Enriching single-arm clinical trials with external controls: possibilities and pitfalls

Jérôme Lambert, Etienne Lengliné, Raphaël Porcher, Rodolphe Thiébaut, Sarah Zohar, Sylvie Chevret

REVIEW ARTICLE

blood advances

Enriching single-arm clinical trials with external controls: possibilities and pitfalls

Jérôme Lambert,[1,2] Etienne Lengliné,[3] Raphaël Porcher,[4,5] Rodolphe Thiébaut,[6,7] Sarah Zohar,[8,9] and Sylvie Chevret[1,2]

[1]Biostatistical Department, Hôpital Saint-Louis, Assistance Publique–Hôpitaux de Paris, Paris, France; [2]Epidemiology and Clinical Statistics for Tumor, Respiratory, and Resuscitation Assessments (ECSTRRA) Team, UMR1153, INSERM, Université Paris Cité, Paris, France; [3]Department of Hematology, Hôpital Saint-Louis, Assistance Publique–Hôpitaux de Paris, Paris, France; [4]Center for Clinical Epidemiology, Hôtel-Dieu, Assistance Publique–Hôpitaux de Paris, Paris, France; [5]The Institut national de la recherche agronomique (INRAE), Université Paris Cité, INSERM, CRESS-UMR1153, Paris, France; [6]Medical Information Department, Centre Hospitalier Universitaire Bordeaux, Bordeaux, France; [7]University of Bordeaux, INRIA SISTM, Bordeaux, France; [8]Centre de Recherche des Cordeliers, Université Paris Cité, Sorbonne Université, INSERM, Paris, France; and [9]Inria, HeKA, Inria Paris, Paris, France

For the past decade, it has become commonplace to provide rapid answers and early patient access to innovative treatments in the absence of randomized clinical trials (RCT), with benefits estimated from single-arm trials. This trend is important in oncology, notably when assessing new targeted therapies. Some of those uncontrolled trials further include an external/synthetic control group as an innovative way to provide an indirect comparison with a pertinent control group. We aimed to provide some guidelines as a comprehensive tool for (1) the critical appraisal of those comparisons or (2) for performing a single-arm trial. We used the example of ciltacabtagene autoleucel for the treatment of adult patients with relapsed or refractory multiple myeloma after 3 or more treatment lines as an illustrative example. We propose a 3-step guidance. The first step includes the definition of an estimand, which encompasses the treatment effect and the targeted population (whole population or restricted to single-arm trial or external controls), reflecting a clinical question. The second step relies on the adequate selection of external controls from previous RCTs or real-world data from patient cohorts, registries, or electronic patient files. The third step consists of choosing the statistical approach targeting the treatment effect defined above and depends on the available data (individual-level data or aggregated external data). The validity of the treatment effect derived from indirect comparisons heavily depends on careful methodological considerations included in the proposed 3-step procedure. Because the level of evidence of a well-conducted RCT cannot be guaranteed, the evaluation is more important than in standard settings.

Introduction

In oncology, new classes of anticancer agents have become an increasingly available and promising treatment option in several cancer indications, looking for precision cancer treatment.[1] The development of these innovative therapies, such as molecularly-targeted agents, has led to an important modification in the evaluation process of cancer drugs, with an apparent need to improve the speed and efficiency of drug development. This has changed the way tolerance[2] and antitumor activity[3] are

Submitted 19 October 2022; accepted 14 December 2022; prepublished online on *Blood Advances* First Edition 19 December 2022; final version published online 27 September 2023. https://doi.org/10.1182/bloodadvances.2022009167.

assessed in clinical trials, especially for early-stage trials. In contrast to the standard and separated phase I-II-III trials, accelerating clinical research with fewer patients involved and reduced costs may appear justified from the perspectives of both patients and public health.[4] To this aim, single-arm trials are growingly reported as the sole basis for evaluating the efficacy of cancer drugs, mostly based on a surrogate end point,[5] and this impacts the whole approval pathway.[6] This observation is in line with the implementation of accelerated approval mechanisms by regulatory agencies such as the Food and Drug Administration -breakthrough therapy designation and European Medicines Agency -accelerated assessment. However, the approval of those therapies is based on weak or limited evidence.[7,8] This is one of the reasons Health Technology Assessment bodies struggle to approve the reimbursement of these treatments associated with weak evidence compared with the gold standard. This was notably exemplified with immune checkpoint inhibitors, where 9 of the 10 accelerated approvals involved single-arm trials with the response rate as the main end point.[6,9] However, the effect size of new molecules is mostly small, based on poorly relevant outcomes such as tumor response,[10] though, in most settings, it has not been demonstrated that improving response yields an improvement in survival. The open nature of the design may introduce additional classification biases.[11,12] This may explain why no benefit in overall survival has been demonstrated so far for many oncology drugs.[5]

Besides the study of drugs for registrational purposes, it is often reported that randomized clinical trials (RCT) may not be feasible or practical for rare diseases and biomarker-specific selected populations of more common diseases owing to ethical considerations, the requirement of large sample sizes, and extended durations of time.[10,11] However, contrarily to situations of quasi-deterministic disease evolution, where nearly 0 or 100% of patients respond, relying on the observed "before-after" patient status to define a treatment effect is well known to be biased.[13]

To handle the variability in the disease course as well as the unobserved effects of being enrolled in a trial, the measure of treatment effect requires to be relative to a control group. Thus, to increase the level of evidence in these uncontrolled settings, the use of external controls has been promoted.[14] Such indirect comparisons are being growingly reported.[15-18] However, as recently reported,[19] they require careful implementation of innovative statistical methods accounting for between-group variation and selection biases, depending on the availability and nature of external data.[20] Although many authors warned against the misuse of each approach and methodological issues from the use of external controls,[21-24] none have detailed the whole process, including the underlying assumptions for leveraging those data.[24]

In this paper, we aim to provide some guidance for clinicians, investigators, manufacturers, and all stakeholders, highlighting the main issues of such external incorporations into single-arm trial data, and distinguishing a 3-step process (Figure 1). First, the specifications of key attributes or "estimands", in line with the

