Article

CLINICAL TRIALS

Clinical Trials
1–10
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1740774520944855
journals.sagepub.com/home/ctj
SAGE

Summarising salient information on historical controls: A structured assessment of validity and comparability across studies

Anthony Hatswell[1,2], Nick Freemantle[3], Gianluca Baio[1], Emmanuel Lesaffre[4] and Joost van Rosmalen[5]

Abstract

Background While placebo-controlled randomised controlled trials remain the standard way to evaluate drugs for efficacy, historical data are used extensively across the development cycle. This ranges from supplementing contemporary data to increase the power of trials to cross-trial comparisons in estimating comparative efficacy. In many cases, these approaches are performed without in-depth review of the context of data, which may lead to bias and incorrect conclusions.

Methods We discuss the original 'Pocock' criteria for the use of historical data and how the use of historical data has evolved over time. Based on these factors and personal experience, we created a series of questions that may be asked of historical data, prior to their use. Based on the answers to these questions, various statistical approaches are recommended. The strategy is illustrated with a case study in colorectal cancer.

Results A number of areas need to be considered with historical data, which we split into three categories: outcome measurement, study/patient characteristics (including setting and inclusion/exclusion criteria), and disease process/intervention effects. Each of these areas may introduce issues if not appropriately handled, while some may preclude the use of historical data entirely. We present a tool (in the form of a table) for highlighting any such issues. Application of the tool to a colorectal cancer data set demonstrates under what conditions historical data could be used and what the limitations of such an analysis would be.

Conclusion Historical data can be a powerful tool to augment or compare with contemporary trial data, though caution is required. We present some of the issues that may be considered when involving historical data and what (if any) statistical approaches may account for differences between studies. We recommend that, where historical data are to be used in analyses, potential differences between studies are addressed explicitly.

Keywords

Historical control, power prior, propensity scoring, matching-adjusted indirect comparison

Background

Clinical investigations of novel interventions are usually performed as randomised controlled trials (RCTs) against placebo or a relevant comparator. Although RCTs in which all patients are randomised to either the intervention or control treatment are the standard approach for comparing efficacy, there is interest in making use of historical data to reduce the need for contemporary controls. The historical data can, in principle, increase the power of tests for treatment efficacy and improve precision of the estimates. In some cases (e.g. in rare/orphan diseases), the use of historical data

[1]Department of Statistical Science, University College London, London, UK
[2]Delta Hat Limited, Nottingham, UK
[3]Institute of Clinical Trials and Methodology, University College London, London, UK
[4]L-Biostat, KU Leuven, Leuven, Belgium
[5]Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands

Corresponding author:
Anthony Hatswell, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK.
Email: ahatswell@deltahat.co.uk

is necessary to obtain a sufficiently powered analysis, due to the limited number of patients available for randomisation. In principle, both intervention and control patients of previous studies can be combined with data of the current study. Previous studies typically do not feature exactly the same intervention and control treatment, and especially the intervention may change between trials. Therefore, the focus lies on combining historical control patients with patients randomised to the control in the current study. The combination of historical and randomised controls to supplement the analysis of contemporary or future clinical trials was popularised by Pocock.[1] The main concern with the use of historical data is the possibility that the historical studies differ in characteristics with the current study, due to, for example, improvement in supportive care over time, differences in patient selection, and between-centre differences. Indeed, previous work has demonstrated substantial bias when comparing the outcomes of comparisons made using observational data to RCTs conducted in the same population.[2–7] This finding was also quantified in a simulation study.[8]

To overcome the issues discussed above, Pocock suggested criteria that should be met for the historical data to be deemed ‘acceptable’. These criteria are presented in Figure 1. Provided the criteria are met, the suggestion is that the historical data may be combined with contemporary randomised controls, thereby reducing the number of patients required in the control arm. Since the use of historical controls was originally proposed, their application has proliferated beyond the pooling of placebo data from sequential studies to include areas, such as sample size calculations, synthesis of published historical data, and comparative efficacy analyses for treatments, which are granted a marketing authorisation on the basis of uncontrolled studies.[9,10] Additional relevant criteria for the acceptability of historical data are drift, exchangeability, and conditional exchangeability.[11] Drift can be defined as a bias (difference in underlying model parameters) between the historical data and current data, which can arise due to general improvements in supportive care over time. Exchangeability implies that the historical studies and the current study constitute a random sample from a population of studies, whereas conditional exchangeability implies that the studies constitute a random sample from a population of studies only after accounting for differences in patient characteristics. Conditional exchangeability applies more generally than exchangeability. Under these criteria, meta-analytic approaches for combining the historical and current controls may be applied. The use of historical data has proven to be a useful tool for studies on orphan diseases,[12,13] whereas it remains controversial in the primary analysis of a phase III RCT.

