Contemporary Clinical Trials Communications 16 (2019) 100486

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: http://www.elsevier.com/locate/conctc

Review article

Choosing primary endpoints for clinical trials of health care interventions

Charlie McLeod [a,b,c,*], Richard Norman [d], Edward Litton [b,e], Benjamin R. Saville [f,g], Steve Webb [e,h], Thomas L. Snelling [a,c,d,i]

[a] Wesfarmers Centre for Vaccines and Infectious Diseases, Telethon Kids Institute, University of Western Australia, Nedlands, Australia
[b] School of Medicine, University of Western Australia, Nedlands, Australia
[c] Infectious Diseases Department, Perth Children's Hospital, Nedlands, Australia
[d] School of Public Health, Curtin University, Bentley, Australia
[e] St John of God Hospital, Subiaco, Australia
[f] Berry Consultants, Austin, TX, United States
[g] Vanderbilt University Department of Biostatistics, Nashville, TN, United States
[h] School of Population Health and Preventive Medicine, Monash University, Clayton, Australia
[i] Menzies School of Health Research, Tiwi, Australia

ARTICLE INFO

Keywords:
Endpoint determination
Surrogate
Biomarkers
Outcome assessment
Clinical trials
Research design

ABSTRACT

The purpose of late phase clinical trials is to generate evidence of sufficient validity and generalisability to be translated into practice and policy to improve health outcomes. It is therefore crucial that the chosen endpoints are meaningful to the clinicians, patients and policymakers that are the end-users of evidence generated by these trials. The choice of endpoints may be improved by understanding their characteristics and properties. This narrative review describes the evolution, range and relative strengths and weaknesses of endpoints used in late phase trials. It is intended to serve as a reference to assist those designing trials when choosing primary endpoint (s), and for the end-users charged with interpreting these trials to inform practice and policy.

1. Introduction: Purpose of clinical trials and why endpoint selection is important

The purpose of late phase trials is to generate evidence to guide decision-making in clinical practice and in policy. In this regard, clinicians, patients, and policymakers are all end-users of clinical trial evidence. Randomised clinical trials represent a gold standard for generating evidence, as they are the least biased way of measuring and comparing treatment effects [1].

Many outcomes occur among trial participants [2]; some outcomes occur *because* of an intervention or because of the absence of one, some outcomes may be *modified* by an intervention (for example, time to event or severity), while many more outcomes occur unaffected by an intervention. Outcome(s) selected for evaluation must address the trial objective(s) and should be acknowledged as meaningful to end-users. For an outcome to be meaningful, it should reflect or describe how a person feels, functions and survives [3]. Endpoints are the specific measures of these outcomes [2]. If end-users are going to make decisions based on measured differences in one or more endpoints between treatment groups, they must understand what those differences are; but endpoints have properties and characteristics that have strengths and limitations that are critical to their interpretation.

It is a responsibility of those who design and conduct trials to choose endpoints which will influence decision-making by clinicians and policymakers. Endpoint selection is a complex process. End-users bring differing needs and perspectives. Poor selection of endpoints makes interpretation and implementation of findings difficult or impossible, limits evidence synthesis, and thereby diminishes the value of the research, resulting in wasted use of resources [4].

A single endpoint may not capture the important effects of an intervention to the satisfaction of all end-user groups, so multiple endpoints are usually selected, which are categorized as *primary, secondary* or *tertiary*. Primary endpoint(s) are typically efficacy measures that address the main research question [3]. Secondary endpoints are generally not sufficient to influence decision-making alone, but may support the claim of efficacy by demonstrating additional effects or by supporting a causal mechanism [2]. If tertiary endpoints are nominated, they typically capture outcomes that occur less frequently or which may

* Corresponding author. Infectious Diseases Department, 15 Hospital Avenue, Nedlands, WA, 6009, Australia.
E-mail addresses: charlie.mcleod@health.wa.gov.au (C. McLeod), richard.norman@curtin.edu.au (R. Norman), ed.litton@health.wa.gov.au (E. Litton), ben@berryconsultants.net (B.R. Saville), steven.webb@monash.edu (S. Webb), Tom.Snelling@telethonkids.org.au (T.L. Snelling).

https://doi.org/10.1016/j.conctc.2019.100486
Received 1 June 2019; Received in revised form 29 October 2019; Accepted 9 November 2019
Available online 12 November 2019

be useful for exploring novel hypotheses [3].

The primary aim of this review is to summarise the range of clinical and non-clinical endpoints used in late phase trials and their relative strengths and weaknesses. The secondary aims are to describe their evolution and consider which characteristics of endpoints are valuable for evaluating treatment effects. This review is intended to serve as a reference to assist researchers when choosing primary endpoints, and for the end-users of clinical trial data tasked with translating this evidence into clinical practice or policy. Early phase trials may have a more proximal aim such as establishing proof-of-principle, trial feasibility, or assessing the mechanistic effects of an intervention; this review does not discuss endpoints relevant to these types of trials. Further, whilst we recognise that statistical and regulatory considerations are also important factors weighing into overall endpoint selection, a detailed analysis of these topics is beyond the scope of this review.

2. Methods

We developed a two-strand search method to address our research questions, incorporating the following search terms using a Boolean strategy, including papers published up to October 2018: "Endpoint determination", "surrogate, biomarker, combination, individuali?*, multiple or composite," "end?point," "Outcome Assessment (Health Care)," "Research Design" and "Clinical trials." This search was executed in Medline (Medline and Epub Ahead of Print, In-Process and other non-indexed citations 1946-) and Embase (Embase & Classic 1947-) and limited to articles written in English. Registration guidelines issued by the Food and Drug Administration and European Medicine Association (EMA) were also examined using the same keywords. Additional articles were identified through citation review of selected articles and some clinical examples were drawn from the authors' experience. Our full search strategy (including additional limits) is detailed in Appendix A and B. The search was performed by a single reviewer (CM) and findings are reported by narrative synthesis.

