Invited review: Maximizing value and minimizing waste in clinical trial research in dairy cattle: Selecting interventions and outcomes to build an evidence base

Jan M. Sargeant,1* Annette M. O’Connor,2 Stephen J. LeBlanc,1 and Charlotte B. Winder1
1Department of Population Medicine, University of Guelph, Guelph, ON N1G 2W1 Canada
2Department of Large Animal Clinical Sciences, College of Veterinary Medicine, Michigan State University, East Lansing 48824

ABSTRACT

Clinical trials are a valuable study design for evaluating interventions when it is ethical and feasible for investigators to randomly allocate study animals to intervention groups. Researchers may choose to evaluate the comparative efficacy of intervention groups for their effect on outcomes that are relevant to the specific objectives of their trial. However, the results across multiple trials on the same intervention and with the same outcome should be considered when making decisions on whether to use an intervention, because the results of a single trial are subject to sampling error and do not reflect all biological variability. The objective of this review was to provide an overview of important concepts when selecting intervention groups and outcomes within a randomized controlled trial, and when building a body of evidence for intervention efficacy across multiple trials. Empirical evidence is presented to highlight that integrating and interpreting the efficacy of an intervention across trials is hindered by a lack of replication of interventions across trials. Inconsistency in the outcomes and their measurement among trials also limits the ability to build a body of evidence for the efficacy of interventions. The development of core outcome sets for specific topic areas in dairy science, updated as necessary, may improve consistency across trials and aid in the development of a body of evidence for evidence-based decision-making.

Key words: experiment, clinical trial, interventions, outcomes

INTRODUCTION

Research is essential to acquiring new knowledge and is a pillar of evidence-based decision-making. The annual investment in biomedical research totals billions of dollars globally and involves millions of individuals (Macleod et al., 2014). However, it has been estimated that 85% of biomedical research in humans is wasted (Chalmers and Glasziou, 2009). Reasons for research wastage include addressing research questions that are not relevant, inadequate study design and methods, inaccessibility of research reports, and biased or unusable results.

Research related to dairy cattle and dairy products also represents a substantial investment; in 2021 alone, over 950 primary research articles were published in the Journal of Dairy Science. The extent of research wastage in dairy science is unknown. However, in an editorial summarizing the results of 10 systematic reviews related to management or preventive antibiotics in multiple livestock species, including dairy cattle, the authors noted consistent issues across the included trials related to a lack of replication of interventions, inconsistent outcomes, incomplete reporting, and design concerns related to the potential for bias (Sargeant et al., 2019). All of these factors contribute to research wastage. Thus, a reflection on issues related to potential research wastage and how the dairy research community can maximize the value of research is warranted.

Research wastage is probably an issue for all types of research, and the potential causes and solutions are broad. However, one area where the return on investment in research can be increased is by the selection of interventions and outcomes in clinical trials. Although these topics are of relevance to multiple animal species, we focus here on trials conducted in dairy cattle to match the mandate and audience of the Journal of Dairy Science. Interventions would include pharmaceutical products, biologics such as vaccines, ration formulations, management practices, and so on. Well-conducted randomized controlled trials provide the highest level of evidence among the primary research designs for evaluating interventions when it is ethical and feasible to randomly allocate study subjects to intervention groups (Burns et al., 2011; Sargeant et
The aims of this review are to provide an overview of important concepts when selecting intervention groups and outcomes within a trial, to review evidence from the published literature on whether interventions and outcomes are being selected effectively to build a body of evidence, and to propose solutions to allow researchers to maximize the value of clinical trials in dairy cattle.

