Combining the Target Trial and Estimand Frameworks to Define the Causal Estimand: An Application Using Real-World Data to Contextualize a Single-Arm Trial

Lisa V. Hampson\textsuperscript{a}, Jufen Chu\textsuperscript{b}, Aiesha Zia\textsuperscript{a}, Jie Zhang\textsuperscript{b}, Wei-Chun Hsu\textsuperscript{c}, Craig S. Parzynski\textsuperscript{c}, Yanni Hao\textsuperscript{b}, and Evgeny Degtyarev\textsuperscript{a}

\textsuperscript{a}Novartis Pharma AG, Basel, Switzerland; \textsuperscript{b}Novartis Pharmaceuticals Corporation, East Hanover, NJ; \textsuperscript{c}Genesis Research, Hoboken, NJ

ABSTRACT
Single-arm trials (SATs) may be used to support regulatory submissions in settings where there is a high unmet medical need and highly promising early efficacy data undermine the equipoise needed for randomization. In this context, patient-level real-world data (RWD) may be used to create an external control arm (ECA) to contextualize the SAT results. However, naive comparisons of the SAT with its ECA will yield biased estimates of causal effects if groups are imbalanced with regards to (un)measured prognostic factors. Several methods are available to adjust for measured confounding, but the interpretation of such analyses is challenging unless the causal question of interest is clearly defined, and the estimator is aligned with the estimand. Additional complications arise when patients in the ECA meet the inclusion/exclusion criteria for the SAT at multiple timepoints. In this article, we use a case-study of a pivotal SAT to illustrate how a combination of the target trial and the ICH E9(R1) estimand frameworks can be used to define the target estimand and avoid common methodological pitfalls. We also propose an approach to address the challenge of how to define an appropriate time zero for external controls who meet the SAT inclusion/exclusion criteria at several timepoints.

1. Introduction
Randomized controlled trials (RCTs) are considered to be the gold standard for evaluating the efficacy and safety of new treatments and for supporting informed decisions by regulators, payers, physicians, and patients. Randomization ensures that in large samples, the two groups of patients will be balanced with respect to measured and unmeasured prognostic factors and, hence, with respect to their risks of any type of health outcome. Consequently, a comparison between treatment groups provides an unbiased estimate of the causal treatment effect (Collins et al. 2020).

Single-arm trials (SATs) are often performed in oncology to support health-authority approval. A retrospective study of European Marketing Authorizations between 2010 and 2019, identified 22 initial marketing authorizations based on SATs (Tenhunen et al. 2020). A review of U.S. Food and Drug Administration (FDA) oncology and hematology approvals between 2008 and 2016 identified 25% of approvals as being primarily supported by SATs (Zhou et al. 2019).

Real-world data (RWD) are data relating to patients’ health status and/or the delivery of health care which are routinely collected when patients interact with the healthcare system. Therefore, by definition RWD can be obtained from a variety of sources such as electronic health records (EHRs), insurer and/or provider administrative claims data, disease or patient registries, or chart reviews. Real-world evidence (RWE) describes any analysis using RWD (FDA 2018). There has been a long history of using RWD for post-approval safety and effectiveness studies of medical products. More recently, there is growing interest in the use of RWE to support pre-authorization regulatory decisions. Much momentum comes from regulators’ interest in leveraging RWE for decision-making as evidenced by the 21st Century Cure Act in the US, and in Europe the HMA—EMA Joint Big Data Task Force and the European Medicine Agency (EMA) regulatory science 2025 strategic reflection. Existing use cases include leveraging RWD for regulatory submission of supplemental indications and to support full registration after an initial accelerated or conditional approval of a medical product.

The use of patient-level RWD to create an external control arm (ECA) represents an important application as comparisons between a SAT and ECA can be useful to contextualize SAT results. In recent years, the availability of RWD (e.g., EHRs with linked clinico-genomics data) for highly targeted populations with molecular subsets have also facilitated the creation of ECAs.

Naive comparisons between the treated cohort of a SAT and an ECA can yield biased estimates of the causal treatment effect due to systematic differences between groups with respect to measured and unmeasured baseline prognostic factors. Different methodologies have been proposed to adjust for measured confounding and used to support regulatory submissions for new treatments in U.S. and EU (FDA 2019a, 2019b, 2019c).
However, the interpretation of such analyses may be challenging if the causal question of interest is not clearly defined, or the estimator is not aligned with the estimand. Hernán and Robins (2016) highlighted that causal analyses of observational data need to be evaluated with respect to how well they emulate a particular target RCT, that is, a hypothetical RCT that would answer the question of interest if conducted. They outlined a framework to make the target trial explicit by providing a structured approach to define the question of interest and to avoid common methodological pitfalls.

In this article, we discuss the application of the target trial framework when comparing a SAT with an ECA to support regulatory submissions. Section 2 introduces the motivating example of a SAT of tisagenlecleucel in follicular lymphoma. In Section 3, we describe how the indirect comparison of the SAT and ECA was planned using a combination of the target trial and ICH E9(R1) estimand frameworks, while Section 4 elaborates on how we addressed the challenge of defining an appropriate time zero for external controls who meet the SAT eligibility criteria at several timepoints, and how we used the selected data to estimate the causal estimand. We present the results of our case-study in Section 5 before concluding with a discussion of our practical experiences of applying the combined target trial and estimand frameworks and their potential value for drug development in Section 6.

2. Motivating Example

2.1. The ELARA Trial: A Single-Arm Pivotal Phase II Study

Follicular lymphoma is a non-Hodgkin lymphoma that is generally considered indolent, but the disease remains incurable and the majority of patients eventually relapse. Patients with relapsed or refractory (r/r) follicular lymphoma will experience progressively shorter progression-free survival (PFS) to subsequent treatments, which has been shown to decrease from 6.6 years after the first line of therapy (LoT) to 1.5 years and 2.1 years after the second and third lines of therapy, respectively (Link et al. 2019). Hence, r/r follicular lymphoma remains the leading cause of mortality for patients with this disease and represents an unmet medical need (Sarkozy et al. 2019).