3. Results

3.1. Classification of endpoints

Endpoints for late phase trials can be broadly classified as either clinical or non-clinical (see Fig. 1) [3,5].

Clinically meaningful endpoints relate to outcomes which capture how a person feels, functions or survives [3]. These endpoints may be measured objectively or subjectively, and are either (i) reported by clinicians (ClinRO), which involves judgement or interpretation of clinical signs or events (such as stroke, myocardial infarct or cancer remission), (ii) assessed by standardised performance measures (6-min walk test), (iii) patient-reported (PRO), which are directly reported by patients (such as self-reported symptoms or function, or a measure of perceived quality of life) or (iv) observer-reported (ObsRO), such as a parent log of seizure activity in a child [5].

Non-clinical endpoints, including biomarkers, do not relate directly to how a person feels, functions or survives, but are instead objectively measured indicators of a biological or pathogenic process, for example a pharmacological response to a treatment intervention. Biomarkers may include blood tests (for example laboratory measures such as troponin and haemoglobin concentration or serological assays), tissue/fluid analyses (for example histopathological results), imaging results, or physiological measures (for example blood pressure) which are used for diagnostic, prognostic, monitoring (including safety) or predictive purposes [2].

Some endpoints may be clinically important even though they are non-clinical and not meaningful to all end-users (See Fig. 2). Such endpoints do not directly reflect or describe how a patient feels, functions and survives and therefore hold no intrinsic value to patients, but are nonetheless important because they are strongly associated with a meaningful outcome, and therefore compellingly influence clinical decision making, for example a troponin result or a measured blood pressure.

Trial endpoints may be used to derive metrics which are used to further evaluate the impact of an intervention, for example from a population or policy perspective, such as number needed to treat or harm, or the incremental cost per quality-adjusted-life-years gained. These metrics are important from a societal, and consequently translation perspective [3].

3.2. Surrogate endpoints

Surrogates are those endpoints that do not directly measure how a person feels, functions or survives, but which are so closely associated

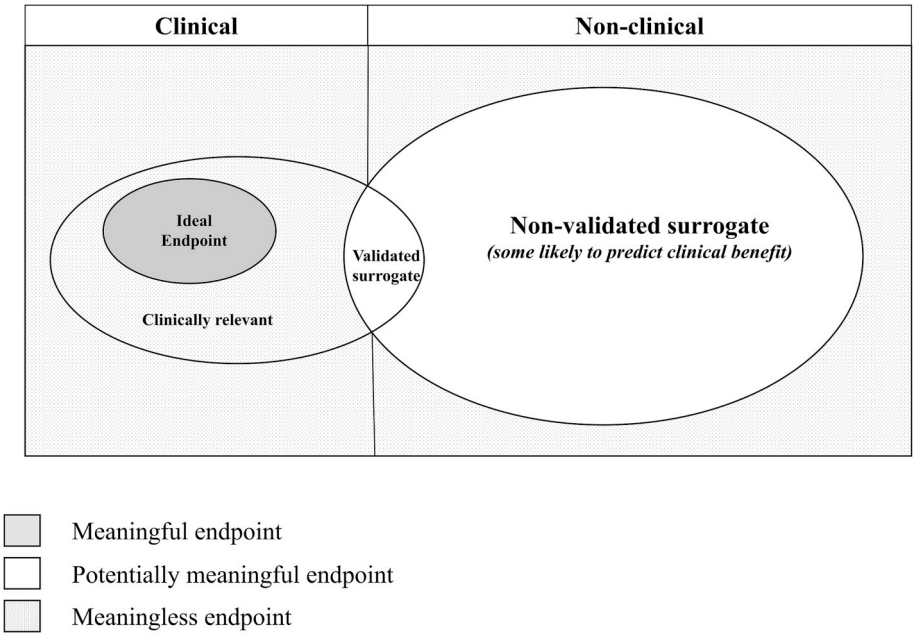

Fig. 1. Classification of endpoints.

