Beyond Clinical Trials: Using Real World Evidence to Investigate Heterogeneous, Time-Varying Treatment Effects

Isabel Chien* IC390@CAM.AC.UK
University of Cambridge

Cliff Wong CLIFF.WONG@MICROSOFT.COM
Microsoft Research

Zelalem Gero ZELALEMGERO@MICROSOFT.COM
Microsoft Research

Jaspreet Bagga JABAGGA@MICROSOFT.COM
Microsoft Research

Risa Ueno RISA.UENO@MICROSOFT.COM
Microsoft Research

Richard E. Turner RET26@CAM.AC.UK
University of Cambridge

Roshanthi K. Weerasinghe ROSHANTHI.WEERASINGHE@PROVIDENCE.ORG
Providence Health

Brian Piening BRIAN.PIENING@PROVIDENCE.ORG
Providence Health

Tristan Naumann TRISTAN@MICROSOFT.COM
Microsoft Research

Carlo Bifulco CARLO.BIFULCO@PROVIDENCE.ORG
Providence Health

Hoifung Poon HOIFUNG@MICROSOFT.COM
Microsoft Research

Javier Gonzàlez Hernàndez GONZALEZ.JAVIER@MICROSOFT.COM
Microsoft Research

Abstract

Randomized controlled trials (RCTs), though essential for evaluating the efficacy of novel treatments, are costly and time-intensive. Due to strict eligibility criteria, RCTs may not adequately represent diverse patient populations, leading to equity issues and limited generalizability. Additionally, conventional trial analysis methods are limited by strict assumptions and biases. Real-world evidence (RWE) offers a promising avenue to explore treatment effects beyond trial settings, addressing gaps in representation and providing additional insights into patient outcomes over time. We introduce TRIALSCOPE-X and TRIALSCOPE-XL, machine learning pipelines designed to analyze treatment outcomes using RWE by mitigating biases that arise from observational data and addressing the limitations

* Work completed while intern at Microsoft Research

of conventional methods. We estimate causal, time-varying treatment effects across heterogeneous patient populations and varied timeframes. Preliminary results investigating the treatment benefit of Keytruda, a widely-used cancer immunotherapy drug, demonstrate the utility of our methods in evaluating treatment outcomes under novel settings and uncovering potential disparities. Our findings highlight the potential of RWE-based analysis to provide data-driven insights that inform evidence-based medicine and shape more inclusive and comprehensive clinical research, supplementing traditional clinical trial findings.

1. Introduction

Clinical trials are conducted to establish the impact of novel treatments as compared to a control, typically the standard-of-care or a placebo. Treatments that progress development beyond early safety and efficacy trials are investigated in larger randomized controlled trials (RCTs). A well-designed RCT controls for possible confounders and biases and is thus considered the gold standard for deriving causal treatment effects (Concato et al., 2000). However, trial eligibility criteria may be unnecessarily restrictive or arbitrary (Kim et al., 2015), deviating from the realities of medical practice (Stephenson, 2020). Historical biases have excluded under-represented populations such as women and people of color (Mccarthy, 1994; Cho et al., 2021). This has lead to health equity harms, as trial populations are often not representative of those receiving treatment in practice (Averitt et al., 2020; Chien et al., 2022). For example, women experience increased adverse effects across various drug classes (Unger et al., 2022; Zopf et al., 2008; Zucker and Prendergast, 2020).

Real-world evidence (RWE) derived from electronic health records (EHRs) are a valuable resource for investigating treatment effects beyond the confines of clinical trials. RWE includes clinical data on individuals who may have been excluded from or not adequately represented in trials, as well as those treated under diverse circumstances. Typically, additional trials are required to explore treatment effects under different settings. Keytruda, a prominent immunotherapy drug, has undergone extensive study through various KEYNOTE trials examining a diverse set of combination therapies and eligible cohorts (Oncology, 2019). Clinical trials are costly and time-consuming (Spall et al., 2007); in the US, the average cost of an RCT ranges from $11.5 million (in dermatology) to $52.9 million (in pain and anesthesia) (Sertkaya et al., 2016). With RWE, researchers can assess treatment effects across a broader set of parameters and use data-driven insights to guide the planning of future studies or inform medical practice. RWE-based analysis could uncover disparate treatment outcomes, prompting further investigation.

Analysis of treatment effects using RWE is complicated by the difficulties of parsing EHR data, complexities introduced by observational data, and limitations of conventional analysis methods. The predominant approach for estimating treatment effects from trials is the *Cox proportional hazards model (Cox PH)* (Cox, 1972), which computes a *hazard ratio (HR)*, interpreted as the relative risk of hazard between compared treatments. Controversy surrounds the interpretation of this quantity as causal, due to often violated required assumptions and built-in biases (Hernán, 2010; Martinussen et al., 2020). Real-world data also introduces biases such as confounding, necessitating adjustments to obtain causal estimates. EHR data often takes the form of unstructured text, which is challenging to parse into structured data (Tayefi et al., 2021). Previous efforts to utilize RWE to investigate trial outcomes have relied on manually curated structured databases, constructed

via costly labor-intensive labeling of unstructured EHR data by experts (Liu et al., 2021). Advances in large language models (LLMs), particularly for biomedical data, may enable accurate curation of unstructured clinical data. We build on the trial emulation framework, trialscope (González et al., 2023), leveraging its automatic data curation process and extending its capabilities of analyzing treatment effects beyond trial settings.

In this work, we introduce trialscope-x and trialscope-xl, pipelines designed to investigate outcomes of novel treatments using RWE. RWE offers insights into treatment effects beyond trial settings from two key aspects: (1) heterogeneous effects on patients ineligible for trials, and (2) longitudinal effects extending beyond trial durations. trialscope-x facilitates trial emulation across varied durations, while trialscope-xl models time-varying treatment effects. Figure 1 summarizes our aims. Our contributions are as follows:

- We clearly detail the complex problem setting of estimating comparative treatment effects from RWE.
- We present ML pipelines, trialscope-x and trialscope-xl, for estimating treatment effects from raw EHR data. We include options for addressing both existing flaws in standard analysis methods and biases introduced by real-world data.
- We document challenges faced and decisions made when employing EHR data for analysis as compared to curated clinical trials data.
- We apply both pipelines to the emulation and further analysis of keynote-042, a large, two-arm randomized controlled trial investigating immunotherapy for the treatment non-small cell lung cancer. We investigate expanded participant eligibility criteria and treatment effects over time. We present compelling findings that demonstrate the value of using RWE to assess clinical trials outcomes.

While the complexities of RWE make it difficult to draw definitive conclusions on treatment effects, our pipelines can be used to triage areas of interest for further study. We demonstrate that RWE can uncover outcome disparities related to protected attributes (e.g., age) as well as contradictions with clinical trials findings that warrant further investigation.

Generalizable Insights about Machine Learning in the Context of Healthcare

In this study, we thoroughly detail the challenges involved in estimating treatment effects from RWE and the limitations of conventional analysis methods. This discussion serves as a roadmap for future researchers interested in this domain. Furthermore, we introduce ML pipelines tailored for the analysis of treatment effects using EHR data, which can be readily adapted for exploring other clinical trials outcomes. Insights gained from these analyses can inform hypotheses for future trial design and pinpoint areas warranting further scrutiny. We present preliminary results from the application of our pipelines to a large EHR dataset, showcasing the usefulness of further investigating patient outcomes using RWE. To our knowledge, ours is the first work that leverages RWE to gain further insights into treatments investigated in clinical trials, addressing widely acknowledged flaws in standard estimation methods, and examining time-varying treatment effects.

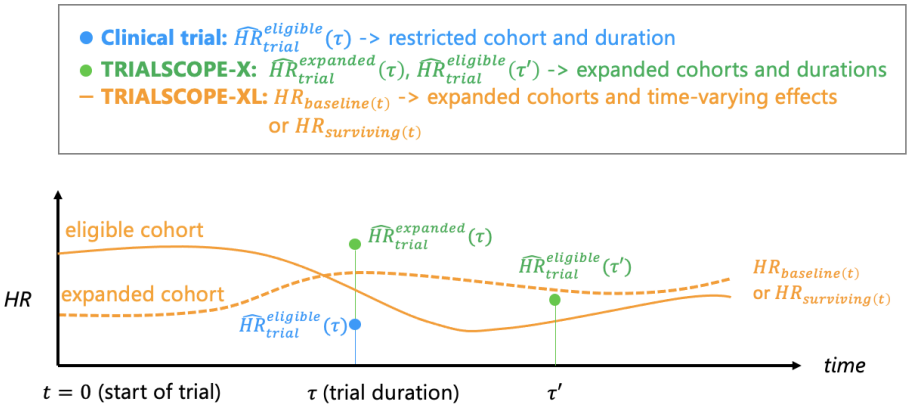


Figure 1: RCTs (blue) report a single HR, $\hat{HR}_{trial}(\tau)$, over the trial duration, τ. However, researchers may be interested in time-specific HRs. With RWE, TRIALSCOPE-X (green) can be used to estimate $\hat{HR}_{trial}(\tau)$ for different trial durations τ and cohorts. TRIALSCOPE-XL (orange) can be used to estimate time-varying treatment effects $HR_{baseline}$ or $HR_{surviving}$ over different cohorts.

2. Related Works

Clinical trials data is collected under strict experimental conditions. Observational data, collected in real-world practice, serves as a crucial resource for investigating causal treatment effects in the absence of RCTs (Hansford et al., 2023; Hernán et al., 2022), especially when RCTs are deemed ethically or logistically challenging. Hernán and Robins (2016) propose a widely-referenced framework advocating for the analysis of observational data through the lens of *target trial emulation*, where observational data is used to construct a hypothetical RCT for estimating causal effects. We adopt their perspective on target trial emulation. We extend the framework TRIALSCOPE, introduced by González et al. (2023), which utilizes automatically curated RWE from unstructured EHR for target trial emulation; results demonstrate its effectiveness in replicating outcomes of several non-small cell lung cancer trials. We introduce TRIALSCOPE-X and TRIALSCOPE-XL as direct expansions of TRIALSCOPE, aiming to explore patient outcomes beyond rigid trial settings and offer deeper insights into treatment effects dynamics over time. A related study by Liu et al. (2021) explores RWE for data-driven design of trial eligibility criteria, presenting a method for projecting trial outcomes with modified eligibility criteria based on estimated importance. However, our approach differs in our focus on accurately estimating treatment effects for both modified eligibility criteria and over time, an aspect typically overlooked in trial findings and not addressed by Liu et al. (2021). Furthermore, their reliance on manually curated datasets hinders scalability, contrasting with our automated ML data curation process.