The ELARA trial is an ongoing, single-arm, global, multicenter, Phase II trial to determine the efficacy and safety of tisagenlecleucel (a novel cell therapy) in adult patients with r/r follicular lymphoma who were r/r to second or later line therapy (Fowler et al. 2020; Schuster et al. 2019). A total of 98 patients were enrolled, of whom 97 were infused with tisagenlecleucel. A study schema is presented in Figure 1.

During the national regulatory authority protocol review process, an EMA rapporteur highlighted the importance of having an ECA with patient-level data to support the review of any potential submission based on the ELARA trial. Consequently, RWE was derived using two pre-specified external data sources to provide stakeholders with evidence on the magnitude of the effect of tisagenlecleucel efficacy compared with the standard of care (SoC). The first was a noninterventional retrospective cohort study based on chart review (ReCORD-FL; for more details see below); data were collected from academic centers in Europe and North America using an electronic data collection form. The second data source was a noninterventional study using electronic health records from the US Flatiron Health Research Database (FHRD). The FHRD is a nationwide US EHR-derived de-identified database originating from approximately 280 cancer clinics (~800 sites of care) (Flatiron Health Database 2020). Given the differences in how these RWD were generated and their geographical spread, it was decided not to combine them but rather use them to create two separate ECAs. The totality of the data from the FHRD and ReCORD-FL was expected to support a comprehensive efficacy assessment of tisagenlecleucel in r/r follicular lymphoma patients. For ease of presentation, in this manuscript we focus on the use of the ReCORD-FL dataset to create an ECA for ELARA. A similar approach was applied to the FHRD to create an ECA; the results of the comparison of that ECA with ELARA are provided elsewhere (Hao et al. in preparation).

2.2. Dataset Used to Create an External Control Arm for ELARA

A non-interventional retrospective chart review study, titled A Retrospective Cohort Study of Treatment Outcomes Among Adult Patients with Refractory or Relapsed Follicular Lymphoma (ReCORD-FL) (hereafter referred to as "ReCORD") was conducted to provide comparative, contextual evidence to the existing data on the efficacy of tisagenlecleucel from ELARA. Patient-level data were collected from centers in Europe and North America (Salles et al. 2021). To obtain an adequate sample size and to include patients treated with different regimens reflecting evolving management strategies, data were collected from patients with r/r follicular lymphoma treated between 1998 and 2020. No initial diagnoses before January 1, 1998 were permitted as a key treatment for r/r follicular lymphoma, rituximab, was approved by the FDA and EMA in 1997 and 1998, respectively. Wherever feasible, the ReCORD study adopted the same eligibility criteria as ELARA. As of cutoff date of December 31, 2020, 187 patients fitting the study eligibility criteria were
enrolled. For each eligible patient, key baseline and clinical variables were collected or derived at the start of each LoT from first diagnosis onward, including the patient’s best clinical response to the treatment line, dates of progression, death, or start of new anticancer therapy.

3. Defining the Target Estimand for Studies with a RWD-based ECA

3.1. Motivation for Combining the Target Trial and ICH E9(R1) Estimand Frameworks

The process of designing an ECA for a SAT begins with specifying the causal question that we want to answer with the comparison of these two non-randomized groups. Shaping a relevant and feasible question requires input from several stakeholders, including clinicians and statisticians, and there are several frameworks which can structure this process.

The PICO aid (Richardson et al. 1995) plays an important role in evidence-based medicine, often informing the search criteria for quantitative systematic reviews and how selected studies are combined (Thomas et al. 2021). The PICO acronym itemizes four key elements of a clinical question about treatments: patient Population, Intervention, Control, and Outcome(s). Extensions such as PICOT and PICOS go further to make explicit the role of Time (i.e., length of follow-up) and Study design, respectively, in a research question. Clearly, the PICO elements overlap with the attributes of an estimand as defined by the ICH E9(R1) estimand framework (ICH 2019). However, we prefer the estimand framework since this goes even further to capture how the effect of a treatment (vs. control) will be summarized, and outlines strategies for handling intercurrent events which either affect the presence or interpretation of relevant data. The estimand framework is relevant for RCTs as well as SATs (ICH 2019) and clinical trials run through registries (EMA 2021).

Beyond clinical trials, the ENCePP Guide on Methodological Standards in Pharmacoepidemiology advises on how to set the research question for non-interventional studies (ENCEPP 2021). A number of frameworks and tools have also been proposed for studies generating RWE (Wang et al. 2021; Gatto, Reynolds, and Campbell 2019). Section 1 of this article highlighted the target trial framework (Hernán and Robins 2016) as a tool that can be used to ensure the design and analysis of an observational study is aligned with a well-posed causal question. This framework guides the researcher to carefully define key elements of the protocol of the “target” RCT that would answer the causal question of interest. Several of these elements specify ICH E9(R1) estimand attributes such as patient population, treatment strategies, outcome variable and summary measure. Additional protocol sections such as “follow-up period” facilitate careful consideration of how to define start of follow-up (also referred to as “time zero”) for patients; while this may be straightforward for RCTs, since start of follow-up usually coincides with time of randomization, defining time zero is often more challenging for non-interventional studies making secondary use of RWD. However, the target RCT protocol defined by Hernán and Robins (2016) does not include a section explicitly detailing how intercurrent events would be handled. Several authors have proposed combining the target trial and ICH E9(R1) estimand frameworks (Hampson et al. 2023; Polito et al. 2021) for clinical studies using RWD. In particular, Hampson et al. (2023) propose extending the protocol of the target RCT protocol to include a section on handling intercurrent events, in order to facilitate a more granular approach to handling intercurrent events and communication of these strategies. Since many clinical trial teams are already familiar with the estimand framework through its application to RCTs, we decided to adopt this combination of the target trial and estimand frameworks for the ELARA trial. We describe this approach in more detail below.

