HOW TO INTERPRET A JOURNAL ARTICLE

www.vetlit.org

Simon Cook BSc BVSc MVetMed DipACVECC DipECVECC FHEA MRCVS

Accessing and reading veterinary literature is important, but being able to interpret it is not necessarily straight forward. **Common myths** regarding scientific literature include:

- All journals are reliable/reputable – **sadly not**
- If something is published it must be accurate – **sadly not**
- You can skip the materials and methods – **absolutely not!**
- If something is statistically significant then it is clinically significant – **sadly not!**
- A finding needs an associated *P*-value to be considered important – **definitely not!**
- Your conclusions from the data should be in line with the study authors – **nope!**

When you can critically evaluate the literature, you may well find you reach a different conclusion to that of the study investigators.

It is **common** for the title and abstract of an article to misrepresent the data actually being presented, and there is often huge scope for interpretation. Skim reading abstracts absolutely has its place (!) but acting upon something you read, or forming a relevant opinion on the data, requires a thorough appraisal, most importantly of the materials and methods.

Common findings in the **materials and methods** that limit the degree to which the data can be interpreted include:

The population

- Is it a first opinion or referral population sample?
- Is the illness severity similar to those you encounter in your practice?
- Is it consistent with previously published data or are these atypical cases?
- If patients were excluded – what effect has this had on the remaining population sample? Does it still represent the population the authors intended?

Data collection

- Was data retrospectively collected with multiple assumptions made?
- Was it prospectively collected and more likely to be accurate and complete?
- Have clear inclusion criteria and definitions been provided? Eg. “Animals with vomiting were included…” – Does this mean vomiting anywhere in the clinical history? Only in the preceding 24 hours? The population samples created are very different depending on the detail.
- “Animals with AKI were included…” - what was the definition of AKI for the purpose of the study? Was it appropriate?
- “Animals with comorbidities were excluded…” – what was a comorbidity defined as and how does that affect the remaining sample of animals?

Statistics

Everybody’s favourite.

Especially in observational and retrospective data, it’s not possible to assume that differences between groups are due (only) to the variable of interest. Yet, this is what many statistical tests assume.

A *P*-value is not the be-all and the end-all, especially with retrospective data; it is not possible to cater for multiple ways in which data may be biased. The most classical example is outcome between two groups treated differently. Outcome is influenced by: illness severity, finances, clinical communications, institution, level of care, comorbidities etc etc. None of which we can control for, or even capture data on, easily. So, differences in outcome when a population sample is divided by treatment group can be troublesome to interpret.

Another example:

A retrospective study evaluates timing of enteral nutrition in a sample of animals with a particular disease. Animals fed sooner had a better outcome. It would be inappropriate to conclude that feeding early **results** in a better outcome.

Perhaps in less affected or more stable animals, which were always going to do better clinically, it was possible to start nutrition sooner?

Results

In retrospective studies, a **correlation** between two variables is often reported, or an outcome is reported to be associated with a particular variable. It is crucial to recognise that an identified **association** or **correlation** in a retrospective study does not equal **causation**.

For example, a study documents that shorter antimicrobial prescription lengths were associated with a better outcome. It can't be inferred that shorter courses improve outcome. It's perhaps more likely that less sick animals were prescribed shorter courses.

Another study documents a strong correlation between ice cream sales and shark attacks. Do ice cream sales cause shark attacks? Possibly(!) though this absolutely cannot be inferred. There is much more likely another variable involved (eg. temperature) which could be causal.

In the examples above, inappropriate conclusions would be that banning long antimicrobial courses and ice creams would be necessary to improve outcomes and limit shark attacks. Prospective studies are needed to investigate these types of associations because they are able to **control** all other variables.

If a new test is proposed – what is it compared to? and in what population is it tested?

An example: How useful is a diagnostic test for FIP in a selection of young, purebred cats with pyrexia, marked hyperglobulinaemia, proteinaceous peritoneal effusion and chorioretinitis? Using a new test to diagnose FIP in those cats is less useful than accurately diagnosing it in a selection of less classical cases; those with mild hyperglobulinaemia only, no effusion, which are little bit older or that are not pyrexic – cases in which other differential diagnoses are still on the table. Reporting a 100% sensitivity and specificity is of limited value if all the test has done is distinguish extremely classical cases, from healthy controls, for example.

Does the study have enough power to support the conclusions?

You may notice that more studies now include power analyses or, more precisely, sample size calculations. These calculations define the number of subjects it would take to document a biologically plausible difference between groups.

The closer (in number) to the whole population you study, the closer you get to the truth, or a genuine difference between two populations. But how many do you actually need?

An example: you aim to detect a difference in survival with a novel treatment. If you expect the improvement in survival to be marked, eg. reducing mortality from 90% to 40%, you will not need large sample sizes. If you expect the difference to be more subtle, eg. a reduction in mortality from 90% to 86%, you will require much larger sample sizes.

The concept of a study being underpowered is one reason why statistically significant differences between groups can be clinically irrelevant.

If a study is underpowered, but a 'significant difference' is detected, there is a real potential for that to reflect a **type I error** – a significant difference that is detected where there is not a real difference.

More broadly though, underpowered studies are at risk of failing to detect the difference that is present. They run the risk of a **type II error**.

The truth is that many veterinary studies are underpowered. This is not a criticism, it's just necessary to recognise that even where statistical significance is documented, there may not be sound conclusions to draw.

Type I vs Type II error

Type I error
Rejecting a **null hypothesis** that is actually true in the population.
ie. Finding a difference that isn't there; a **false positive** conclusion.

Type II error
Failing to reject a null hypothesis that is false in the population

ie. Failing to find the difference, a **false negative** conclusion…

Questions to help frame your discussion of an article

What is the (or is there a…) knowledge gap?
Did the study design create the population sample(s) necessary to address the hypothesis?
Are the conclusions supported by the data presented?
Did the authors achieve their original aims?
How does the study contribute to your understanding or practice?

Some articles here that will help you explore (and contribute to) the literature

What is the value of statistical testing of observational data?
Veterinary Surgery 2022. Open access
https://onlinelibrary.wiley.com/doi/10.1111/vsu.13845

An introduction to power and sample size estimation.
Emergency Medicine Journal 2003
https://emj.bmj.com/content/20/5/453

Statistics at Square One – British Medical Journal. (old resource, still extremely useful)
https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one

Scientific writing and editorial policies and procedures of the Journal of Veterinary Internal Medicine
JVIM 2024. Open access
https://onlinelibrary.wiley.com/doi/10.1111/jvim.16987

The cost of dichotomising continuous variables
BMJ 2006. Open access
https://www.bmj.com/content/332/7549/1080.1