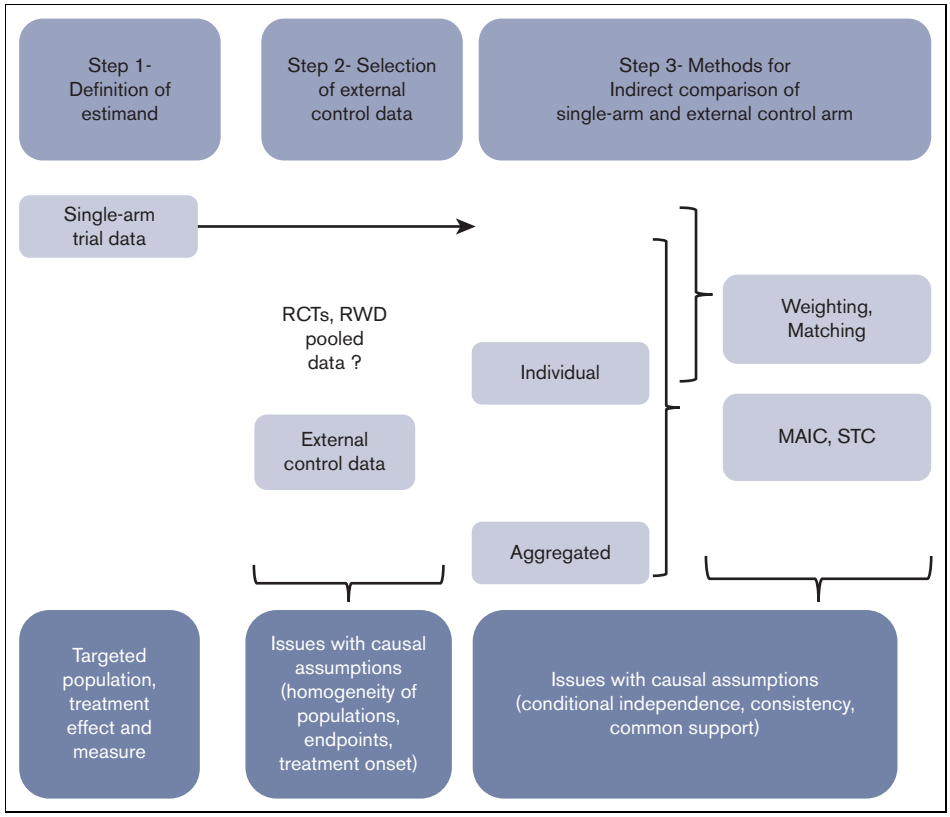

Figure 1. Schematic 3-step process to be applied when incorporating external control data into single-arm trial data to maximize the validity of indirect comparisons.

objectives, should be defined according to the principles of such "emulated" target trials. Second, the selection of the controls should consider the various sources of external controls to adequately mimic the lacking randomized experiment while avoiding substandard control arms. Specific statistical considerations may arise, according to the data type and characteristics. The last step consists of the indirect comparison itself, based on different methods according to the available data and the targeted treatment effect. A motivating example is used to illustrate this 3-step process.

Illustrating example

As an illustrative example, we used ciltacabtagene autoleucel (CARVYKTI; Janssen Biotech, Inc., Horsham, PA) approved by the Food and Drug Administration in February 2022 for the treatment of adult patients with relapsed or refractory multiple myeloma (RRMM) after 3 or more prior lines of therapy, including a proteasome inhibitor (PI), an immunomodulatory agent (IMiD), and an anti-CD38 monoclonal antibody. The pivotal trial was CARTITUDE-1 (NCT03548207), a multicenter, phase 1b/2 open-label, single-arm, clinical trial conducted in the United States between July 2018 and October 2019.[25] A total of 113 patients with RRMM, with at least 3 prior lines of therapy including a PI, an IMiD, and an anti-CD38 monoclonal antibody, and disease progression on or after the last regimen were enrolled. Among the 113 enrolled patients, 97 (85.8%) patients who received ciltacabtagene autoleucel (cilta-cel) were included in the analysis. The efficacy was established based on the overall response rate (ORR) as the main end point, estimated at 97% (95% confidence interval [CI], 91.2-99.4). However, RRMM, especially the triple-class-refractory disease, is an extremely active area of research, in which many drugs that may act as pertinent comparators have been proposed. Indeed, in that population, many drugs from distinct classes have been approved by the FDA, including monoclonal antibodies such as belantamab mafodotin,[26] isatuximab,or teclistamab,[27] small molecule inhibitors/modulators such as selinexor,[28] or melphalan flufenamide,[29] or other CAR T cells such as idecabtagene vicleucel (ide-cel)[25] (Figure 2). We will show how indirect comparisons can be performed and findings can be achieved on the relative efficacy of cilta-cel.

Step 1- definition of estimands

An estimand is a precise description of a treatment effect reflecting a clinical question that should inform study design and analysis under 5 attributes: target population, treatment, end point, intercurrent events, and population level summary of the treatment effect measured against some valid comparator. First described for RCTs,[30] its principles can be easily extended to observational studies.[31,32]

Rarely, the treated and control populations can be assumed similar, owing to similar eligibility criteria, time period, and the sites of enrolment.[33] To overcome this issue, down weighting the external control data allows to decrease the level of evidence from the external source to be addressed using either power prior models[34-36] or meta-analytic approaches.[22]

However, most of the time, populations differ in characteristics that may also affect the outcome, these are termed "confounders" (Box 1). Ignoring those differences will lead to misleading inferences owing to confounding bias.[37] Indeed, any differences in outcomes could no longer be attributed to differences in treatments but rather to confounders.

Thus, reaching a balance in confounder population is at the core of causal inference in observational studies. Regression models providing estimates of the treatment effect adjusted on prognostic factors have been long used for that purpose. However, they do not ensure a balance of prognostic variables across groups, notably, where their values widely differ across groups; in these areas of nonoverlap, estimates are extremely sensitive to model choices. Thus, rather than focusing on the outcome model (by introducing both treatment and confounders to predict the outcome), one may focus on the treatment model through the propensity score (PS), that is, the probability of being in the treatment group, conditional on the set of observed confounders.[38] Then, individuals are given individual "balancing" weights,[31] derived from their PS, to under- or overrepresent the characteristics of their treatment group compared with the other group (Figure 3). Under different assumptions of conditional independence, consistency, and common support (Box 2), valid estimators of the treatment effect can be directly derived from the weighted data. The main advantage of the propensity score is to separate the treatment model and the

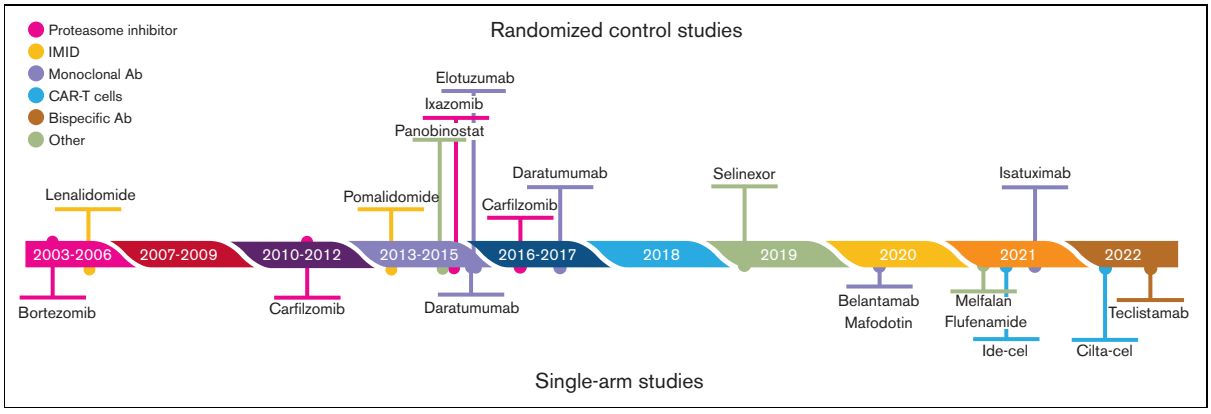

Figure 2. Timeline of the drugs approved by the FDA for the treatment of patients with RRMM.

