







Making Statistics Clinically Meaningful

J L Peacock ^{1,2}, P J Peacock ³, E Horváth-Puhó ¹, J C Madan ², M R Karagas ²,
H T Sørensen ^{1,2}

¹Department of Clinical Epidemiology, Center for Population Medicine, Aarhus University and Aarhus University Hospital, Aarhus, Denmark;

²Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA; ³Children's Emergency Unit, Great Western Hospitals NHS Foundation Trust, Swindon, UK

Correspondence: J L Peacock, Department of Clinical Epidemiology, Center for Population Medicine, Aarhus University Hospital and Aarhus University, Olof Palmes Allé 43-45, Aarhus N, DK-8200, Denmark, Email janet.peacock@clin.au.dk

Abstract: The growth in evidence-based medicine clearly benefits patient care but it is important that evidence is accessible and usable. Translating research results into usable, meaningful information can be challenging and impede the implementation of robust evidence. There are many aspects to making the statistical parts of a research study actionable for practicing clinicians and policy makers. These include ensuring that the right questions are asked, the study is designed appropriately, and results are transparent and applicable to the clinical setting. In this perspectives-style article we offer guidance on these considerations by highlighting several approaches that we have found effective for improving the interpretability and practical application of statistical findings. Our overarching aim is to stimulate interdisciplinary dialogue throughout the research process. Specifically, we discuss the interpretation of p-values, effect estimates, differences between means, scaling regression coefficients, unadjusted/adjusted estimates, Minimal Clinically Important Difference, absolute and relative risk, and suggest how clinical meaning can be enhanced by presenting the same information in different but complementary ways. We conclude with a recommendation that study teams prioritize interdisciplinary discussions around clinical meaningfulness throughout our research studies to maximize their clinical impact.

Keywords: evidence-based medicine, biostatistics, implementation, clinical practice, public health policy

Introduction

Clinical epidemiology is a basic science that supports clinicians in making informed prevention, diagnostic and treatment decisions.¹ A core principle of clinical epidemiology is the use of robust statistical methods alongside transparent interpretation and communication of statistical results. These results need to be useful and relevant, not only for clinicians making diagnostic and treatment decisions but also for population health researchers informing public health policy and disease prevention strategies. Many studies are statistically sound but are still difficult to translate into clinical decision-making because the results are not presented in a clinically interpretable way.

In this article we provide an informed perspective with illustrative examples rather than a comprehensive review, to discuss how applying statistical thinking at each stage of the research process can enhance the clinical meaningfulness of statistical findings. We use examples from different study designs to illustrate these issues considering approaches to presenting results and providing examples using i) simulation and ii) published data from observational studies based on primary data collection and registry data, as well as randomized clinical trials (RCTs). Through these examples, we illustrate strategies for maximizing the usefulness and impact of statistical results. The principles illustrated in this article are relevant to most quantitative research studies and complement work in the translational statistics space.² Our aim is to stimulate dialogue among clinicians, epidemiologists, statisticians and data scientists about how statistics, in the broadest sense, can be made more clinically relevant. An online document provides additional reading for the unreferenced statistical terms used in this article ([supplement 1](#)).

What Do We Mean by Clinically Meaningful?

Evidence-based medicine is the present-day term for the application of scientific evidence including principles of clinical epidemiology to the care of patients.^{1,3} Due to the growth of evidence-based medicine, *evidence* is readily available, often in large quantities that are difficult to synthesize and interpret. This challenge becomes particularly evident when clinicians seek to translate research findings into clinical decision-making in the era of personalized medicine. For example, once a diagnosis has been established, the clinician and patient must collaborate to determine the most appropriate care plan for that individual patient, weighing expected benefits against potential side effects. Doing so requires evidence that is interpretable, relevant and applicable to the clinical context. However, interpretation is complicated by the varied nature of clinical outcomes, which range from “hard” objective endpoints or outcomes such as mortality, to “soft” subjective endpoints such as symptom relief.⁴

It is important that results are accessible to clinicians who often work in fast-paced environments and, particularly for generalists, may treat patients with wide-ranging symptoms and diagnoses. Key findings therefore need to be presented clearly. When clinicians are unfamiliar with specific statistical concepts, there is a shared responsibility to foster learning and clarify what information is needed to enable results to be translated into the clinical framework.

While it is arguably the responsibility of statisticians/data scientists to present results in a way that aligns with clinical needs, and the responsibility of clinicians to understand the underlying medical concepts, it is essential that the questions asked and the results produced are relevant to the clinical need and therefore *clinically meaningful*. For these reasons it is essential that clinicians, epidemiologists and statisticians/data scientists work together throughout the planning, information-gathering, processing and dissemination stages to maximize the usefulness of the information produced.

Clinical Meaningfulness: The Research Process

In this section, we take a whistlestop tour through the traditional research process from formulating a research question, to selecting an appropriate study design, to conducting statistical analyses. We seek to highlight key principles that help ensure the statistics are clinically meaningful (Figure 1). We do not focus primarily on hypothesis-generating studies, although we return to them later in the article. From herein we refer to “statisticians and data scientists” as “statisticians” for conciseness.