3. Problem Setting

3.1. Clinical trials design

Clinical trials range from early-phase dose-finding studies to larger confirmatory trials evaluating efficacy, which are typically RCTs. RCTs are regarded as the gold standard strategy for determining the causal effect of an intervention on an outcome (Sibbald and Roland,

1998). By design, RCTs mitigate possible biases – distortions of the truth due to some systematic error. Individuals meeting specified *eligibility criteria* (e.g. diagnostic status) are randomly assigned to treatment (investigated intervention) or control (usually standard-of-care or placebo) groups (Kendall, 2003). Eligibility criteria are crucial for participant safety (by excluding those who face unacceptably high risk) and analysis of treatment efficacy (by reducing possible confounding) (Kim et al., 2015). However, there is widespread concern of overly restrictive criteria, which impacts the generalizability of results and limits patient participation, leading to failed trials (Jin et al., 2017; Kim et al., 2017). The common practice of adopting eligibility criteria from previous trials, sometimes arbitrarily, may exacerbate this issue (Kim et al., 2015). These concerns have prompted calls for modernizing and broadening eligibility criteria through data-driven approaches (ASCO et al., 2011).

3.2. Analysis of treatment effects

3.2.1. Estimands of interest

Clinical researchers seek to compare the relative impact of treatments on health outcomes, such as disease progression or mortality. Estimands of interest include survival models characterizing treatment event processes and causal contrasts comparing the relative effects of treatments. These include the survival function, $S(t|a) = \mathbb{P}(T > t|A = a)$, representing the probability of survival past time t, and the hazard function, $h(t|a)$, representing the instantaneous event rate at time t conditioned on survival until t: $h(t|a) = \lim_{dt\to 0} \frac{\mathbb{P}(t\leq T<t+dt|T\geq t, A=a)}{dt}$. $S(t)$ and $h(t)$ are defined in a continuous-time setting. T represents the event (outcome) time. The widely-used causal contrast *hazard ratio (HR)* compares hazard functions of the investigated treatment, $a = 1$ and the control, $a = 0$, such that $HR(t) = \frac{h(t|a=1)}{h(t|a=0)}$. Clinical trials typically report average treatment effects (ATE), representing contrasts in patient outcomes over the entire trial population. However, interest is growing in conditional average treatment effects (CATE), where treatment effects are stratified by patient covariates, x. CATE are crucial when covariates act as *effect modifiers*, affecting treatment outcomes differently across heterogeneous populations. The conditional HR is defined as $HR(t|a, x) = \frac{h(t|a=1,x)}{h(t|a=0,x)}$.

3.2.2. Current practices and challenges

Currently, the most widely used method for computing HRs is the *Cox proportional hazard model (Cox PH)* (Cox, 1972), a semi-parametric model that assumes a hazard function of the form $h(t|a) = h_0(t)\exp(\beta \cdot a)$. The Cox PH assumes that the baseline hazard, $h_0(t)$, remains consistent across treatment groups, with treatment-specific hazard functions differing only by a constant scaling factor, $\exp(\beta \cdot a)$. This *proportional hazards assumption* requires that treatment-specific hazard functions are proportional and that treatment coefficients β remain constant over time. Estimation of the Cox PH involves a partial likelihood assuming *censoring at random*, where censoring is independent of treatment assignment. Though popular due to the straightforward definition of the hazard ratio ($HR = \exp(\beta)$), the Cox PH suffers from limitations, including its reliance on the proportional hazards assumption.

The proportional hazards assumption is often violated, particularly when treatments have varied effects over time, yet many clinical studies do not verify this assumption (Bellera et al., 2010). For example, treatments investigated in KEYNOTE-042, which we adopt as a

case study, exhibit non-proportional hazards, evidenced by crossing survival curves (Mok and et al., 2019). While the HR should represent a time-dependent quantity, the Cox PH yields a constant HR, which reflects a weighted average of the HRs over the investigated period (Stensrud and Hernán, 2020), potentially obscuring time-varying treatment effects. The HR given by the Cox PH is also subject to an inherent selection bias, known as *survivorship bias* (Hernán, 2010; Hernán et al., 2016a; Martinussen, 2021), occurs in both RCTs and RWE. Causal treatment effect estimation relies on *exchangeability*, where the counterfactual risk (of some outcome) is the same across comparison populations. However, if treatment affects outcome, the distributions of the surviving treatment-specific populations deviate over time, rendering the two groups *non-exchangeable* with each other and with the baseline population (Hernan and Robins, 2023). In such cases, treatment effect contrasts, like HRs derived from the Cox PH, cannot be considered causal. To our knowledge, there is limited research in methods that address survivorship bias in HR estimation.

3.2.3. Biases introduced by real world data

Observational data introduces additional biases that can disrupt exchangeability between treatment groups and obscure true treatment effects. These biases also cause covariate shifts impacting survival model estimation. *Confounding* is when treatment assignment and patient outcome share a common cause, widespread in real-world data as patient health characteristics are likely to influence treatment. For example, sicker patients may be given riskier, experimental treatments. In RCTs, this effect is mitigated with treatment assignment randomization that prevents patient characteristics from influencing treatment assignment. *Immortal time bias* can occur in observational studies when analysis includes intervals of time where outcomes cannot occur (Suissa, 2008; Lévesque et al., 2010). This issue emerges when analysis is not carefully framed with respect to treatment start times, treatment assignment, and eligibility criteria (Hernán et al., 2016b). *Selection bias* occurs when the analysis population is conditioned on a common cause (or effect) of both treatment and outcome (Hernan and Robins, 2023). This includes *informative censoring*, where patient covariates may affect both presence of censoring (due to loss to follow-up or deviation from treatment strategy) and outcome. For example, sicker patients may be assigned to riskier treatments with higher likelihood of adverse effects, which may then cause patients to drop out of treatment. Analysis that includes only uncensored patients may be biased. *Missing data bias* is a form of selection bias that occurs when individuals with missing data are excluded from analysis (Hernan and Robins, 2023); for example, individuals in the EHR may be missing information on relevant biomarkers which are not commonly measured in practice, but used in experimental settings. Biases introduce difficulties with identifying appropriate treatment cohorts and modelling treatment effects and therefore require careful consideration during analysis.

4. Methods

4.1. Defining causal contrasts

Despite known flaws in conventional HR estimates, we prioritize estimation of HRs from RWE to enable comparison to existing clinical trial outcomes. Using trialscope-x, we can

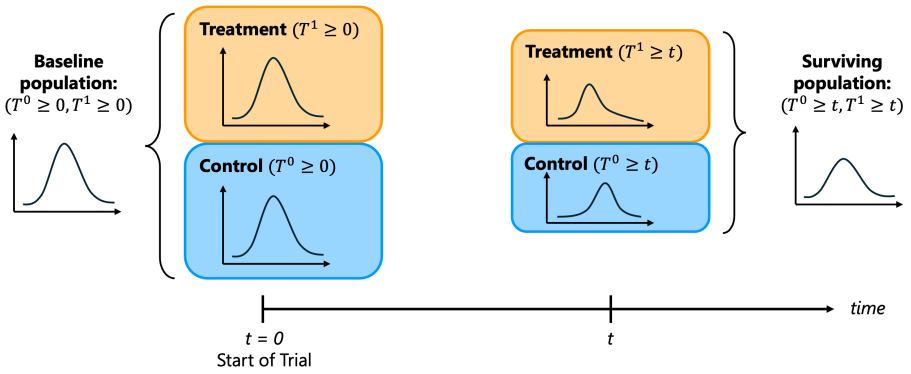


Figure 2: We depict the consequences of HTEs in a RCT: (1) group covariate distributions shift differently by treatment, (2) proportion of survival, which depends on both covariates and assigned treatment, differ by treatment. This results in non-comparable treatment and control groups due to lack of exchangeability (3.2.2). In this example, covariates are effect modifiers on treatment. The hazard rate $h(t|a, x)$ is conditional on treatment a and covariates x, determining survival. The plots show covariate distributions: the x-axis represents some covariate x that affects survival, the y-axis shows the proportion of the surviving population. At the start of the trial, the baseline populations are distributed identically in each treatment group. By time t, the covariate distributions differ due to HTEs. The colored boxes represent the overall number of surviving patients. At the start of a trial, the treatment and control groups are assigned an equal number of patients. In the scenario depicted, by time t, the control group has fewer survivors. T^a is the time-to-event of the patient assigned to treatment a (3.2.1). The *baseline population*, representing the covariate distribution of patients at the outset of a trial, is used in the estimate of $HR_{baseline}(t)$, while $HR_{surviving}(t)$ is conditioned on the *surviving population* at timepoint t.

compute HRs consistent with clinical trial definitions ($\hat{HR}_{trial}$). TRIALSCOPE-XL addresses the limitations of $\hat{HR}_{trial}$, providing causally interpretable HRs, as defined below. T^a refers to the event time observed if the patient is given treatment a. Each HR definition is conditioned on a distinct population group, illustrated in Figure 2:

- **HR as estimated in trials:** Typically, RCTs report a single HR derived from the Cox PH model, representing the entire trial duration. However, this estimate is only valid if treatment effects remain constant over time and adhere to the proportional hazards assumption. The intended HR estimate is defined as follows:

$$HR_{trial}(t) = \frac{\lim_{dt \to 0} \mathbb{P}(t \leq T^1 < t + dt | T^1 \geq t)}{\lim_{dt \to 0} \mathbb{P}(t \leq T^0 < t + dt | T^0 \geq t)} \tag{1}$$

  However, the Cox PH yields a constant, $\hat{HR}_{trial}(\tau) = \exp(\beta)$, where β represented estimated coefficients. If treatment effects are time-varying, but not modified by covariates, $\hat{HR}_{trial}(\tau)$ is a weighted average of time-specific HRs over the trial period rather than an instantaneous HR. We use the input τ, denoting trial duration, distinguishing

it from t, denoting time point. If treatment effects are time-varying based on covariate effect modifiers, $\hat{HR}_{trial}(\tau)$ still represents a weighted average of time-specific HRs, but these HRs also lack causal interpretation due to the non-exchangeability between surviving treatment-specific populations ($T^0 \geq t$ and $T^1 \geq t$). Using TRIALSCOPE-X, we emulate standard trial conditions and estimate a single HR (over trial duration), also providing the option to extend this emulation to different durations τ.

- **HR with respect to surviving population:** We introduce a modified definition of the HR, also referred to as the *causal hazard ratio* in previous literature (Martinussen, 2021). This quantity is similar to the proposed *survival average causal effect (SACE)* (Rubin, 2006; Tchetgen, 2014), which estimates treatment effects conditioned on the population of "always-survivors," individuals expected to survive irrespective of treatment received. $HR_{surviving}(t)$ is also conditioned on the population expected to survive under both treatments, denoted as $(T^0 \geq t, T^1 \geq t)$.

$$HR_{surviving}(t) = \frac{\lim_{dt \to 0} \mathbb{P}(t \leq T^1 < t + dt | T^0 \geq t, T^1 \geq t)}{\lim_{dt \to 0} \mathbb{P}(t \leq T^0 < t + dt | T^0 \geq t, T^1 \geq t)} \tag{2}$$

  This quantity is difficult to estimate as it relies on the satisfaction of strong identifiability assumptions, namely that all covariates that impact patient outcomes are observed and measured (Martinussen, 2021; Tchetgen, 2014).