3.2. Applying the Target Trial and Estimand Frameworks to the ELARA Trial

Table 1 illustrates how the combined target trial and estimand frameworks were applied to design a real-world (RW) based ECA for the ELARA trial. Broadly speaking, for each key efficacy endpoint, the question of interest selected by the team was the marginal causal effect of prescribing tisagenlecleucel versus SoC in the patient population who participated in the ELARA trial; in common terminology, this is an “average treatment effect on the treated” (ATT). The left-hand column of Table 1 defines the protocol of the target RCT that we would ideally perform to answer this question. Key points to note are that since we are targeting an ATT, the eligibility criteria of the target RCT are of those of ELARA. We aim to emulate an active controlled randomized trial comparing the tisagenlecleucel strategy evaluated in ELARA (and illustrated in Figure 1) with the SoC per physician’s choice. In this indication, the number of prior LoT at baseline is a well-understood strong prognostic variable. In terms of the key efficacy endpoints, the most relevant intercurrent event in the target RCT would be initiation of a new anticancer therapy. For complete response (CR), if a patient hasn’t achieved response prior to starting a new therapy they would be regarded as a nonresponder. For PFS, initiation of a new therapy would be handled through a hypothetical strategy where a patient would be censored at the time of initiating a new anticancer therapy if no progression or death is observed before the new anticancer therapy. Meanwhile, for overall survival (OS), it is handled using a treatment policy approach, recording the time to all-cause mortality regardless of whether a patient deviates from their randomized strategy. The causal contrasts are difference in probabilities (CR) and hazard ratios (PFS and OS).

The right-hand columns of Table 1 summarize the RCT that we can emulate given the data available from ELARA and the external controls. Comparing this with our target RCT highlights limitations of the RWD. In particular, while we can emulate the target population based on key ELARA eligibility criteria as envisioned in future drug label criteria, not all criteria were captured in the RWD (e.g., ECOG or lab results were not recorded or incomplete). However, based on clinical assessment it was concluded that the impact of these criteria on the adequacy of the ECA is likely to be limited and does not prevent meaningful comparisons. Another potential limitation is that CR and progression in the RWD were assessed per real-world
**Table 1.** Application of the target trial and estimand frameworks to design an ECA based on the ReCORD study and the comparison of this with the ELARA trial cohort.

| Component | Target RCT to be emulated | RCT that can be emulated using ELARA and external RWD |
|-----------|----------------------------|-----------------------------------------------------|
| **Objective** | To evaluate the efficacy of tisagenlecleucel as compared to the current SoC as measured by complete response (CR), overall survival (OS), and progression free survival (PFS) | ELARA | ReCORD cohort |
| **Patient Population** | Eligibility criteria of ELARA | Same as target RCT | Eligibility criteria of ELARA that are feasible to implement in a retrospective assessment of the RWD. In addition, patients with r/r follicular lymphoma should have been treated between 1998 and 2020. |
| **Treatment** | Optional bridging therapy and lymphodepleting therapy followed by tisagenlecleucel infusion vs SoC | Tisagenlecleucel infusion after optional bridging therapy and lymphodepleting therapy | Current SoC including anti-CD20 monoclonal antibodies, alkylating agents, radiotherapies and PI3K inhibitors (e.g., idelalisib) |
| **Treatment assignment** | Block randomized (unblinded) to either tisagenlecleucel arm or SoC arm. | Emulate simple randomization with no blinding |
| **Variables** | CR and progression are assessed per Lugano classification 2014 \(^a\) until patient initiates a new anticancer treatment. CR is best overall response. If no response is reported before time of new therapy, patient is regarded as non-responder. PFS is time to the first documented progression or death from any cause. Initiation of a new anticancer therapy would be handled through a hypothetical strategy; if a patient does not have progression or death prior to initiation, PFS would be censored at the time of initiation. OS is time to death from any cause regardless of treatment | CR and OS as target RCT. For comparison with ReCORD, PFS is performed by considering new anticancer therapy as an event. | CR and progression were based on real-world response criteria. A subgroup analysis for patients treated after January 01, 2014 was conducted as year of introduction of Lugano response criteria. In ReCORD, progression dates were unavailable for many patients, so consider new anticancer therapy as an event for PFS. OS same as in target RCT. |
| **Start and end of follow-up** | Date of randomization, regarded as date of prescription. Assessment of CR and PFS is based on the first LoT received after randomization. Follow-up for OS continues until time of death, loss to follow-up or study end. | Date of enrollment into ELARA, regarded as date of prescription. Assessment of CR and PFS is based on the first LoT received after study enrollment. Follow-up for OS continues until time of death, loss to follow-up or study end. | One eligible LoT per patient is systematically selected based on the highest propensity score to be in ELARA; the start date of this selected LoT defines time zero (or start of follow-up) for each ReCORD patient; see Section 4 for more details. Assessment of CR and PFS is based on a patient's selected LoT. Follow-up for OS continues until time of death, loss to follow-up or study end. |
| **Intercurrent event(s)** | Initiation of new anticancer therapy before progression. CR: ICE reflected in Variable PFS: Hypothetical strategy (reflected in Variable) OS: Treatment policy strategy (reflected in Variable) | Same as target RCT for CR and OS PFS comparison: Composite strategy |
| **Performed comparison** | Relative benefit after prescribing tisagenlecleucel vs SoC | Relative benefit after prescribing tisagenlecleucel vs SoC after being treated with SoC assessed by weighting by odds, as described in Section 4. |
| **Causal effect** | Effect of prescribing tisagenlecleucel vs SoC in patients meeting ELARA inclusion/exclusion criteria | The effect of prescribing tisagenlecleucel vs SoC is assessed in patients who participated in ELARA |
| **Causal contrast (i.e., Summary measure)** | Binary endpoints: Difference in marginal response probabilities on tisagenlecleucel vs SoC Time-to-event (TTE) endpoints: Marginal HR | Same as target RCT |
| **Analysis** | Binary endpoints: Difference in response rate between arms TTE endpoints: HR obtained from Cox regression | Binary endpoints: Difference in weighted proportion of responders on tisagenlecleucel vs SoC. TTE endpoints: HR obtained from a weighted Cox regression model For both types of analysis, weights are calculated using the methodology described in Section 4. 95% CIs are calculated using bootstrap resampling. |