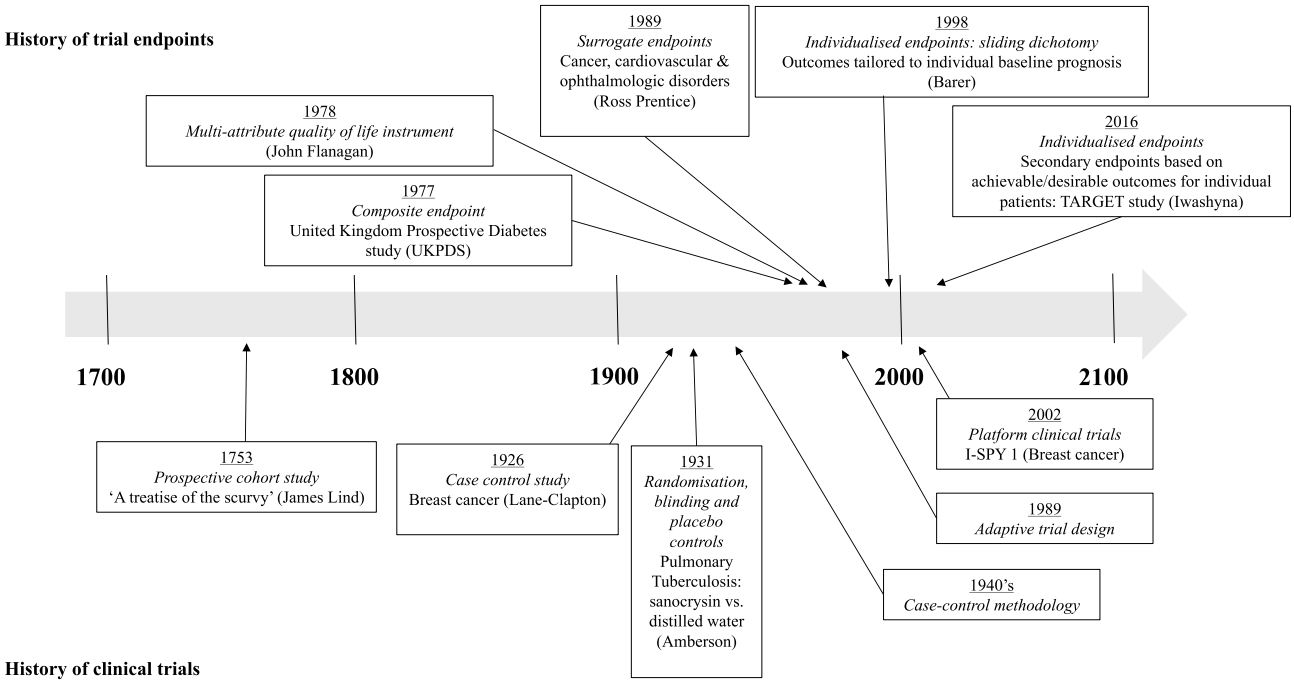

Fig. 2. Evolution of clinical trials and historical use of endpoints.

with a clinically meaningful endpoint that they are taken to be a reliable substitute for them [2]. The quality of a surrogate endpoint is therefore determined by the extent to which a treatment effect on that surrogate corresponds to a treatment effect against one or more clinically meaningful outcomes. Conceptually, the best surrogate endpoints directly measure causal intermediaries of the effect of an intervention on a clinically meaningful outcome, where essentially all effects on that outcome are mediated through that intermediary, and where there is little attenuation between the effect of a treatment on the intermediary and the intermediary's effect on the clinically meaningful outcome [5]. Surrogates which do not causally influence the meaningful outcome may still be statistically associated with it for a given treatment in a given context, but this association may not generalise well to other clinical contexts, populations or interventions.

A validated surrogate is one which reliably captures a treatment effect against one or more clinically meaningful endpoints, bearing in mind that the strength of this association may be context dependent, and reliability cannot be inferred unless there are multiple randomised, controlled trials of interventions that have the same or similar effect on both the surrogate and the clinically meaningful end-point [5]. The US Food and Drug Administration (FDA) provides a list of validated and likely surrogates [6], for example HbA1c is listed as a marker of risk of long term microvascular complications in type 2 diabetes mellitus. However, *unvalidated* surrogates are sometimes selected for lack of a validated surrogate, and there is no standardised process or agreed criteria that must be met for validation.

Prentice first described the criteria for scientific validation [7], proposing the surrogate should be statistically correlated with the clinical outcome of interest, and also fully capture the effect of the intervention on the outcome. The latter criterion has been critiqued as being too stringent [5]. Fulfilment of the Prentice criteria requires an understanding of the causal pathways of disease and the effects of an intervention on this pathway, and such complexities might never be confidently understood entirely. Surrogates typically only capture 'on-target' effects, that is effects that are anticipated based on our understanding of the causal pathway of the disease process; 'off-target' effects of an intervention lie outside this causal pathway, are therefore unanticipated, and may not be captured by a surrogate, but may nonetheless impact importantly (positively or negatively) on the meaningful outcome [5]. Alternative approaches for validation of surrogates have been described elsewhere [8,9].

Fleming & deMets warn that even if surrogates correlate with an outcome of interest, they may fail to predict clinical endpoints through one of three mechanisms [10]. The first is failure of the surrogate to lie on the causal disease pathway. An example is the use of laboratory measures to evaluate the impact of HIV treatment in pregnancy to reduce mother to child transmission of HIV infection [10]. The maternal CD4 count and HIV viral load are both statistically correlated with the risk of transmission in untreated women; low CD4 count and high viral load are both associated with increased risk. HIV viral load, which measures the amount of circulating virus in the mother's blood, is thought to lie on the causal pathway between treatment and transmission because circulating virus is thought to be a prerequisite for transmission. Any treatment that reduces the maternal viral load can therefore reasonably be expected to reduce the risk of transmission. The CD4 count however, which measures the status of the mother's immune system, may not be causally related to transmission. Instead, high viral load in untreated women causes low CD4 count, so the association between low CD4 count and risk of transmission may be confounded by the higher viral loads in women with low CD4 count. This means that treatments that impact on CD4 count (and not the viral load) may not influence the risk of transmission. HIV viral load is therefore *prima facie* a more reasonable surrogate than the CD4 count for capturing the effect of maternal interventions on risk of mother to child transmission.

The second reason for failure of a surrogate is the existence of more than one causal pathway impacting on the outcome, where the surrogate lies on one pathway only [11]. In the above example, maternal viral load might only be a reasonable surrogate for mother-to-child-transmission for those treatments that mediate their protective effects by inhibiting viral replication. Caesarean section is also protective against transmission, but through alternative pathways, presumably by decreasing exposure of the newborn to maternal blood and secretions. Maternal viral load would not be expected to be a useful surrogate in that context.

Thirdly, the intervention may produce off-target effects that impact on the measured outcome [11]. The Cardiac Arrhythmia Suppression Trial (CAST) was designed to test the hypothesis that suppression of

asymptomatic or mildly symptomatic ventricular arrhythmias with anti-arrhythmic agents (flecanide or encainide) would reduce the risk of death or cardiac arrest requiring resuscitation in survivors of myocardial infarction [12]. Although the pilot study for this trial found these agents suppressed arrhythmias adequately in the target population [13], mortality increased 3-fold in the CAST owing to effects of these drugs on mortality through alternative pathways, possibly through unanticipated pro-arrhythmic effects [12], prompting withdrawal of these drugs from the market [5].