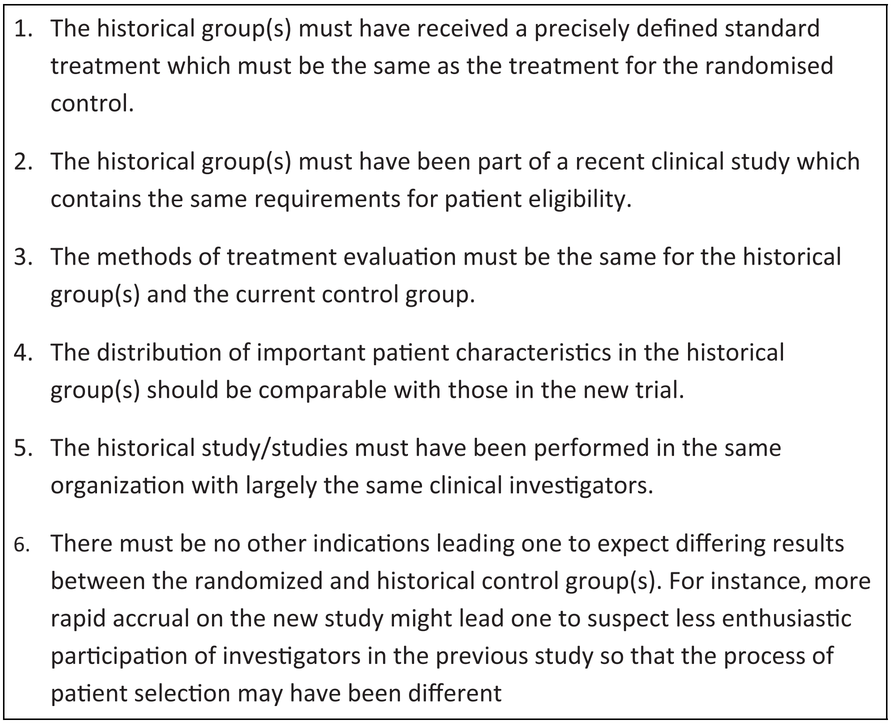
Historical controls may originate from a previous study of the same group/institution, from the literature, or from other sources (for example, patient registries). However, such data only occasionally meet the criteria set out originally by Pocock in an exact manner, which specify that the historical studies must have been conducted by the same investigators, have similar patient characteristics, and must have been done in roughly the same time period. Since such cross-study comparisons are of considerable interest, one could make use of techniques, such as propensity score matching,[14] among others,[15] which allow some of these barriers to be addressed. Depending on the reason for the use of historical controls and the purpose of the analysis, some of the Pocock criteria could be relaxed with the use of appropriate methodology without introducing bias. Likewise where multiple studies are available, there may be a desire to use all studies, but with different weightings.

Given the developments in statistical methodology, trial design, and medical research practice since the 1976 paper of Pocock, the objective of this article is to describe an updated tool for assessing the similarity between historical and contemporary data. With this aim in mind, we develop a set of questions which may help identify areas that differ between studies (whether in design, patients, or setting). We also discuss potentially appropriate statistical methods by which historical data may still be utilised in case data sets are not perfectly aligned. The application of the resulting tool is illustrated with an example of cetuximab compared to standard care in colorectal cancer. Our hope is that by highlighting the relevant issues (and signposting methodologies) the tool enables analysts to better understand the issues with historical data and justify the choice of methodology used for analysis.

Methods

We aim to identify the relevant different aspects between historical and current studies that should be taken into account when historical data are considered to be included in the analysis. Pocock’s original criteria serve as the starting point for the development of the tool. In addition, we considered the statistical assumptions implicit in the use of historical data, and which differences between studies could result in violations of those assumptions. These statistical considerations are described in Supplemental Appendix 1, together with the resulting insights used to augment the proposed tool.

In order to ascertain whether historical data set(s) may be sufficiently similar for use alongside current data (for instance, with some form of combining), the process by which outcomes were obtained should be understood. This means there is a need to compare the treatments given and the circumstances of the studies–for example, patient characteristics. We therefore divide



1. The historical group(s) must have received a precisely defined standard treatment which must be the same as the treatment for the randomised control.
2. The historical group(s) must have been part of a recent clinical study which contains the same requirements for patient eligibility.
3. The methods of treatment evaluation must be the same for the historical group(s) and the current control group.
4. The distribution of important patient characteristics in the historical group(s) should be comparable with those in the new trial.
5. The historical study/studies must have been performed in the same organization with largely the same clinical investigators.
6. There must be no other indications leading one to expect differing results between the randomized and historical control group(s). For instance, more rapid accrual on the new study might lead one to suspect less enthusiastic participation of investigators in the previous study so that the process of patient selection may have been different

Figure 1. Acceptability conditions proposed by Pocock regarding the use of a historical control group.

the relevant issues into three areas (outcome measurement, study and patient characteristics, and disease process and intervention effects) and discuss each in turn. These issues are illustrated statistically in Supplemental Appendix 1 and described in non-technical language below. These areas are subsequently used to derive questions that can help identify the relevant differences between studies.

Outcome measurement

One of the changes that can occur over time is the way outcomes are measured. For example, as technology has advanced, tests have become more sensitive, and definitions have evolved – for example, the widely used Common Terminology Criteria for Adverse Events (CTCAE) are now at version 4.0. There have also been changes in the types of outcomes used in studies over time, for instance, moving from response rates to median survival, and then subsequently beyond to endpoints such as restricted mean survival time.[16] Any differences in how outcomes are measured/defined (or in the type of outcome) should therefore be understood and accounted for – for example, by reanalysis of the contemporary trial or mapping between endpoints – even if outcomes are named similarly, for example, partial response, the definition may have changed over time. If the differences between studies cannot be bridged, the historical data may need to be discarded.