regular practice and not following the Lugano classification 2014 criteria (Cheson et al. 2014) as used in ELARA. Regarding this, a subgroup analysis for patients treated after January 01, 2014, to coincide with the introduction of the Lugano response criteria, was conducted. Moreover, this limitation does not impact OS. A third limitation is that while the date of treatment initiation
is available for all patients, the date of prescription is unknown for ReCORD patients. However, in the context of currently available therapies and a serious condition such as follicular lymphoma, the lag between prescription and start of treatment is likely to be short, meaning the date of treatment initiation will be an adequate approximation for the date of prescription. Since the date of initiation of all treatments is recorded for ReCORD patients, key intercurrent event of initiation of a new anticancer therapy is captured. We can apply the same strategy as would be used in the target RCT to handle this event when comparing the ELARA and external controls with regards to CR and OS. However, progression dates were unavailable for many patients in ReCORD, so the comparison of PFS was performed by considering a new anticancer therapy as an event for patients in both ReCORD and ELARA. If no progression, death or new anticancer therapy occurred, patients are censored at the last contact date.

We followed a two-step procedure to emulate the treatment assignment and follow-up strategies of the target RCT. In Step 1, we used a propensity-score based approach to select one eligible LoT per ReCORD patient. In Step 2, we applied a weighting-by-odds approach to the dataset comprising the ELARA patients and the ReCORD patients at their selected LoT to mimic randomization and estimate the causal estimand. More details on this emulation approach are provided in the next section.

4. Emulating the Treatment Assignment and Follow-Up Strategies of the Target RCT

4.1. Rationale for Selecting One LoT per ReCORD Patient

As mentioned in Section 2, the ELARA trial was open to patients who were r/r to a second or later line of systemic therapy, with no upper limit placed on the number of previous LoT. The median number of previous LoT on entry to ELARA was 4 (range: 2–13). In the ReCORD cohort, 80% of patients had received 2 previous LoT at the time of first meeting all eligibility criteria (range: 2–5 lines).

Another key difference between the ELARA and ReCORD patients is that CR and PFS were assessed in the trial only for a single LoT, whereas longitudinal data were available on these endpoints for ReCORD patients who were followed and remained eligible across multiple LoT. These differences in data structure are common for SATs in late-stage r/r disease with RW-based ECAs.

If observational data from the same source were to be used to emulate both arms of a target RCT and patients would meet the eligibility criteria at multiple LoT, valid approaches to LoT selection would include selecting a single eligible line or using data from all eligible LoT to emulate multiple nested RCTs (Hernán and Robins 2016). However, our case study differs from that setting because there are systematic differences in data capture between the ELARA trial and the RWD. In the case of a SAT in late-line therapy with RW external controls, Backenroth (2021) evaluates several approaches for defining start of follow-up for the external controls with respect to bias for the treatment effect; these approaches include: selecting the last eligible LoT; selecting a line at random from a patient’s eligible LoT; and using information on all eligible LoT. To avoid bias, the probability of line selection should not depend on the total number of (eligible) LoT a patient received before being censored or experiencing the event of interest (Backenroth 2021), a condition that is violated if, say, we select the last eligible LoT for each ReCORD patient or randomly select from a ReCORD patient’s eligible LoT. To avert this limitation, we devised the selection approach that will be described in Section 4.2. For ease of analysis and interpretation, it was decided to mimic the structure of the ELARA dataset by selecting one eligible LoT per ReCORD patient. This is also closely aligned with the structure of a future RCT that would compare responses on tisagenlecleucel and SoC on the first received LoT in the trial. While this approach sacrifices some information, it has the advantage of avoiding correlations between repeated measurements on the same RW patient and ensures individual trial participants and external controls contribute similar amounts of statistical information to the final analysis. The challenge then is how to select one eligible LoT per ReCORD patient.

4.2. Novel Approach to Select One Eligible LoT per ReCORD Patient

First we pooled the ELARA cohort and the ReCORD cohort (comprising observations on all eligible LoT) and used this to derive for each patient at each eligible LoT a value eij, defined as the conditional probability patient i was prescribed tisagenlecleucel at their jth eligible LoT given the values of their covariates at the time of starting this LoT. Here, a tisagenlecleucel prescription is synonymous with enrollment into the ELARA trial. While there are parallels between the definition of eij and a propensity score (PS), as an anonymous reviewer highlighted the structure of the longitudinal ReCORD data and the use of the eijs to facilitate line selection differ from a traditional PS application. To avoid confusion, we therefore prefer to refer to each eij as a similarity measure (SM), since a large value of eij for the ith ReCORD patient implies that at the start of line j they were similar to the ELARA cohort in terms of their baseline covariates.

To illustrate the definition of the eijs, suppose a ReCORD patient labelled patient i* received third, fourth and fifth lines of treatment but only met the ELARA eligibility criteria at the third and fourth lines. Then, we calculated a SM for patient i* at each of these two eligible LoT, denoted by eij1 (corresponding to line 3, patient i*’s first eligible LoT), and eij2 (corresponding to line 4, their second eligible LoT). Note that since ELARA patients were followed-up for progression across one LoT, each ELARA patient is associated with only one SM.