- **HR with respect to baseline population:** We also present a modified definition of the HR conditioned on the baseline population, regardless of survival. This adjustment aims to establish comparable treatment groups. However, estimation challenges arise due to the absence of data for deceased or censored patients. This quantity represents the HR at time t for the baseline population at the trial's outset, while $HR_{surviving}(t)$ reflects the HR at time t for individuals expected to survive until that point.

$$HR_{baseline}(t) = \frac{\lim_{dt \to 0} \mathbb{P}(t \leq T^1 < t + dt | T^0 \geq 0, T^1 \geq 0)}{\lim_{dt \to 0} \mathbb{P}(t \leq T^0 < t + dt | T^0 \geq 0, T^1 \geq 0)} \tag{3}$$

4.2. TRIALSCOPE-X: Clinical trial emulation and extension

With the TRIALSCOPE-X pipeline (Figure 3), we emulate clinical trial outcomes by estimating $\hat{HR}_{trial}(\tau)$, where τ represents trial duration. We can explore extended outcomes by varying trial durations τ and investigating varied sets of eligibility criteria. TRIALSCOPE-X provides initial validation for our RWE approach by enabling direct comparison with published trial results. This is particularly helpful when long-term outcomes are reported, such as in KEYNOTE-042, which published a five-year extension. We can also investigate outcomes for patients ineligible for trials but treated in practice. TRIALSCOPE-X places us otherwise in the same context as a clinical trial, allowing us to answer the following question: what might the outcome have been if these patients were included in the trial?

1. **Data processing.** We automate the curation of structured and unstructured textual data, including scanned notes, from the EHR dataset, using LLMs tailored for biomedical language processing (González et al., 2023; Preston et al., 2023). This process

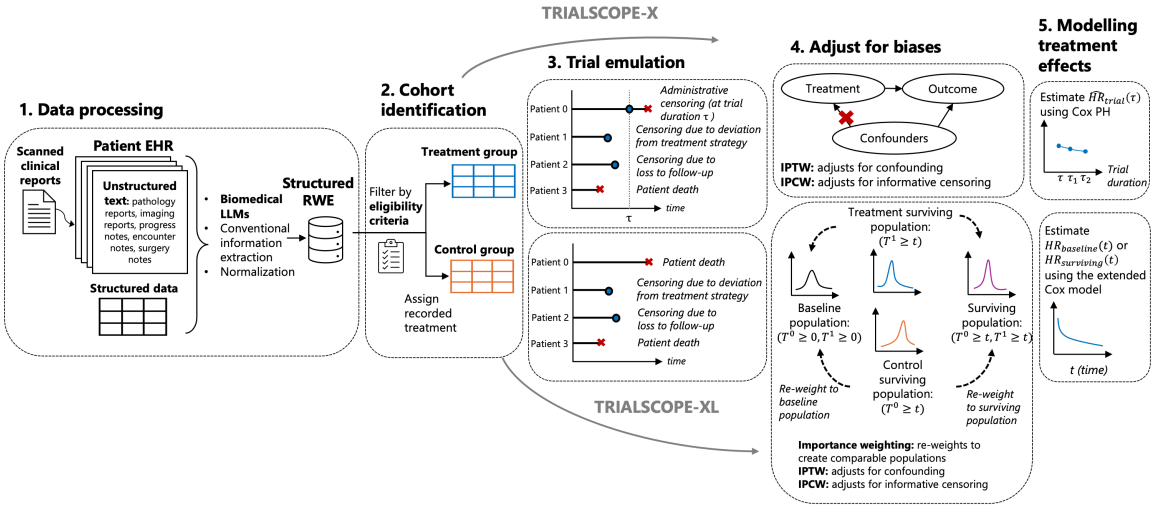


Figure 3: The TRIALSCOPE-X pipeline is used for estimation of the HR as typically reported by clinical trials, $\hat{HR}_{trial}$, and the TRIALSCOPE-XL pipeline is used for estimation of the time-varying HRs $HR_{baseline}$ or $HR_{surviving}$. Pipelines diverge at step 3.

converts relevant patient covariates such as demographics, biomarkers, gene variants, lab tests, as well as diagnoses, treatment information, and dates, into structured data. By leveraging ML models, our pipeline is scale-ble to large EHR datasets, eliminating the need for manual annotation. Further details are provided in Section 5.1.

2. **Cohort identification.** After curating and structuring the EHR data, we establish treatment cohorts according to the target treatments and eligibility criteria. We must carefully account for satisfaction of eligibility criteria while allowing for flexibility to due variations in observational data. This involves various challenges, such as the identification of prior and concurrent treatments, managing combination therapies, and determining associated health data, biomarkers, and diagnoses based on time stamps. For an in-depth discussion, refer to Section 5.3.

3. **Trial emulation.** We use the extracted observational treatment and control cohorts to emulate a hypothetical clinical trial. In experiments, we emulate documented trials, exploring modifications to eligibility criteria and trial duration. Participants meeting eligibility criteria and receiving the treatments are selected. In practice, patients may deviate from the investigated treatment strategy and receive conflicting treatments that can bias study results. Because the exclusion of such patients can introduce selection bias (Hernán et al., 2016a), we include all patients who meet the eligibility criteria at their identified treatment start time. Those who deviate from the assigned treatment strategy or are lost to follow-up are censored. Patient follow-up may extend indefinitely in real-world data, but for emulation purposes, outcomes beyond a specified trial duration, τ, are censored to prevent immortal time bias (Section 3.2.3). With RWE, we can analyze outcomes across various hypothesized trial durations, including extensions beyond the original trial duration.

4. **Adjust for biases.** Real world data may be subject to additional biases (Section 3.2.3). We detail our approach to mitigating biases in Section 4.2.1.

5. **Modelling treatment effects.** For each investigated target trial duration, we train a Cox PH model, incorporating IP weights from the previous step, to estimate the HR over the trial duration, defined in Section 4.1 as $\hat{HR}_{trial}(\tau)$. In TRIALSCOPE-X, we aim to emulate published trials and so adopt the standard methodology (Cox PH), accounting only for real world biases and modifying only eligibility criteria and trial duration in experiments.

4.2.1. Adjusting for biases from real-world evidence

With TRIALSCOPE-X, we aim to emulate standard clinical trial conditions and extend analyses over different (1) trial durations and (2) participant eligibility criteria. Therefore, we address the biases arising from real-world evidence (Section 3.2.3), but do not yet address the limitations of conventional practices (Section 3.2.2). While challenges of RWE are mitigated through careful data processing and filtering decisions (Section 5.3), and others via a principled trial emulation procedure (Section 4.2), the issues of confounding and censoring bias require modelling adjustments. We employ inverse probability of treatment weighting (IPTW) and inverse probability of censoring weighting (IPCW) to create pseudo-populations where confounders no longer impact treatment assignment and covariates no longer impact censoring, as detailed in Appendix B.

4.3. TRIALSCOPE-XL: Estimating treatment effects over time

We introduce TRIALSCOPE-XL (Figure 3), which can be used to estimate causal time-varying treatment effects. With TRIALSCOPE-X, we replicate trial findings through emulation at original trial duration, allowing us to validate our approach against reported trial outcomes. We can also extend analyses to differing trial durations and eligibility criteria. However, this approach faces two major flaws as discussed in Section 3.2.2: (1) restrictive Cox PH assumptions and (2) survivorship bias in HR estimation ($\hat{HR}_{trial}$). Steps 3-5 of TRIALSCOPE-XL differ from TRIALSCOPE-X and addresses these flaws. With TRIALSCOPE-XL, we can estimate $HR_{surviving}$ and $HR_{baseline}$, defined in Section 4.1:

3. **Trial emulation.** As in TRIALSCOPE-X, we censor patients who deviate from the assigned treatment strategy or are lost to follow-up. However, unlike TRIALSCOPE-X, here we do not perform administrative censoring to emulate the conclusion of a trial. We aim to properly model treatment effects over time and therefore include all recorded outcomes into the model.

4. **Adjust for biases.** Previously, we used IP weighting to address biases from real-world data by creating a pseudo-population in which covariates do not affect treatment assignment or censoring (Hernan and Robins, 2023). However, this does not account for survivorship bias; the two treatment groups are not exchangeable in the presence of HTEs, so the estimated HR is not causally valid. Survivorship bias occurs when the identifiability assumption of *conditional exchangeability* is violated over time as

treatment-specific population distributions deviate from each other and from the baseline. We adjust for this issue in TRIALSCOPE-XL, where we use importance weighting to create comparable treatment groups, accounting for covariate shift over time. Importance weighting (different from IP weighting) is used to re-weight a distribution to match a target distribution (Shimodaira, 2000). Figure 2 depicts two options: $HR_{surviving}$, where the target population is the pool of survivors of either treatment (also known as the *risk set*), and $HR_{baseline}$, where the target population is the initial baseline population. In practice, particularly for severe diseases, estimation of $HR_{surviving}$ is more challenging as the surviving population dwindles over time. Calculation of $HR_{surviving}$ and $HR_{baseline}$ incorporates importance weights alongside IP weights to account for confounding and selection biases. As discussed in Section 4.1, both definitions offer useful contextualizations of the HR. Note that valid causal estimates require the additional assumption of *no unmeasured effect modifiers*, where all patient covariates that impact outcomes are observed.

5. **Modelling (time-varying) treatment effects.** As discussed in previous sections, the Cox PH is flawed due to its requirement of the proportional hazards assumptions as well as its misinterpretation as a time-dependent quantity, when in fact it results in a constant HR which is a weighted average of time-dependent HRs over the trial period (Hernán, 2010). We instead adopt the *extended Cox model* (also known as the *time-varying Cox model*) in order to estimate truly time-specific hazard ratios (Tian et al., 2005). Further details are provided in Section 4.3.1.

4.3.1. The extended Cox model

The Cox PH assumes proportional hazards over time, differing only by a constant-time scaling factor, β (Section 3.2.2). However, we aim to address time-varying treatment effects where the relative impact of treatments change over time. One approach is to incorporate time-dependent regression coefficients into the Cox model, such that $h(t|a) = h_0(t)\exp(g(\beta,t)\cdot a)$, where $g(\beta,t)$ is a specified continuous function of time and β is a vector of coefficients (Tian et al., 2005; Thomas and Reyes, 2014). We can model $g(\beta,t)$ with a simple time function, such that $g(\beta,t) = \beta\cdot g(t)$, so that the hazard function can be factored into $h(t|a) = h_0(t)\exp(\beta\cdot g(t)\cdot a) = h_0(t)\exp(\beta\cdot a(t))$ where $a(t) = g(t)\cdot a$, so that the problem of time-varying coefficients can be converted to one of time-varying covariates (Zhang et al., 2018). A common time function is a simple logarithmic function, $g(t) = 1+\log(t)$. The Cox likelihood can be generalized to accommodate for time-varying covariates during inference (Cai and Sun, 2003; Tian et al., 2005; Thomas and Reyes, 2014). In practice, we generate time-varying covariates $a(t) = g(t)\cdot a$ by applying $g(t)$ to the long format of our dataset, which contains an entry per sample, per timepoint.