Box 1. Glossary

- **Confounder**: Variable that affects both the treatment choice and the outcome. Ignoring those variables in the comparison of treatment groups achieves a "confounding-by-indication" bias.
- **External control** is defined as a group of patients external to the study that differs from an internal control group consisting of patients from the same population assigned to a different treatment.
- **Immortal time bias** refers to the difference in estimation achieved by differences in the selection time of the treatment groups that favor the treated who "survived" up to the actual administration of the treatment. This can be controlled for by carefully defining the time to selection in both groups and should be as similar as possible.
- **Propensity Score (PS):** The propensity score is the probability that a patient would receive the treatment of interest based on the pretherapeutic characteristics of the patient, the treating clinician, and the clinical environment. PS methods are used to reduce bias in estimating treatment effects and allow investigators to reduce the likelihood of confounding when analyzing nonrandomized, observational data. Under several assumptions, such methods allow causal treatment effects to be provided.
- **Real World Evidence (RWE)** is clinical evidence of the usage and potential benefits or risks of a medical product derived from an analysis of RWD.
- **Real World Data (RWD)** are data relating to a patient's health status and/or the delivery of health care routinely collected from a variety of sources.

outcome model. Modelling the treatment probability further forces one to think about the imbalances in covariates before estimating the treatment effect.

When comparing single-arm vs external control groups, these methods could be used. However, the target population should first be defined, as this definition impacts the definitions of weights and the targeted treatment effect (Table 1). Indeed, one may focus on the average treatment effect (ATE) in the population represented by the combined single-arm and external control groups that would be observed by switching every unit in the whole population from one treatment to the other, the average treatment effect in the treated (ATT), obtained by only switching the treated to the control group; or the average treatment effect in the control (ATC).

For instance, when evaluating the benefit of cilta-cel over some pertinent comparator in the patients with RRMM, the ATE, corresponding to switching every unit in the study population from the comparator to cilta-cel and reciprocally, may result in the effect of an infeasible intervention. In contrast, choosing the ATT targets the treated population, that is, those included in the single-arm trial and attempts to answer "what would have been the ORR of the patients treated with cilta-cel, had they all received the comparator instead?". This may be the estimand of interest in this setting, and it was mostly used in the published indirect comparisons of cilta-cel against standard treatment.[39-41] The ATC provides the alternate answer to "what should have been the ORR in the patients from the comparator group had they received cilta-cel instead?". Such an estimand was used to assess the benefit of cilta-cel against active comparators, though not reported as such.[42,43]

Step 2- selection of the external control data

Then, one may look for external, sometimes called "synthetic",[44] controls. In line with the objective, the closeness of the external population with the targeted population should be first required to avoid the risk of substantial biases. This could be evaluated using the acceptability criteria proposed by Pocock.[33] The selection of external controls should use predefined eligibility criteria for the inclusion of studies to ensure patient similarity, relevant end points, and pertinent comparators.

External controls could be directly selected from pertinent and efficacious active arms from previously completed RCTs[20] or reconstituted from real-world data (RWD).[45]

When external controls are selected from RCTs, it is likely that the potential comparator has been sponsored by another firm, so only aggregated data are available. Pooled data from previous RCTs could also be used as external controls, as exemplified by the FDA that approved a synthetic control generated from more than 22 000 previous studies to be used in a phase III glioblastoma cancer trial.[46]

When no available controls from previous trials are available, controls can be selected from RWD, including observational cohorts, registries, or electronic health records (EHR),[47] as well as claims and prescription data.[48] Although primary end points may be difficult to match in RWD and clinical trials, it is not the case in cancer where the date of death is usually reported in the EHR or any administrative registry. To control for the potential effects of time and center, an adequate selection of both should be considered first.[49] The closeness of populations is of particular concern in the observational setting in which the choice of treatment based on a patient's disease status achieves a "confounding-by-indication" bias. In many chronic diseases, there is also no obvious single timepoint for treatment decisions.[49] Thus, when the population differs in terms of the time of treatment decision-making, "immortal time bias" or "time-lag bias" could be additionally introduced.[49] Once sources of control data are found, their validity should be measured by assessing the risk of bias. As reported recently, based on publicly available FDA reviews of medical products, most reasons why RWD did not contribute to regulatory decision-making relied on a lack of a prespecified study design and analysis as well as data reliability and relevancy concerns.[50]

In the cilta-cel example, several indirect comparisons in patients with RRMM were secondarily published, as summarized in Table 2. They first used conventional treatment as the comparator of interest, with data obtained from long-term follow-up of previous clinical trials,[39] or multicenter retrospective studies,[41] and RWD.[40] However, the clinical relevance of such a "standard treatment" group may be questioned because of targeting a very heterogenous and frail population that may not be a candidate for CART-cell therapy. Moreover, the use of retrospective studies and RWD raises the issue of data quality (data do not undergo the same level of quality checks as in the trial), resulting in the selection, measurement, and attrition biases. Last, CAR T cells are administered after a variable period on potentially selected patients. This raises concerns about the comparison with those cohorts, with different start dates of follow-up.[51]