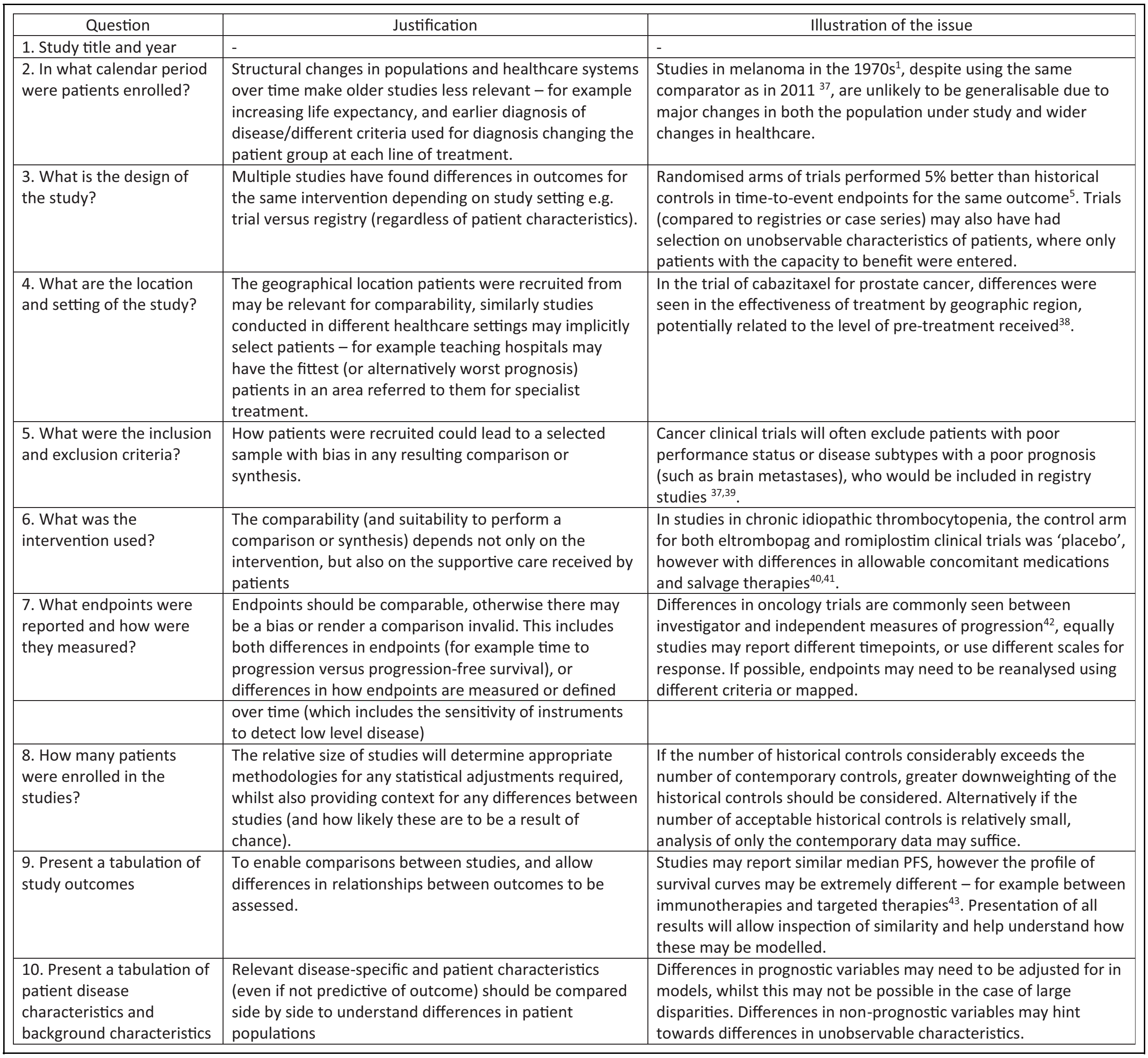
Study and patient characteristics

In order to combine the historical data with current data, the inputs to the process by which outcomes are generated, that is, the interaction of the disease process and the mechanism of action of any interventions, must remain similar. In practice, this implies that the study inclusion criteria and the individual patient characteristics should be assessed for similarity between the historical studies and the current study. If these are not fully aligned, for example, if the historical patients and the current patients exhibit different characteristics, statistical methods accounting for this imbalance across studies may be required (which are likely to, correctly, increase the uncertainty around estimates).

If there is little overlap in patient characteristics or if there are structural differences in inclusion criteria, the historical studies and current study may be incompatible, such that no statistical adjustment method would be able to overcome the unquantifiable bias. An example here would be if some but not all studies required patients to first complete a ‘wash-out’ period from their previous treatment. This implies that only patients who survive to the beginning of the study (a form of immortal time bias)[17] are included in the results.

Disease process and intervention effects

Furthermore, the way in which outcomes are achieved must be similar across studies. However, differences



Question	Justification	Illustration of the issue
1. Study title and year	-	-
2. In what calendar period were patients enrolled?	Structural changes in populations and healthcare systems over time make older studies less relevant – for example increasing life expectancy, and earlier diagnosis of disease/different criteria used for diagnosis changing the patient group at each line of treatment.	Studies in melanoma in the 1970s[1], despite using the same comparator as in 2011 [37], are unlikely to be generalisable due to major changes in both the population under study and wider changes in healthcare.
3. What is the design of the study?	Multiple studies have found differences in outcomes for the same intervention depending on study setting e.g. trial versus registry (regardless of patient characteristics).	Randomised arms of trials performed 5% better than historical controls in time-to-event endpoints for the same outcome[5]. Trials (compared to registries or case series) may also have had selection on unobservable characteristics of patients, where only patients with the capacity to benefit were entered.
4. What are the location and setting of the study?	The geographical location patients were recruited from may be relevant for comparability, similarly studies conducted in different healthcare settings may implicitly select patients – for example teaching hospitals may have the fittest (or alternatively worst prognosis) patients in an area referred to them for specialist treatment.	In the trial of cabazitaxel for prostate cancer, differences were seen in the effectiveness of treatment by geographic region, potentially related to the level of pre-treatment received[38].
5. What were the inclusion and exclusion criteria?	How patients were recruited could lead to a selected sample with bias in any resulting comparison or synthesis.	Cancer clinical trials will often exclude patients with poor performance status or disease subtypes with a poor prognosis (such as brain metastases), who would be included in registry studies [37,39].
6. What was the intervention used?	The comparability (and suitability to perform a comparison or synthesis) depends not only on the intervention, but also on the supportive care received by patients	In studies in chronic idiopathic thrombocytopenia, the control arm for both eltrombopag and romiplostim clinical trials was 'placebo', however with differences in allowable concomitant medications and salvage therapies[40,41].
7. What endpoints were reported and how were they measured?	Endpoints should be comparable, otherwise there may be a bias or render a comparison invalid. This includes both differences in endpoints (for example time to progression versus progression-free survival), or differences in how endpoints are measured or defined over time (which includes the sensitivity of instruments to detect low level disease)	Differences in oncology trials are commonly seen between investigator and independent measures of progression[42], equally studies may report different timepoints, or use different scales for response. If possible, endpoints may need to be reanalysed using different criteria or mapped.
8. How many patients were enrolled in the studies?	The relative size of studies will determine appropriate methodologies for any statistical adjustments required, whilst also providing context for any differences between studies (and how likely these are to be a result of chance).	If the number of historical controls considerably exceeds the number of contemporary controls, greater downweighting of the historical controls should be considered. Alternatively if the number of acceptable historical controls is relatively small, analysis of only the contemporary data may suffice.
9. Present a tabulation of study outcomes	To enable comparisons between studies, and allow differences in relationships between outcomes to be assessed.	Studies may report similar median PFS, however the profile of survival curves may be extremely different – for example between immunotherapies and targeted therapies[43]. Presentation of all results will allow inspection of similarity and help understand how these may be modelled.
10. Present a tabulation of patient disease characteristics and background characteristics	Relevant disease-specific and patient characteristics (even if not predictive of outcome) should be compared side by side to understand differences in patient populations	Differences in prognostic variables may need to be adjusted for in models, whilst this may not be possible in the case of large disparities. Differences in non-prognostic variables may hint towards differences in unobservable characteristics.