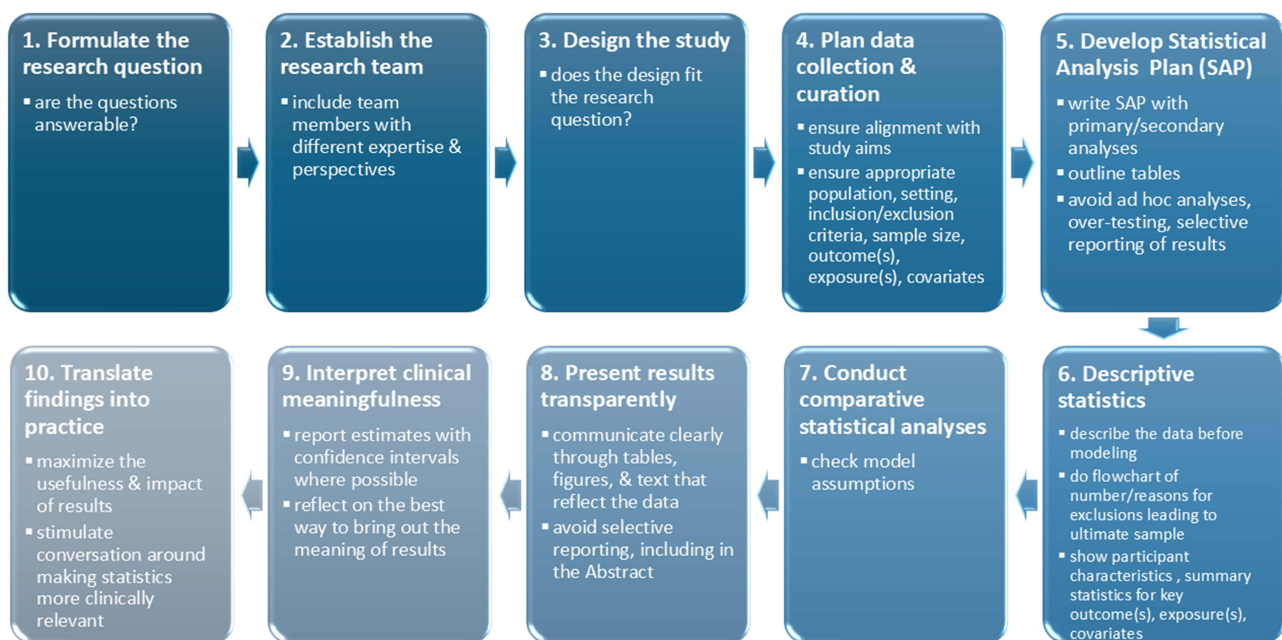


Figure 1 Summary of the research process.

At the design and planning stage, it is important to review the study aims and research questions to ensure that they are clearly defined and answerable. The next step is to determine whether the proposed study design addresses these questions and whether it is practical and feasible to implement. Following this, a data collection protocol must be developed if primary data are required, or a data curation protocol if existing data, such as a registry or administrative dataset, will be used. Data collection must align closely with the study aims and include appropriate outcome(s), exposures, covariates, population, setting, inclusion/exclusion criteria, and sample size considerations. Whether planning new data collection or assessing the feasibility of using existing data, there is a balance between obtaining sufficient data to address the research questions and collecting additional data that may be useful later. Expanding the data collection may, however, decrease data quality and increase missingness, eg. through participant fatigue, non-response, or added time and resource demands. If introducing new data collection tools or procedures, piloting them is advisable. Finally, data checking is vital, preferably during the study period, so that any issues can be identified and protocols adjusted while changes are still possible.

Before analyzing the data, it is good practice to write a statistical analysis plan (SAP) that specifies the primary and secondary analyses and includes outline tables. This helps prevent ad hoc analyses, excessive testing or data-dredging, and selective reporting of results. These SAPs are sometimes published with the study protocol⁵ or registered online⁶ before analysis to ensure transparent pre-planned methods. The SAP is developed collaboratively by the project team members, including clinicians, epidemiologists and statisticians, and is a useful bridge between clinical insight and methodological thinking. Revisions to a SAP may be needed for several reasons including: i) if preliminary descriptive analyses reveal very small cell sizes that could compromise confidentiality and legislation, ii) if unexpected missing data are identified in routine or registry sources, iii) if recruitment and/or outcome rates are different from those expected in the study population or iv) if new information comes to light that affects the design or analysis. Whenever the SAP is changed, it is important to document a clear chronology of changes. This may involve a formal amendment if the original protocol or SAP have been published⁷ to ensure ongoing transparency.

It is important to describe the data before doing any modeling. This includes presenting a flowchart showing who was excluded and ultimately included, summarizing participant characteristics and reporting key outcomes, exposures, and covariates. Such information enables clinicians to understand who participated in the study and helps them assess how applicable the findings may be to their own patient populations. When conducting analyses, it is important to examine the comparability of people with missing or excluded data relative to those included, and to verify that model assumptions are met. When appropriate, and particularly in large datasets, stratified descriptive analyses can be valuable before undertaking multivariable modeling. These analyses can reveal patterns of effect modification and confounding that may be obscured in regression models. This approach strengthens the validity and interpretability of subsequent regression models and enables researchers to engage more deeply with the underlying data structure.

When reporting results, we address the study aims by communicating clearly through tables, figures, text, that are true reflections of the data.⁸ Abstracts, in particular, are prone to selective reporting.⁹ If possible, results should be presented as estimates with measures of precision such as confidence or credible intervals, and p-values when they aid interpretation. Care is needed to avoid drawing inappropriate causal conclusions from observational data. All these aspects of design are crucial to producing work that is clinically interpretable. Nonetheless, we often spend limited time considering how best to convey the meaning of results and what they mean in practice. These issues will be the focus of the remainder of this article.

Clinical Meaningfulness: Interpreting Results – P-Values and Estimates

P-Values

It has long been considered good practice to report estimates with confidence intervals (or credible intervals) where possible, rather than relying on p-values as a measure of the size of difference or strength of association.¹⁰ The p-value is a measure of the weight of evidence for or against a null hypothesis being tested but is often presented and interpreted as a dichotomous entity, “significant” versus “non-significant”, using an arbitrary cut-off, eg $p < 0.05$, to define what is meant by “significant”. This practice is tantalizingly attractive in facilitating fast decision-making but is deeply flawed since

a p-value provides no information about the strength of effect and so is of limited use in summarizing and interpreting a study's findings. Another problem with the sole use of p-values to evaluate study results is that the p-value is influenced by the sample size, n , as well as the actual effect size. Thus, the same effect size can yield very different p-values depending on the value of n . For example: if Pearson's correlation coefficient = $+0.3$, $n=30$ then $p=0.11$, whereas the same correlation, $+0.3$ with $n=200$ has $p<0.001$ (Figure 2A and Figure 2B). This property of p-values, that they are related to the study size as well as the estimate, means that "not significant" cannot be interpreted as implying that there is "no difference" or "no association", as is commonly and incorrectly done. Further, the dichotomization of results as significant/non-significant that imply important/non-important is a factor contributing to the observed excess of statistically significant findings in the published literature, ie publication bias.