3.3. Endpoint characteristics: what is ideal?

Conceptually, an ideal endpoint should be a valid and applicable measure of how a patient feels, functions or survives [2] and be perceived by end-users of the research as having meaning and value. To be valid, an endpoint should capture the outcome of interest accurately (measure what is intended), precisely (with minimal error or uncertainty) and consistently with repeated measurements [14]. This is easiest to achieve when the outcome of interest can be measured directly, such as death. An ideal endpoint should also be measured easily, without additional risk, at low cost, at minimal inconvenience to the patient [15], and, if possible, captured as part of routine data collected as part of clinical care. Death is one example of an endpoint for interventions of highly fatal conditions which fulfils all these criteria, including the fact that this endpoint is meaningful to all end-user groups. For the majority of conditions where death is rare, or where survival may be associated with significant suffering or disability, death will not capture all relevant and meaningful outcomes.

Standardization of endpoints is increasing through the development and adoption of core outcome sets [16]. Core outcomes are the effect(s) of a health intervention which are agreed as being important to end-users, including patients. A core outcome set (COS) is a minimum agreed list of outcomes that should be measured and reported in trials [17]. COS are disease-, population- and/or intervention-specific, however there is often considerable overlap between outcomes selected across different research domains given that outcomes are likely to be important irrespective of the underlying disease process. Guidelines are available to inform development of core outcomes sets and identification of optimal methods for outcome measurement [16,17]. Patients, clinicians, policy-makers, industry representatives, and members of the public may be involved in the development of core outcome sets depending on existing subject matter knowledge, the rationale for development, and feasibility constraints [16–18].

3.4. Evolution of endpoints to capture different treatment effects

It may be helpful to consider when and why the use of different endpoint types has evolved over time; this is summarized in Fig. 2 [19–22]. Because interventions impact patients in different ways and may have more than one consequence (positive or negative), decision making around the use of an intervention should consider the net benefit versus risk [17]. Increasingly complex endpoints have evolved in parallel to advances in trial design and data capture in order to assess multiple important effects of an intervention in aggregate, or to determine whether the intervention is likely to have a net benefit to a patient overall. In the current era of patient-centered healthcare, individualised endpoints have also been recently proposed as a framework for evaluating personally defined risk and benefit [23].

No endpoint type is universally better than all others, but rather, the different characteristics and properties of each type make them better suited for use in different contexts; this is considered in further detail below. A summary of the strengths and limitations of various types of endpoints described in the literature are presented in Table 1 [14,19,21,24].

3.5. Multiple and combination primary endpoints

Multiple or combination primary endpoints may be required to capture the aggregate risk-benefit effect of an intervention [3,25]. This may be considered when multiple disparate outcomes have comparable importance, if each of those outcomes are individually rare, or if no consensus can be reached regarding which is most important [3].

3.6. Multiple endpoints

Multiple endpoints can be chosen and evaluated separately, such that a significant treatment effect against any one of the endpoints may be taken as evidence of efficacy. This approach may be useful in diseases that have multiple sequelae, where improvement in any pre-specified endpoint is clinically meaningful even in the absence of improvement in any other [3,19]. Because the risk of type 1 error increases with every additional endpoint assessed, appropriate statistical adjustments for multiplicity are generally needed to contain the risk of a false positive trial result; regulatory authorities are particularly focussed on this issue and have given guidance on managing this risk [3].

Multiple primary endpoints become ‘co-primary’ if an effect on multiple outcomes is required to demonstrate proof of efficacy [3]. An example of co-primary endpoints includes both cognitive and functional assessments in studies of Alzheimer’s disease [3,26] in which for a treatment to be considered efficacious it must demonstrate a beneficial effect on both cognition and function. There is no risk from multiplicity when co-primary endpoints are used [3]; conversely, the power of a study is typically diminished by the requirement to demonstrate significant efficacy against more than one endpoint, unless those endpoints are highly correlated.

3.6.1. Combination endpoints

Combination endpoints may be either composite or multi-component [3,19].

3.6.1.1. Composite endpoints. Some trials combine measures of multiple outcomes (such as death and major morbidity events) into a single measure of effect, or composite endpoint [3]. This helps to avoid the multiplicity issues inherent when multiple endpoints are assessed separately [14,21]. Composite endpoints are sometimes used to aggregate the total benefit when the goal of therapy is to prevent or delay a number of important but uncommon clinical events [21]. One example is a composite endpoint which comprises any of death, myocardial infarction, stroke or revascularisation in cardiovascular trials [3].

The value of composites is influenced by the relative importance of its components. The components of a standard composite endpoint are implicitly ascribed equal weight. If the components do not have comparable importance (for instance death and revascularisation) [27] the trial results may be difficult to interpret and less useful for end-users unless the size or direction of the treatment effect against each component is uniform. Individual components of the composite must be reported separately (as secondary endpoints) in addition to the overall result, but there may be insufficient power to determine the treatment effect for each component. An additional limitation of composite endpoints is that repeated (and possibly more serious) events are ignored [28].

There are two broad approaches to the analysis of composite endpoints. The ‘first combine and then compare’ method involves combining the components into a single composite endpoint and then comparing the frequency or rate of the composite between treatment and placebo groups [21,29]. The second method, to ‘first compare and then combine’ or the ‘win-ratio’ approach, is gaining traction as an alternative method which helps to account for heterogeneity in treatment effects across the component outcomes [30]. This involves matching pairs of patients in the treatment and placebo arms based on

Table 1
Strengths and limitations of trial endpoints.