Figure 2. Motivating example: Cetuximab in metastatic colorectal cancer.

may arise for several reasons. For example, changes in supportive care over time may improve outcomes, even if the intervention remains unchanged. Alternatively, interventions may change over time as their usage is refined, or protocols in different centres may prescribe a different usage. For example, the use of stem cell transplantation has been continually refined over time, with improvements in both short-term and long-term mortality.[18]

It is also possible that the impact of the disease changes over time – either as screening improves resulting in earlier diagnosis (and patients with a more favourable outcome) or as a new treatment is introduced at early stage disease, meaning patients who do progress have much more severe disease subtypes. It is also possible that the disease itself changes in the case of pathogens, which may evolve over time or become resistant to treatments (as seen with influenza, bacterial infection, and HIV).

Application to illustrative example: cetuximab in metastatic colorectal cancer

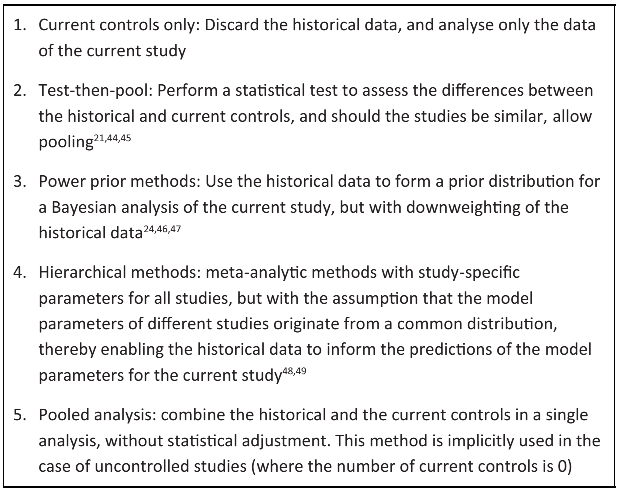
To illustrate and evaluate the proposed framework, it was applied to a motivating example of a historical controlled study conducted by Annemans et al.[19] of cetuximab in metastatic colorectal cancer (mCRC) (Figure 2). Cetuximab was licenced on the basis of the BOND study (see Cunningham et al.[20]), an RCT of 329 patients which compared cetuximab plus irinotecan versus cetuximab monotherapy in 11 centres across Europe. This study design addressed the benefit of combination therapy, which was shown to provide a higher response rate than monotherapy (22.9% versus 10.8%), but did not provide an estimate of the efficacy of cetuximab versus standard care. To make such a comparison for the benefit of reimbursement agencies, Annemans et al.[19] conducted a retrospective review of patient notes to identify untreated patients from the largest centres in BOND. This investigation was conducted in the three largest centres, enrolling 66 patients who received standard care in the same centres, but outside the period the cetuximab trial was ongoing. The tool was therefore used to evaluate the comparison of these two studies.

Results

Using Pocock's criteria as a starting point, we updated and more elaborately described the issues that are relevant for incorporating historical data, which yielded the proposed tool for assessing historical data that is presented in Figure 3. In this figure, each issue is highlighted with a reference to the literature illustrating the issue.

The areas of importance are the patient characteristics, precise intervention used (including supportive care), outcome measurement, and patient selection to the study. The statistical/methodological rationale for the question as well as an example of how studies may be non-comparable and the impact on any comparison(s) are given alongside each question. The tool in Figure 3 may be used to identify areas where historical controls differ from contemporary data, and as a result whether statistical adjustments may be required, or where there are such important differences that only narrative comparisons are warranted.

The items in the tool do have a large degree of overlap with the criteria proposed by Pocock; however, they are different in nature. Rather than proposing a set of dichotomous criteria that must all be met to allow for pooling or synthesis of the studies, our proposed tool instead seeks to present relevant data and study design aspects and to identify and quantify differences, giving a more nuanced picture. This is in line with the aim of comparing between studies and allowing a judgement to be made of the appropriate next steps. The main additions we make to the original criteria pertain to the study design and the patient selection, which in the original Pocock criteria were taken to be placebo arms from the same centre. The tool also asks for the study results to be presented for comparison, as these may indicate differences between studies (if the same intervention is used in both arms) or be required for the use of statistical methods such as 'test then pool' as a next step.[21]



1. Current controls only: Discard the historical data, and analyse only the data of the current study
2. Test-then-pool: Perform a statistical test to assess the differences between the historical and current controls, and should the studies be similar, allow pooling[21,44,45]
3. Power prior methods: Use the historical data to form a prior distribution for a Bayesian analysis of the current study, but with downweighting of the historical data[24,46,47]
4. Hierarchical methods: meta-analytic methods with study-specific parameters for all studies, but with the assumption that the model parameters of different studies originate from a common distribution, thereby enabling the historical data to inform the predictions of the model parameters for the current study[48,49]
5. Pooled analysis: combine the historical and the current controls in a single analysis, without statistical adjustment. This method is implicitly used in the case of uncontrolled studies (where the number of current controls is 0)

Figure 3. Questions and justifications regarding important items for historical controls.

Illustrative example: cetuximab in mCRC

The study by Annemans et al. appears to be of high quality in attempting to identify similar patients to those from the clinical trials, from the same centres and date range – although imperfect, it is a pragmatic attempt to estimate the outcomes of standard care without availability of RCT or network of trials.