P-values therefore incorporate information about the size of effect and the sample size but they do not in themselves adjust for confounding and/or systematic biases which are common in non-randomized designs - an "adjusted p-value" usually has a different meaning to an "adjusted estimate". P-values can therefore give a false sense of certainty even when substantial bias may be present.

For all of these reasons, p-values cannot be interpreted as a measure of the size or clinical relevance of a study's effect, and their use as such is highly misleading. P-values cannot be used as a measure of clinical meaningfulness.¹⁰⁻¹⁴

Estimates

However, the interpretation of estimates of effect size can also be challenging – how large must an effect be to be interpretable as "large"? Cohen has provided one way to handle this by proposing a standardized effect size measure, $d = \text{mean}/\text{standard deviation}$, with descriptors such that an effect size of 0.8 is considered "large", 0.5 is "medium" and 0.2 is "small", assuming the data follow a normal distribution.¹⁵ Cohen's d is commonly applied and has clear merit when little information is available. However, it can be problematic when comparing populations, since a small effect may have substantially greater clinical impact in a sick or vulnerable population than in a healthy one.¹⁶ This can be illustrated using the example of birthweight. A mean difference of 180 g between infants born to non-smoking versus smoking women may have limited clinical implications among full-term infants, but the same difference may be far more consequential among preterm infants, who are already smaller and more vulnerable.

This issue has implications for the use of *Minimal Clinically Important Difference (MCID)*, in sample size calculations. The MCID represents the smallest effect size that corresponds to a clinically meaningful change in the selected

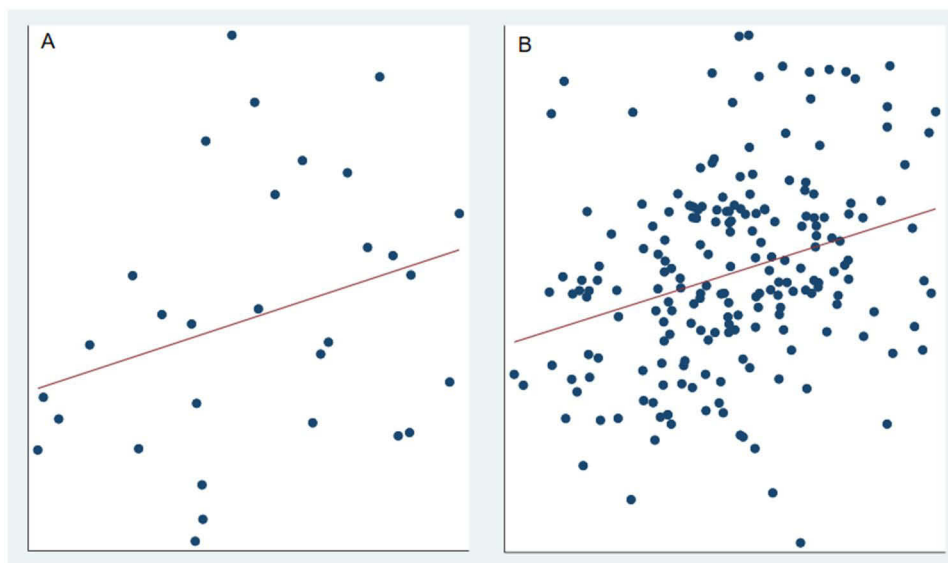


Figure 2 (A and B) Two datasets, same correlation, different sample size, different p-value. Datasets were simulated to have Pearson's correlation = $+0.3$. (A) $r = +0.3$, $n=20$, $p=0.11$. (B) $r = +0.3$, $n=200$, $p<0.001$.

patient outcome. However, the same MCID may not be appropriate across different populations. For example, in systems with optimized discharge processes where prolonged length of stay is not necessary, even a modest reduction in hospital length of stay can have substantial system-level benefits. In a busy general hospital, freeing beds more quickly improves patient flow from the emergency department and enhances care for future patients through a more efficient use of resources, without detriment to current patients.¹⁷

Interpreting Differences Between Groups in Mean Outcome for Continuous Data

When a study outcome is continuous, we often summarize the data using measures of centrality and spread such as mean, median, standard deviation, inter-quartile range. However, comparing two or more groups, and interpreting between-group differences in mean values can be challenging from a clinical perspective. Common strategies to aid interpretation include observing the direction of effect, ranking of the sizes of differences across several groups, comparing observed differences with those previously reported by others, and using the p-value to define whether a difference is meaningful. These approaches, particularly the reliance on p-values,¹¹ are, as mentioned, of limited value in providing clinical meaning. For these reasons researchers sometimes dichotomize continuous outcomes to improve interpretability. Ideally, dichotomization employs a clinically relevant cut-point to identify persons at high risk. The proportions of high-risk participants can then be compared directly as risk differences, or as risk ratios. Some cut-points derive from established diagnostic thresholds, such as using sustained systolic blood pressure ≥ 130 mmHg or diastolic blood pressure ≥ 80 mmHg to classify hypertension. These thresholds may, however, be somewhat arbitrary. For example, during the COVID-19 pandemic, temperature criteria differed slightly between the UK and the US due to their reliance on Celsius vs Fahrenheit scales and rounding practices.

A problem may arise when an analysis based on a dichotomized outcome replaces one based on the original continuous measure. Dichotomization discards information, obscures underlying relations and reduces statistical power. For these reasons statisticians strongly advise against it.^{18,19} A more robust strategy is to present the dichotomized analysis as a complementary secondary analysis alongside the primary analysis of the continuous outcome. Further examples are provided in more detail below.