Endpoint	Strengths	Limitations
Singular		
Clinically observed	Routinely collected information Typically well-accepted approach by scientific community	Doesn't consider that an intervention may impact on the patient in different ways May need supportive secondary analyses to be persuasive
Surrogates	Reduction in sample size Shorter trial duration Decreased cost of trial Accelerated approval/dissemination of effective therapies	May fail to predict clinically meaningful endpoints May not be sensitive to change at all stages of disease Validation process often challenging Therapeutic advances may alter the validity of the surrogate measure Cost Reproducibility may be problematic Utility limited to early phase trials (1/2)
Multiple or combination		
Multiple primary	Useful if more than one important outcome exists & demonstration of 1 is enough to support clinical efficacy	Adjustment for Type 1 error is required. Hard to interpret if results occur in different directions
Co-primary	Useful if demonstration of two or more outcomes is necessary to establish clinical benefit	Adjustment for Type 2 error is required
Composites	Improves statistical efficiency and precision. Increases power (reduces sample size requirement). Ability to measure small effects. Lower cost. Earlier trial completion	Implementation may be complex and resource-intensive. Components may be inappropriately combined or reported. May be difficult to interpret study findings and determine which of the component endpoints are impacted by the intervention; the effect is often smallest for the most important component and biggest for the less important components. Prone to post-hoc analyses/bias. Key data often missing or unclear. Can lose meaning if components of composite move in opposite directions
Multi-component endpoints	Allows single evaluation of numerous components without creating multiplicity issues	Individual components may not have clear meaning. If components aren't concordant, study power may be compromised
Weighted endpoints	More complex/robust evaluation of the effectiveness of treatment intervention(s) that considers the relative importance of components	Process of assigning weights not standardised, and can involve lengthy processes. May be costly
Endpoints that are participant specific	Best reflects clinical decision-making. Theoretically would represent the gold-standard for informing personalised, evidence-based medicine. May result in increased power to detect real treatment effects for patients	Complex; logistically difficult. Generalisability of trial results may be limited. Requires large data capture

their risk of experiencing the outcome of greatest importance included in the composite (such as death), and examining component outcomes in a prioritised fashion. This creates an implicit weight between the component outcomes of the composite, but doesn't consider their exact weighting; if assessment of the first outcome included in the composite results in a tie (or doesn't occur in either group), the second most important outcome is evaluated. The number of 'wins' versus 'losses' is then compared between groups to calculate the win ratio [31]. Pocock et al. [31] have applied this method to the CHARM trial results, which evaluated use of an ACE inhibitor compared to placebo in chronic heart failure using a composite which prioritised the evaluation of death over hospitalisation.

3.6.2. *Multi-component endpoints*

A multi-component endpoint combines numerous pre-specified component outcomes into a single score or rating which is calculated using a multi-attribute instrument, where the scores for each attribute may be either weighted or unweighted [3,32]. In contrast to composite endpoints, the components in a multi-component endpoint may not be meaningful when analysed individually, and all components must be assessed for each participant and contribute to the overall score. Unweighted multi-attribute instruments effectively assign equal importance to all items, and an overall score is obtained by simply summing the responses, such as psychometric assessments that measure cognitive ability.

3.7. *Weighting and utility*

The individual components included in composite and multi-component endpoints often don't have comparable importance. Weighted analysis has been proposed as one method for overcoming this issue [33], where the weights are intended to reflect the relative importance of an individual outcome relative to others [20].

Weights may be assigned by expert judgement (obtained through a Delphi panel process, for example) [32] or elicited using either 'stated' or 'revealed' preference methods [34]. Stated preferences are derived from decisions made by individuals when confronted with realistic, hypothetical choice scenarios. Time-trade-off, standard gamble, visual analogue scales (where a specific health state is rated on a scale from 0 to 100), and discrete choice experiments (DCEs) are examples of stated preference techniques [34,35]. Revealed preference methods assign weights to outcomes based on observed choices made by individuals in real-life scenarios. Most obtain individual patient preference information, although disability-adjusted life years (DALY) is one method which reflects weighting of health outcomes obtained at a population level [36].

Weighted composite endpoints have been used extensively in cardiovascular research. One example is a post hoc analysis of the DELTA trial [37]. This study evaluated the impact of either percutaneous coronary interventions (PCI) or coronary artery bypass grafting (CABG) in patients with left main coronary artery disease. Using a primary composite endpoint incorporating death, myocardial infarction, cerebrovascular accident (CVA) and target vessel revascularisation showed CABG to be superior to PCI in 1204 propensity-matched patients at 3 years. Weighting the component outcomes according to clinical significance, however (with death considered worse than CVA, followed by MI and finally revascularisation) found no significant difference between revascularisation strategies.

Utilities are sometimes applied to individual outcomes (including health state descriptions) in which the health state outcome is converted to a utility measure. Utilities attempt to quantify the desirability or value of an outcome or health state, and specifically how much better/worse one is over another [32]. Assigning a utility value to a range of possible health outcomes enables calculation of a single overall utility score for each participant which can then be aggregated over all participants in each study arm.

Discrete choice experiments (DCEs) are one method for calculating utility, and are relatively new to health, but growing over time [38]. DCEs present participants with a series of hypothetical scenarios, called choice sets, asking them to make decisions about preferred health outcome states. This process requires respondents to make trade-offs between different aspects health-related status (like benefit versus drug toxicity) which enables the quantification of the relative importance of outcomes, which is superior to other techniques which simply rank or rate them [39]. Quality of life tools, which capture an individual's subjective assessment of their physical, mental and social wellbeing are one example of utility instruments [34,35,39]; they may be generic, such as the EurolQol five-dimensional questionnaire (EQ-5D) and the six-dimensional health state short form (SF-6D) [40] or disease-specific, such as the Cystic Fibrosis Questionnaire-revised [41].