For the purposes of fitting a SM model, each patient at each eligible LoT was defined to have an exposure outcome equal to 1 if they were prescribed tisagenlecleucel, and 0 otherwise. Therefore, each ELARA patient has one such outcome imputed (which takes the value 1), while each ReCORD patient may have several such exposure outcomes imputed (all of which take the value 0). One could model the ‘exposure data’ generated in this way using a mixed effects logistic regression, with baseline covariates represented by fixed effects in the linear predictor and a random subject effect included for each ReCORD
patient to capture the correlation between their longitudinal data. While such a model would precisely capture the structure of the ELARA and RECORD data, one would need to write bespoke code to fit it and there is the potential for convergence issues if few RECORD patients are followed across multiple eligible LoT. With these challenges in mind, for simplicity we modeled the $e_{ij}$ using the fixed-effects logistic regression shown below:

$$\log \left( \frac{e_{ij}}{1 - e_{ij}} \right) = \mu + \sum_{k=1}^{10} \beta_k X_{ijk}$$

where $X_{ijk}$ is the value of the $k$th covariate for the $i$th patient measured at the time of initiating their $j$th eligible LoT. Ten covariates were pre-specified by clinical experts before seeing the outcome data for RECORD patients including age (continuous variable), region (Europe or North America), gender (Female or Male), history of autologous HSCT (Yes or No), number of prior lines of therapy (>4 lines or 2-4 lines), disease stage at initial diagnosis (Stage I, II, III, or VI), time from initial diagnosis to initiation of treatment (continuous variable), sites of nodal involvement (>4 or ≤4), double refractory status (Yes or No), and progression of disease within 24 months of first line treatment (POD24;Yes or No).

The logistic regression was fitted using a generalized estimating equations (GEE) approach, using a robust sandwich variance estimator to ensure the standard errors of model parameters reflect the correlation between longitudinal measurements. However, in what follows, only the point estimates of the $e_{ij}$ (denoted by $\hat{e}_{ij}$) were used to inform LoT selection.

For each RECORD patient, we selected the eligible LoT associated with their highest estimated SM. For example, suppose the estimated SMs for RECORD patient $i^*$ mentioned above were $\hat{e}_{i1} = 0.56$ (given the values of their 10 covariates at the time of starting line 3, their first eligible LoT) and $\hat{e}_{i2} = 0.75$ (given the values of their 10 covariates at the time of starting line 4, their second eligible LoT). Then, line 4 would be the selected LoT for patient $i^*$, and we would define time zero as the calendar time at which they started this LoT. Intuitively, we interpret the selected LoT as the line at which the RECORD patient is most similar (in terms of their baseline covariates, including number of previous lines of therapy) to the ELARA cohort. We note that for patients eligible at multiple LoT, the selection process is not directly influenced by a patient’s number of eligible lines; instead, selection is more closely linked with our objective of trying to minimize confounding between groups. In addition, our proposed approach preserves the sample size of the RECORD cohort.

### 4.3. Emulating Randomization to Estimate the Causal Estimand

Causal inferences were based on the dataset formed by pooling the ELARA cohort with the RECORD cohort at their selected LoT. Note that even after selecting one LoT per RECORD patient, imbalances between groups at time zero will likely still persist. To mitigate these differences and thus emulate randomization, we derived one PS per patient. This PS is denoted by $p_i$ and defined as the conditional probability patient $i$ was assigned tisagenlecleucel given their baseline covariates at their selected LoT. We modeled the $p_i$ using a logistic regression with main effects for the 10 covariates specified in Section 3.2. This model was fitted using least squares estimation, since all patients contributed only one observation to the analysis. Estimates of the $p_i$ were then used in a PS-based analysis to adjust for confounding. In Section 3, we defined the target estimand as the average effect of prescribing tisagenlecleucel as compared to SoC in those who participated in the ELARA trial. To estimate this, we reweighted patients to create tisagenlecleucel and SoC groups with covariate distributions identical to the distribution seen in the ELARA trial. Clearly, we did not need to reweight the ELARA cohort to create a tisagenlecleucel group fulfilling this objective. With this in mind, if patient $i$ in our final analysis dataset participated in the ELARA trial, they were assigned a weight of $1$; otherwise they were assigned a weight equal to $\frac{\hat{p}_i}{(1 - \hat{p}_i)}$. This “weighting-by-odds” strategy creates an external control group with the same covariate distribution as the ELARA cohort.

We then used these weights to compute estimates of causal effects by comparing weighted aggregate statistics (e.g., the difference in the weighted proportion of responders for binary endpoints). Alternatively, we extracted treatment effect estimates from (semi-)parametric models fitted allocating patients different weights, for example, estimating the log-HR by fitting a weighted Cox regression with treatment as the only regressor and treating the weights $\frac{\hat{p}_i}{(1 - \hat{p}_i)}$ assigned to RECORD patients as replication weights. The standard error of the causal treatment effect estimator should account for the fact that patient weights are estimated from the data. Austin (2016) compared three approaches (naïve-model based variance estimator, robust sandwich-type variance estimator, and bootstrap) for estimating the variance of a treatment effect estimate obtained from an inverse probability treatment weighting analysis. The author found that the bootstrap estimator provides the most robust estimate of the standard error of the hazard ratio from a Cox proportional-hazards model. In this case-study we used nonparametric bootstrap to estimate the standard error based on 10,000 bootstrap samples (Efron and Tibshirani 1993) randomly drawn (with replacement) from the dataset obtained after selecting one LoT per RECORD patient.

### 5. Application to the ELARA Trial

In this section, we provide the application of the methods described in the previous section to the ELARA trial using the RECORD data as an ECA. Since our primary aim was to use this ECA to contextualize the ELARA trial and understand the magnitude of the effect of tisagenlecleucel versus SoC, we summarize the results of comparative analyses with point estimates and confidence intervals rather than $p$-values.