**SELECTION OF INTERVENTIONS**

**Defining the Trial Purpose**

Before selecting the specific comparison group or groups for the intervention of interest, researchers must determine the purpose of the trial. Trials may be designed to evaluate a hypothesis related to superiority (whether the intervention of interest is superior to another intervention), equivalence (whether the intervention has the same efficacy as another intervention), or noninferiority (whether the intervention is no worse than the comparison group; Christensen, 2007; Roumeliotis et al., 2020). Because these designs differ, it is critical to have sound and explicit hypothesis at the outset, to consider whether 2-sided or 1-sided hypothesis testing is appropriate for the hypothesis, and to explain the rationale for this choice explicitly in the paper. An intervention that is equivalent or noninferior to another intervention may be a viable alternative if, for example, there are fewer side effects, shorter milk withholding period, or lower cost. The null hypothesis for a trial will differ depending on the trial purpose. For a superiority trial, the null hypothesis is no difference between the intervention groups. For an equivalence trial, the null hypothesis is that the intervention groups differ by at least a prespecified amount. For a noninferiority trial, the null hypothesis is that the intervention of interest is superior to another intervention by more than a prespecified amount. For a noninferiority trial, the null hypothesis is that the intervention groups differ by at least a prespecified amount. For a noninferiority trial, the null hypothesis is that the intervention groups differ by more than a prespecified amount (Freise et al., 2013; Roumeliotis et al., 2020). This will result in different sample size calculations, with superiority trials typically having smaller sample sizes and equivalence trials having the largest sample sizes (Christensen, 2007; Roumeliotis et al., 2020). The study purpose also affects the analysis and the interpretation of the trial results. Intention to treat analysis means retaining all subjects that were assigned to an intervention in the intervention group where they were allocated, even if they did not complete the experimental intervention exactly as intended, whereas per-protocol analysis includes only subjects that received the planned intervention. Intention to treat analysis may better reflect real-world conditions and expected effects, given imperfect compliance with interventions. In either case, exclusion criteria should be explicit in the manuscript and the numbers excluded for each reason reported. Intention to treat analysis is the preferred approach for superiority trials, with both per-protocol and intention to treat analysis recommended for equivalence and noninferiority trials (Christensen, 2007; Roumeliotis et al., 2020).
Based on sample size calculations and inference in most trials, it might reasonably be assumed that the purpose of most trials is to evaluate superiority. However, this aspect of trial design is poorly reported in the literature. For instance, a recent systematic review and network meta-analysis included 39 trials evaluating preventive antibiotics at dry-off for reducing new IMI at calving (Winder et al., 2019a). Network meta-analysis is an extension of meta-analysis wherein the relative efficacy of multiple interventions can be estimated across studies, including treatments that were not compared directly in the primary trials (Hu et al., 2019). Of the 39 included trials in Winder et al. (2019a), none included an explicit statement that the purpose was to evaluate superiority, 1 specified equivalence, and 4 trials had the stated purpose to evaluate noninferiority. The purpose of the trial was not stated in 34 of the 39 published trials.

**Defining the Intervention Type**

Comparison groups in clinical trials include untreated groups, placebo or sham intervention groups (nonactive controls), or alternative intervention groups (e.g., the current diet, practice, or standard of care, or a different dose or method of administration of the intervention of interest). The type of comparison group that is appropriate will depend on the purpose of the trial. When a placebo or sham treatment is used as the comparison group, the results are interpreted as whether the intervention of interest is better than, equivalent to, or noninferior to no treatment. Although it is possible to use a nonactive control for any trial purpose, it only makes sense to use an untreated or nonactive comparison group if the trial purpose is to evaluate superiority. In situations where interventions have consistently (i.e., across multiple trials) shown an inability to outperform an untreated group, the use of a placebo group is appropriate. However, end users are often interested in knowing what intervention to use among several choices (i.e., the relative efficacy of available interventions), rather than whether to intervene at all. There also may be ethical and economic concerns with comparing a new intervention to an untreated or nonactive treatment group if an alternative or established intervention has been consistently found to be effective (Mann and Djulbegovic, 2003; IHC, 2020). Given that decision-makers may be more interested in the relative efficacy of available intervention options, it may be tempting to design a trial comparing an intervention of interest to an alternative treatment (i.e., head-to-head comparisons of active ingredients or interventions). However, in the absence of evidence that the alternative treatment is superior to an untreated group, the value of head-to-head comparisons is limited. If the new intervention is found to be superior, equivalent, or noninferior to the alternative treatment, it is necessary to know whether the alternative treatment is superior to no treatment. Otherwise, it is possible that both the new and alternative interventions are highly effective or that neither is better than an untreated group (Temple and Ellenberg, 2000; Freise et al., 2013). For example, in a network meta-analysis on the efficacy of antibiotics to treat clinical mastitis caused by *Staphylococcus* species (Winder et al., 2019d), a single trial comparing intramammary cloxacillin and ampicillin to parenteral penethamate was identified. There were no other trials identified that compared either intramammary cloxacillin and ampicillin to an untreated group or parenteral penethamate to an untreated group. Therefore, if intramammary cloxacillin and ampicillin was found to be equivalent (for instance) to parenteral penethamate, it is possible that neither intervention is effective (i.e., superior to no treatment).