Interpreting Differences in Mean Lung Function z-Scores

Z-scores are commonly used to standardize measurements, such as lung function, for age/sex/height. This allows a person's lung function measurement to be interpreted relative to the value expected in a healthy individual of the same age/sex/height. To illustrate the challenges of interpreting mean z-scores and suggest a way forward, we use data from a cohort study that examined differences in a portfolio of lung function measures in teenage boys and girls born extremely preterm.²⁰ The lung function measures were analyzed as z-scores and showed better mean z-scores in girls than boys for 11 of the 16 measures, although only four of these differences were statistically significant. In order to help interpret the differences in mean z-scores, the difference in percentage below the lower limit of normal (LLN; below 5th percentile) was calculated for all measures. For illustration, we only present results on FEV₁ (Table 1).

By presenting the outcome both as mean z-scores and as the percentage of individuals below the LLN, we can see that an adjusted mean difference of 0.36 z-score units corresponds to 23.3% of boys versus 13.4% of girls falling below the LLN. This difference in percentages was easier to interpret and considered to be clinically important.

We note that in this example, a distributional approach was used to calculate the percentages below LLN rather than dichotomizing the raw data. The distributional approach has the advantage of preserving the precision and power of the estimated difference in percentages by using a function of the normal distribution in a similar way to the calculation of quantiles.^{21,22}

Interpreting the Meaning of Regression Coefficients

How we present regression coefficients affects the ease with which we can infer their clinical meaning. Simple steps include ensuring units are specified and that the direction of any differences is stated (as in Table 1) and described in chapter 10 of Peacock, Kerry and Balise.⁸ Beyond that, it is helpful to consider the range of both exposures and outcomes

Table 1 Presenting FEV₁ z-Score Analyses as Means and % Below Lower Limit of Normal (LLN)²⁰

	Males	Females
Mean FEV ₁ z-score	-1.04	-0.49
Adjusted difference in means (M-F) (95% CI)	-0.36 (-0.65 to -0.06)	
% below LLN	23.3%	13.4%
Adjusted difference in % below LLN (M-F) (95% CI)	9.9% (3.6% to 16.3%)	

Notes: FEV₁: Forced Expiratory Volume in one second; LLN: lower limit of normal lung function; M=males, F=females; model adjusted for antenatal steroid use (y/n), birth weight, oxygen dependency at 36 wk PMA (y/n), postnatal steroids (y/n), major neonatal neurological impairment (y/n), mode of ventilation at birth (HFOV/CV), age at follow-up assessment, reached puberty (y/n), passive, or active exposure to smoking, presence of a smoker in the home (y/n), clustering due to presence of multiple births. This table is a portion of Table 2 in "Lung function and respiratory outcomes in teenage boys and girls born very prematurely". Harris C, Zivanovic S, Lunt A, Calvert S, Bisquera A, Marlow N, Peacock JL, Greenough A. *Pediatric Pulmonology*. 2020;55: 0.682–689²⁰ © 2020 Wiley Periodicals, Inc. Used with permission.

to see whether the coefficient would benefit from being scaled. Rather than reporting the effect per one-unit change in exposure—the default produced by regression models—researchers can present coefficients in terms of a “scaled” change that reflects a meaningful or realistic difference in exposure. This can highlight the practical impact of the exposure and facilitate comparison with other studies. The example below illustrates how scaling can help.

Scaling Estimates to Improve Interpretation

An analysis of data from the New Hampshire Birth Cohort Study compared how well two tobacco biomarkers, urinary cotinine and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL) in pregnant women, predicted outcomes of pregnancy.²³ Cotinine and NNAL were positively skewed and were analyzed on the log_e scale. The resulting coefficients were scaled to a one standard deviation increase in each of log cotinine and log NNAL to aid interpretation. Five continuous birth outcomes were analyzed: birthweight (g and z-score), gestational age (weeks), crown-heel length, and head circumference (z-score). The scaling allowed direct comparison of the predicted value of the two biomarkers, i.e. a comparison across the rows of Table 2. The paper also included two binary outcomes, small-for-gestational age, and preterm birth. The odds ratios for those were also similarly scaled.

Table 2 Regression Coefficients Scaled to One Standard Deviation Change in Exposure²³

Outcome	Regression Coefficient ^a (95% Confidence Interval)	
	Cotinine, ng/mL (log _e)	NNAL ^b , pmol/mL (log _e)
Birthweight ^c , gram	-55.5 (-93.2 to -17.8)	-57.8 (-96.6 to -18.5)
Birthweight z-score ^c	-0.11 (-0.18 to -0.04)	-0.11 (-0.18 to -0.04)
Gestational age ^c , weeks	-0.11 (-0.24 to 0.01)	-0.06 (-0.19 to 0.07)
Crown-heel length z-score ^c	-0.11 (-0.22 to -0.003)	-0.10 (-0.21 to 0.02)
Head circumference z-score ^c	-0.04 (-0.12 to 0.04)	-0.03 (-0.11 to 0.05)

Notes: a All models include the following covariates: maternal age, log_e body mass index, maternal education (high school vs beyond high school), parity (0 vs 1+); b NNAL= 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol; c regression coefficients scaled to one standard deviation increase in log_e cotinine (2.520) or log_e NNAL (1.012) as appropriate. This table is reproduced from "Assessing tobacco smoke exposure in pregnancy from self-report, urinary cotinine and NNAL: a validation study using the New Hampshire Birth Cohort Study". Peacock JL, Palys TJ, Halchenko Y, Sayarath V, Takigawa CA, Murphy SE, Peterson LA, Baker ER, Karagas MR. *BMJ Open* 2022;12:e054535. doi:10.1136/bmjopen-2021-054535²³ © 2022 with permission from BMJ Publishing Group Ltd.