5. Data

5.1. EHR Dataset

We apply our clinical trials analysis pipelines to RWE derived from a Providence Health EHR dataset. Providence Health & Services is a major health care system operating across several states in the United States. This work was performed under the guidance of

an Institutional Review Board (IRB)-approved research protocol (Providence protocol ID 2019000204) and was conducted in compliance with Human Subjects research and clinical data management procedures–as well as cloud information security policies and controls–administered within Providence Health. All study data were integrated, managed and analyzed exclusively and solely on Providence-managed cloud infrastructure. All study personnel completed and were credentialed in training modules covering Human Subjects research, use of clinical data in research, and appropriate use of IT resources and IRB-approved data assets. The dataset used consists of EHR data from around 3.3 million patients, including about 1 million cancer patients. Electronic medical records are processed through the TRIALSCOPE pipeline (González et al., 2023), which processes scanned reports and unstructured notes and then structured clinical text using a combination of biomedical language models and conventional information extraction systems. The TRIALSCOPE pipeline extracts relevant patient covariates (including demographic, medical, and genetic data in some cases) alongside diagnoses, diagnoses date, treatments, and treatment dates and exhibits high accuracy in extracting relevant attributes (González et al., 2023).

5.2. Target trial selection

Following González et al. (2023), we focus on completed phase III trials for non-small cell lung cancer (NSCLC). KEYNOTE studies offer the largest RWE-extracted cohorts due to widespread use of the immunotherapy drug pembrolizumab (brand name Keytruda). PD-L1 expression is considered a biomarker for pembrolizumab efficacy and is measured by tumor proportion score (TPS), categorized by high ($TPS \geq 50\%$), low ($50\% > TPS \geq 1\%$), positive ($TPS \geq 1\%$), and negative ($TPS < 1\%$) expression. Previous studies indicate that higher PD-L1 expression correlates with increased treatment benefit from pembrolizumab (Mok and et al., 2019), though these findings are controversial, with other studies reporting conflicting results (Zhao et al., 2018; Xu et al., 2019). We focus on emulation and extension of KEYNOTE-042 (Mok and et al., 2019), which investigates pembrolizumab as a monotherapy for previously untreated NSCLC patients and reports additional 5-year trial outcomes (de Castro Jr et al., 2023). Appendix A.1 details our trial selection process.

5.3. Cohort selection

KEYNOTE-042 is a randomized phase III clinical trial comparing first-line pembrolizumab monotherapy with standard-of-care chemotherapy for *untreated*, metastatic NSCLC patients with a PD-L1 TPS of $\geq 1\%$ (Mok and et al., 2019). The chemotherapy control is platinum-based doublet chemotherapy, which is a combination therapy consisting of the platinum-based drug Carboplatin plus either Paclitaxel or Pemetrexed (in this trial). Trial eligibility includes adults ($\geq$ 18 years) without sensitizing EGFR mutation or ALK translocation, an Eastern Cooperative Oncology Group (ECOG) score of 0 or 1, and life expectancy $\geq$ 3 months. We construct an *eligible* cohort, where we include all original trial eligibility criteria to the best of our ability, and an *expanded* cohort, where many of the criteria are removed. Table 2 shows the eligibility criteria used. Outcomes are reported by PD-L1 strata; we employ an additional stratum consisting of patients with no explicit negative PD-L1 expression (not-negative). Summary statistics for the high and positive PD-L1 strata of the *eligible* and *expanded* cohorts are shown respectively in Table 1 and 5, with low

Table 1: Baseline characteristics of *eligible* RWE cohort with PD-L1 TPS strata of high ($\geq 50\%$) and positive ($\geq 1\%$, includes high *and* low).

	Pembrolizumab		**Chemotherapy**	
PD-L1 TPS	High (n=151)	Pos (n=204)	High (n=58)	Pos (n=126)
Age (years)	70.3 ± 10.4	71.1 ± 10.4	65.9 ± 9.2	66.4 ± 9.3
<65	52 (34%)	65 (32%)	29 (50%)	60 (48%)
Male	73 (48%)	104 (51%)	29 (50%)	55 (44%)
Female	78 (52%)	100 (49%)	29 (50%)	71 (56%)
Race/ethnic group				
Asian	10 (7%)	15 (7%)	3 (5%)	6 (5%)
White or Caucasian	127 (84%)	169 (83%)	47 (81%)	107 (85%)
Other	14 (9%)	20 (10%)	8 (14%)	13 (10%)
ECOG score				
0	28 (19%)	31 (15%)	17 (29%)	39 (31%)
1	123 (81%)	173 (85%)	41 (71%)	87 (69%)
Smoking status				
Current/Former	91 (60%)	123 (60%)	44 (76%)	98 (78%)
Never	60 (40%)	81 (40%)	14 (24%)	28 (22%)
Tumor hist. features				
Squamous	29 (19%)	38 (19%)	9 (16%)	27 (21%)
Non-squamous	122 (81%)	166 (81%)	49 (84%)	99 (79%)
Disease status				
Locally advanced (III)	31 (21%)	42 (21%)	19 (33%)	45 (36%)
Metastatic (IV)	120 (79%)	162 (79%)	39 (67%)	81 (64%)

and not-negative PD-L1 strata respectively in Table 4 and 6. Kaplan-Meier survival curves for the PD-L1 positive strata of both eligible and expanded cohorts are in Figure 10; the crossing survival curves indicate non-proportional hazards, also seen in the original trial.

5.3.1. Identifying treatment cohorts

In KEYNOTE-042 (Mok and et al., 2019), pembrolizumab is administered as a monotherapy, while the control, platinum-based doublet chemotherapy, is a combination treatment consisting of two drugs. For the eligible cohort, we aim to closely match the record trial eligibility criteria with some flexibility for RWE. It is particularly difficult to ensure that therapies are first-line, to identify combination treatments, and to identify conflicting concurrent treatments. Our data processing pipeline extracts treatment start dates, but the EHR may lack treatment end dates. Combination therapies are identified if two qualifying drugs have treatment start dates that overlap at least 2 weeks (allowing for discrepancies in the genuine start treatment start date versus prescription date). In KEYNOTE-042, both treatments are first-line therapies; individuals must not have received prior treatment for NSCLC. Individuals are included if no conflicting treatments start before the investigated treatment. Conflicting treatments include previous chemotherapy or immunotherapy (fur-

ther details in Table 2). Some individuals may be administered a conflicting treatment after the start of the investigated treatment and are therefore considered to deviate from the assigned treatment strategy. These individuals are included in analysis but censored at the time of treatment deviation to avoid selection bias (Hernán et al., 2016b).

5.3.2. Missing data

KEYNOTE-042 requires that participants have positive (low or high) PD-L1 expression. However, we extract significantly fewer control chemotherapy patients with a recorded positive PD-L1 TPS from the EHR dataset; the vast majority of otherwise eligible chemotherapy patients lack PD-L1 readings (see Table 3 for sample sizes by PD-L1 strata). The presence (or lack thereof) of a PD-L1 reading may be informative; PD-L1 expression is viewed as a predictive biomarker for the efficacy of immunotherapy and current treatment guidelines for pembrolizumab monotherapy require that patients have positive PD-L1 expression (U.S. Food and Drug Administration; European Medicines Agency). There may be underlying factors affecting the patients who have positive PD-L1 readings but do not receive any form of immunotherapy; for example, a clinician could decide a patient has other risk factors that still preclude them from immunotherapy. Strictly filtering chemotherapy patients based on positive PD-L1 may introduce selection bias, while not filtering may include patients with negative PD-L1, which also affects analysis. Although PD-L1 expression is not thought to be predictive of chemotherapy efficacy, it may correlate with patient prognosis (Pawelczyk et al., 2019). In addition, a number of (otherwise) eligible pembrolizumab patients in the EHR are missing PD-L1 values despite positive expression being a treatment requirement, suggesting possible abnormality in their situation. Given these complications, we experiment with cohorts including those with explicit positive PD-L1 (Sections 6.1-6.4) and cohorts including all (otherwise) eligible patients without recorded negative or indeterminate PD-L1 (Section 6.5). Tables 4 and 6 show baseline summary statistics for the PD-L1 *not-negative* stratum of the eligible and expanded cohorts respectively.

6. Experiments

In our experiments, we apply TRIALSCOPE-X and TRIALSCOPE-XL to a large EHR dataset from Providence Health to explore the treatment benefit of pembrolizumab, as in KEYNOTE-042. We first assess the efficacy of TRIALSCOPE-X in replication of KEYNOTE-042 (Mok and et al., 2019), including five-year outcomes (de Castro Jr et al., 2023). We then explore variations in eligibility criteria and modelling of time-varying treatment effects. The eligibility criteria used for the *eligible* and *expanded* cohorts are detailed in Table 2.

6.1. Target trial emulation: validation

Using TRIALSCOPE-X, we first replicate KEYNOTE-042, which had an initial duration of 750 days (Mok and et al., 2019), and later reported 5-year outcomes (de Castro Jr et al., 2023). We maintain the original trial eligibility criteria, constructing an *eligible cohort*, and adjust for biases from observational data (Section 4.2.1). Results of this target trial emulation are shown in Figure 4: blue points (●) represent the emulation HR, $\hat{HR}_{trial}$, for different trial durations τ. The emulation closely matches reported results at $\tau = 750$ and $\tau = 1825$ (5

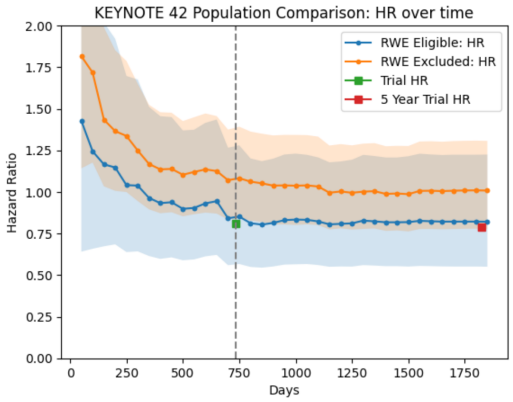


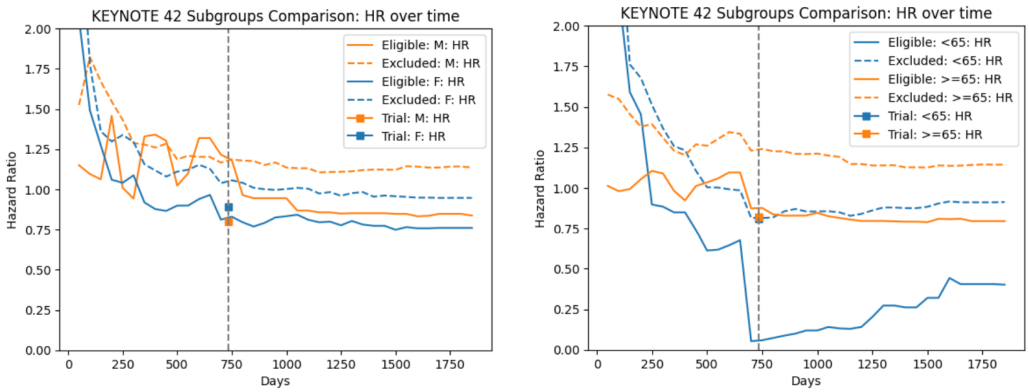
Figure 4: Target trial emulation over possible trial durations, τ, of the *eligible* cohort (blue points ●), and *expanded* cohort (orange points ●). The green square (■) is the reported HR from KEYNOTE-042 and the red square (■) is the reported 5-year HR. RWE emulation of the eligible cohort almost exactly matches reported HRs.

years), demonstrating that TRIALSCOPE-X can accurately emulate reported trial outcomes with RWE. We can also estimate hazard ratios over various trial durations τ.