An efficient approach, if researchers aim to compare active interventions without consistent evidence of efficacy for either, would be to add a third arm that was a placebo arm to a trial comparing 2 active interventions. This allows the researcher to confirm that the alternative treatment is superior to no intervention. Such an approach would maximize the value of the trial. Without the third arm, if the interventions are found to be equivalent, the results will be uninterpretable and wasted. Therefore, the type of comparison group not only affects the interpretation of a single trial, but is important in building a body of evidence on the relative efficacy of multiple intervention options.

**Describing the Intervention**

Regardless of the intervention groups that are compared, the trial report should describe each intervention group in sufficient detail so that another researcher or person using the intervention could replicate the intervention (O’Connor et al., 2010). The information that is needed to allow replication will vary based on the type of intervention. For instance, for pharmaceutical interventions, it is recommended that the compound name, concentration, dose, delivery matrix, frequency of administration, and route of administration are all reported (Sargeant et al., 2010). For nutrition studies, the basal diet, nature and amount of treatments, and method of feed delivery should be described in sufficient detail to allow others to exactly reproduce the experiment. Recommendations are available for reporting and analysis of reproduction trials in dairy cattle (Lean et al., 2016). Reporting of interventions is complete for most dairy trials; for example, an evalu-
Linking Interventions Across Trials

Building a body of evidence not only requires that research is replicated, but also that information can be summarized across trials of the same and different interventions for the same outcome. Summarizing across trials also assumes that the populations and environments are similar enough to make such synthesis valid. For instance, if the same intervention is evaluated against the same outcome for dairy cattle housed in large dry-lot herds in a hot climate in one trial and in dairy cattle housed indoors in small herds in a cold climate in another trial, it may not be meaningful to combine the results across these trials. Nonetheless, when the populations and environments are appropriately similar, summarizing the results across studies may be done informally by reading the results of multiple trials, or formally using pairwise meta-analysis for comparisons between 2 interventions (Sargeant and O’Connor, 2020) or network meta-analysis to evaluate relative (comparative) efficacy of all available intervention options (Hu et al., 2019; Hu et al., 2020). To include the results of a trial in a network meta-analysis, the trial needs to include at least one intervention group that corresponds to at least one intervention group in another trial. To illustrate this concept, Figures 1 to 3 are network diagrams created using data from 3 recently published network meta-analyses on topics related to mastitis in dairy cattle (Winder et al., 2019a,c,d). In these diagrams, each circle (node) represents an intervention. A line joining 2 nodes indicates that a comparison between these 2 interventions was evaluated in at least one trial. The thickness of the lines reflects how many comparisons between those interventions were used in the network. Labels on each circle reflect a different acronym for a given intervention group, with NAC representing a nonactive control group.

Figure 1 uses data collected as part of a network meta-analysis on the comparative efficacy of antimicrobial treatment in dairy cattle at dry-off for IMI at calving (Winder et al., 2019a). This figure represents a fully connected network, where all interventions can be compared with each other through direct or indirect pathways. For example, tilmicosin (TIL) was directly compared with a NAC in one or more trials as indicated by the line connecting these interventions. However, TIL was not directly compared with novobiocin in any of the trials included in the network, as shown by the lack of a line directly connecting these interventions. However, TIL may be indirectly compared with novobiocin through a common comparison group in trials evaluating these interventions, for instance, based on comparisons of each of these products to a NAC. Therefore, because all the interventions are connected to each other either directly or indirectly, and there is a robust number (n = 36) of trials evaluating the efficacy of antimicrobial interventions on IMI, it is possible to evaluate the comparative efficacy of antibiotic options across the body of evidence. Network meta-analysis increases the value of investment in the original research because not only is the comparison from each trial available, but multiple comparisons and rankings between interventions can be obtained from the network of evidence, which provides additional information to the end users.
Figure 2, from a network meta-analysis on the comparative efficacy of teat sealants to prevent IMI at calving (Winder et al., 2019c), shows a network where most interventions are linked directly or indirectly to each other. However, there was one trial where neither of the 2 interventions that were evaluated (penicillin alone versus which is a teat sealant plus penicillin) were included in any other published trial. In this example, we can estimate the efficacy of penicillin alone compared with teat sealant plus penicillin, with the caveats related to a single trial results and sampling error. However, it is not possible to estimate the relative efficacy of these interventions to any other intervention in the larger network. The information “return on investment” of this trial would have been improved greatly by including an intervention group that had been evaluated in at least one other trial. Even if based on only one trial, end users would have been able to obtain estimates.
of comparative efficiency for penicillin alone and teat sealant plus penicillin to all other interventions. As the study stands, no such information can be obtained and no other value can be extracted from the investment in the research. It is recognized that including additional intervention groups is associated with additional cost. However, it is hoped that a strong argument for the importance of an additional group to the aim of building an evidence base and maximizing the value of the research investment will persuade research funders to provide the needed resources.