Using \log_2

We note that sometimes researchers use logs to base 2, instead of base e, as we have here, for transforming skew exposure data. The mathematical effect of the transformation is unchanged but there is a useful interpretation of the regression coefficient on the \log_2 scale as the change in outcome associated with a doubling of the exposure. For example, Signes-Pastor 2022²⁴ used a \log_2 transformation of total urinary arsenic and reported that a doubling of total arsenic was associated with decreases in childhood cognitive abilities.

Interpreting Results Analyzed with the Outcome Analyzed on Logarithmic Scale

Logarithmic transformation is often applied to positively skewed outcomes such as serum creatinine to fulfill the requirement of a normal distribution when using a *t* test or normally distributed residuals when fitting a regression model. It is best practice to back-transform the data after analysis to get back to the natural scale and provide summary data that are more easily interpreted. In the case of a *t* test, the difference of means when back-transformed is the ratio of the geometric means, for example, comparing baseline serum creatinine in patients with peripheral vascular disease in survivors and those who died:

“Geometric means were: patients who survived, 97 compared to those who died, 108. The ratio of geometric means was 0.90, 95% CI: 0.79 to 1.02”.⁸

Note that the null hypothesis value is now 1, so the results indicate that mean serum creatinine values were 10% lower among survivors. The 95% CI indicates that the true population value might be as great as 21% lower or as high as 2% higher. See Peacock, Balise and Kerry, chapter 8 for more details.⁸

Interpreting Unadjusted and Adjusted Estimates

When a regression analysis includes covariates, it is often useful to report both the unadjusted estimates and covariate-adjusted estimates. We show below an extract from a table of the unadjusted and adjusted relative risks of any upper and any lower respiratory tract infection before and during the COVID-19 pandemic using data from the New Hampshire Birth Cohort Study.²⁵ Adjustment for several key covariates reduced the relative risk though not by much, strongly suggesting a real and sizable impact of stay-at-home orders in the State (Table 3).

Figure 3 provides another example, using data from the Danish National Patient Registry covering all Danish hospitals to investigate risk factors for VTE following a cancer diagnosis in a competing risks analysis.²⁶ This figure takes the tabulated data from the paper and displays them in a forest plot, making it easy to compare the unadjusted and adjusted subdistribution hazard ratios. The graphical presentation shows that after adjustment, most estimates move towards the null and become more precise (with narrower CIs). This simple presentation helps interpretation by clearly illustrating the impact of confounding on the estimates and thereby allowing the reader to judge whether the association is real.

Table 3 Caregiver Reported Upper and Lower Respiratory Tract Infections (RTI) Requiring a Doctor Visit in Children Age 0–11 Years Before and During the COVID-19 Pandemic²⁵

Outcome	Unadjusted RR ^a (95% CI)	Adjusted RR ^b (95% CI)
Any upper RTI	0.46 (0.39 to 0.54)	0.51 (0.42 to 0.64)
Any Lower RTI	0.52 (0.37 to 0.71)	0.61 (0.41 to 0.91)

Notes: a RR is relative risk; b adjusted for mother's age at enrolment, educational level (less than 11th grade, high school graduate, junior college graduate, college graduate, any post-graduate schooling), smoking status during pregnancy (yes/no), parity (0, 1, 2+), sex of infant, gestational age (continuous), birth weight (continuous), breast feeding (yes/no) and child daycare attendance (yes/no), season, time interval between interviews, age of the children at reporting. This table is reproduced from “Did children's symptoms and infections decline during the COVID-19 pandemic? A comparison of parental reports before and during the pandemic from a birth cohort study in New Hampshire, USA”. Peacock JL, Diaz-Coto S, Sayarath V, Madan J, Karagas M. *BMJ Public Health* 2023;1:e000069. doi:10.1136/bmjph-2023-000069²⁵ © 2023 with permission from BMJ Publishing Group Ltd.

