is the one outlined in the framework for descriptive research, and you can see it briefly touched on in the supplementary materials to [this](https://royalsocietypublishing.org/doi/full/10.1098/rspb.2020.2815) paper also previously referenced.

Two more points. First, there is therefore no clear definition of what “independent effects” produced by this blind, multivariable approach mean. As discussed, they are not valid causal effects but neither do they describe real-world associations, so how can you interpret them and make use of them? I’m not clear and there seems to be little guidance in the literature, which suggests they may actually be fairly meaningless despite their ubiquity. See [this](https://arxiv.org/ftp/arxiv/papers/2309/2309.06668.pdf) excellent paper for further discussion on these issues explicitly in relation to regression modelling. Second, as touched on in [this](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9383568/) excellent paper (which makes the case for more emphasis and focus on descriptive research), confounding is a causal concept that does not apply, or is not defined, in a descriptive epidemiological setting. Hence, if your research question is descriptive you should never be attempting to adjust for confounding. Therefore, do make use of these references, particularly the framework, to justify your approach in any papers, as you are otherwise likely to get reviewers asking why you haven’t adjusted everything for everything like in most papers!

Finally, I would like to be clear that I have certainly used this blind, multivariable adjustment approach many times, including in the recent past, often with only a vague research question in mind, because I was not aware of all these problems and this better guidance until more recently. You will also find huge, prominent studies regularly doing this too. See for example these comment pieces [here](https://journals.lww.com/epidem/fulltext/2021/01000/the_quest_for_timely_insights_into_covid_19_should.39.aspx) and [here](https://journals.lww.com/epidem/fulltext/2021/01000/comment_on_williamson_et_al___opensafely___the.38.aspx) that hack a Nature paper on COVID-19 “risk factors” to bits for making this and other mistakes. Note that this study was done by a large team of prominent authors (although apparently no-one with any expertise in causal inference) and gained huge media traction at the time, and probably led to actual policy decisions based on almost certainly very biased results. So please don’t feel bad if you have used this problematic approach before but take this as an opportunity to move beyond it.

### Missing data

The final key analysis consideration when addressing descriptive questions is how you deal with any missing data. See [this](https://stefvanbuuren.name/fimd/) excellent online web-book for detailed guidance on how to robustly deal with missing data, and see section 1.1. and 1.2. for a brief overview of the different types of missing data. Robustly dealing with missing data involves quite advanced techniques, primarily multiple imputation, that may not be feasible for you though. However, at the very least you should do the following.

1) In your analysis methods explain how you dealt with missing data. Even if you simply removed any cases with missing data on required variables (also known as a complete-case analysis) explain this, and in your discussion be transparent that by doing so your results will almost always be biased, unless the data were “missing completely at random” (which is very rare).

2) Report the level of missing data (the number of cases and the percentage) in relation to your obtained sample size. Specifically, report the number and percentage of sampled units that participated in the study and the number and percentage that did not, and the reasons for this refusal. I would recommend using a flow chart to do this, as described by the [STROBE](https://journals.lww.com/epidem/fulltext/2007/11000/strengthening_the_reporting_of_observational.28.aspx) guidance for observational studies.

3) Report the level of missing data present in every single analysis. Specifically, for every result presented, whether that is the description of a characteristic or a measure of association, it should be possible for readers to understand how many of the sampled units were included in the analysis that produced that result. I would recommend making use of rows in tables and footnotes to present these details.

## The importance of descriptive research

This final short section does not contain any guidance but I thought was worth including. Descriptive research has arguably been viewed (primarily in academia) as less important than causal research, or solely an initial step on the way to addressing a causal question. However, addressing descriptive research questions can be extremely important for policy and practice as they allow policy and decision makers to better understand the scale of a problem and how the scale of the problem varies by key factors. Just think about the importance of all the descriptive research that occurred during the COVID-19 pandemic to describe morbidity and mortality rates and how they varied by time, area, and socio-demographic characteristics etc (note that most of this work was not academic though). Descriptive questions may of course also help to guide future research into a causal question that is not well understood.

A very compelling paper arguing for more emphasis and rigor in the teaching and practice of descriptive research is [here](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9383568/). In short, consider giving more importance to your descriptive research questions when planning your research, particularly if they are actually the only feasible questions at the stage your at or with the resources available.

# Causal research questions and how to address them

To follow…

1. This definition is adapted from the one given in a recent framework for descriptive epidemiology (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10144679/). [↑](#footnote-ref-1)
2. Also called relationships, correlates or correlations. [↑](#footnote-ref-2)
3. Note that a population in the general quantitative sense can be any collection of observational units, i.e. units you collect data from, whether that is individuals or non-human entities like health facilities. [↑](#footnote-ref-3)
4. You may also find this paper useful (even though it’s from the field of law!): https://static1.squarespace.com/static/59a7762f2994ca11765ff510/t/5a98803cc83025780f221070/1519943741689/Kaufman-2017-statistics.pdf [↑](#footnote-ref-4)