Finally, Figure 3 illustrates a highly disconnected network, where many of the interventions are unconnected across trials, because they lack a common comparison group. This figure is from a network meta-analysis of the comparative efficacy of antimicrobials for the treatment of clinical mastitis in lactating dairy cattle (Winder et al., 2019d), for the outcome of bacteriologic cure of coagulase-negative Staphylococcus species. This figure shows that although trials have been conducted using 17 different intervention groups, at most only 7 intervention groups can be compared for relative ef-
ficiency. Based on the trials illustrated in this figure, it is not possible to build a body of evidence on the comparative efficacy of available antimicrobials for treating clinical mastitis.

To maximize the value of the research investment, and to minimize research wastage, all trials would ideally include at least one intervention group that has been evaluated in a previous trial. Considering this broader context in designing studies can improve the situation. Although network meta-analysis are available for a few interventions in dairy science, there are few such syntheses in this discipline. To our knowledge, there are 5 published network meta-analyses in dairy cattle that include intervention network diagrams (Jacobs et al., 2019; Winder et al., 2019 a,c,d; Nobrega et al., 2020). As the number of published network meta-analyses increases, it would be useful to have a repository where intervention networks can be made available, recognizing that they would need to be updated as additional studies in the field become available. Such a repository
could be used by researchers planning trials to aid in selecting comparison groups that will allow a body of evidence to be built.

**SELECTION OF OUTCOMES**

Selection of one or more outcomes to compare between intervention groups is a key decision during the design of a trial. However, “outcomes” need to be considered and defined on several levels. First, researchers need to identify the relevant outcome domains; in dairy research, these could include health, productivity, welfare, or product attributes. Within each of the outcome domains that are included in a trial, researchers will need to select one or more conceptual outcomes. Conceptual outcomes describe the concept that the researcher wishes to evaluate, but conceptual outcomes cannot be directly measured. Operational outcomes can be directly measured and must be identified for each conceptual outcome. Once the operational outcome or outcomes are identified, the researcher will need to identify appropriate outcome measures. An example will serve to illustrate these concepts. Suppose that a researcher wishes to evaluate the usefulness of a new product to treat diarrhea in calves. They may be interested in evaluating the outcome domain of health. Within the domain of health, the conceptual outcome may be recovery. This conceptual outcome could be operationalized as a return to normal fecal consistency or as normal hydration status, or both. For the operational outcome of return to normal fecal consistency, the outcome measure may be a specified value using a 4-level fecal scoring scale. Regardless, all outcome measures should include a consideration of what is measured, how it is measured, who conducts the measurement, and the time for measuring the outcome (time at risk).

**Considerations When Selecting an Outcome Measure**

When selecting an outcome measure, the researcher will need to determine whether to use a direct or indirect measure and whether the outcome is measured subjectively or objectively. In healthcare, direct outcomes are referred to as clinical outcomes, although that terminology may not be relevant for all areas of research in dairy science. A clinical outcome reflects how an animal feels, functions, or survives (McLeod et al., 2019; FDA 2021). Clinical outcomes in dairy trials would include measures of morbidity and mortality. Mortality is an objectively measured outcome (i.e., can be measured impartially and consistently, assuming that euthanasia is not a component), but measuring morbidity generally involves judgment and so is considered a subjectively measured outcome. Biomarkers (e.g., plasma concentrations of metabolites or hormones; Hailemariam et al., 2014) are objectively measured indicators of a biological or pathological process, although they are not a direct measure of how an animal feels, functions, or survives. Surrogate outcomes are outcomes that are sufficiently closely associated with a clinical outcome to be used as a proxy (McLeod et al., 2019). For instance, SCC has been used as a proxy for IMI.