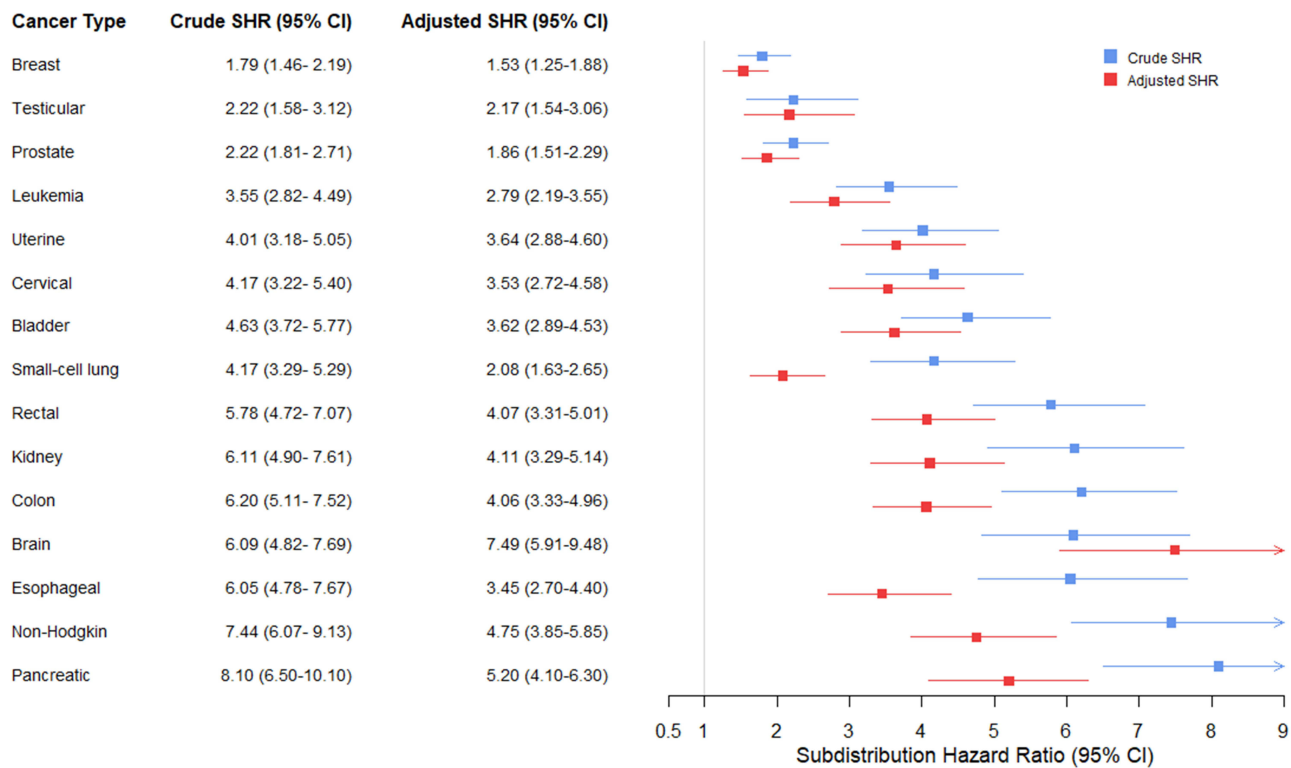


Figure 3 Unadjusted and adjusted subdistribution hazard ratios for risk factors for venous thromboembolism (VTE) following cancer diagnosis in a competing risks analysis.²⁶ The following covariates were adjusted for: age, sex, prior VTE (y/n), cancer stage (0-IV solid cancers), cancer type (categorized for VTE risk according to Khorana score classification).

There are many other examples of the challenges in uncovering the clinical meaning of regression analyses such as the modeling and interpretation of trends in regression models²⁷ among others.

Using and Interpreting Minimal Clinically Important Differences

In this final section we re-visit MCIDs which we introduced earlier. MCID is commonly used in study design to inform sample size calculations so that the target sample size is sufficient to be able to detect a clinically relevant change in a chosen patient outcome. This is important so that the findings from a study are most likely to inform clinicians and impact clinical decision making and patient outcomes. MCIDs are used in a range of study designs, including equivalence trials, which aim to demonstrate that two interventions produce clinically similar effects within a predefined margin of equivalence, and in non-inferiority trials that aim to demonstrate that a new intervention is not worse than a current intervention by more than a prespecified non-inferiority margin. MCIDs are also used as a benchmark against which an observed effect can be compared in order to determine whether or not it is clinically relevant.

Illustrating Clinical Impact in Vulnerable versus Healthy Populations

While MCIDs are useful for interpreting the clinical relevance of statistical results, challenges arise when applying an MCID derived from one population to another with different characteristics. In particular, it has been shown that a fixed exposure effect will have a greater clinical impact in vulnerable populations than in healthier ones.^{16,28} This has implications for interpreting study findings in sub-populations defined by sociodemographic factors, as illustrated using data from over 28,000 mother/child pairs from the US Environmental influences on Child Health Outcomes (ECHO) study.²⁸

A simulation study using ECHO data examined the impact of an exposure that reduced mean birthweight by the same amount across sub-populations defined by sociodemographic categories. Below, we show the results for a 50g reduction in mean birthweight stratified by race. When the effect size is expressed as the change in mean birthweight, the effect is

identical across groups (by design). However, when the effect is translated into its impact on the percentage of infants with low birthweight (LBW), the effect size becomes larger in the more vulnerable groups—those with lower mean birthweight. Figure 4 illustrates this, using data from this study showing that the effect size as LBW % points varies directly by race. For example among White participants, where the mean birthweight is approximately 3373g, the effect on LBW is around one percentage point. In contrast, among Black participants, whose mean birthweight is considerably lower at 3131g, the effect on LBW doubles to around two percentage points.

Risks

Relative vs. Absolute Risks

Absolute risk varies with the underlying prevalence of disease in the study population. An intervention that reduces risk by 50%, ie $RR=0.5$, will look very different in absolute terms if baseline risk is 2% versus 20%, ie $2\% \rightarrow 1\%$, $20\% \rightarrow 10\%$. If these differences are not taken into account, comparisons of clinical relevance across populations can be misleading. A well-known example is the association between CT-scan radiation exposure and subsequent cancer risk. Even if the increase in risk for an individual child undergoing a single CT scan is small, widespread use of CT scans in children with head injuries could lead to a meaningful increase in cancer incidence at the population level. For these reasons, it is important to report both relative and absolute risks to maximize the information available to make clinical decisions at both the individual patient and population level.

Odds ratios (ORs) are often used as proxies for relative risks (RRs), but caution is needed since unless the event in question has very low prevalence (rare disease assumption), the RR and OR calculated from the same set of data have different values, with the OR taking values farther from the null.

NNT/NNH

Number Needed to Treat (NNT) and Number Needed to Harm (NNH) are complementary measures that quantify the clinical impact of an intervention. NNT represents the average number of patients who must receive the intervention to prevent one additional adverse outcome compared with a control group. NNH reflects the number of patients who need to be treated for one additional harmful event to occur. Both measures are derived from absolute risk differences and provide an intuitive way to balance benefits and risks in clinical decision-making. A low NNT suggests high treatment efficacy, whereas a low NNH signals a higher likelihood of harm. Interpreting these measures together is essential for evaluating the net clinical benefit of a therapy. The risk–benefit ratio, often expressed as NNH divided by NNT, provides