3.8. Participant specific endpoints

Significant heterogeneity exists in individual preferences for outcome states, even among those with the same disease and similar baseline health states [3]. End-users may want to individualise treatment decisions according to personal characteristics including disease stage and comorbidities, the availability or lack of treatment options, values and beliefs, and financial considerations [14]. Consequently, it may be difficult to directly apply trial evidence to inform patient management if the patient at hand differs demographically or clinically from average trial participants, or if the endpoints selected for the trial are of secondary or of no importance to the patient. For example, a trial which reports the efficacy of a drug for return to work, may have little applicability for a retired patient who desires a return to independent living.

Individualised endpoints for participants in a trial may provide a framework for evaluating personally-defined risk and benefit. Two approaches have been proposed. The first [42] employs a sliding dichotomy which defines treatment success for a given trial participant based on what experts deem to be achievable and desirable given their baseline disease status and prognosis. This allows patients to be enrolled into a trial across a spectrum of baseline health and disease severity (such as stroke severity), with all participants able to contribute to the analysis.

Iwashyna et al. [23] has built on this prognostic stratification by proposing use of a 'values clarification instrument' to elicit preferences from prospective participants in a trial for a range of possible health outcome states. These preferences might be combined algorithmically prior to randomisation to determine which endpoints are both achievable and most desirable to patients and treatment success could then be defined as the realization of one or more of these endpoints. This approach might be difficult to apply in trials in acute care settings, but might be applicable in non-acute settings, like trials of interventions in chronic diseases like cystic fibrosis.

4. Discussion

The range and complexity of endpoint types available for use over time has increased alongside the evolution of increasingly complex trial designs. Increasing recognition of the need to engage different end-user groups in endpoint determination is an important step towards improving the value of research that is conducted and the likelihood of translation of research findings. Whilst significant developments have occurred in this field, clearly the optimisation and selection of endpoints for clinical trials remains a science in evolution.

Whilst the CONSORT statement provides guidance about how to report outcomes for randomised controlled trials [43], no guidance is available to inform optimal selection of endpoints, nor methodology available to quantify endpoints as best, and so endpoint selection will likely remain the prerogative of those who design and sponsor trials. The exception to this is for trials to support licensure of new therapeutic products, where the regulator may be directly involved in nominating the endpoints. Such trials should provide evidence to guide practice, however registration endpoints do not always capture outcomes that are clinically meaningful to patients or clinicians, and consequently may fail to achieve this goal. There is increasing recognition of the need for regulatory authorities to consider patient preference information when stipulating endpoints [3].

Outside the context of licensure, an end-point can be regarded as having some value if there is a history of end-users changing practice or policy for trials that have that end-point. Where there is no history there may be value in pre-trial surveys or focus groups of end-users to establish that proposed end-points would be regarded as sufficient to change practice or policy.

Failure to involve all end-user groups in discussion about endpoint suitability can compromise the translation of the results of a trial into clinical practice. We believe this is common and an important oversight. Assumptions made during the design of clinical trials can contribute to this situation. Firstly, clinicians often make incorrect assumptions about patient preferences and what patients value [44]. Secondly, the broader community of clinicians and policy-makers are often not engaged to establish that a selected end-point meets their requirements. While it can be difficult to select endpoints that will satisfy all end-users, at least understanding the perspectives and priorities of these groups will help avoid the use of end-points that are misaligned with the trial's objectives. In trials examining optimal treatment of pulmonary exacerbations of cystic fibrosis for example, clinicians might most value the impact of treatment on measurements of lung function (change in $FEV_{1.0}$ from baseline) [unpub, Snelling], patients may attach greater significance to the effect of treatment on functional status or quality of life, while policymakers might prioritise the cost-effectiveness or externalities like antimicrobial resistance implications of specific treatment regimens. An empiric understanding of alternative endpoint types can assist in the design of trials to meet the requirements of all end-users. Achieving consensus between disparate groups of end-users about meaningful outcomes is more likely to occur when research questions arise in the context of goal-orientated care, where treatment is administered based on what outcomes are considered achievable and desirable to patients [44].

Endpoints may also be limited to the extent to which they are context- (including timing), patient- and intervention (including drug class)- specific [14]. Endpoints may not be applicable in settings that do not have the capability of performing the selected outcome measurement(s). Endpoints might not reliably measure outcomes in all individuals included in the study population, which produces ceiling and floor effects that must be considered [45]. For example, spirometry is a valuable measure of lung function in adults but cannot be reliably performed in children less than 6 years old or in patients with end-stage respiratory failure. Endpoints may also not consistently detect the effect of a treatment intervention across all stages of disease. Forced expiratory volume in 1 s ($FEV_{1.0}$), a marker of lung function, and radiological studies are both insensitive markers to change in lung disease early in cystic fibrosis, for example [14]. Conversely the 6-min walk test is not applicable for patients with muscular dystrophy who are already confined to a wheelchair. Drugs may impact on the same outcome via different pathways; for example, anti-arrhythmic and lipid-lowering agents impact on cardiovascular mortality through different mechanisms. This means use of surrogates across different classes of drugs, even when used for a similar purpose, cannot be assumed to be appropriate [5].

Where surrogates are used as endpoints, it may be unclear what degree or duration of effect corresponds with a clinically meaningful effect [5]. While it is widely agreed that lowering blood pressure is causally associated with reduction in the risk of cardiovascular death, it may not be easy to translate exactly how a given reduction in blood pressure over a given period of time translates into a quantifiable reduction in mortality.