Another type of outcome measure is a composite endpoint (outcome), which involves combining multiple related outcomes into a single measure (Oyama et al., 2017; Vetter and Mascha, 2017). Examples in dairy trials include all-cause morbidity or mortality, or IMI with any pathogen. Composite endpoints can be used to increase statistical power for rare outcomes or to describe overarching outcomes that are relevant (e.g., no disease vs. any disease). However, there are caveats; for instance, if an intervention increased morbidity due to metritis, but decreased morbidity due to mastitis, then the results of any changes to all-cause morbidity would be difficult to interpret. Combining lameness and dystocia under “all-cause morbidity” risks blurring conditions with different causes. Alternatively, if an intervention decreased IMI due to one pathogen, but had no effect on other pathogens, the use of a composite “all pathogen IMI” endpoint could mask an important finding.

**Determining the Number of Outcomes to Include in a Trial**

Investigators must decide how many outcomes to include in a trial. In dairy trials, multiple outcome domains (e.g., health and productivity and welfare) may be relevant for decision-making around the use of an intervention. An individual trial may then include one or more conceptual and operational outcomes within each of the selected domains. Finally, an operational outcome may be measured in multiple ways within the same trial. For example, “mastitis in early lactation” may be measured as clinical mastitis during the first 7 d postcalving or as IMI within the first 7 d postcalving. Multiple outcomes were included in almost all (91/100) clinical trials related to health and production in livestock (Sargeant et al., 2009), with a mean of 9.5 (range 1 to 41) outcomes per trial. Wareham et al., (2017) reported that the median number of outcomes in trials in cattle (dairy or beef) was 5.5, with a range of 1 to 36. Although it may increase the efficiency of data collection, having multiple outcomes has important implications for study design and analysis.

Outcomes for field trials (as opposed to proof of concept experiments) should be selected to provide the
information needed to aid in making decisions regarding the use of an intervention. However, the use of too many outcomes can lead to a lack of focus and increases the probability of type I errors; that is, concluding that there is a significant difference when the observed difference is due to chance (Tukey 1977; Vetter and Mascha, 2017). When multiple outcomes are included, the probability of type I errors can be substantial. The probability of at least one type I error can be calculated as 

$$1 - (1 - 0.05)^k$$

where \( k \) = the number of outcome measures, assuming that a conventional \( P \)-value of 0.05 is used to define statistical significance. Using the minimum (1), mean (9.5), and maximum (41) outcomes from 100 trials referenced above (Sargeant et al., 2009), a cut-point for statistical significance of 0.05, and assuming that the outcomes are independent, the probability of at least one type I error would be 5, 38.6, and 87.8% for a trial with the minimum, mean, and maximum number of outcomes, respectively. It should be noted that outcomes in a trial often may not be independent; for instance, if a trial found a significant association between an intervention and the prevalence of IMI, one also might expect that the trial would find an association between the intervention and SCC if both outcomes were included in the trial. Another (related) aspect of multiple outcomes is the false impression that additional information is gained from related outcomes and that such evidence provides “more” proof of effect. Using the previous example, a trial may use IMI and SCC as outcomes; if both outcomes are “significant,” this might be misconstrued as more evidence. However, as SCC might be considered a proxy for IMI, really no new information is gained. If resources are limited, more value from the trial might have been gained if the resources were spent on obtaining information about a truly different outcome. However, if measuring outcomes is not expensive or onerous, measuring and reporting the results of multiple outcomes does increase the chances that the results could be included in a meta-analysis, not withstanding the potential for type I errors. This underscores the need to choose outcomes thoughtfully in light of clear biological hypotheses.

**Reporting the Outcomes and Their Measurement**

Comprehensive reporting of all outcomes, including details of their measurement and analysis, is essential to interpreting the results of a trial. Expert consensus-based recommendations for reporting of outcomes are available for trials in livestock in the REFLECT statement and checklist (O’Connor et al., 2010; Sargeant et al., 2010). In an evaluation of the comprehensiveness of reporting in livestock trials, the measurement of all outcomes was described in 79 of 100 trials (Sargeant et al., 2009). This implies that details of the measurement of all outcomes was not provided in approximately one-fifth of published trials. Clearly, this is an issue for individuals wishing to understand and interpret the results of a trial or apply the intervention. Ensuring comprehensive reporting of outcomes (and other aspects of a trial) is the responsibility of the author. However, there also is a role for peer-reviewers and journal editors to enforce comprehensive reporting. The *Journal of Dairy Science* requires inclusion of an appropriate reporting guidelines checklist for articles submitted to the journal for peer-review (https://www.journalofdairyscience.org/reportingguidelines).