Using our proposed tool, it is apparent that, although the largest centres from the Cunningham et al.'s study were used as the source of control data, there may be differences between these studies and the other centres enrolled – due to their size or their location (Belgium, France, and Italy) being unrepresentative. To present a valid comparison between combination therapy and standard care, a useful next step would be to inspect the subsample of the Cunningham et al.'s study using only the centres enrolled in the Annemans et al.'s study. The patient characteristics for the Annemans et al.'s study are not given in the publication even at the aggregate level – to be confident in the use of these data as a historical control, the distribution of patient characteristics should be shown.

Provided there is good overlap between the study characteristics and patient characteristics within the studies, the use of the Annemans et al.'s data as a historical control would seem appropriate, provided it is used with suitable statistical techniques (which may involve either matching or weighting methods). To perform any statistical adjustments in this example, access would be required to the patient-level data from both studies. The main limitation of such a comparison is also highlighted through the use of our proposed table, which clearly identifies that only overall survival (OS) is available in the Annemans et al.'s data set. Comparisons on the primary endpoint of the Cunningham et al.'s study (response rate) or time to progression are not possible without the use of further assumptions.

Appropriate statistical methods

A variety of methods has been proposed to account for differences in observed patient characteristics between historical and contemporary data. These methods include the aforementioned use of not only propensity scores but also meta-regression,[22] matching-adjusted indirect comparisons,[23] simulated treatment comparisons,[24] and other regression-based techniques[25]– each of which have varying data requirements.

If there is a desire to perform a combined analysis of historical and contemporary data, an appropriate statistical method or methodology to adjust for study-specific effects and other potentially unobservable differences between studies is required; see Wadsworth et al.[26] for a systematic review. A selection of techniques that perform adjustments for between-study differences is presented in Figure 4, with the ordering of techniques based on how stringent the required assumptions regarding comparability are. These techniques range from simply discarding the historical data to 'naïve' pooling of historical and contemporary evidence. The former of these methods is appropriate if there are irreconcilable differences between historical and current controls, while the latter is acceptable only if the historical and current controls originate from essentially the same study – a criterion that seldom applies.

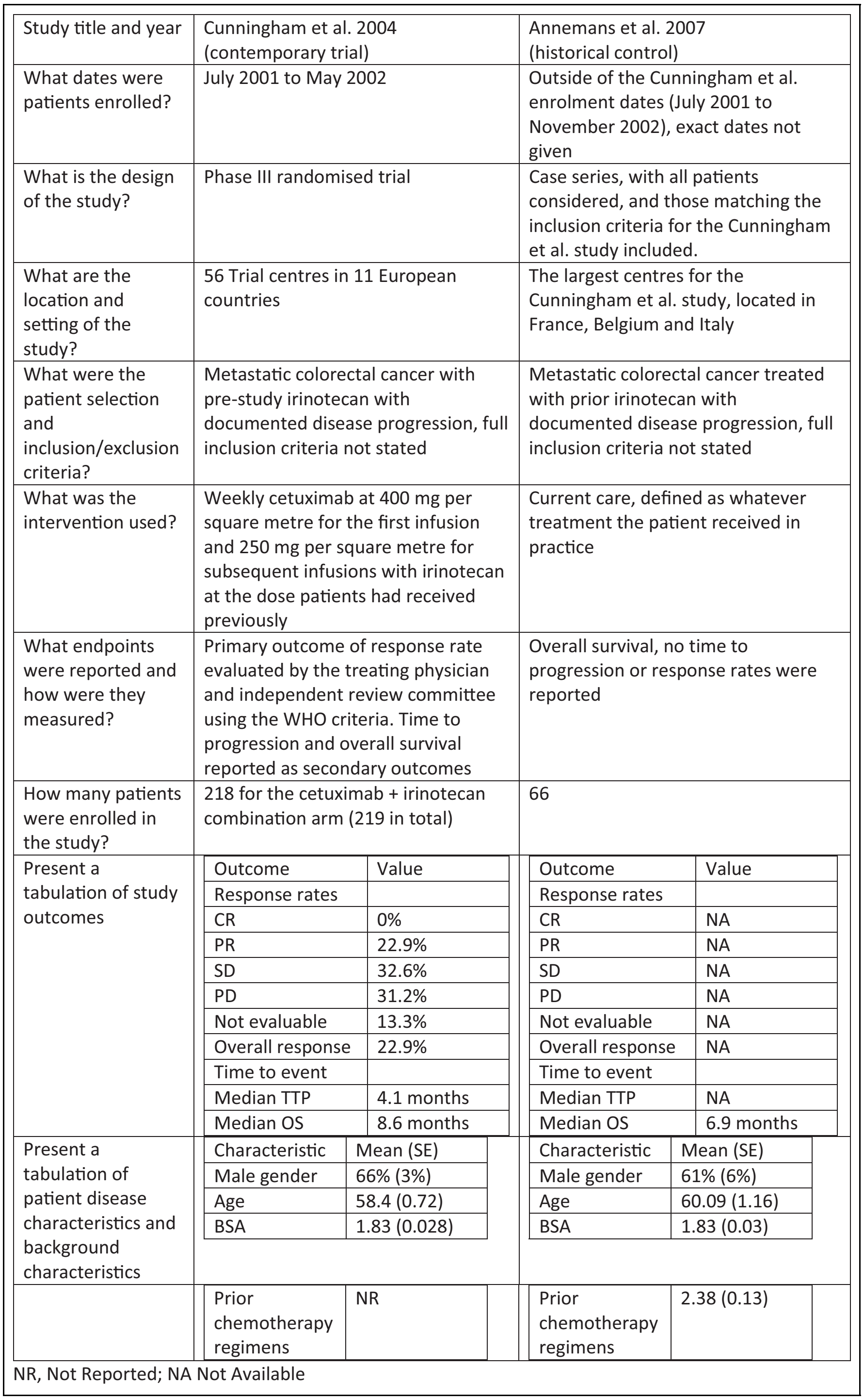
The remaining three highlighted methods (test then pool, power priors, and hierarchical methods) can be used in intermediate situations, where there are no insurmountable differences in outcome measures, patient characteristics, and disease process/intervention effect, but some adjustment is still necessary. Most implementations of these methods perform 'dynamic borrowing', that is, the information in the historical controls is given a lower weight in the analysis depending on the size of the observed differences between historical and contemporary controls – the more similar the data, the higher weight assigned to the historical data. Recent simulation studies have shown that the three intermediate methods (test-then-pool, power prior, and meta-analytic methods) may suitably account for between-study differences and may also be used to control the type I error rate. In particular, the meta-analytic methods appear to lead to the largest reduction in bias and ability to control the type I error rate.[27] New variants and adaptations of these methods are frequently proposed such as a test-then-pool approach based on equivalence tests,[28] extensions of the power prior to data with multiple historical studies, and combinations of these methods.[29] However, none of these methods are guaranteed to control the type I error rate at the nominal 5% level.[30] While these techniques are aimed at accounting for between-study heterogeneity (which is unobserved), in the case of observed differences, for example, patient characteristics, these methods may need to be augmented either with adjustment techniques or by including the relevant patient characteristics as covariates in the statistical model.