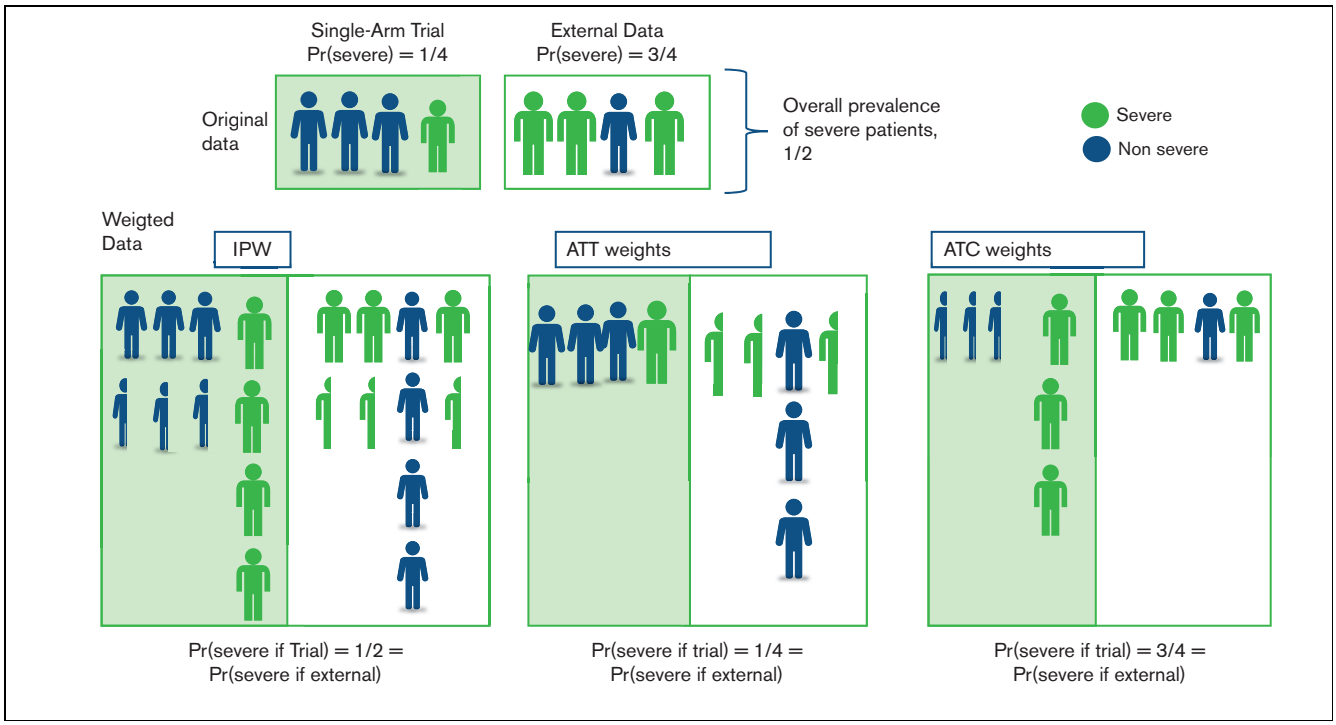

Figure 3. Schematic representation of how data are weighted according to an estimand. Suppose the original sample from the single-arm trial differs from the external controls in terms of patient severity, with 1 severe case over 4 in the trial compared with 3 over 4 in the external data. The objective is to modify the pooled data to obtain 2 groups where the proportion of severe cases is similar. Most methods are based on the PS, which is the probability of each patient being in the trial, conditional on their severity. In this setting, each severe case is given a PS of 1/4, whereas each nonsevere case is given a PS of 3/4. IPW consists of inversely weighting each individual in the original sample according to their probability of being in the original group, that is, for the treated, the individual contribution of each patient is divided by their PS (thus resulting in adding 1/3 of a fictive patient for each nonsevere patient and 3 fictive individuals for severe cases), while in the external group, this value is divided by 1 minus their PS (thus adding 1/3 of a fictive patient for each severe patient and 3 fictive individuals for each nonsevere case). This yields a weighted sample where the proportion of severe cases is similar in both groups (1/2) and differs from that in both original groups. ATT weights consist of using all individuals from the single-arm trial (weight of 1) and weighting each individual in the external sample by the odds of being in the trial. This results in odds of (1/4)/(3/4) = 1/3 in nonsevere cases and (3/4)/(1/4) = 3 in severe cases, reaching a ¼ prevalence of severe cases in the pooled weighted data set, that is, observed in the originally treated patients from the trial. ATC weights are conversely computed, with a weight of 1 for each patient from the external sample, whereas patients from the single-arm trial are given a weight of (3/4)/(1/4) (severe cases) or (1/4)/(3/4) (nonsevere cases). The resulting prevalence of severe cases is now that of the original external control group, that is, 3/4.

More recently, 2 indirect comparisons focused on more pertinent active comparators, recently approved by the FDA at the time ciltacel was proposed (Figure 2), namely belantamab mafodotin and melphalan flufenamide, each assessed from a single-arm trial or selinexor, using RCT data[42] and ide-cel, another CAR T-cell therapy.[43] Given that the data of these control groups were prospectively recorded in clinical trials, it likely improved the control of other sources of bias compared with RWD.

Step 3- methods for indirect comparisons of single-arm and external control arms

Last, an indirect comparison of the single-arm trial and the external control arm should be performed using appropriate statistical methods, and underlying assumptions should be checked. Such methods mostly depend on whether the control data have been measured at the individual level or aggregated level.

Individual-level external control data. The availability of individual-level data for both groups allows the PS to be estimated to balance the confounders of the treated (trial) group and the (external) control group using weighting or matching (Table 1). When the external individual-level data are obtained from observational data, additional weights may be used to incorporate the decreased level of evidence of the controls.[52]

The most common approach to estimate the inverse probability of treatment weights (IPWs) is to estimate the PS through logistic regression, ideally including all the true confounders, then directly defining weights for both the treated and control population. Such weights target the ATE of the underlying population defined by the combination of the treated and untreated groups (Figure 3). Unfortunately, the "convenience" sample defined by the pool of the trial sample and the external controls, does not always represent a population of scientific interest, in contrast to surveys from which such methods have been derived. To focus on the treated population and estimate the ATT, only control patients are given a weight depending on the odds of being treated whereas treated patients are given a unit weight.

For both types of weights, the challenge associated with extreme propensities has been identified as a primary downside of weighting, with no clear definition of the resulting ambiguous target

Box 2. Causal assumptions

- **Consistency** relates the observed outcome to potential outcomes that would be observed under each treatment compared and forms the underlying statistical framework for the approach. Consistency is generally assumed to be part of the causal model itself but also implies that the treatments to be compared are well defined and that there are no "hidden" versions of those, which may be arguable for external controls who may receive different treatments. In this case, consistency should be considered more at a distributional level, ie, the distribution of different versions of the "treatment" in the population.
- **No interference** is defined as the effect of a treatment on the outcome of an individual and is not affected by the other individuals being treated. It can be generally accepted for external controls, in particular because they are often selected from existing cohorts, registries or electronic health records, and would be unaffected by a limited-sample size trial being conducted, possibly in different locations or time periods.
- **No unmeasured confounding** indicates that the covariates measured for the trial participants and external controls comprise all those that are likely to affect the outcome and differ between groups. This assumption is more challenging, since it requires in practice that all relevant prognostic factors are recorded for both participants in the trial and external controls. Additionally, factors that may affect outcomes such as center-specific characteristics, socioeconomic variables, environmental factors, standard of care, or health systems may not be available for either the trial participants or the external controls.
- **Positivity or common support** indicates that all individuals have a nonnull probability of receiving either treatment. External controls have virtually no chance of receiving the experimental treatment, but one should determine whether controls could have received the experimental treatment given their individual characteristics had they been followed-up in an institution participating in the trial. This is not limited to a trial's eligibility criteria, but one should also examine other potential confounders. For instance, if the aforementioned factors are recorded but the standard of care or center expertise differed between the controls and treated patients, this may violate positivity. Moreover, if the standard of care or center expertise differed between the controls and treated patients, this may also violate positivity.

population.[53] Methods that address nonoverlap, such as trimming or downweighting data in regions of poor data support, excluding or censoring weights at some extreme percentiles, change the estimand so that inference cannot target the population of interest. Thus, balancing weights has been proposed as a simple way to define, based on specific tilting functions, individual weights, and the resulting target population,[54] as it integrates most approaches, including PS matching.[38] Recently, "overlap weights" were proposed to focus on the population for which observed confounders have been adequately balanced (Table 1). Finally, it should be noted that all those weighted samples differ in terms of the target population, as illustrated in the observed patient characteristics, either close to those of the pooled groups, of the treated, the controls, or the overlapping sample (Figure 2). In all cases, the exchangeability of the restructured groups should be measured, using simple measures such as standardized mean difference (SMD) which should be below 10% (as a rule-of-thumb) or any other distances.[55]

In the indirect comparisons of cilta-cel vs observational cohorts or RWD,[39-41] individual patient data were available to estimate PS from multivariable logistic models, then using either matching[41] or weighting,[39,40] to estimate the ATT. However, none of these comparisons fulfilled all those "quality" requirements (Table 2). Notably, confounders included in the propensity score were not fully reported or did not include all expert knowledge of true confounders. All analyses failed to reach a clearcut exchangeability of groups, with reported persistent imbalances (either not detailed or with SMDs above 15% for several confounders). This resulted in a risk of bias for the estimated cilta-cel effect.