In trials with multiple outcomes, there may be some outcomes that are associated with the intervention and others that are not. It is essential that results are reported for all evaluated outcome measures regardless of their statistical significance. Failure to do so can lead to bias due to selective outcome reporting (Higgins et al., 2021). It is commonly required that protocols for human medical trials are publically registered before commencement of a trial (for example, at https://clinicaltrials.gov/). Comparisons of outcomes reported in trial protocols and in subsequent publications for human trials suggest that outcomes with statistically significant results are more likely to be reported (Dwan et al., 2013). Unfortunately, protocols for dairy trials are seldom publicly available; for example, the American Veterinary Medical Association animal health studies database (https://ebusiness.avma.org/aahsd/study_search.aspx) does not include any protocols specific to dairy trials. Therefore, is it not possible to determine the extent to which selective outcome reporting might be an issue for dairy trials. However, if outcome measures associated with nonsignificant findings are excluded from trial reports, it can lead to an exaggeration of the effectiveness of an intervention in the published literature. It also is possible that failing to report nonsignificant findings could lead to additional trials being conducted on ineffective interventions, which is a waste of available research funding resources. It may be reasonable to report the details of some outcomes in supplementary materials to keep publications concise and to make all results permanently accessible.

A final consideration related to the reporting of outcomes is the use of both relative and absolute outcome measurements. This pertains to how the results of the statistical analysis of trial outcome measures are presented. Relative effects are those that compare one intervention to another in a single measure, such as risk ratio (RR), hazard ratio, or odds ratio. Absolute risk measures, such as absolute risk and risk difference, describe the likelihood that an outcome will occur (Noordzij et al., 2017; Oyama et al., 2017). The advan-
Importance of Defining the Primary Outcome and Ensuring Sufficient Power

A primary outcome should be identified for all trials. The primary outcome is used to calculate the sample size and should be the outcome of most relevance to decision-making by the target audience (Moher et al., 2012). This also provides justification for the use of the animals for experimentation. If a study has very low power, then resources are devoted (wasted) to a question that cannot be answered. Given the bias toward publishing significant results, low power studies are less likely to be published and therefore the results wasted. Defining primary versus secondary outcomes enables the reader to distinguish between outcomes for which the study was powered to detect meaningful differences as statistically significant (primary) versus outcomes that were also measured but for which the study was not specifically powered to identify meaningful differences as statistically significant. If more than one outcome is critical to the trial objectives, it is possible to have more than one primary outcome. If there is more than one primary outcome, researchers should conduct sample size calculations for all primary outcomes and use the largest calculated sample size for the trial (Sargeant et al., 2010). Identifying the primary outcome is an area that needs to be improved in dairy trials. Of 91 trials published in the Journal of Dairy Science in 2017 where there was more than one outcome included in the trial, the primary outcome was only identified in only 4 trials (Winder et al., 2019b). This is lower than for trials published in selected veterinary journals in 2013 (19.3% of trials) and for trials published in selected human medical journals in 2013 (98.3% of trials; Di Gerolamo and Meursinge Reynders, 2016).

Selecting a primary outcome, and conducting a sample size calculation for that outcome, involves identifying the difference between intervention groups that the researchers consider to be “meaningful” (aka the equivalence margin), as well as the desired confidence level and statistical power (Stevenson, 2021). Meaningful differences may be defined based on economic or animal welfare considerations, or other factors. If the sample size is not large enough, meaningful differences may not be detected as statistically significant. Given the potential for publication bias against studies without statistically significant results, this can lead to research wastage if the results are not published. Alternatively, too large a sample size wastes resources that could be used elsewhere, may unnecessarily expose animals to an ineffective intervention, and may detect statistically significant differences that are not important.