We expand our analysis to also include individuals excluded from the original trial due to strict eligibility criteria. Table 2 outlines the criteria for this *expanded* cohort, which maintains the requirements for diagnosis, age, PD-L1 TPS, and conflicting concurrent treatments. Figure 4 displays HRs for trial emulations over different trial durations for the expanded cohort with orange points (●). The HRs of the expanded cohort consistently exceed those of the eligible cohort across all trial durations. At the original trial duration, $\tau = 750$, the excluded cohort HR is > 1, indicating better outcomes in the control chemotherapy cohort than the pembrolizumab cohort. This outcome may stem from explainable factors, such as unmeasured confounders in the expanded cohort that result in worse pembrolizumab outcomes. However, it may also indicate potential health inequities: pembrolizumab may truly be less effective than chemotherapy over a more representative patient population. Restrictive eligibility criteria is known to limit generalizability (Kim et al., 2015), in particular through the exclusion of patients with certain risk factors or comorbid conditions due to safety concerns (Jin et al., 2017). Consequently, trial results are limited to generally healthier patients who may not be representative of the broader patient population, who are still given these treatments. This discrepancy underscores the importance of further investigation into treatment effectiveness across diverse patient groups, despite the challenges of adopting real-world evidence.

6.2. Target trial emulation: implications for health equity

We estimate conditional HRs by population subgroups to investigate possible disparities between the eligible and expanded populations. In Figure 5(a), we stratify cohorts into male/female groups, observing that the expanded cohort males experience worse outcomes, with a large discrepancy at longer trial durations, than observed in females. The artifacts in the eligible, male estimate may be due to lack of data, or could indicate time-varying treatment effects. Over a longer duration of observation, while pembrolizumab outcomes (as compared to chemotherapy) for the expanded population are worse for both males and



(*a*) Stratification by female (blue) and male (orange). KEYNOTE-042 report HRs for female (blue square ■) and male (orange square ■) groups.

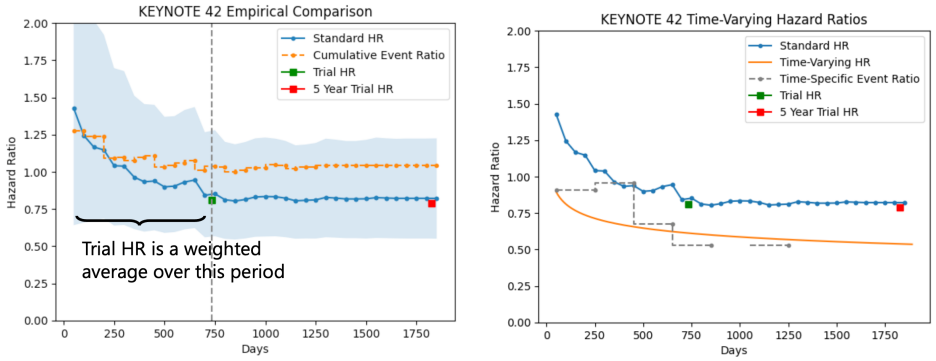
(*b*) Stratification by age groups: < 65 (blue) and ≥ 65 (orange). KEYNOTE-042 repot HRs for < 65 (blue square ■) and ≥ 65 (orange square ■).

Figure 5: $\hat{HR}_{trial}$ estimates over trial durations for the eligible cohort (—) and the expanded cohort (- -), stratified by covariate subgroups.

females, there is a much larger gap in the male group. While these results may stem from unmeasured confounders or analysis flaws, they may also indicate a health disparity. Exclusion from trials risks harm to any population segment. Figure 5(b) shows age-based stratification, indicating a significant disparity between age groups (comparing the blue to lines), where the < 65 HR is much lower, and a large disparity between the eligible/expanded cohorts by age group (comparing the solid and dashed lines). This contrasts with KEYNOTE-042 findings (Mok and et al., 2019), where the reported subgroup HRs are near identical, suggesting age-based outcome differences in real-world evidence, regardless of eligibility.

6.3. Uncovering time-varying treatment effects

In previous sections, we use TRIALSCOPE-X to estimate $\hat{HR}_{trial}$ for trial emulation and extension to different trial durations and eligibility criteria. However, the Cox PH represents a weighted average of time-specific HRs over the trial duration τ and cannot model dynamic treatment effects (Section 3.2.2). Figure 6(a) illustrates that $\hat{HR}trial$ roughly approximates the empirical cumulative event ratio, as computed directly from the dataset. This ratio signifies the cumulative number of deaths in the risk set (surviving and uncensored population) up to the target trial duration, τ. Note: while $\hat{HR}trial$ adjusts for real-world biases (confounding and informative censoring), the raw empirical cumulative event ratio remains uncorrected for potential biases. Other drawbacks of the Cox PH include the proportional hazards assumption and inherent survivorship bias (Section 3.2.2). To address this, we propose TRIALSCOPE-XL, which employs the time-varying Cox model alongside IP weighting to estimate time-varying HRs relative to the baseline population. In our experiments, we calculate $HR_{baseline}$ as described in Section 4.1. Figure 6(b) shows the time-varying HR, $HR_{baseline}$, alongside the Cox PH estimate, $\hat{HR}_{trial}$, and bucketed, time-specific empirical event ratio estimates. Notably, $HR_{baseline}$ better aligns with the time-specific empirical event ratio estimates, which reflect the mortality rate within the risk set in 200-day in-



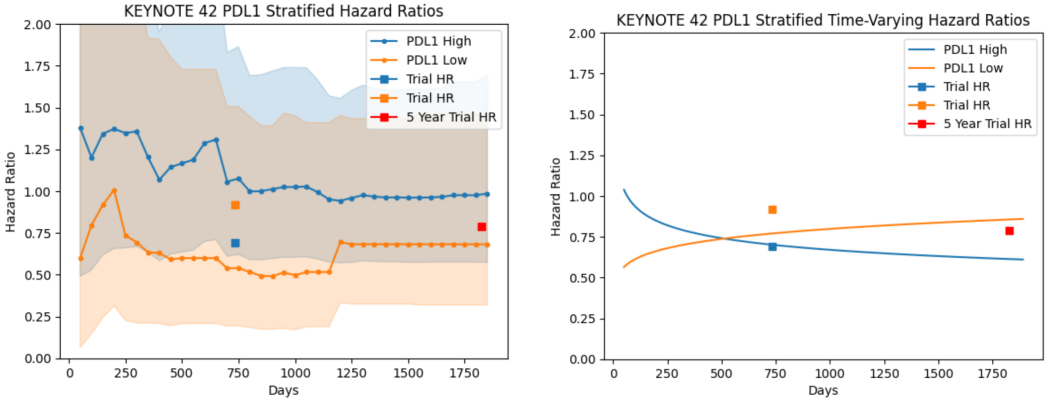
(a) (b)

Figure 6: *(a):* $\hat{HR}_{trial}(\tau)$ (blue points ●) estimated with Cox PH represents a weighted average of HRs over the trial duration τ, rather than a truly time-specific HR. $\hat{HR}_{trial}$ matches the pattern of the empirical cumulative event ratio (orange dashed line - -). *(b):* The time-varying $HR_{baseline}$ (orange solid line —) better fits the time-specific empirical event ratio (gray dashed line - -).

tervals. In contrast, $\hat{HR}_{trial}$ is more aligned with the empirical cumulative event ratios shown in Figure 6(a). Over time points, $HR_{baseline}$ estimates consistently remain lower than both the $\hat{HR}_{trial}$ estimates and the reported KEYNOTE-042 HR, suggesting fairly beneficial outcomes for pembrolizumab patients based on time-specific HRs. However, it's important to recognize that the time-specific $HR_{baseline}$ requires a distinct interpretation from $\hat{HR}_{trial}$; because clinical trials typically report $\hat{HR}_{trial}$ (representing weighted averages of time-specific HRs), the perception of relative treatment benefit may be tailored to this prevalent interpretation. We argue that $HR_{baseline}$ reflects a more accurate interpretation of a time-specific HR, despite $\hat{HR}_{trial}$ often being misconstrued as such (Martinussen et al., 2020). Nevertheless, this requires a shift in perspective.

6.4. Uncovering time-varying effects: stratified populations

We use TRIALSCOPE-X and TRIALSCOPE-XL to explore time-varying treatment effects in population subgroups. In KEYNOTE-042, patient outcomes are also reported by PD-L1 strata, high and low, finding that pembrolizumab is more effective for patients with high PD-L1 TPS, consistent with previous studies which suggest that tumor PD-L1 expression is linked with pembrolizumab treatment benefit (Sacher and Gandhi, 2016). However, our dataset reveals a curious result. Figure 7(a) shows that our estimates of $\hat{HR}_{trial}$ demonstrate an inverse effect: higher HRs for patients with high PD-L1 than low PD-L1. Using a time-varying Cox model to estimate $HR_{baseline}$ (Figure 7(b)), we mostly recover the expected outcomes of PD-L1 stratification, with an interesting insight: outcomes for low PD-L1 expression patients worsen slightly over time, while those for high PD-L1 expression patients improve, leading to better long-term outcomes. This suggests the possibility of time-varying treatment effects that may be obscured by standard methods, although unmeasured confounding could also explain this behavior. For example, patients with high PD-L1 expression in the control chemotherapy group (who, strangely, did not receive immunotherapy) may have had other risk factors. In the US and Europe, pembrolizumab is only approved as a first-line

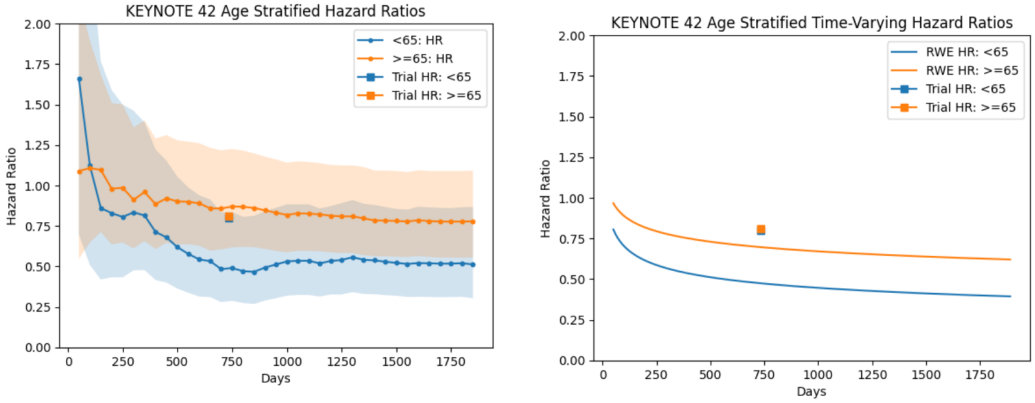


(a) $\hat{HR}_{trial}$ estimates over trial durations (b) $HR_{baseline}$ estimates over time

Figure 7: Conditional HRs by PD-L1 TPS strata of high (blue line —) and low (orange line —) show distinct time-varying patterns. KEYNOTE-042 reports HRs for PD-L1 high (blue square ■) and low (orange square ■).

monotherapy for late-stage NSCLC with positive PD-L1 expression (U.S. Food and Drug Administration; European Medicines Agency). Further investigation is warranted due to the controversy surrounding the association of PD-L1 expression levels and the efficacy of pembrolizumab, with some studies reporting treatment benefit even with negative PD-L1 expression (Zhao et al., 2018; Xu et al., 2019).