Aggregated external control data. When control data are derived from clinical trials not sponsored by the manufacturer's product of the single-arm trial, it is not uncommon for only published aggregate data to be available. In this setting, only summary measures of both the confounders and outcomes are at most available. Notably, for time-to-event data, some types of individual-level data can be extracted from published Kaplan–Meier curves using digitization,[56] but individual-level data on confounders would still not be obtained. To address such aggregated control data, population-adjusted indirect comparisons have been proposed, the 2 most popular methods being matching-adjusted indirect comparison (MAIC)[57] and simulated treatment comparison (STC).[58]

Table 1. Targeted population, weights, and estimands

Method for controlling confounders	Weights for treated, untreated	Target population	Estimand
Inverse weighting	$\frac{1}{e(x)}, \frac{1}{(1-e(x))}$	Combined from the treated and untreated	ATE
	$1, \frac{e(x)}{(1-e(x))}$	Treated population	ATT
	$\frac{1-e(x)}{e(x)}, 1$	Control population	ATC
	1-e (*x*), e (*x*)	Overlapping population	ATO
	$\frac{1\,(a<e(x)<1-a)}{e(x)}, \frac{1\,(a<e(x)<1-a)}{(1-e(x))}$	Trimming population	Not specified
Matching	$\frac{Min\,(e(x),1-e(x))}{e(x)}, \frac{Min\,(e(x),1-e(x))}{(1-e(x))}$	Matching population	ATT
Matching-adjusted indirect comparison	$\frac{1-e(x)}{e(x)}, 1$	Control population	ATC

$e(x) = PS = Pr(T = 1|V)$ is the propensity score, where T = 1 for the single-arm treatment group, T = 0 for the external control group, and V is the set of observed confounders in both groups. ATO, average treatment effect in the overlap population.

Table 2. Illustration of the 3-step assessment on the main indirect comparisons of celta-cel against comparators. Bold cells indicate the main issues of the performed comparisons. The following 5 main confounders were considered and ranked as major confounders by experts: Refractory status, cytogenetic profile, R-ISS stage, plasmacytomas, and time to progression on last prior line

	Step 1- Estimand		Step 2- External source of data		Step 3- Methods of comparison			
Indirect comparison	**Main objective**	**Comparator**	**Type**	**Source**	**Propensity score**	**Method**	**Balance diagnostics, common support**	**Estimation of effect**
Merz 2021	ATT	**Standard treatment heterogeneity**	IPD	**Retrospective German RWD database risk of bias**	**9 confounders* cytogenetic and plasmacytomas missing**	IPW	**Undetailed "remaining imbalances" (SMD >0.20)**	Weighted analyses with robust variance
Weisel 2022	ATT	**Physician choice heterogeneity**	IPD	Follow-up of trial data (POLLUX, CASTOR, EQUULEUS)	8 confounders†	IPW	**Mean SMD reduced from 0.33 to 0.16**	Weighted analyses with robust variance
Costa 2022	ATT	**Conventional treatment heterogeneity**	IPD	**Retrospective study Risk of bias**	**16 confounders‡ Plasmacytomas missing**	Matching 1:1, no replacement, caliper 0.05	**SMD between 0.10 and 0.20 (ASCT, refractory to carfilzomib, penta-drug refractory)**	Stratified/weighted analyses
Weisel 2022	ATC not explicitly reported	Belantamab mafodotin	Aggregate	One-arm (2.5 mg/kg dose) of the 2-arm trial data (DREAMM-2) ECOG 0-2	**4 confounders § Time to progression on the last regimen missing**	Unanchored MAIC	**ESS = 39 (60% reduction) no report of weight distribution**	Weighted analyses
	ATC not explicitly reported	Selinexor-DXM	Aggregate	**RCT data (mITT of STORM-2) Penta-exposed ECOG 0-2**	**4 confounders § Time to progression on the last regimen missing**	Unanchored MAIC	**ESS = 73 (25% reduction) No report of weight distribution**	
	ATC not explicitly reported	Melphalan-flufenamide-DXM	Aggregate	A subset of Single-arm trial data (HORIZON) received ≥2 prior LOTs ECOG 0-2	**3 confounders refractory status missing**	Unanchored MAIC	**ESS = 85 (12% reduction) no report of weight distribution**	
Martin 2022	ATC not explicitly reported	Ide-cel	Aggregate	Single-arm trial data (KarMMa)	**4 confounders § Time to progression on the last regimen missing**	Unanchored MAIC	**Skewed distribution of weights ESS: 46%-57% reduction**	Weighted analyses failure times measured from cells infusion

ASCT, allogeneic stem cell transplantation; DXM, dexamethasone; ECOG, Eastern Cooperative Oncology Group; ISS, international staging system; LOT, line of treatment; MM, multiple myeloma.

*Age, sex, refractory status, R-ISS stage, time to progression on last prior line, number of prior LOTs, average duration of prior lines, years since diagnosis, ECOG status.

†Age, refractory status, ISS stage, cytogenetic profile, time to progression on last regimen, plasmacytoma, number of prior LOTs, years since MM diagnosis.

‡Age, sex, race/ethnicity (white vs other), ISS stage 3 (vs 1, 2, or unknown), time from diagnosis to index date, number of prior LOT, prior autologous stem cell transplant, presence of high- risk cytogenetic abnormalities in any prior sample [t(4;14), t(14;16), del(17p)], refractoriness to bortezomib or ixazomib, refractoriness to carfilzomib, refractoriness to lenalidomide, refractoriness to pomalidomide, refractoriness to anti-CD38 monoclonal antibody, triple-class refractoriness, penta-drug exposure (to bortezomib or ixazomib plus carfilzomib plus lenalidomide plus pomalidomide plus anti-CD38 monoclonal antibody), and penta-drug refractoriness.

§Refractory status, cytogenetic profile, R-ISS stage, plasmocytomas.