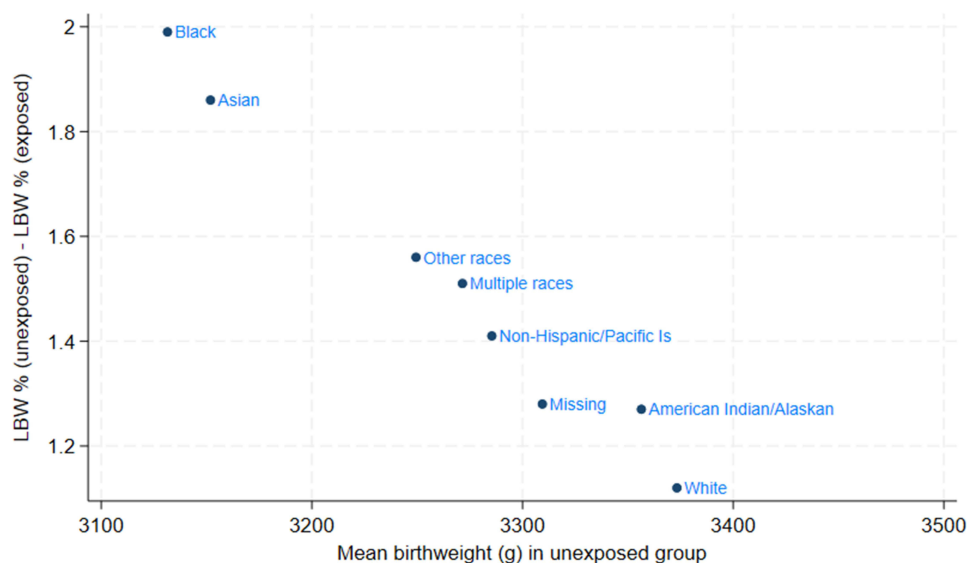


Figure 4 Effect an exposure reducing mean birthweight by 50 g on LBW stratified by Race.²⁸

Table 4 NNT and NNH from RCT of Treatment for Gestational Diabetes Mellitus on Pregnancy Outcomes²⁹

Outcome	Proportion in Treated Group	Proportion in Routine-Care Group	NNT (95% CI)	NNH (95% CI)
Any serious perinatal complication	7/506 0.014	23/524 0.044	34 (20 to 103)	
Admission to neonatal unit	357/506 0.706	321/524 0.613		11 (7 to 29)

a single metric summarizing whether the expected benefit outweighs the potential harm; values greater than 1 indicate a favorable balance toward benefit. An example is given below from a RCT of the effect of treatment of gestational diabetes mellitus on pregnancy outcomes.²⁹ The main outcome was any serious perinatal complications (any vs none: death, shoulder dystocia, bone fracture, nerve palsy). NNT and NNH (Table 4) have been calculated from the results given in the paper.

Other Issues and Considerations

Here, we comment briefly on the challenges involved in deriving clinical meaning from high-dimensional ‘omics studies. These studies may be hypothesis-generating rather than hypothesis-testing which brings challenges for interpretation and reproducibility in the presence of multiple testing. Commonly approaches to adjust for multiple testing include Bonferroni, False Discovery Rate, and methodological work on adapting these approaches to complex ‘omics datasets is an active area of research (see Ebrahimpour³⁰). A detailed review is beyond the scope of this article. The interpretation of small effects has been discussed earlier and is raised by Breton et al in relation to epigenetic studies in children’s environmental health where small observed effects may be considered meaningful when supported by consistency across studies.³¹

There are many other topics that could be included in a comprehensive review of clinical meaningfulness including the Bayesian decision framework for communicating clinical evidence and causal inference but these are beyond the scope of this article.

Finally, ensuring clinical meaningfulness is just the first step in translating robust evidence into change in practice.³²

Conclusions

In this paper, we have described and illustrated ways in which clinical meaning can be affected. First, it is challenging to infer robust clinical meaning where there are deficiencies in study design, data collection/measurement, or data analyses. However, even if these aspects are all adequately handled, sound clinical meaning can be obscured by avoidable issues in how results are presented for example inadequately annotated tables and figures, or poor formatting. As we have shown here, presenting the same information in complementary but different ways can often enhance clinical meaningfulness and support clearer interpretation.

Improving clinical meaningfulness takes time, for example to discuss, review, and revise results, and sometimes this step is overlooked because of time-pressures. We recommend that study teams prioritize interdisciplinary discussions of clinical meaningfulness from the outset and throughout the research process to ensure that the clinical impact of our work is maximized.

Acknowledgments

Tables 1–3 contain material from our own publications. Permission to reproduce each was obtained using RightsLink. The original publication has been fully cited in each case and a permissions and copyright statement made as specified by the publisher. All figures included are our own work, were generated for this paper and have not been published in this form elsewhere.

Disclosure

Henrik Toft Sørensen has made paid evaluations for University of Oslo, the Norwegian Research Council, the Independent Research Fund Denmark, and the European Research Council. He also reports that his institution, The Department of Clinical Epidemiology, Aarhus University, receives funding for other studies in the form of institutional research grants to (and administered by) Aarhus University. The Department of Clinical Epidemiology, Aarhus University confirms that none of these studies have any relation to the present study. No other authors have any conflicts of interest in this work.