One way to evaluate the potential effect of statistical power in dairy trials is to consider the minimum detectable RR from a sample of published dairy trials. The minimum detectable difference is the smallest difference between intervention groups (expressed as a RR) that could have been detected as statistically significant, given the baseline prevalence of the outcome (for a binary outcome) and the sample size used in the trial. We calculated the minimum detectable RR post hoc for each of 39 trials included in a systematic review and network meta-analysis of preventive antibiotics at dry-off for reducing new IMI at calving (Winder et al., 2019a). Using the epi.sscompb Function in epiR (Stevenson et al., 2022), the minimum detectable RR was calculated using the proportion of dairy cows with new IMI at calving in the placebo group (baseline risk), the total sample size, power of 0.8, and confidence of 0.95. (noting that power of 0.8 and confidence of 0.95 are based on convention and are not fixed values). If there were more than 2 intervention groups, then 2 groups were randomly selected with the total sample size corresponding to the sum of the 2 selected groups. The median baseline risk across the 39 trials was 17.9% (range among trials: 2.6 to 57.1%) and the median total sample size was 457 (range: 28 to 3,837). The distribution of minimum detectable RR for the incidence risk of IMI at calving is shown in Figure 4. The median for the minimum detectable RR was 1.64 (range: 1.21 to 5.37). This example illustrates that many trials have sufficient sample size to identify relatively small differences between intervention groups. However, some minimum detectable RR values were large. Using the highest minimum RR (5.37), and the median baseline risk of 17.9% as the risk in the placebo group, the treated group would need to have an incidence risk of approximately 96% for that difference to be detected as statistically significant. It is likely that an intervention could be considered to be effective with a much smaller difference from the placebo group.
Inconsistency of Outcomes Across Trials

Although individual researchers select outcomes of interest to their specific research question, building a body of evidence for the efficacy of an intervention requires the same outcomes to be evaluated across trials. Trial results from multiple studies can be combined statistically using systematic review and meta-analysis, which improves the precision of estimates of effects related to intervention efficacy (Higgins et al., 2021). Combining the results from multiple trials also increases confidence that differences between intervention groups are not the result of sampling error (Valentine et al., 2011). This argument can be extended to trials examining different interventions for the same disease or condition of interest, where network meta-analysis can be used to combine the results of trials evaluating different interventions for the same outcome to estimate the relative efficacy of the intervention options (Hu et al., 2019). However, combining results from multiple trials requires that the trials include the same outcomes and outcome measures, and that the populations and environments are similar enough that combined results would be meaningful. Differences between trials in the outcome measurements, case definitions for disease outcomes, or time at risk (i.e., follow up time) can prevent comparison. These concepts can be illustrated using examples from the literature. The first relates to differences in the outcome measurement using different laboratory methods for bacteriological culture of milk. There were 36 clinical trials included in a systematic review and network meta-analysis on the efficacy of antimicrobial treatments at dry-off to reduce the incidence of IMI at calving (Winder et al., 2019a). The authors of 10 trials specifically stated that IMI was diagnosed by bacteriological culture using National Mastitis Council methodology (unpublished data, C. B. Winder), with the remainder either not explicitly stating the method used or using a different methodology. The use of different diagnostic methods makes it difficult to combine results across studies, as the sensitivity and specificity of the diagnostic methods may differ. These inconsistencies mean that results across studies may not be comparable, and not all studies can be included in meta-analysis. Therefore, the value of the information has not been maximized.

Classification of reproductive tract health in dairy cows continues to evolve. A small-group consensus paper (Sheldon et al., 2006) attempted to standardize definitions for metritis and endometritis using the evidence available at the time. Metritis is characterized by fetid uterine discharge, which is associated with uterine bacterial dysbiosis (Galvão et al., 2019). The consequences of metritis and its response to therapy may or may not depend on whether the cow has a fever at diagnosis (Giuliodori et al., 2013; Figueiredo et al., 2021). In this case, there is no clearly established case definition (Sannmann et al., 2012; de Lima, 2020). This uncertainty need not preclude design of valid trials and synthesis of results across trials if the outcomes are clearly defined and reported; here, detailing how, when, and how frequently both uterine discharge and fever were assessed, and the inclusion criteria for assignment to intervention groups is essential.