MAIC is a reweighting method similar to IPW that targets the control population. Its principle is to reweigh the individual-level data such that the mean characteristics of the treated population are balanced with those of the controls, with weights estimated from the PS of being treated. The resulting target population is that of the external data set, thus, allowing the estimation of the ATC (Table 1). Notably, the PS cannot be estimated as usual given the lack of individual patient data for the controls, but alternate methods can be used.[59] It is then important to evaluate the distribution of weights, which should be centered around 1. If there are too many participants being allocated near zero or very high weights, the comparability of groups is questioned, with increased uncertainty of the results. The effective sample size (ESS) can also be computed as a measure of information provided by the weighted data set. A small ESS, relative to the original sample size, is an indication that the weights are highly variable and that the estimate may be unstable. In STC, individual-level data are used to model the relationship between predictors and outcome of the single-arm trial, and then the model is used to estimate the outcomes in external controls.

Both MAIC and STC rely on the strong assumption of a constant absolute treatment effect at any level of the effect modifiers and prognostic variables and that all effect modifiers and prognostic variables have been observed, otherwise, the estimates are biased.[58] Thus, providing information on the likely biases resulting from unobserved prognostic factors and effect modifiers distributed differently across the trials is mandatory. Such indirect comparisons require additional recommendations. First, evidence that absolute outcomes can be predicted with sufficient accuracy in relation to the relative treatment effect should be provided. Moreover, the choice of the outcome scale is critical and should be justified because the effect modifier status is scale specific. An important limitation is that MAIC or STC is only able to provide estimates in the target population represented by the external comparator population and not that of the single-arm trial of interest. For any other target population, a supplementary assumption, the shared effect modifier, is needed.[58]

Two unanchored MAICs were published to compare the effect of cilta-cel with active pertinent comparators from single-arm clinical trials.[42,43] Only the 97 patients infused by cilta-cel were selected. Except when compared with other CAR T cells, a potential selection bias of the treatment group can be suspected, given the 16 patients who could not be reinfused owing to disease progression (n = 2), death (n = 9), or patient withdrawal (n = 5), were excluded.[25] None of the MAICs included the 5 "true" confounders selected by the experts, so the underlying assumption of no unmeasured confounders is possibly violated. Moreover, the distribution of the weights and the weighted baseline characteristics were not fully reported, whereas the reduction in the effective sample size of the cilta-cel–treated population was relatively high, from 46% to 60%, resulting in the ESS being down to 39 (Table 2). It indicates that there may be poor overlap between the study populations, violating the underlying assumption of common support (illustrating the potential selection bias described above), again resulting in a high risk of bias.

Discussion and perspectives

The provision of rapid answers when evaluating a new treatment outside the standard phase I-III strategy is becoming increasingly important.[60] Currently, the use of single-arm clinical trials as the sole source of evidence provided by pharmaceutical firms to obtain, at least temporaririly, drug approvals, is accepted by regulatory agencies for populations or individuals with certain indications. This is also widely used by academics when evaluating interventions in rare cancer subgroups or combination therapies.[61] This may appear contradictory to the statistical literature reporting its many sources of bias since the early 1980s.[62]

There could be some ways of improving the value of data and thus increasing the utility of single-arm trials.[63] Thus, to decrease the uncertainty of such uncontrolled trials, comparisons using external controls have been growingly reported in oncohematology, for instance, in acute lymphoblastic leukemia,[15] large B-cell lymphoma,[64] anaplastic lymphoma,[17] follicular lymphoma,[18] metastatic nonsmall-cell lung cancer,[65] endometrial cancer,[16] and glioblastoma.[66] Such indirect comparisons require a complex implementation to be valid, as recently reported.[67] In the specific setting of single-arm trials, we aimed to report how to enhance the evidence from such trials by incorporating and leveraging external data as a "synthetic" control arm to mimic the lacking "head-to-head" comparison. Thus, we provided some guidance for incorporating such external controls by defining a 3-step process to stop the sequence whenever a target or underlying assumption could not be satisfied. First, the target population, pertinent comparator, and measure of the treatment effect should be clearly delineated. Second, the selection of the target controls should be carefully and adequately performed with respect to the population, end point and treatment decision. Indeed, using controls from previous RCT or other trials is likely different than defining controls from RWD, from which selection of pertinent patients raises issues, notably concerning the immortal time bias and reverse causation issues. This raises the issue of sharing individual patient data so that the secondary use of available health data should be promoted, which begins by encouraging secure and facilitated access to those data by researchers worldwide, as proposed by the American Society of Hematology's Research Collaborative.[45] Last, the method of analysis should be justified based on the type of available data and on the underlying target population and the therapeutic question of interest (eg, to treat all patients or not?). The use of external controls finally entails merging different sources of data, which may complicate the verification of causal assumptions and not adequately control for confounding factors, which is a necessary but not sufficient framework for valid estimation of treatment effect. Indeed, although treatment groups achieved by random allocation are exchangeable in terms of all (observed or not) prognostic covariates and treatment-effect modifiers, PS methods could only rely on the observed confounders, their main limitation, even if the analysis is well conducted. Nevertheless, well-conducted indirect comparisons may generate hypotheses for new trials regarding pertinent comparators and thus may appear as an option while or before an RCT is conducted.

In all cases, especially given the risk that analyses would be data-driven and adapted ad hoc, the statistical analysis plan for such incorporation should be publicly issued before the analysis, and only external controls recruited after that publication should be used in the comparisons in a similar approach as in registered reports.[68] The principled framework of emulating a target trial combining the principles of clinical trials and causal methods to control for confounding appears particularly adequate in this situation.[69,70]

We mostly considered methods derived from propensity scores, although other approaches could also be considered, such as *g*-computation,[71] or "double-robust" or "augmented IPW" estimators.[72] To the best of our knowledge, these approaches have not been used for regulatory approval with external controls but remain promising alternatives. Other issues, such as time-dependent biases, may exist as well.[49] How to adequately control for time-dependent biases with external controls is still an open issue.

In summary, when reporting results from a single-arm trial, the provision of some external comparison to controls is often reported, with the aim to obtain marketing authorization. In all cases, it should be adequately done and reported to provide evidence. It should be kept in mind that such indirect comparisons aim to mimic the lacking randomized clinical trials. Only respect for the proposed 3-step guidance may provide a correct level of evidence, although it cannot be guaranteed that it will reach the level of a well-conducted RCT.

Authorship

Contribution: S.C. was responsible for supervision and project administration and visualization; and all authors worked on the conceptualization, data curation, methodology, formal analysis, resource writing of the original draft, review, and editing of the manuscript.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

ORCID profiles: J.L., 0000-0001-7086-9295; E.L., 0000-0003-0965-6615; R.P., 0000-0002-5277-4679; R.T., 0000-0002-5235-3962; S.Z., 0000-0002-8429-2340; S.C., 0000-0001-6449-4730.

Correspondence: Sylvie Chevret, Biostatistics and Medical Information Service (SBIM)-Saint Louis Hospital, 1 Ave Claude Vellefaux 75010 Paris, France; email: sylvie.chevret@u-paris.fr.