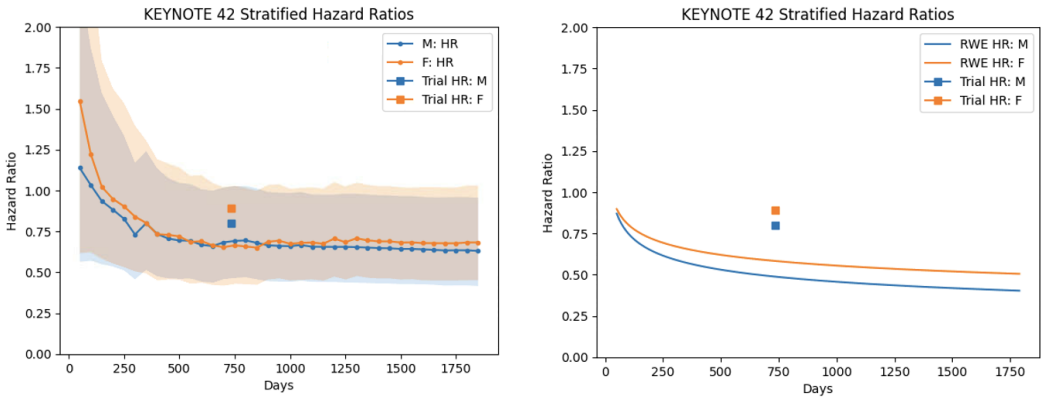
6.5. Uncovering time-varying effects: implications for health equity



(a) $\hat{HR}_{trial}$ estimates over trial durations (b) $HR_{baseline}$ estimates over time

Figure 8: Conditional HRs by age groups < 65 (blue line —) and ≥ 65 (orange line —) for the eligible cohort show a noticeable disparity in treatment outcomes. KEYNOTE-042 reports similar HRs for < 65 (blue square ■) and ≥ 65 (orange square ■).

With RWE, we uncover potential health inequities, especially through modelling of time-varying treatment effects. Figure 8 shows age-stratified treatment outcomes of the eligible population (including all PD-L1 not-negative): $\hat{HR}_{trial}$ estimates in Figure 8(a) and



(*a*) $\hat{HR}_{trial}$ estimates over trial durations (*b*) $HR_{baseline}$ estimates over time

Figure 9: Conditional HRs by male (blue —) and female (orange —) groups for the eligible cohort show no disparity in (a), but an increasing disparity over time in (b). KEYNOTE-042 reports HRs for males (blue ■) and females (orange ■).

time-varying $HR_{baseline}$ estimates in Figure 8(b). Both models reveal significant outcome disparities between age groups, contrasting with the trial where outcomes were nearly identical (Mok and et al., 2019). This suggests differences in real-world treatment by age that may not occur in controlled clinical settings, warranting further investigation. In Figure 9, we stratify the population by biological sex, observing similar values of $\hat{HR}_{trial}$ (Figure 9(a)) over male/female subgroups. However, we notice a slight discrepancy of $HR_{baseline}$ (Figure 9(b)), where time-varying HRs are higher in the female group, with this difference amplifying over time. This finding is more consistent with the reported KEYNOTE-042 HRs by biological sex, where the female group exhibited a slightly higher HR. It is possible that the time-varying HR model captures treatment effects that are obscured by the standard Cox model, even with corrections for real-world biases. When navigating RWE complexities, employing a better specified model may uncover relevant treatment effects.

7. Discussion

Our work demonstrates the utility of leveraging RWE through the TRIALSCOPE-X and TRIALSCOPE-XL pipelines to investigate treatment effects, using KEYNOTE-042 as a case study. By replicating trial conditions and adjusting for real-world biases, we successfully reproduce the reported HRs from KEYNOTE-042, both at the original trial duration and the extended 5-year follow-up. This fidelity underscores the potential of RWE to complement traditional clinical trial findings and provide valuable insights into treatment efficacy over extended durations. Additionally, we expose and address the limitations of conventional methods for the estimation of heterogeneous treatment effects over time, uncovering possible time-varying treatment effects. By expanding our analysis to include individuals excluded from the original trials, we reveal potential health disparities stemming from strict eligibility criteria. The expanded cohort consistently exhibited higher HRs as compared to the eligible cohort, raising important questions about the generalizability and fairness of trial results. While the purpose of strict eligibility criteria is to ensure safety of participants

and study validity, it may lead to the inadvertent exclusion of certain demographic or clinical subgroups, leading to biased or incomplete conclusions about treatment effectiveness. Our results highlight the potential of RWE-driven methodologies to enhance our understanding of treatment effectiveness over time in heterogeneous populations, guide clinical decision-making in real-world settings, and uncover potential health disparities.

However, we acknowledge several limitations of our study. Firstly, while RWE provides a rich source of data, it comes with inherent biases and limitations. Observational data lacks the randomization of clinical trials, leading to potential confounding and selection biases. Despite efforts to mitigate these biases through IP weighting and careful cohort selection, residual biases may still influence our results. Missing data and inaccuracies within the EHR dataset may further impact the validity of results. Secondly, the emulation of clinical trials using RWE necessitates careful consideration of eligibility criteria and treatment protocols. While efforts were made to align with the original KEYNOTE-042 trial, certain criteria, such as treatment start and end dates, may be challenging to ascertain accurately from EHR data. This could introduce inaccuracies in cohort selection, potentially influencing the estimated treatment effects. Additionally, our reliance on retrospective data inherently restrict our ability to prospectively control for variables or account for unmeasured factors. While efforts were made to adjust for known confounders and biases, the presence of unobserved variables or unmeasured confounding factors cannot be fully addressed. This highlights the need for cautious interpretation of results; our findings can be adopted to guide future trial design and research areas, but cannot be interpreted as fact. We note the importance of considering broader societal factors, such as access to healthcare, socioeconomic status, and structural barriers, that may influence treatment outcomes beyond the scope of clinical trials. By recognizing and addressing these systemic inequities, we can work towards a more inclusive and equitable healthcare system that prioritizes the needs of all patient populations.

Acknowledgments

Isabel Chien and Richard E. Turner are supported by an EPSRC Prosperity Partnership EP/T005386/1 between Microsoft Research and the University of Cambridge.

References

ASCO et al. Accelerating progress against cancer: Asco's blueprint for transforming clinical and translational cancer research. American Society of Clinical Oncology, 2011.

Peter C. Austin and Elizabeth A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34:3661 – 3679, 2015. URL https://api.semanticscholar.org/CorpusID:14478957.

Amelia J Averitt, Chunhua Weng, Patrick Ryan, and Adler Perotte. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. *NPJ digital medicine*, 3(1):67, 2020.

Carine A Bellera, Gaëtan MacGrogan, Marc Debled, Christine Tunon De Lara, Véronique Brouste, and Simone Mathoulin-Pélissier. Variables with time-varying effects and the

cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC medical research methodology*, 10:1–12, 2010.

Zongwu Cai and Yanqing Sun. Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics*, 30(1):93–111, 2003.

Nicholas C. Chesnaye, Vianda S Stel, Giovanni Tripepi, Friedo W. Dekker, Edouard L. Fu, Carmine Zoccali, and Kitty J. Jager. An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal*, 15:14 – 20, 2021. URL https://api.semanticscholar.org/CorpusID:244632797.

Isabel Chien, Nina Deliu, Richard E. Turner, Adrian Weller, Sofía S. Villar, and Niki Kilbertus. Multi-disciplinary fairness considerations in machine learning for clinical trials. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022. URL https://api.semanticscholar.org/CorpusID:248863076.

Leslie Cho, Amanda R. Vest, Michelle L. O'Donoghue, Modele O Ogunniyi, Amy A. Sarma, Kara J. Denby, Emily S. Lau, Jeanne E. Poole, Kathryn J. Lindley, and Roxana Mehran. Increasing participation of women in cardiovascular trials: Jacc council perspectives. *Journal of the American College of Cardiology*, 78 7:737–751, 2021. URL https://api.semanticscholar.org/CorpusID:236999087.

John Concato, Nirav Shah, and Ralph I Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England journal of medicine*, 342(25):1887–1892, 2000.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Gilberto de Castro Jr, Iveta Kudaba, Yi-Long Wu, Gilberto Lopes, Dariusz M Kowalski, Hande Z Turna, Christian Caglevic, Li Zhang, Boguslawa Karaszewska, Konstantin K Laktionov, et al. Five-year outcomes with pembrolizumab versus chemotherapy as first-line therapy in patients with non–small-cell lung cancer and programmed death ligand-1 tumor proportion score¿= 1% in the keynote-042 study. *Journal of clinical oncology*, 41 (11):1986, 2023.

European Medicines Agency. Keytruda: European medicines agency. https://www.ema.europa.eu/en/medicines/human/EPAR/keytruda. Accessed: April 12, 2024.

Leena Gandhi, Delvys Rodríguez-Abreu, Shirish Gadgeel, Emilio Esteban, Enriqueta Felip, Flávia De Angelis, Manuel Domine, Philip Clingan, Maximilian J Hochmair, Steven F Powell, et al. Pembrolizumab plus chemotherapy in metastatic non–small-cell lung cancer. *New England journal of medicine*, 378(22):2078–2092, 2018.

Javier González, Cliff Wong, Zelalem Gero, Jass Bagga, Risa Ueno, Isabel Chien, Eduard Orakvin, Emre Kiciman, Aditya Nori, Roshanthi K. Weerasinghe, Rom S. Leidner, Brian Piening, Tristan Naumann, Carlo B Bifulco, and Hoifung Poon. Trialscope: A unifying causal framework for scaling real-world evidence generation with biomedical

language models. *ArXiv*, abs/2311.01301, 2023. URL https://api.semanticscholar.org/CorpusID:264935423.

Harrison J. Hansford, Aidan G. Cashin, Matthew D. Jones, Sonja A. Swanson, Nazrul Islam, Susan R. G. Douglas, Rodrigo R. N. Rizzo, Jack J. Devonshire, Sam A. Williams, Issa J. Dahabreh, Barbra A. Dickerman, Matthias Egger, Xabier Garcia-Albeniz, Robert M. Golub, Sara Lodi, Margarita Moreno-Betancur, Sallie-Anne Pearson, Sebastian Schneeweiss, Jonathan A. C. Sterne, Melissa K Sharp, Elizabeth A. Stuart, Miguel A. Hernán, Hopin Lee, and James H. McAuley. Reporting of observational studies explicitly aiming to emulate randomized trials. *JAMA Network Open*, 6, 2023. URL https://api.semanticscholar.org/CorpusID:263152047.