References

- Fletcher GS, Lippincott W, Wilkins. *Clinical Epidemiology: The Essentials*. 6th ed. Philadelphia: Wolters Kluwer; 2021.
- McCabe GP, Newell J. The art of translational statistics. *Stat Us*. 2022;11(1):e519.
- Guyatt G, Rennie D, Meade M, Cook D; American Medical Association. *Users' Guides to the Medical Literature. A Manual for Evidence-Based Clinical Practice*. Third edition ed. McGraw-Hill Education Medical: New York; 2015.
- Rothman. *Modern Epidemiology*. 4th ed. Wolters Kluwer; 2021.
- Bearne L, Galea Holmes M, Biele J, et al. Motivating Structured walking Activity in people with Intermittent Claudication (MOSAIC): protocol for a randomised controlled trial of a physiotherapist-led, behavioural change intervention versus usual care in adults with intermittent claudication. *BMJ Open*. 2019;9(8):e030002. doi:10.1136/bmjopen-2019-030002
- Blondeau L. STATISTICAL ANALYSIS PLAN: Protocol number: MHIPS-003 Colchicine Cardiovascular Outcomes Trial COLCOT. 2019. Available from: https://cdn.clinicaltrials.gov/large-docs/94/NCT02551094/SAP_000.pdf. Accessed May 27, 2026.
- Stringer D, Gardner LM, Peacock JL, et al. Update to the study protocol, including statistical analysis plan, for the multicentre, randomised controlled OuTSMART trial: a combined screening/treatment programme to prevent premature failure of renal transplants due to chronic rejection in patients with HLA antibodies. *Trials*. 2019;20(1):476. doi:10.1186/s13063-019-3602-2
- Peacock JL, Kerry SM, Balise RR. *Presenting Medical Statistics From Proposal to Publication*. 2 ed. Oxford: Oxford University Press; 2017.
- Peacock PJ, Peters T, Peacock JL. How well do structured abstracts reflect the articles they summarize. *Eur Sci Edit*. 2009;35:3–5.
- Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J Clin Res Ed*. 1986;292(6522):746–750. doi:10.1136/bmj.292.6522.746
- Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337–350. doi:10.1007/s10654-016-0149-3
- Rothman KJ. Disengaging from statistical significance. *Eur J Epidemiol*. 2016;31(5):443–444. doi:10.1007/s10654-016-0158-2
- Lang JM, Rothman KJ, Cann CI. That confounded P-value. *Epidemiology*. 1998;9(1):7–8. doi:10.1097/00001648-199801000-00004
- Wasserstein RL, Lazar NA. The ASA's statement on P-values: context, process, and purpose. *Am Stat*. 2016;70(2):129–131. doi:10.1080/00031305.2016.1154108
- Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, N.J: L. Erlbaum Associates; 1988.
- Peacock JL, Lo J, Rees JR, Sauzet O. Minimal clinically important difference in means in vulnerable populations: challenges and solutions. *BMJ Open*. 2021;11(11):e052338. doi:10.1136/bmjopen-2021-052338
- Hirani R, Podder D, Stala O, Mohebbpour R, Tiwari RK, Etienne M. Strategies to reduce hospital length of stay: evidence and challenges. *Medicina*. 2025;61(5):922.
- Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006;332(7549):1080. doi:10.1136/bmj.332.7549.1080
- Senn S. Disappointing dichotomies. *Pharm Stat*. 2003;2(4):239–240. doi:10.1002/pst.90
- Harris C, Zivanovic S, Lunt A, et al. Lung function and respiratory outcomes in teenage boys and girls born very prematurely. *Pediatr Pulmonol*. 2020;55(3):682–689. doi:10.1002/ppul.24631
- Peacock JL, Sauzet O, Ewings SM, Kerry SM. Dichotomising continuous data while retaining statistical power using a distributional approach. *Stat Med*. 2012;31(26):3089–3103. doi:10.1002/sim.5354
- Sauzet O, Breckenkamp J, Borde T, et al. A distributional approach to obtain adjusted comparisons of proportions of a population at risk. *Emerg Themes Epidemiol*. 2016;13:8. doi:10.1186/s12982-016-0050-2
- Peacock JL, Palys TJ, Halchenko Y, et al. Assessing tobacco smoke exposure in pregnancy from self-report, urinary cotinine and NNAL: a validation study using the New Hampshire Birth Cohort Study. *BMJ Open*. 2022;12(2):e054535. doi:10.1136/bmjopen-2021-054535
- Signes-Pastor AJ, Romano ME, Jackson B, et al. Associations of maternal urinary arsenic concentrations during pregnancy with childhood cognitive abilities: the HOME study. *Int J Hyg Environ Health*. 2022;245:114009. doi:10.1016/j.ijheh.2022.114009
- Peacock JL, Diaz-Coto S, Sayarath V, Madan J, Karagas M. Did children's symptoms and infections decline during the COVID-19 pandemic? A comparison of parental reports before and during the pandemic from a birth cohort study in New Hampshire, USA. *BMJ Public Health*. 2023;1(1):e000069. doi:10.1136/bmjph-2023-000069
- Mulder FI, Horvath-Puho E, van Es N, et al. Venous thromboembolism in cancer patients: a population-based cohort study. *Blood*. 2021;137(14):1959–1969. doi:10.1182/blood.2020007338
- Bellavia A, Murphy SA. Linearity in regression models: meaning, implications, and how to handle it. *Circulation*. 2025;152(25):1739–1741. doi:10.1161/CIRCULATIONAHA.125.073647
- Peacock JL, Coto SD, Rees JR, et al. Do small effects matter more in vulnerable populations? An investigation using environmental influences on child health outcomes (ECHO) cohorts. *BMC Public Health*. 2024;24(1):2655. doi:10.1186/s12889-024-20075-x
- Crowther CA, Hiller JE, Moss JR, et al. Effect of treatment of gestational diabetes mellitus on pregnancy outcomes. *N Engl J Med*. 2005;352(24):2477–2486. doi:10.1056/NEJMoa042973
- Ebrahimpoor M, Menezes R, Xu N, Goeman JJ. Multiple testing of mix-and-match feature sets in multi-omics. *Stat Med*. 2026;45(1–2):e70367. doi:10.1002/sim.70367

31. Breton CV, Marsit CJ, Faustman E, et al. Small-magnitude effect sizes in epigenetic end points are important in children's environmental health studies: the children's environmental health and disease prevention research center's epigenetics working group. *Environ Health Perspect*. 2017;125(4):511–526. doi:10.1289/EHP595
32. Rubin R. It takes an average of 17 years for evidence to change practice—the burgeoning field of implementation science seeks to speed things up. *JAMA*. 2023;329(16):1333–1336. doi:10.1001/jama.2023.4387

Clinical Epidemiology

Publish your work in this journal

Clinical Epidemiology is an international, peer-reviewed, open access, online journal focusing on disease and drug epidemiology, identification of risk factors and screening procedures to develop optimal preventative initiatives and programs. Specific topics include: diagnosis, prognosis, treatment, screening, prevention, risk factor modification, systematic reviews, risk & safety of medical interventions, epidemiology & biostatistical methods, and evaluation of guidelines, translational medicine, health policies & economic evaluations. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use.

Submit your manuscript here: <https://www.dovepress.com/clinical-epidemiology-journal>

Dovepress

Taylor & Francis Group