References

1. Pleasance E, Bohm A, Williamson LM, et al. Whole-genome and transcriptome analysis enhances precision cancer treatment options. *Ann Oncol*. 2022;33(9):939-949.
2. Le Tourneau C, Diéras V, Tresca P, Cacheux W, Paoletti X. Current challenges for the early clinical development of anticancer drugs in the era of molecularly targeted agents. *Target Oncol*. 2010;5(1):65-72.
3. Kummar S, Gutierrez M, Doroshow JH, Murgo AJ. Drug development in oncology: classical cytotoxics and molecularly targeted agents. *Br J Clin Pharmacol*. 2006;62(1):15-26.
4. Zelner J, Riou J, Etzioni R, Gelman A. Accounting for uncertainty during a pandemic. *Patterns (NY)*. 2021;2(8):100310.
5. Kim C, Prasad V. Cancer drugs approved on the basis of a surrogate end point and subsequent overall survival: an analysis of 5 years of us food and drug administration approvals. *JAMA Intern Med*. 2015;175(12):1992-1994.
6. Beaver JA, Pazdur R. "Dangling" accelerated approvals in oncology. *N Engl J Med*. 2021;384(18):e68.
7. Naci H, Davis C, Savović J, et al. Design characteristics, risk of bias, and reporting of randomised controlled trials supporting approvals of cancer drugs by European Medicines Agency, 2014-16: cross sectional analysis. *BMJ*. Published online September 18, 2019:l5221. https://doi.org/10.1136/bmj.l5221
8. Hatswell AJ, Freemantle N, Baio G. The effects of model misspecification in unanchored matching-adjusted indirect comparison: results of a simulation study. *Value Health*. 2020;23(6):751-759.
9. Beaver JA, Pazdur R. The wild west of checkpoint inhibitor development. *N Engl J Med*. 2022;386(14):1297-1301.
10. Muchtar E, Gertz MA, LaPlant BR, et al. Phase 2 trial of ixazomib, cyclophosphamide, and dexamethasone for previously untreated light chain amyloidosis. *Blood Adv*. 2022;6(18):5429-5435.
11. Ribeiro TB, Colunga-Lozano LE, Araujo APV, Bennett CL, Hozo I, Djulbegovic B. Single-arm clinical trials that supported FDA accelerated approvals have modest effect sizes and at high risk of bias. *J Clin Epidemiol*. 2022;148:193-195.
12. Saccà L. The uncontrolled clinical trial: scientific, ethical, and practical reasons for being. *Intern Emerg Med*. 2010;5(3):201-204.
13. Sedgwick P. Before and after study designs. *BMJ*. 2014;349:g5074.
14. Davi R, Mahendraratnam N, Chatterjee A, Dawson CJ, Sherman R. Informing single-arm clinical trials with external controls. *Nat Rev Drug Discov*. 2020;19(12):821-822.
15. Ribera JM, García-Calduch O, Ribera J, et al. Ponatinib, chemotherapy, and transplant in adults with Philadelphia chromosome–positive acute lymphoblastic leukemia. *Blood Adv*. 2022;6(18):5395-5402.
16. Mathews C, Lorusso D, Coleman RL, Boklage S, Garside J. An indirect comparison of the efficacy and safety of dostarlimab and doxorubicin for the treatment of advanced and recurrent endometrial cancer. *Oncologist*. 2022;27(12):1058-1066.
17. Smith S, Albuquerque de Almeida F, Inês M, Iadeluca L, Cooper M. Matching-adjusted indirect comparisons of lorlatinib versus chemotherapy for patients with second-line or later anaplastic lymphoma kinase-positive non-small cell lung cancer. *Value Health*. 2022;16. S1098-3015(22)02098-8.
18. Salles GA, Schuster SJ, Dreyling M, et al. Efficacy comparison of tisagenlecleucel vs usual care in patients with relapsed or refractory follicular lymphoma. *Blood Adv*. 2022;6(22):5835-5843.
19. Collignon O, Schritz A, Spezia R, Senn SJ. Implementing historical controls in oncology trials. *Oncologist*. 2021;26(5):e859-e862.

20. Goring S, Taylor A, Müller K, et al. Characteristics of non-randomised studies using comparisons with external controls submitted for regulatory approval in the USA and Europe: a systematic review. *BMJ Open*. 2019;9(2):e024895.
21. Burcu M, Dreyer NA, Franklin JM, et al. Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms. *Pharmacoepidemiol Drug Saf*. 2020;29(10):1228-1235.
22. Schmidli H, Häring DA, Thomas M, Cassidy A, Weber S, Bretz F. Beyond randomized clinical trials: use of external controls. *Clin Pharmacol Ther*. 2020; 107(4):806-816.
23. Wang C, Berlin JA, Gertz B, et al. Uncontrolled extensions of clinical trials and the use of external controls–scoping opportunities and methods. *Clin Pharmacol Ther*. 2022;111(1):187-199.
24. Yap TA, Jacobs I, Baumfeld Andre E, Lee LJ, Beaupre D, Azoulay L. Application of real-world data to external control groups in oncology clinical trial drug development. *Front Oncol*. 2022;11:695936.
25. Berdeja JG, Madduri D, Usmani SZ, et al. Ciltacabtagene autoleucel, a B-cell maturation antigen-directed chimeric antigen receptor T-cell therapy in patients with relapsed or refractory multiple myeloma (CARTITUDE-1): a phase 1b/2 open-label study. *Lancet Lond Engl*. 2021;398(10297):314-324.
26. Lonial S, Lee HC, Badros A, et al. Belantamab mafodotin for relapsed or refractory multiple myeloma (DREAMM-2): a two-arm, randomised, open-label, phase 2 study. *Lancet Oncol*. 2020;21(2):207-221.
27. Moreau P, Garfall AL, van de Donk NWCJ, et al. Teclistamab in relapsed or refractory multiple myeloma. *N Engl J Med*. 2022;387(6):495-505.
28. Chari A, Vogl DT, Gavriatopoulou M, et al. Oral selinexor-dexamethasone for triple-class refractory multiple myeloma. *N Engl J Med*. 2019;381(8): 727-738.
29. Olivier T, Prasad V. The approval and withdrawal of melphalan flufenamide (melflufen): Implications for the state of the FDA. *Transl Oncol*. 2022;18: 101374.
30. Ratitch B, Goel N, Mallinckrodt C, et al. Defining efficacy estimands in clinical trials: examples illustrating ich e9(r1) guidelines. *Ther Innov Regul Sci*. 2020;54(2):370-384.
31. Li H, Wang C, Chen W, et al. Estimands in observational studies: Some considerations beyond ICH E9 (R1). *Pharm Stat*. 2022;21(5):835-844.
32. Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I; "on behalf of" the topic group Causal Inference (TG7) of the STRATOS initiative. Formulating causal questions and principled statistical answers. *Stat Med*. 2020;39(30):4922-4948.
33. Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis*. 1976;29(3):175-188.
34. Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical Commensurate and Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials. *Biometrics*. 2011;67(3):1047-1056.
35. Brard C, Hampson LV, Gaspar N, Le Deley MC, Le Teuff G. Incorporating individual historical controls and aggregate treatment effect estimates into a Bayesian survival trial: a simulation study. *BMC Med Res Methodol*. 2019;19(1):85.
36. Roychoudhury S, Neuenschwander B. Bayesian leveraging of historical control data for a clinical trial with time-to-event endpoint. *Stat Med*. 2020;39(7): 984-995.
37. Dron L, Golchi S, Hsu G, Thorlund K. Minimizing control group allocation in randomized trials using dynamic borrowing of external control data – An application to second line therapy for non-small cell lung cancer. *Contemp Clin Trials Commun*. 2019;16:100446.
38. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
39. Weisel K, Martin T, Krishnan A, et al. Comparative efficacy of ciltacabtagene autoleucel in cartitude-1 vs physician's choice of therapy in the long-term follow-up of POLLUX, CASTOR, and EQUULEUS clinical trials for the treatment of patients with relapsed or refractory multiple myeloma. *Clin Drug Investig*. 2022;42(1):29-41.
40. Merz M, Goldschmidt H, Hari P, et al. Adjusted comparison of outcomes between patients from CARTITUDE-1 versus multiple myeloma patients with prior exposure to PI, Imid and anti-CD-38 from a german registry. *Cancers*. 2021;13(23):5996.
41. Costa LJ, Lin Y, Cornell RF, et al. Comparison of cilta-cel, an anti-BCMA CAR-T cell therapy, versus conventional treatment in patients with relapsed/ refractory multiple myeloma. *Clin Lymphoma Myeloma Leuk*. 2022;22(5):326-335.
42. Weisel K, Krishnan A, Schecter JM, et al. Matching-adjusted indirect treatment comparison to assess the comparative efficacy of ciltacabtagene autoleucel in CARTITUDE-1 versus belantamab mafodotin in DREAMM-2, selinexor-dexamethasone in STORM part 2, and melphalan flufenamide-dexamethasone in HORIZON for the treatment of patients with triple-class exposed relapsed or refractory multiple myeloma. *Clin Lymphoma Myeloma Leuk*. 2022;22(9):690-701.
43. Martin T, Usmani SZ, Schecter JM, et al. Updated results from a matching-adjusted indirect comparison of efficacy outcomes for ciltacabtagene autoleucel in CARTITUDE-1 versus idecabtagene vicleucel in KarMMa for the treatment of patients with relapsed or refractory multiple myeloma. *Curr Med Res Opin*. 2023;39(1):81-89.
44. Seeger JD, Davis KJ, Iannacone MR, et al. Methods for external control groups for single arm trials or long-term uncontrolled extensions to randomized clinical trials. *Pharmacoepidemiol Drug Saf*. 2020;29(11):1382-1392.
45. Wood WA, Marks P, Plovnick RM, et al. ASH Research Collaborative: a real-world data infrastructure to support real-world evidence development and learning healthcare systems in hematology. *Blood Adv*. 2021;5(23):5429-5438.
46. Spinner J. Medidata synthetic control arm lands FDA approval for cancer trial. 19 November 2020. Accessed 4 January 2023. https://www.outsourcing-pharma.com/Article/2020/11/19/Synthetic-control-arm-lands-FDA-approval-for-cancer-trial