Roy S Herbst, Paul Baas, Dong-Wan Kim, Enriqueta Felip, José L Pérez-Gracia, Ji-Youn Han, Julian Molina, Joo-Hang Kim, Catherine Dubos Arvis, Myung-Ju Ahn, et al. Pembrolizumab versus docetaxel for previously treated, pd-l1-positive, advanced non-small-cell lung cancer (keynote-010): a randomised controlled trial. *The Lancet*, 387(10027): 1540–1550, 2016.

M.A. Hernan and J.M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC monographs on statistics & applied probability. Taylor & Francis, 2023. ISBN 9781315374932. URL https://books.google.co.jp/books?id=FPkN0AEACAAJ.

Miguel A. Hernán. The hazards of hazard ratios. *Epidemiology*, 21 1:13–5, 2010. URL https://api.semanticscholar.org/CorpusID:40559995.

Miguel A. Hernán and James M. Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183 8:758–64, 2016. URL https://api.semanticscholar.org/CorpusID:40032199.

Miguel A Hernán, Brian C Sauer, Sonia Hernández-Díaz, Robert Platt, and Ian Shrier. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of clinical epidemiology*, 79:70–75, 2016a.

Miguel A. Hernán, Brian C. Sauer, Sonia Hernández-Díaz, Robert William Platt, and Ian Shrier. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of clinical epidemiology*, 79:70–75, 2016b. URL https://api.semanticscholar.org/CorpusID:37188377.

Miguel A. Hernán, Wei Wang, and David E. Leaf. Target trial emulation: A framework for causal inference from observational data. *JAMA*, 328 24:2446–2447, 2022. URL https://api.semanticscholar.org/CorpusID:254627380.

Susan Jin, Richard Pazdur, and Rajeshwari Sridhara. Re-evaluating eligibility criteria for oncology clinical trials: analysis of investigational new drug applications in 2015. *Journal of clinical oncology*, 35(33):3745, 2017.

Jason Kendall. Designing a research project: randomised controlled trials and their principles. *Emergency Medicine Journal*, 20:164 – 168, 2003. URL https://api.semanticscholar.org/CorpusID:5628578.

Edward S Kim, David Bernstein, Susan G Hilsenbeck, Christine H Chung, Adam P Dicker, Jennifer L Ersek, Steven Stein, Fadlo R Khuri, Earle Burgess, Kelly Hunt, et al. Modernizing eligibility criteria for molecularly driven trials. *Journal of Clinical Oncology*, 33 (25):2815–2820, 2015.

Edward S Kim, Suanna S Bruinooge, Samantha Roberts, Gwynn Ison, Nancy U Lin, Lia Gore, Thomas S Uldrick, Stuart M Lichtman, Nancy Roach, Julia A Beaver, et al. Broadening eligibility criteria to make clinical trials more representative: American society of clinical oncology and friends of cancer research joint research statement. *Journal of Clinical Oncology*, 35(33):3737, 2017.

Myung S Kim and Vinay Prasad. Pembrolizumab for all. *Journal of Cancer Research and Clinical Oncology*, 149(3):1357–1360, 2023.

Linda E Lévesque, James A. Hanley, Abbas Kezouh, and Samy Suissa. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ : British Medical Journal*, 340, 2010. URL https://api.semanticscholar.org/CorpusID:46672786.

Ruishan Liu, Shemra Rizzo, Sam Whipple, Navdeep Pal, Arturo López Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William B. Capra, Ryan Copping, and James Zou. Evaluating eligibility criteria of oncology trials using real-world data and ai. *Nature*, 592:629 – 633, 2021. URL https://api.semanticscholar.org/CorpusID:233183554.

Torben Martinussen. Causality and the cox regression model. *Annual Review of Statistics and Its Application*, 2021. URL https://api.semanticscholar.org/CorpusID:244218042.

Torben Martinussen, Stijn Vansteelandt, and Per Kragh Andersen. Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis*, 26:833 – 855, 2020. URL https://api.semanticscholar.org/CorpusID:220501349.

Charles R. Mccarthy. Historical background of clinical trials involving women and minorities. *Academic Medicine*, 69:695–8, 1994. URL https://api.semanticscholar.org/CorpusID:31999343.

Tony Shu Kam Mok and Yi-Long Wu et al. Pembrolizumab versus chemotherapy for previously untreated, pd-l1-expressing, locally advanced or metastatic non-small-cell lung cancer (keynote-042): a randomised, open-label, controlled, phase 3 trial. *The Lancet*, 393:1819–1830, 2019. URL https://api.semanticscholar.org/CorpusID:93004086.

Medicine Matters Oncology. At a glance: The KEYNOTE lung cancer trials. https://oncology.medicinematters.com/non-small-cell-lung-cancer/pembrolizumab/at-a-glance-the-keynote-lung-cancer-trials/13445200, 2019. Accessed: April 8, 2024.

Konrad Pawelczyk, Aleksandra Piotrowska, Urszula Ciesielska, Karolina Jablonska, Natalia Glatzel-Plucinska, Jedrzej Grzegrzolka, Marzenna Podhorska-Okolow, Piotr Dziegiel, and Katarzyna Nowinska. Role of pd-l1 expression in non-small cell lung cancer and their

prognostic significance according to clinicopathological factors and diagnostic markers. *International journal of molecular sciences*, 20(4):824, 2019.

Luis Paz-Ares, Alexander Luft, David Vicente, Ali Tafreshi, Mahmut Gümüş, Julien Mazières, Barbara Hermes, Filiz Çay Şenler, Tibor Csőszi, Andrea Fülöp, et al. Pembrolizumab plus chemotherapy for squamous non–small-cell lung cancer. *New England Journal of Medicine*, 379(21):2040–2051, 2018.

Sam Preston, Mu Wei, Rajesh Rao, Robert Tinn, Naoto Usuyama, Michael Lucas, Yu Gu, Roshanthi Weerasinghe, Soohee Lee, Brian Piening, et al. Toward structuring real-world data: Deep learning for extracting oncology information from clinical text with patient-level supervision. *Patterns*, 4(4), 2023.

Martin Reck, Delvys Rodríguez-Abreu, Andrew G Robinson, Rina Hui, Tibor Csőszi, Andrea Fülöp, Maya Gottfried, Nir Peled, Ali Tafreshi, Sinead Cuffe, et al. Pembrolizumab versus chemotherapy for pd-l1–positive non–small-cell lung cancer. *New England Journal of Medicine*, 375(19):1823–1833, 2016.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994. URL https://api.semanticscholar.org/CorpusID:120769390.

Donald B. Rubin. Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Statistical Science*, 21:299–309, 2006. URL https://api.semanticscholar.org/CorpusID:15520248.

Adrian G Sacher and Leena Gandhi. Biomarkers for the clinical use of pd-1/pd-l1 inhibitors in non–small-cell lung cancer: a review. *JAMA oncology*, 2(9):1217–1222, 2016.

Aylin Sertkaya, Hui-Hsing Wong, Amber Jessup, and Trinidad Beleche. Key cost drivers of pharmaceutical clinical trials in the united states. *Clinical Trials*, 13(2):117–126, 2016.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 2000. URL https://api.semanticscholar.org/CorpusID:9238949.

Bonnie Sibbald and Martin Roland. Understanding controlled trials: Why are randomised controlled trials important? *BMJ*, 316:201, 1998. URL https://api.semanticscholar.org/CorpusID:35105860.

Harriette Gillian Christine Van Spall, Andrew Toren, Alex Kiss, and R. A. Fowler. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA*, 297 11:1233–40, 2007. URL https://api.semanticscholar.org/CorpusID:1593975.

Mats J Stensrud and Miguel A Hernán. Why test for proportional hazards? *Jama*, 323 (14):1401–1402, 2020.

Joan Stephenson. Fda offers guidance for boosting diversity in clinical trials. *JAMA health forum*, 1 11:e201434, 2020. URL https://api.semanticscholar.org/CorpusID:229475237.

Samy Suissa. Immortal time bias in pharmaco-epidemiology. *American journal of epidemiology*, 167 4:492–9, 2008. URL https://api.semanticscholar.org/CorpusID:24324551.

Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1549, 2021.

Eric J. Tchetgen Tchetgen. Identification and estimation of survivor average causal effects. *Statistics in Medicine*, 33:3601 – 3628, 2014. URL https://api.semanticscholar.org/CorpusID:2376959.

Laine Thomas and Eric M Reyes. Tutorial: survival estimation for cox regression models with time-varying coefficients using sas and r. *Journal of Statistical Software*, 61:1–23, 2014.

Lu Tian, David Zucker, and LJ Wei. On the cox model with time-varying regression coefficients. *Journal of the American statistical Association*, 100(469):172–183, 2005.

Joseph M. Unger, Riha Vaidya, Kathy S. Albain, Michael L LeBlanc, Lori M. Minasian, Carolyn Gotay, N. Lynn Henry, Michael J. Fisch, Shing M. Lee, Charles David Blanke, and Dawn L. Hershman. Sex differences in risk of severe adverse events in patients receiving immunotherapy, targeted therapy, or chemotherapy in cancer clinical trials. *Journal of Clinical Oncology*, 40:1474 – 1486, 2022. URL https://api.semanticscholar.org/CorpusID:246556420.

U.S. Food and Drug Administration. Fda expands pembrolizumab indication for first-line treatment of nsclc with tps ¿= 1. https://www.fda.gov/drugs/fda-expands-pembrolizumab-indication-first-line-treatment-nsclc-tps-1. Accessed: April 12, 2024.

Yangyang Xu, Bing Wan, Xi Chen, Ping Zhan, Yuan Zhao, Tianli Zhang, Hongbing Liu, Muhammad Zubair Afzal, Said Dermime, Steven N Hochwald, et al. The association of pd-l1 expression with the efficacy of anti-pd-1/pd-l1 immunotherapy and survival of non-small cell lung cancer patients: a meta-analysis of randomized controlled trials. *Translational Lung Cancer Research*, 8(4):413, 2019.

Zhongheng Zhang, Jaakko Reinikainen, Kazeem Adedayo Adeleke, Marcel E Pieterse, and Catharina GM Groothuis-Oudshoorn. Time-varying covariates and coefficients in cox regression models. *Annals of translational medicine*, 6(7), 2018.

Qiuling Zhao, Ruixiang Xie, Shen Lin, Xiang You, Xiuhua Weng, et al. Anti-pd-1/pd-l1 antibody therapy for pretreated advanced or metastatic nonsmall cell lung carcinomas and the correlation between pd-l1 expression and treatment effectiveness: an update meta-analysis of randomized clinical trials. *BioMed research international*, 2018, 2018.

Yurdaguel Zopf, Christina Rabe, Antje Neubert, Karl Günter Gassmann, Wolfgang Rascher, Eckhart G. Hahn, Kay Brune, and Harald Dormann. Women encounter adrs more often than do men. *European Journal of Clinical Pharmacology*, 64:999–1004, 2008. URL https://api.semanticscholar.org/CorpusID:21829393.