A third example relates to inconsistency in time at risk, with data from a systematic review and network meta-analysis on the efficacy of teat sealants at dry-off (Winder et al., 2019c). Table 1 illustrates the time at risk for “new IMI at calving” as the outcome from the 27 trials included in this review. The results show that time at risk for detecting IMI was quite variable, with most times at risk represented by only a single trial. This is problematic, because the probability of a new IMI differs based on the time at risk, both cumulatively
and because mastitis risk is commonly greater soon after calving. Therefore, combining results from these trials may provide a summary effect measure that is misleading.

Core Outcome Sets

Individual researchers can identify the outcome measures used in previous trials to ensure that their trial includes some of the same outcome measures. However, a more efficient solution to both inconsistency of outcome measures and selective outcome reporting might be for researchers in a specific topic area to agree on which outcomes to include in all trials on that topic. This approach is being applied in human healthcare through the creation of core outcome sets (Williamson et al., 2012, 2017). Core outcome sets represent an agreed minimum set of outcomes and outcome measures that should be reported in all trials that are conducted on a specific disease or condition and are created by consensus of experts. Researchers may choose to include additional outcomes in a trial but also should include all core outcomes to allow results across trials to be combined (Williamson et al., 2017; Webbe et al., 2018). Developing a core outcome set involves defining the topic (e.g., a health or other outcome domain, or a type of intervention and conceptual outcome), evaluation of the existing literature to determine what outcome domains, conceptual outcomes, operational outcomes, and outcome measures have previously been used in the topic area, and a consensus process to determine which of these to include in the core outcome set (Prinsen et al., 2016; Williamson et al., 2017). Guidelines are available for developing core outcome sets (Williamson et al., 2017). The nature and number of core outcomes will vary with the topic area and will include a consideration of feasibility, relevance, precision and accuracy of measurements, and the probability of type I errors associated with multiple outcomes. Core outcome sets need to be updated periodically as new outcome measurements are developed and validated or as disease definitions are modified.

In 2018, there were 410 core outcome sets available for human healthcare topic areas, including areas as diverse as cancer, urology, and child health (Gargon et al., 2019). To date, there are only 2 core outcome sets in veterinary medicine, one for chronic kidney disease in cats (Doit et al., 2021) and one for dermatitis therapies in dogs (Olivry et al., 2018). There is precedent for consensus processes in research design in dairy sciences. The National Mastitis Council produced guidance documents for standardized milk culture methods for mastitis pathogens (Middleton et al., 2014). Researchers in dairy cattle reproduction have recommended outcome measures that provide a foundation for development of a core set of outcomes (Lean et al., 2016). Some traditional measures of reproductive performance are biased and should be abandoned (e.g., calving interval and services per conception exclude cows that failed to become pregnant). Other measures need to be reported fully for valid interpretation. For example, the proportion of animals pregnant at first insemination should be reported with a valid assessment of time to first insemination and description of numbers and reasons for exclusion of animals.

Creating core outcome sets for dairy topics would be useful for individual researchers planning trials and would be helpful when synthesizing the results from multiple trials via meta-analysis or network meta-analysis. Developing a body of evidence across multiple trials will allow more evidence-informed decision-making, which will benefit not only dairy producers and dairy cattle. This would maximize the utility of the research investment.

CONCLUSIONS

The results of a research project are used for many years after publication; further, the approach taken by researchers to select interventions and outcomes can substantially increase the utility of those results years later. When study results are used in secondary ways (such as meta-analysis, network meta-analysis, or clinical guidelines), this adds value to the original investment in research. The intervention groups and outcomes to include will depend on the interests of the researcher and the specific objectives of the trials.
However, the results of multiple trials on a topic should inform evidence-based decision-making. Therefore, a lack of common intervention or outcome groups with other experiments greatly reduces the value of a trial and may represent wasted resources. The value of the research investment can be maximized by ensuring that at least one intervention group is common with another published trial, and that common outcome measures are used across trials. Core outcome sets can increase the utility of research by allowing comparison and synthesis of results across trials.

ACKNOWLEDGMENTS

Partial financial support for this project was provided through a University of Guelph Research Leadership Chair (Sargeant; Guelph, ON, Canada). The authors have not stated any conflicts of interest.

REFERENCES


ORCIDS
Jan M. Sargeant https://orcid.org/0000-0003-2420-1766
Annette M. O’Connor https://orcid.org/0000-0003-0604-7822
Stephen J. LeBlanc https://orcid.org/0000-0003-2027-7704
Charlotte B. Winder https://orcid.org/0000-0002-7314-3657