The 'naïve' pooling of studies is the final method in the list, which implies that no kind of statistical adjustment is made for differences between historical and current controls. We believe that there are applications where pooling data may be appropriate, however, only if the strictest criteria for comparability are met, and between-study differences can virtually be ruled out. Potential examples of such applications are the example given in the 1976 paper (sequential studies in a single centre, if appropriate) and some studies used for US Food and Drug Administration approvals. At times, two RCTs are conducted in parallel with essentially the same protocol and time frame, but in different centres – should these circumstances be met, patients may be allocated by the instrument of geography into one study or the other. Even in such cases (effective allocation by time or geography) between-study differences should be investigated and naïve pooling should be used with extreme caution.

Discussion

The questions identified in our tool lead to a summary of relevant information of the historical controls. From this information, we can identify where data are suitable for pooling (i.e. rarely), where they could be used with statistical adjustment, or alternatively where the review may highlight that an informative comparison between studies is not possible. The questions included in the table represent key areas where studies may differ, collated based on Pocock's original criteria, statistical theory, and the experience of the authors. The novelty of this tool is that it provides a framework for systematically comparing a historical study to a current study, using both the observed data and the study designs, while leaving necessary room for debate. Except for the work of Pocock, few authors have discussed or proposed clear criteria for choosing historical data. See the work of Lim et al.[31,32] for a broader review of general principles to consider when selecting and incorporating historical control data.

Because a single major difference can render a comparison between studies inappropriate (in line with the general experience from quality scores for clinical studies), we do not suggest to use the tool to score trials in an objective and quantitative manner. However, we



Study title and year	Cunningham et al. 2004 (contemporary trial)		Annemans et al. 2007 (historical control)	
What dates were patients enrolled?	July 2001 to May 2002		Outside of the Cunningham et al. enrolment dates (July 2001 to November 2002), exact dates not given	
What is the design of the study?	Phase III randomised trial		Case series, with all patients considered, and those matching the inclusion criteria for the Cunningham et al. study included.	
What are the location and setting of the study?	56 Trial centres in 11 European countries		The largest centres for the Cunningham et al. study, located in France, Belgium and Italy	
What were the patient selection and inclusion/exclusion criteria?	Metastatic colorectal cancer with pre-study irinotecan with documented disease progression, full inclusion criteria not stated		Metastatic colorectal cancer treated with prior irinotecan with documented disease progression, full inclusion criteria not stated	
What was the intervention used?	Weekly cetuximab at 400 mg per square metre for the first infusion and 250 mg per square metre for subsequent infusions with irinotecan at the dose patients had received previously		Current care, defined as whatever treatment the patient received in practice	
What endpoints were reported and how were they measured?	Primary outcome of response rate evaluated by the treating physician and independent review committee using the WHO criteria. Time to progression and overall survival reported as secondary outcomes		Overall survival, no time to progression or response rates were reported	
How many patients were enrolled in the study?	218 for the cetuximab + irinotecan combination arm (219 in total)		66	
Present a tabulation of study outcomes	Outcome	Value	Outcome	Value
	Response rates		Response rates	
	CR	0%	CR	NA
	PR	22.9%	PR	NA
	SD	32.6%	SD	NA
	PD	31.2%	PD	NA
	Not evaluable	13.3%	Not evaluable	NA
	Overall response	22.9%	Overall response	NA
	Time to event		Time to event	
	Median TTP	4.1 months	Median TTP	NA
	Median OS	8.6 months	Median OS	6.9 months
Present a tabulation of patient disease characteristics and background characteristics	Characteristic	Mean (SE)	Characteristic	Mean (SE)
	Male gender	66% (3%)	Male gender	61% (6%)
	Age	58.4 (0.72)	Age	60.09 (1.16)
	BSA	1.83 (0.028)	BSA	1.83 (0.03)
	Prior chemotherapy regimens	NR	Prior chemotherapy regimens	2.38 (0.13)

NR, Not Reported; NA Not Available

Figure 4. Recognised statistical methods for the use of historical data.

hope that in answering the questions presented, a strong basis for a decision on the comparability of studies may be achieved. When compared with Pocock's original criteria, the questions ask for details of the studies, instead of asking if they are the same. This is in keeping with the different objective, of understanding the similarity of studies for further analysis – we then highlight a variety of techniques that may aid the analyst in conducting such pooling. While there is an unavoidable need for judgement of similarity, and when different methods would be appropriate, we would hope that this at least makes such decisions explicit, rather than implicit.