Selecting endpoints which meet the conceptual ideal is challenging

and may be not be possible, forcing researchers to compromise and make pragmatic decisions about those selected. Further, health outcomes identified as meaningful to end-users may be abstract, and therefore may not be directly measurable. In this regard, the availability (or absence) of appropriate scales of measurement for outcome evaluation may be an important factor that drives endpoint selection [23].

5. Conclusions

Optimisation and selection of endpoints for clinical trials is an evolving field. Given the purpose of late phase trials is to inform clinical practice and policy, endpoints should measure outcome(s) which are meaningful to end-users that reflect or describe how a patient feels, functions or survives. Understanding the range of endpoints available for use and the context in which they have arisen together with their strengths and limitations will help inform end-users when selecting endpoints for late phase trials.

Future work should focus on streamlining processes for identifying prioritised outcomes for different end-user groups across different research domains and on developing a methodology for qualifying endpoints as best. There is a need for universally agreed guidelines to inform optimal selection and reporting of endpoints; such guidelines should emphasise the importance of endpoints being suited to the trial purpose and participants and acceptable to relevant end-user group(s). Justification of the selection of endpoints in all trials should be reported. Such guidelines may be beneficial for end-users and help reduce research waste.

Funding

This work was supported by a Perth Children's Hospital Foundation Grant (#9757). CM is supported by fellowship grants from the National Health and Medical Research Council (NHMRC; GNT1150996), the Wesfarmers Centre, and the Perth Children's Hospital Foundation (#9772). TS is supported by a Career Development Fellowship from the NHMRC (GNT1111657). Sponsors had no role in the design, collection or synthesis of data, the writing of this review, or the decision to submit for publication.

Declaration of competing interest

The author(s) declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgements

The authors wish to thank senior librarians Karen Jones and Samantha Blake located at the J. Robin Warren Library (UWA, Perth) for their guidance which informed the search strategy employed in this review.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2019.100486.