Included in the analysis were a total of 97 patients (out of 98 enrolled) from ELARA and 143 patients (out of 187 recruited) from the RECORD study who had complete data on key baseline covariates; this sample consisted of 326 LoT for 143 RECORD patients. As described in Section 4, to mimic the structure of the ELARA dataset, a propensity score-based approach was used to select one LoT per RW patient.
After line selection, we observed that the median number of prior lines of therapy in ReCORD is close to ELARA (3 vs. 4), with 23.1% of ReCORD patients having received more than four prior LoT (vs. 28.9% in ELARA). However, imbalances between groups still persisted in many baseline covariates. Thus, as described in Section 4, the weighting by odds approach was used to adjust for potential confounding. Table 2 shows the comparison of baseline variables between ReCORD patients (at their selected LoT) and ELARA patients before and after ReCORD patients are weighted by their odds of being in ELARA. Standardized mean differences (SMD) between the two cohorts were assessed for pre-weighted and post-weighted data. The weighting by odds analysis appears to be reasonable to adjust for imbalances of measured baseline covariates between groups as suggested by absolute SMDs less than 25% (Stuart, Lee, and Leacy 2013). The estimated propensity scores at the selected LoT

Table 2. Baseline variables for ReCORD and ELARA before and after weighting at selected LoT.

| Baseline variable                        | ELARA | Before weighting | After weighting |
|-----------------------------------------|-------|------------------|-----------------|
| Statistics                              |       | ReCORD           |                 |
| n                                       | 97    | 143              |                 |
| Mean(SD)                                | 56.5  | (10.40)          | 60.1 (11.72)    |
| Median                                  | 58    |                  | 56.1 (11.52)    |
| Min–Max                                 | 29–73 |                  | 25–86           |
| Age at treatment initiation category—n (%) |       |                  |                 |
| <65 years                               | 73    | (75.3)           | 89 (62.2)       |
| ≥65 years                               | 24    | (24.7)           | 54 (37.8)       |
| Gender—n (%)                            |       |                  |                 |
| Female                                  | 33    | (34.0)           | 61 (42.7)       |
| Male                                    | 64    | (66.0)           | 82 (57.3)       |
| Region—n (%)                            |       |                  |                 |
| Europe                                  | 44    | (45.4)           | 90 (62.9)       |
| RoW                                     | 53    | (54.6)           | 53 (37.1)       |
| Prior Auto-HSCT—n (%)                   |       |                  |                 |
| Yes                                     | 36    | (37.1)           | 53 (37.1)       |
| No                                      | 61(62.9)|              | 90 (62.9)       |
| Number of previous lines of systemic treatment |       |                  |                 |
| n                                       | 97    | 143              |                 |
| Mean(SD)                                | 3.9   | (1.78)           | 3.7 (2.05)      |
| Median                                  | 4     |                  | 3               |
| Min–Max                                 | 2–13 |                  | 2–10            |
| Disease stage at initial FL diagnosis—n (%) |       |                  |                 |
| Stage I                                 | 6     | (6.2)            | 10 (7.0)        |
| Stage II                                | 13    | (13.4)           | 13 (9.1)        |
| Stage III                               | 21    | (21.6)           | 26 (18.2)       |
| Stage IV                                | 57    | (58.8)           | 94 (65.7)       |
| Months between initial FL diagnosis and initiation of treatment |       |                  |                 |
| n                                       | 97    | 143              |                 |
| Mean(SD)                                | 77.3  | (56.33)          | 72.1 (48.53)    |
| Median                                  | 66.2  |                  | 61.7            |
| Min-Max                                 | 6.4–355.4 |              | 2.8–255         |
| n                                       | 97    | 143              |                 |
| Mean(SD)                                | 77.3  | (56.33)          | 72.1 (48.53)    |
| Median                                  | 66.2  |                  | 61.7            |
| Min-Max                                 | 6.4–355.4 |              | 2.8–255         |
| Number of nodal involvement at treatment initiation—n (%) |       |                  |                 |
| ≤4                                      | 39    | (40.2)           | 74 (51.7)       |
| >4                                      | 58    | (59.8)           | 69 (48.3)       |
| Double refractoryb—n (%)                |       |                  |                 |
| Yes                                     | 66    | (68.0)           | 97 (67.8)       |
| No                                      | 31    | (32.0)           | 46 (32.2)       |
| POD24c—n (%)                            |       |                  |                 |
| Yes                                     | 61    | (62.9)           | 86 (60.1)       |
| No                                      | 36    | (37.1)           | 57 (39.9)       |

SMD: standard mean difference; FLIPI: Follicular Lymphoma International Prognostic Index; Auto-HSCT: Autologous hematopoietic stem cell transplantation.

S M D : standard mean difference; F L I P I : Follicular Lymphoma International Prognostic Index; A u t o - H S C T : Autologous hematopoietic stem cell transplantation.

a Sample size after weighting (i.e., sum of weights) was 99 for the ReCORD study and effective sample size was 95.

b Double refractory status is defined as patients failing to respond or experiencing relapse within 6 months to both a prior anti-CD20 antibody and a prior alkylating agent.

c POD24 status is defined as patients failing to respond or experiencing relapse within 24 months to the first-line anti-CD20 mAb containing therapy.
for ReCORD patients and ELARA patients are shown in the upper panels of Figure 2, with a smoothed estimate of the density overlaid by exposure groups. The lower panel of Figure 2 repeats this comparison but this time weighting ReCORD patients by their odds of being enrolled in ELARA. The weighting by odds method appears to have adequately balanced the propensity score distribution between these two groups.

Table 3 shows the estimates of the causal effect of prescribing tisagenlecleucel versus SoC before and after weighting by odds. Figure 3 shows the Kaplan-Meier curves for OS and PFS after weighting. The median follow-up time (defined as time to death or last follow-up date) was 15 months for ELARA, and 22 months in the weighted sample for ReCORD (at the selected LoT). For comparative analyses, the Kaplan-Meier and Cox regression results are based on survival data within the first 24 months, and patients with survival data beyond 24 months were censored at 24 months. A clinically meaningful and consistent improvement for all endpoints was observed before and after weighting in ELARA vs ReCORD. The results from subgroup analysis for patients treated after January 01, 2014 are generally consistent with results in the main analysis presented above.