The proposed criteria do not yield an unequivocal decision of whether the historical data are to be considered 'acceptable'. However, the criteria should enable researchers to make an assessment of whether there is a risk of drift (systematic bias in the historical data compared to the current data) and whether the assumption of exchangeability or conditional exchangeability is reasonable. Exchangeability is an assumption that is relatively difficult to demonstrate and communicate, whereas drift (bias) seems more straightforward. Nevertheless, we believe that both drift and exchangeability are useful criteria for assessing historical data and choosing an appropriate analysis method. Previous simulation work showed that some dynamic borrowing methods are relatively robust as long as the historical data are (conditionally) exchangeable, whereas even a small systematic bias in the historical data threatened the performance of these methods and led to inflated type I error rates.[27] Drift due to improvements in care over time is also an important concern in practice. In a meta-analysis by Snyders et al.,[33] who looked at 63 trials of docetaxel in lung cancer (enrolling over 10,000 patients), it was found that outcomes improved each year by a mean of 0.3% in objective response rate (ORR), 0.5% for progression-free survival (PFS), and 0.9% in OS. These changes over time imply that historical data that are more than say 10 years old for docetaxel in this indication should be avoided. The rate of drift in other diseases will differ, but the potential for change is something the analyst should be aware of.

The application of the tool is illustrated by our motivating example, where it highlights the differences between studies and leads naturally to the next steps required for comparisons to be drawn. In the majority of cases, these next steps will consist of further statistical analysis which may be as simple as trimming data sets to ensure entry criteria are similar or may involve more complex modelling as mapping between outcomes and weighting of patients. Where multiple studies are available, more complex techniques may be needed such as those highlighted in Figure 4. The tool may also highlight that the differences between studies are too great, so that some or all historical studies should be omitted from the analysis.

The main advantage of the proposed tool is that in a side-by-side comparison, differences between studies are highlighted, and their importance can be discussed as opposed to simply referred to as a 'historical control'. The tool can then be presented as a single table or figure in a journal article or evidence dossier for submission to regulatory bodies or payers. Its use would also not represent additional burden, as should historical data be used, it is reasonable to expect the detail of the comparison to be provided (if anything, this burden should reduce).

The main limitation of the proposed tool is that it is not possible to know whether there exists a bias even in trials that appear superficially similar. This is particularly the case in the case of uncontrolled studies – in such instances, it is not possible to compare control arms to assess between-study variation, which has been highlighted in the literature as an area of concern[34] and for future research.[25] Despite this limitation, uncontrolled studies with carefully selected historical controls should be preferable to uncontrolled studies that compare observed outcomes to an assumed outcome so that sampling variability in the historical data is ignored, or, even worse, when the response rate under the null hypothesis lacks a clear data-based justification.[35] Similarly, there may be other biases present; for instance, a sponsor may only commission a Phase 3 study following promising results in Phase 2 – a finding which has been seen with medical interventions, in general.[36] Beyond these limitations, we would also note that the particular tools that should be used for statistical adjustments are not specified and do rely on judgement as there is no easily quantifiable decision rule or flowchart to choose the type of statistical technique.

While the use of historical controls does not represent the highest level of data on the 'evidence pyramid', in many circumstances, such as rare cancers, orphan diseases, and some Class 3 medical devices, their use is necessary. In other cases, appropriate use of historical data may be seen to be more ethical by reducing the need to expose patients to control treatments and reducing the cost of trials. The tool we present may help improve the quality and appropriateness of such historical comparisons in a variety of settings. With appropriate use of historical data (collected at great financial and human cost), the quality of decision-making by clinicians, regulators, and payers could improve, ultimately leading to better patient outcomes.

Declaration of conflicting interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Anthony Hatswell https://orcid.org/0000-0003-1129-326X
Joost van Rosmalen https://orcid.org/0000-0002-9187-244X

Supplemental material

Supplemental material for this article is available online.

References

1. Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis* 1976; 29(3): 175–188.