47. Tan K, Bryan J, Segal B, et al. Emulating control arms for cancer clinical trials using external cohorts created from electronic health record-derived real-world data. *Clin Pharmacol Ther*. 2022;111(1):168-178.

48. Cave A, Kurz X, Arlett P. Real-world data for regulatory decision making: challenges and possible solutions for europe. *Clin Pharmacol Ther*. 2019;106(1):36-39.

49. Suissa S. Single-arm trials with historical controls: study designs to avoid time-related biases. *Epidemiology*. 2021;32(1):94-100.

50. Mahendraratnam N, Mercon K, Gill M, Benzing L, McClellan MB. Understanding use of real-world data and real-world evidence to support regulatory decisions on medical product effectiveness. *Clin Pharmacol Ther*. 2022;111(1):150-154.

51. Lin X, Lee S, Sharma P, George B, Scott J. Summary of US Food and Drug Administration chimeric antigen receptor T-cell biologics license application approvals from a statistical perspective. *J Clin Oncol*. 2022;40(30):3501-3509.

52. Bonander C, Humphreys D, Degli Esposti M. Synthetic control methods for the evaluation of single-unit interventions in epidemiology: a tutorial. *Am J Epidemiol*. 2021;190(12):2700-2711.

53. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187-199.

54. Li F, Thomas LE. Addressing extreme propensity scores via the overlap weights. *Am J Epidemiol*. 2022;191(6):1140-1151.

55. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28(25):3083-3107.

56. Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12(1):9.

57. Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *Pharmacoeconomics*. 2010;28(10):935-945.

58. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Making*. 2018;38(2):200-211.

59. Phillippo DM, Dias S, Elsada A, Ades AE, Welton NJ. Population adjustment methods for indirect comparisons: a review of national institute for health and care excellence technology appraisals. *Int J Technol Assess Health Care*. 2019;35(03):221-228.

60. Johnson JR, Ning YM, Farrell A, Justice R, Keegan P, Pazdur R. Accelerated approval of oncology products: the food and drug administration experience. *J Natl Cancer Inst*. 2011;103(8):636-644.

61. Foster JC, Freidlin B, Kunos CA, Korn EL. Single-arm phase II trials of combination therapies: a review of the CTEP experience 2008–2017. *JNCI J Natl Cancer Inst*. 2020;112(2):128-135.

62. Spodick DH. The randomized controlled clinical trial. *Am J Med*. 1982;73(3):420-425.

63. Glassman RH, Kim G, Kahn MJ. When are results of single-arm studies dramatic? *Nat Rev Clin Oncol*. 2020;17(11):651-652.

64. Banerjee R, Midha S, Kelkar AH, Goodman A, Prasad V, Mohyuddin GR. Synthetic control arms in studies of multiple myeloma and diffuse large B-cell lymphoma. *Br J Haematol*. 2022;196(5):1274-1277.

65. Menefee ME, Gong Y, Mishra-Kalyani PS, et al. Project Switch: Docetaxel as a potential synthetic control in metastatic non-small cell lung cancer (mNSCLC) trials. *J Clin Oncol*. 2019;37(15_suppl):9105.

66. Sampson JH, Achrol A, Aghi MK, et al. MDNA55 survival in recurrent glioblastoma (rGBM) patients expressing the interleukin-4 receptor (IL4R) as compared to a matched synthetic control. *J Clin Oncol*. 2020;38(15_suppl):2513.

67. Xu R, Chen G, Connor M, Murphy J. Novel use of patient-specific covariates from oncology studies in the era of biomedical data science: a review of latest methodologies. *J Clin Oncol*. Published online 8 March 2022:JCO.21.01957. https://doi.org/10.1200/JCO.21.01957

68. Naudet F, Siebert M, Boussageon R, Cristea IA, Turner EH. An open science pathway for drug marketing authorization-Registered drug approval. *PLoS Med*. 2021;18(8):e1003726.

69. Hernán MA, Robins JM. *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC; 2020.

70. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol*. 2016;79:70-75.

71. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011;173(7):731-738.

72. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4):962-973.