6. Discussion

In the ELARA example, members of the clinical trial team found that a combination of the target trial and estimand frameworks was helpful to formulate a well-posed and feasible causal question and to align the estimator with the estimand. Table 1, which compares the protocol of the RCT we want to emulate with that of the one we can emulate, was used in communications with clinical trial team members and health authorities, and facilitated transparent discussion of the feasibility and value of the RW-based ECA. To our knowledge, this is the first time the target trial framework has been used in a regulatory submission with RWE.

Several learnings arose from this case study. First, if planning an ECA, it is best to use the target trial and estimand...
Table 3. Efficacy comparison of ELARA and ReCORD before and after weighting.

|                          | ELARA          | Before weighting | After weighting |
|--------------------------|----------------|------------------|-----------------|
|                          |                | ReCORD           | ReCORD          |
| Response rate            | N = 97         | N = 143          | N = 99<sup>a</sup>|
| CR, 95% CI               | 69.1 (59.8, 78.3) | 39.2 (31.1, 47.2) | 37.3 (26.4, 48.3) |
| Difference in CR, 95% CI | 29.9 (17.7, 42.1) | 31.8 (18.1, 45.3) |
| OS                       |                | N = 97           | N = 99          |
| Kaplan-Meier analysis    | Events/Total (%) N = 97 | 7/97             | 72/143          |
| Median, 95% CI (months)  | 6 months       | 100 (100, 100)   | 88.4 (83.0, 93.7) |
|                          | 12 months      | 96.6 (92.9, 100) | 75.7 (68.3, 83.0) |
|                          | 18 months      | 91.4 (84.6, 98.3) | 71.5 (63.7, 79.3) |
|                          | 24 months      | 87.8 (78.0, 97.6) | 69.7 (61.8, 77.7) |
| Cox proportional hazard model | HR, 95% CI     | 0.25 (0.03, 0.46) | 0.20 (0.02, 0.38) |

OS and PFS are measured relative to enrollment date/treatment start date. All K-M and Cox regression results are based on survival data within the first 24 months (patients with survival data beyond 24 months were censored at Month 24).

<sup>a</sup>Sample size after weighting (i.e., sum of weights) was 99 for the ReCORD study and effective sample size was 95.

Our final key learning is that it is important to align ahead of time with clinical collaborators and health authorities on which baseline covariates are anticipated to be strong prognostic and/or predictive factors, to ensure all relevant patient information is extracted. However, while it is important to follow the same principles for RWD analyses as for clinical trials and prespecify the statistical analysis plan, some flexibility may be required from all stakeholders. For example, if very few RW patients have a baseline characteristic relatively common in the SAT cohort (e.g., prior autologous hematopoietic stem cell transplantation) which is prespecified for inclusion in the PS model, one may need to adopt a reasonable post-hoc approach to handle this.

Two analytical challenges that we haven’t discussed in detail are: (a) how to handle missing data on key baseline covariates or efficacy outcomes when formulating a RW-based ECA; and (b) how to evaluate the robustness of conclusions to unverifiable assumptions. For ELARA, we took a rather simple approach of handling missing data via a complete-case analysis but alternative approaches, such as multiple imputation (Leyrat et al. 2019), could have been considered. In terms of assessing the robustness of results, having a precisely defined estimand ensures sensitivity analyses can be aligned with the question of interest. In the field of epidemiology, a large body of literature exists on quantitative bias assessment and metrics quantifying the sensitivity of conclusions to unmeasured confounding exist for binary and time-to-event endpoints (Lash, Fox, and Fink 2009; VanderWeele and Ding 2017). Future work would develop best practices for sensitivity analyses for ECAs and extend methods to accommodate the range of endpoints we might encounter in clinical trials.

In conclusion, we think combining the target trial and estimand frameworks can be useful when planning a SAT with an ECA as it offers a common language to discuss existing complexities with all relevant stakeholders including health authorities and payers. Experience with these frameworks in clinical studies with RWD is still limited, and their use in the oncology setting is being explored in dedicated industry working groups. More published case-studies will be useful for guiding researchers.

**Acknowledgments**

We are grateful to acknowledge helpful discussions with Hemanth Kanakamedala which informed the development of methods described in this manuscript.
References

Austín, P. C. (2016), “Variance Estimation When Using Inverse Probability of Treatment Weighting (IPTW) with Survival Analysis,” *Statistics in Medicine*, 35, 5642–5653. [6]

Backenroth, D. (2023), “How to Choose a Time Zero for Patients in External Control Arms,” *Pharmaceutical Statistics*, 20, 783–792. [5]

Cheson, B. D., Fisher, R. I., Barrington, S. E., Cavalli, F., Schwartz, L. H., Zucca, E., and Lister, T. A., et al. (2014), “Recommendations for Initial Evaluation, Staging, and Response Assessment of Hodgkin and Non-Hodgkin Lymphoma: The Lugano Classification,” *Journal of Clinical Oncology*, 32, 3059–3068. [4]

Collins, R., Bowman, L., Landray, M., and Petö, R. (2020), “The Magic of Randomization versus the Myth of Real-World Evidence,” *The New England Journal of Medicine*, 382, 674–678. [1]

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, London: Chapman & Hall. [6]

European Medicines Agency European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). (2021), “ENCePP Guide on Methodological Standards in Pharmacoepidemiology.” [3]

European Medicines Agency (2021), “Guideline on Registry-based Studies,” available at https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-registry-based-studies_en-0.pdf [3]