References

[1] B.S. Kramer, J. Wilentz, D. Alexander, J. Burklow, L.M. Friedman, R. Hodes, R. Kirschstein, A. Patterson, G. Rodgers, S.E. Straus, Getting it right: being smarter about clinical trials, PLoS Med. 3 (6) (2006) e144.
[2] Biomarkers, EndpointS and other tools (BEST). http://www.ncbi.nlm.nih.gov/books/NKB338448/.
[3] US Department of Health and Human Services F, Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER): Multiple Endpoints for Clincial Trials: Guidance for Industry, 2017.
[4] D. Moher, Glaszious, Chalmers I: increasing value reducing waste in biomedical research: who's listening? Lancet 387 (2016) 1573–1586.
[5] T.R. Fleming, J.H. Powers, Biomarkers and surrogate endpoints in clinical trials, Stat. Med. 31 (25) (2012) 2973–2984.
[6] Table of drug surrogates that were the basis of drug approval or licensure, in: https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/ucm613636.htm.
[7] R. Prentice, Surrogate endpoints in clinical trials: definition and operational criteria. Stat. Med. 8 (1989) 431–440.
[8] M. Buyse, Statistical validation of surrogate outcome measures, Trials Conf.: Clin. Trials Methodol. 12 (SUPPL. 1) (2011).
[9] D.J. Sargent, S.J. Andrekar, Statistical issues in the validation of prognostic, predictive, and surrogate biomarkers, Clincal 10 (5) (2013) 647–652.
[10] T. Fleming, DD: surrogate endpoints in clinical trials: are we being misled? Ann. Intern. Med. 125 (7) (1996) 605–613.
[11] E.A. Hartung, Biomarkers and surrogate endpoints in kidney disease, Pediatr. Nephrol. 31 (3) (2016) 381–391.
[12] D.L.P. Echt, B. Mitchell, R. Peters, D. Obias-Manno, A. Barker, D. Arendberg, A. Baker, L. Firedman, L. Greene, M. Huther, D. Richardson, Mortality and morbidity in patients receiving encainide, flecainide or placebo: the cardiac arrhythmia suppression trial, N. Engl. J. Med. 324 (12) (1991) 781–788.
[13] Investigators TCAPSC: effects of encainide, flecainide, imipramine and moricizine on ventricular arrhythmias during the year after acute myocardial infarction: the CAPS, Am. J. Cardiol. 61 (8) (1988) 501–509.
[14] F.M. de Benedictis, R. Guidi, S. Carraro, E. Baraldi, Excellence TENo: endpoints in respiratory diseases, Eur. J. Clin. Pharmacol. 1 (67 Suppl) (2011) 49–59.
[15] C. Albera, Challenges in idiopathic pulmonary fibrosis trials: the point on endpoints, Eur. Respir. Rev. 20 (121) (2011) 195–200.
[16] P.R. Williamson, D.G. Altman, H. Bagley, K.L. Barnes, J.M. Blazeby, S.T. Brookes, M. Clarke, E. Gargon, S. Gorst, N. Harman, et al., The COMET handbook, Trials 18 (Suppl 3) (2017) 280, version 1.0.
[17] P.R. Williamson, D.G. Altman, J.M. Blazeby, M. Clarke, D. Devane, E. Gargon, P. Tugwell, Developing core outcome sets for clinical trials: issues to consider, Trials 13 (2012) 132.
[18] C.A. Prinsen, S. Vohra, M.R. Rose, S. King-Jones, S. Ishaque, Z. Bhaloo, D. Adams, C.B. Terwee, Core Outcome Measures in Effectiveness Trials (COMET) initiative: protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set', Trials 15 (2014) 247.
[19] E.A. Buzney, A.B. Kimball, A critical assessment of composite and coprimary endpoints: a complex problem, J. Am. Acad. Dermatol. 59 (5) (2008) 890–896.
[20] A.N. Duc, M. Wolbers, Weighted analysis of composite endpoints with simultaneous inference for flexible weight constraints, Stat. Med. 36 (3) (2017) 442–454.
[21] A.J. Sankoh, H. Li, R.B. D'Agostino, Sr.: use of composite endpoints in clinical trials, Stat. Med. 35 (2) (2016) 319–320.
[22] L.D. Fisher, Advances in clinical trials in the twentieth century, Annu. Rev. Public Health 20 (1999) 109–124.
[23] T.J. Iwashyna, A.M. Deane, Individualizing endpoints in randomized clinical trials to better inform individual patient care: the TARGET proposal, Crit. Care 20 (1) (2016) 218.
[24] S. Snapinn, Some remaining challenges regarding multiple endpoints in clinical trials, Stat. Med. 36 (28) (2017) 4441–4445.
[25] F. JC, A research approach to improving our quality of life, Am. Psychol. 33 (1978) 138–147.
[26] Agency EM: Research and Development. In. Edited by Development Ra. London.
[27] P.W. Armstrong, C.M. Westerhout, Composite end points in clinical research: a time for reappraisal, Circulation 135 (23) (2017) 2299–2307.
[28] S.D. Anker, S. Schroeder, D. Atar, J.J. Bax, C. Ceconi, M.R. Cowie, A. Crisp, F. Dominjon, I. Ford, H.A. Ghofrani, et al., Traditional and new composite endpoints in heart failure clinical trials: facilitating comprehensive efficacy assessments and improving trial efficiency, Eur. J. Heart Fail. 18 (5) (2016) 482–489.
[29] I. Ferreira-Gonzalez, P. Alonso-Coello, I. Sola, V. Pacheco-Huergo, A. Domingo-Salvany, J. Alonso, V. Montori, G. Permanyer-Miralda, Composite endpoints in clinical trials, Rev. Esp. Cardiol. 61 (3) (2008) 283–290.
[30] X. Luo, J. Qiu, S. Bai, H. Tian, Weighted win loss approach for analyzing prioritized outcomes, Stat. Med. 36 (15) (2017) 2452–2465.
[31] S.J. Pocock, C.A. Ariti, T.J. Collier, D. Wang, The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities, Eur. Heart J. 33 (2) (2012) 176–182.
[32] K. Richardson, KM: why are we weighting: do utility weights improve the predictive power of multi attribute utility instruments? in: Economics CfH (Ed.) Monash University, 2012, pp. 1–13.
[33] J.A. Bakal, C.M. Westerhout, P.W. Armstrong, Impact of weighted composite compared to traditional composite endpoints for the design of randomized controlled trials, Stat. Methods Med. Res. 24 (6) (2015) 980–988.
[34] S.F. Wong, R. Norman, T.L. Dunning, D.M. Ashley, P.K. Lorgelly, A protocol for a discrete choice experiment: understanding preferences of patients with cancer towards their cancer care across metropolitan and rural regions in Australia, BMJ Open 4 (10) (2014) e006661.
[35] R.N.R. Viney, M. King, P. Cronin, D. Street, Time trade-off derived EQ-5D weights for Australia, Value Health 14 (2011) 928–936.
[36] D. Hensher, BM: using stated response choice data to enrich revealed preference discrete choice models, Marketting Lett. 4 (2) (1993) 139–151.
[37] D.G.G. Capodanno, S. Buccheri, A. Chieffo, E. Meliga, Computing methods for composite clinical endpoints in unprotected left main coronary artery

revaculatization: a post-hoc analysis of the DELTA registry, JACC Cardiovasc. Imag. 9 (2016) 2280–2288.
[38] M.D.D.D. Clark, S. Petrou, D. Moro, E.W. de Bekker-Grobb, Discrete choice experiments in health conomics: a review of the literature, PharmacoEconomics 32 (9) (2014) 883–902.
[39] E.R.M. Bekker-Grob, K. Gerard, Discrete choice experiments in health economics: a review of the literature, Health Econ. 21 (2012) 145–172.
[40] W.H. Organisation, The world health organization quality of LIfe (WHOQOL), In (2019).
[41] D.R. VanDevanter, S.L. Heltshe, J. Spahr, V.V. Beckett, C.L. Daines, E. C. Dasenbrook, R.L. Gibson, J. Raksha, D.B. Sanders, C.H. Goss, et al., Rationalizing endpoints for prospective studies of pulmonary exacerbation treatment response in cystic fibrosis, J. Cyst. Fibros. 16 (5) (2017) 607–615.
[42] G.D.B.D. Murray, S. Choi, et al., Design and analysis of phase III trials with orered outcomes scales: the concept of the sliding dichotomy, J. Neurotrauma 22 (2005) 511–517.
[43] M. Calvert, J. Blazeby, D.G. Altman, D.A. Recicki, D. Moher, M.D. Brundage, Reporting of patient-reported outcomes in randomized trials, JAMA 309 (8) (2013) 814–822.
[44] S.L.E. Webb, T. Leen, Treatment goals: health care improvement through setting and measuring patient-centred outcomes, Crit. Care Resusc. 15 (2) (2013) 143–146.
[45] J. Abbott, A. Hart, T. Havermans, A. Matossian, L. Goldbeck, C. Barreto, A. Bergsten-Brucefors, T. Besier, P. Catastini, F. Lupi, et al., Measuring health-related quality of life in clinical trials in cystic fibrosis, J. Cyst. Fibros. 2 (10 Suppl) (2011) S82–S85.