2. Thall PF and Simon R. Incorporating historical control data in planning phase II clinical trials. *Stat Med* 1990; 9(3): 215–228.
3. Ioannidis JP, Cappelleri JC and Lau J. Issues in comparisons between meta-analyses and large trials. *JAMA* 1998; 279: 1089–1093.
4. Diehl LF and Perry DJ. A comparison of randomized concurrent control groups with matched historical control groups: are historical controls valid. *J Clin Oncol* 1986; 4(7): 1114–1120.
5. Moroz V, Wilson JS, Kearns P, et al. Comparison of anticipated and actual control group outcomes in randomised trials in paediatric oncology provides evidence that historically controlled studies are biased in favour of the novel treatment. *Trials* 2014; 15: 481.
6. MacLehose RR, Reeves BC, Harvey IM, et al. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000; 4(34): 1–154.
7. Sacks H, Chalmers TC and Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med* 1982; 72(2): 233–240.
8. Tang H, Foster NR, Grothey A, et al. Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. *J Clin Oncol* 2010; 28: 1936–1941.
9. Hatswell AJ, Baio G, Berlin JA, et al. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014. *BMJ Open* 2016; 6: e011666.
10. Hatswell AJ, Freemantle N and Baio G. Economic evaluations of pharmaceuticals granted a marketing authorisation without the results of randomised trials: a systematic review and taxonomy. *Pharmacoeconomics* 2017; 35(2): 163–176.
11. Greenland S and Robins JM. Identifiability, exchangeability and confounding revisited. *Epidemiol Perspect Innov* 2009; 6: 1–9.
12. Weber K, Hemmings R and Koch A. How to use prior knowledge and still give new data a chance. *Pharm Stat* 2018; 17(4): 329–341.
13. Sardella M and Belcher G. Pharmacovigilance of medicines for rare and ultrarare diseases. *Ther Adv Drug Saf* 2018; 9(11): 631–638.
14. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70: 41.
15. Faria R, Alava MH, Manca A, et al. NICE DSU technical support document 17: the use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data, http://nicedsu.org.uk/wp-content/uploads/2016/03/TSD17-DSU-Observational-data-FINAL.pdf (2015, accessed 1 June 2020).
16. Trinquart L, Jacot J, Conner SC, et al. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *J Clin Oncol* 2016; 34: 1813–1819.
17. Lévesque LE, Hanley JA, Kezouh A, et al. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ* 2010; 340: b5087.
18. Sureda A, Arranz R, Iriondo A, et al. Autologous stem-cell transplantation for Hodgkin's disease: results and prognostic factors in 494 patients from the Grupo Espanol de Linfomas/Transplante Autologo de Medula Osea Spanish Cooperative Group. *J Clin Oncol* 2001; 19: 1395–1404.
19. Annemans L, Van Cutsem E, Humblet Y, et al. Cost-effectiveness of cetuximab in combination with irinotecan compared with current care in metastatic colorectal cancer after failure on irinotecan – a Belgian analysis. *Acta Clin Belg* 2007; 62(6): 419–425.
20. Cunningham D, Humblet Y, Siena S, et al. Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *N Engl J Med* 2004; 351: 337–345.
21. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat* 2014; 13(1): 41–54.
22. Berlin JA, Santanna J, Schmid CH, et al. Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med* 2002; 21: 371–387.
23. Signorovitch JE, Sikirica V, Erder MH, et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value Health* 2012; 15(6): 940–947.
24. Caro JJ and Ishak KJ. No head-to-head trial? Simulate the missing arms. *Pharmacoeconomics* 2010; 28(10): 957–967.
25. Phillippo D, Ades AE, Dias S, et al. NICE DSU technical support document 18: methods for population-adjusted indirect comparisons in submissions to NICE. *Decision Support Unit, Scharr, University of Sheffield, Sheffield*, 2016.
26. Wadsworth I, Hampson LV and Jaki T. Extrapolation of efficacy and other data to support the development of new medicines for children: a systematic review of methods. *Stat Methods Med Res* 2018; 27(2): 398–413.
27. van Rosmalen J, Dejardin D, van Norden Y, et al. Including historical data in the analysis of clinical trials: is it worth the effort? *Stat Methods Med Res* 2018; 27: 3167–3182.
28. Li W, Liu F and Snavely D. Revisit of test-then-pool methods and some practical considerations. *Pharm Stat*. Epub ahead of print 14 March 2020. DOI: 10.1002/pst.2009.
29. Liu GF. A dynamic power prior for borrowing historical data in noninferiority trials with binary endpoint. *Pharm Stat* 2018; 17(1): 61–73.
30. Banbeta A, van Rosmalen J, Dejardin D, et al. Modified power prior with multiple historical trials for binary endpoints. *Stat Med* 2019; 38: 1147–1169.
31. Lim J, Walley R, Yuan J, et al. Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: review of methods and opportunities. *Ther Innov Regul Sci* 2018; 52(5): 546–559.
32. Lim J, Wang L, Best N, et al. Reducing patient burden in clinical trials through the use of historical controls: appropriate selection of historical data to minimize risk of bias. *Ther Innov Regul Sci* 2020; 54(4): 850–860.
33. Snyders K, Cho D, Hong JH, et al. Benchmarking single-arm studies against historical controls from non-small cell lung cancer trials – an empirical analysis of bias. *Acta Oncol Stockh Swed*. Epub ahead of print14 October 2019. DOI: 10.1080/0284186X.2019.1674452.

34. Byar DP, Schoenfeld DA, Green SB, et al. Design considerations for AIDS trials. *N Engl J Med* 1990; 323: 1343–1348.
35. Vickers AJ, Ballen V and Scher HI. Setting the bar in phase II trials: the use of historical data for determining 'go/no go' decision for definitive phase III testing. *Clin Cancer Res* 2007; 13: 972–976.
36. Pereira TV, Horwitz RI and Ioannidis JPA. Empirical evaluation of very large treatment effects of medical interventions. *JAMA* 2012; 308: 1676–1684.
37. Robert C, Thomas L, Bondarenko I, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. *N Engl J Med* 2011; 364: 2517–2526.
38. de Bono JS, Oudard S, Ozguroglu M, et al. Prednisone plus cabazitaxel or mitoxantrone for metastatic castration-resistant prostate cancer progressing after docetaxel treatment: a randomised open-label trial. *Lancet* 2010; 376: 1147–1154.
39. Kosary CL, Altekruse SF, Ruhl J, et al. Clinical and prognostic factors for melanoma of the skin using SEER registries: collaborative stage data collection system, version 1 and version 2. *Cancer* 2014; 120(Suppl. 23): 3807–3814.
40. Bussel JB, Cheng G, Saleh MN, et al. Eltrombopag for the treatment of chronic idiopathic thrombocytopenic purpura. *N Engl J Med* 2007; 357: 2237–2247.
41. Kuter DJ, Bussel JB, Lyons RM, et al. Efficacy of romiplostim in patients with chronic immune thrombocytopenic purpura: a double-blind randomised controlled trial. *Lancet Lond Engl* 2008; 371: 395–403.
42. Amit O, Bushnell W, Dodd L, et al. Blinded independent central review of the progression-free survival endpoint. *Oncologist* 2010; 15(5): 492–495.
43. Lee D, Porter J, Hertel N, et al. Modelling comparative efficacy of drugs with different survival profiles: ipilimumab, vemurafenib and dacarbazine in advanced melanoma. *Biodrugs* 2016; 30(4): 307–319.
44. Nikolakopoulos S, van der Tweel I and Roes KCB. Dynamic borrowing through empirical power priors that control type I error. *Biometrics* 2018; 74(3): 874–880.
45. Han B, Zhan J, John Zhong Z, et al. Covariate-adjusted borrowing of historical control data in randomized clinical trials. *Pharm Stat* 2017; 16(4): 296–308.
46. Neuenschwander B, Branson M and Spiegelhalter DJ. A note on the power prior. *Stat Med* 2009; 28: 3562–3566.
47. Ibrahim JG and Chen M-H. Power prior distributions for regression models. *Stat Sci* 2000; 15: 46–60.
48. Schmidli H, Gsteiger S, Roychoudhury S, et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 2014; 70(4): 1023–1032.
49. Schmidli H, Wandel S and Neuenschwander B. The network meta-analytic-predictive approach to non-inferiority trials. *Stat Methods Med Res* 2012; 22: 219–240.