Fowler, N. H., Dickinson, M., Dreyling, M., Martinez-Lopez, J., Kolstad, A., Austin, P. C. (2016), “Variance Estimation When Using Inverse Probability Weighting (IPTW) with Survival Analysis,” *Statistics in Medicine*, 35, 5642–5653. [6]

ICH (2019), “ICH E9(R1) Addendum on Estimands and Sensitivity Analyses in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials,” Available at https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf [3]

ICH (2019), “ICH E9(R1) Addendum on Estimands and Sensitivity Analyses in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials,” Available at https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf [3]

Lash, T. L., Fox, M. P., and Fink, A. K. (2009), *Applying Quantitative Bias Analysis to Epidemiologic Data*, New York: Springer. [9]

Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Sneath, L., Kim, J., Resche-Rigon, M., Carpenter, J. R., and Williamson, E. J. (2019), “Propensity Score Analysis with Partially Observed Covariates: How Should Multiple Imputation be Used?” *Statistical Methods in Medical Research*, 28, 3–19. [9]

Link, B. K., Day, B.-M., Zhou, X., Zelenetz, A. D., Dawson, K. L., Cerhan, J. R., Flowers, C. R., and Friedberg, J. W. (2019), “Second-Line and Subsequent Therapy and Outcomes for Follicular Lymphoma in the United States: Data from the Observational National LymphoCare Study,” *British Journal of Haematology*, 184, 660–663. DOI:10.1111/bjh.15149 [2]

Polito, L., Liang, Q., Pal, N., Mpfou, P., Sawas, A., Humblet, O., Rubbach, K., and Heinzman, D. (2021), “Using the Estimand and the Target Trial Frameworks When Building an External Control Data from Real-World Data: A Case-Study in Oncology,” in *IBS ROE Conference 2021*. Available at https://www.conftool.pro/ibs-roes2021/index.php?page=browseSessions&form_session=2 [3]

Richardson, W. S., Wilson, M. C., Nishikawa, J., and Hayward, R. S. A. (1995), “The Well-Built Clinical Question: A Key to Evidence-based Decisions,” *ACP Journal Club*, 123, A12-3. [3]

Salles, G., Schuster, S. J., Fischer, L., Kuruvilla, J., Patten, P. E., von Tresckow, B., Smith, S. M., Jimenez Ubieto, A. L., Davis, K. L., Nagar, S. P., Zhang, J., Bolu, V., Joussauzy, E., Ramos, R., Wang, Y., and Link, B. K. (2021), “A Retrospective Cohort Study of Treatment Outcomes of Adult Patients with Relapsed or Refractory Low-Grade Follicular Lymphoma (RECORD-FL),” *Blood*, 138, 1349–1349. DOI:10.1182/blood-2021-145870. [2]

Sarkozy, C., Maurer, M. J., Link, B. K., Ghersi, H., Nicolás, E., Thompson, C. A., Traverse-Glehne, A., Feldman, A. L., Allmer, C., Sager, S. L., Ansell, S. M., Habermann, T. M., Bachy, E., Cerhan, J. R., and Salles, G. (2019), “Cause of Death in Follicular Lymphoma in the First Decade of the Rituximab Era: A Pooled Analysis of French and US Cohorts,” *Journal of Clinical Oncology*, 37, 144–152. [2]

Schuster, S. J., Bishop, M. R., Tam, C. S., Waller, E. K., Borchmann, P., McGuirk, J. P., Jager, U., Jaglowski, S., Andreidis, C., Westin, J. R., Fleury, I., Bahanová, V., Foley, S. R., Ho, P. J., Mielke, S., Magenau, J. M., Holte, H., Pantano, S., Paucad, L. B., Avasthi, R., Chu, J., Anak, Ō., Salles, G., and Mazzara, R. T. (2019), “Tisagenlecleucel in Adult Relapsed or Refractory Diffuse Large B-Cell Lymphoma,” *New England Journal of Medicine*, 380, 45–56. [2]

Stuart, E. A., Lee, B. K., and Leacy, F. P. (2013), “Prognostic Score-based Measures Can Be a Useful Diagnostic for Propensity Score Methods in Comparative Effectiveness Research,” *Journal of Clinical Epidemiology*, 66, S84–S90.e1. [7]

Tenhunen, O., Lasch, F., Schiel, A., and Turpeinen, M. (2020), “Single-Arm Clinical Trials as Pivotal Evidence for Cancer Drug Approval: A Reflective Cohort Study of Centralized European Marketing Authorizations between 2010 and 2019,” *Clinical Pharmacology & Therapeutics*, 108, 653–660. [1]

Thomas, J., Kneale, D., McKenzie, J. E., Brennan, S. E., and Bhaumik, S. (2021), “Determining the Scope of the Review and the Questions it will Address,” in *Cochrane Handbook for Systematic Reviews of Interventions Version 6.2*, eds. J. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, V. A. Welch (updated Feb 2021). Available at www.training.cochrane.org/handbook [3]

VanderWeele, T. J., and Ding, P. (2017), “Sensitivity Analysis in Observational Research: Introducing the e-value,” *Annals of Internal Medicine*, 167, 268–274. [9]

Wang, S. V., Pinheiro, S., Hua, W., Arlett, P., Uyama, Y., Berlin, J. A., Bartels, D. B., Kahler, K. H., Bessette, L. G., and Schneeweiss, S. (2021), “STArT-RWE: Structured Template for Planning and Reporting on the Implementation of Real World Evidence Studies,” *British Medical Journal*, 372, m4856. DOI:10.1136/bmj.m4856. [3]

Zhou, J., Vallejo, J., Kluetz, P., Pazdur, R., Kim, T., Keegan, P., Farrell, A., Beaver, J. A., and Sridhara, R. (2019), “Overview of Oncology and Hematology Drug Approvals at US Food and Drug Administration between 2008 and 2016,” *Journal of National Cancer Institute*, 111, dyz130. [1]