Irving Zucker and Brian J. Prendergast. Sex differences in pharmacokinetics predict adverse drug reactions in women. *Biology of Sex Differences*, 11, 2020. URL https://api.semanticscholar.org/CorpusID:219329824.

Appendix A. Additional dataset details

A.1. Target trial selection

Following the work of González et al. (2023) and Liu et al. (2021), we concentrate on completed phase III trials for non-small cell lung cancer (NSCLC). Among these, KEYNOTE studies offered the largest RWE-extracted cohorts due to the widespread use of the immunotherapy drug pembrolizumab, sold under the brand name Keytruda. Pembrolizumab, a humanized antibody that targets the programmed cell death protein 1 (PD-1) receptor of lymphocytes, is the most widely adopted oncologic drug globally (Kim and Prasad, 2023). Previous studies indicate that higher PD-L1 (a PD-1 receptor ligand) expression correlates with increased treatment benefit from pembrolizumab (Mok and et al., 2019), though these findings remain controversial, with other studies reporting conflicting results (Zhao et al., 2018; Xu et al., 2019). PD-L1 expression is measured by tumor proportion score (TPS): $TPS \geq 50\%$ for high expression, $50\% > TPS \geq 1\%$ for low expression, $TPS \geq 1\%$ for positive expression, and $TPS < 1\%$ for negative expression. Some NSCLC KEYNOTE studies explore pembrolizumab as a second-line therapy (Herbst et al., 2016), exclusively for high PD-L1 expression (Reck et al., 2016), or in combination with chemotherapy (Gandhi et al., 2018; Paz-Ares et al., 2018). However, we ruled these out due to difficulties of identifying second-line treatments and multi-drug combination therapies from the EHR and the limited number of suitable chemotherapy patients who satisfy PD-L1 expression criteria. In addition, we aim to emulate long-term trial outcomes; a few KEYNOTE studies, including KEYNOTE-042, report additional 5-year trial outcomes (de Castro Jr et al., 2023). With these factors in mind, we focus on KEYNOTE-042, which investigates pembrolizumab as a monotherapy for previously untreated NSCLC patients.

In the US, pembrolizumab is FDA-approved as a first-line monotherapy for late-stage NSCLC if tumors express positive PD-L1 TPS (U.S. Food and Drug Administration). The European Medicines Agency's (EMA) approval is stricter, requiring high PD-L1 expression (European Medicines Agency). Pembrolizumab is also approved for late-stage NSCLC under other cases, such as second-line therapy, or in combination with chemotherapy drugs (European Medicines Agency). However, in the EHR, we have found cases where first-line pembrolizumab patients lack PD-L1 tests or have negative PD-L1 expression.

Appendix B. Adjusting for biases: inverse probability weighting

Inverse probability weighting (IP weighting) is a statistical technique used to estimate quantities from a population distributed differently from the target inference population (Robins et al., 1994). IP weighting has been widely adopted to adjust for possible confounding and censoring bias in observational studies (Austin and Stuart, 2015; Chesnaye et al., 2021). Confounding occurs when both treatment assignment and patient outcome share a common cause, referred to as a confounder. Inverse probability of treatment weighting (IPTW) is commonly used to adjust for confounding by creating pseudo-populations in which the confounder no longer impacts the treatment assignment (Austin and Stuart, 2015; Hernan and Robins, 2023). A key requirement for confounding adjustment is that the probability of treatment does not depend on the confounders X. With our stabilized weights, we create a pseudo-population in which different people have different probabilities of treatment A, but

that probability of treatment does not depend on the covariates X. For the treated, $A = 1$, the IP weights are the probability of receiving treatment over the probability of receiving treatments given the covariates:

$$SW^{A=1} = \frac{Pr[A = 1]}{Pr[A = 1|X]}, SW^{A=0} = \frac{Pr[A = 0]}{Pr[A = 0|X]} \tag{4}$$

To calculate weights, we fit a simple logistic model to determine the denominator of each equation. Additionally, censoring bias is a form of selection bias that occurs when some patient covariate affects both the presence of censoring and patient outcome. Censoring bias can also be introduced during artificial censoring during assignment of treatment strategies (in Step 3 of the pipeline). Inverse probability of censoring weighting (IPCW) can be used to adjust for informative censoring with a similar principle as IPTW: we create a population where there is no impact of the covariates X on censoring C.

$$SW^{C} = \frac{Pr[C = 0|A]}{Pr[C = 0|X, A]} \tag{5}$$

IPTW and IPCW weights are calculated with a simple logistic model fit to the denominators of the above defined equations. These weights can be combined as $SW = SW^{A} * SW^{C}$ for each sample. Additionally, researchers may wish to calculate conditional average treatment effects based on population subgroups; for example, in male/female subgroups across treatment groups. This can be done in a 2-stage process consisting of 1) stratification by the subgroup variable of interest and 2) IP weighting using all other covariates.

We note that a set of assumptions known as *identifiability conditions* are required for valid causal inference employing IP weighting (Hernan and Robins, 2023) for resolving both issues of confounding and informative censoring. The conditions for IPTW are (1) *consistency*, where the observed outcome is equivalent to the counterfactual outcome under the observed intervention, (2) *conditional exchangeability*, where the average outcome in both treatment groups is equialent conditioned on treatment and measured covariates (this requires that there be *no unmeasured confounders*), (3) *positivity*, where all conditional probabilities of treatment assignment are greater than zero given all patient covariates. For IPCW, analagous assumptions are required where the outcome referenced is instead the censoring outcome rather than the investigated health outcome.

Appendix C. Trial eligibility criteria

Table 2: Eligibility criteria used to construct *eligible* and *expanded* cohorts. We note that this is not the detailed, full set of criteria (which can be seen in the Appendix of Mok and et al. (2019)).

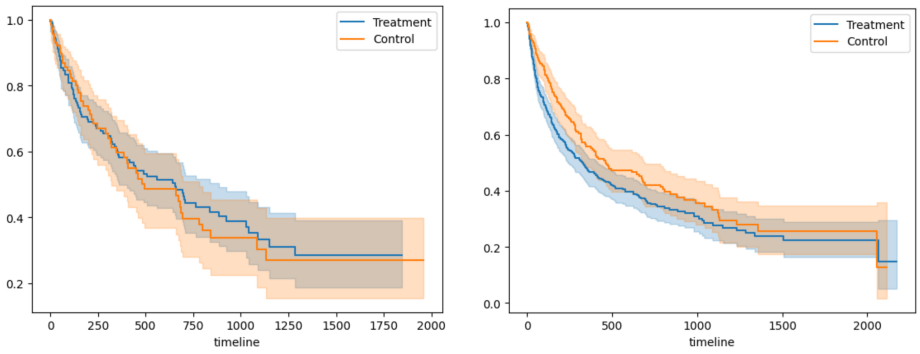
Eligibility Criteria	**Eligible**	**Expanded**	**Notes**
Inclusion Criteria			
Advanced or metastatic NSCLC (stages III and IV)	✔	✔	
Age (≥ 18)	✔	✔	
Life expectancy of at least 3 months			No equivalent measurement in the EHR.
No prior systemic chemotherapy treatment for NSCLC	✔		
ECOG $\in \{0, 1\}$			
Adequate organ function based on lab values	✔		Details in Mok and et al. (2019)
No history of prior malignancy	✔		
PD-L1 positive ($TPS \geq 1\%$) as determined by lab test	✔	✔	
Exclusion Criteria			
Has EGFR sensitizing mutation	✔		
Has ALK translocation	✔		
Has received prior systemic cytotoxic chemotherapy or radiation	✔		
Has received prior therapy with anti-PD-1, anti-PD-L1, anti-PD-L2, anti-CTLA-4 antibodies	✔		
Has known central nervous system metastases	✔		

Appendix D. Cohort summary statistics by PD-L1 strata

Table 3: Distribution of *eligible* and *expanded* cohorts by PD-L1 strata. PD-L1 not-negative consists of all individuals do not have a documented negative or indeterminate PD-L1 expression level. This includes patients with missing PD-L1 values. The treatment groups consist of patients receiving pembrolizumab, and the control groups consist of patients receiving platinum-based doublet chemotherapy, as specified for the KEYNOTE-042 trial. Note that a significant number of chemotherapy patients do not have PD-L1 readings and therefore fall into the not-negative PD-L1 strata. PD-L1 tests are not required for the administration of chemotherapy, nor is PD-L1 considered a predictive biomarker for chemotherapy treatment benefit. Therefore, it is less likely (even unusual) for a patient to have a recorded PD-L1 positive score but no history of immunotherapy.

	Eligible		**Expanded**	
	Treatment (n=306)	Control (n=832)	Treatment (n=822)	Control (n=1605)
Positive ($TPS \geq 1\%$)	204	126	533	299
High ($TPS \geq 50\%$)	151	58	386	132
Low ($50\% > TPS \leq 1\%$)	27	49	83	115
Negative ($TPS < 1\%$)	19	83	60	191
Not-negative	287	749	762	1414

Appendix E. Additional summary statistics and figures



(*a*) Eligible RWE cohort (*b*) Expanded RWE cohort

Figure 10: Kaplan-Meier survival curves for the PD-L1 positive stratum.

Table 4: Baseline characteristics of *eligible* RWE cohort with the PD-L1 TPS low and not-negative strata.

	Pembrolizumab		**Chemotherapy**	
PD-L1 TPS	Low (n=27)	Not Neg (n=287)	Low (n=49)	Not Neg (n=749)
Age (years)	74.4 ± 11.1	70.9 ± 10.8	67.0 ± 9.0	66.9 ± 9.5
<65	7 (26%)	92 (32%)	47 (50%)	325 (43%)
Male	16 (59%)	134 (47%)	18 (37%)	369 (49%)
Female	11 (41%)	153 (53%)	31 (63%)	380 (52%)
Race/ethnic group				
Asian	2 (7%)	22 (8%)	2 (4%)	27 (4%)
White or Caucasian	24 (89%)	230 (80%)	42 (86%)	656 (88%)
Other	1 (4%)	35 (12%)	5 (10%)	66 (9%)
ECOG score				
0	1 (4%)	38 (13%)	17 (35%)	117 (16%)
1	26 (96%)	249 (87%)	32 (65%)	632 (84%)
Smoking status				
Current/Former	14 (52%)	166 (58%)	39 (80)	571 (76%)
Never	13 (48%)	121 (42%)	10 (20%)	178 (24%)
Tumor hist. features				
Squamous	6 (22%)	51 (18%)	12 (24%)	203 (27%)
Non-squamous	21 (78%)	236 (82%)	37 (76%)	546 (73%)
Disease status				
Locally advanced (III)	6 (22%)	58 (20%)	19 (39%)	263 (35%)
Metastatic (IV)	21 (78%)	229 (80%)	30 (61%)	486 (65%)

Table 5: Baseline characteristics of *expanded* RWE cohort with the PD-L1 TPS high and positive strata.

	Pembrolizumab		**Chemotherapy**	
PD-L1 TPS	High (n=386)	Pos (n=533)	High (n=132)	Pos (n=299)
Age (years)	72.2 ± 10.2	72.3 ± 10.3	67.4 ± 8.7	66.8 ± 9.5
<65	106 (27%)	144 (27%)	54 (41%)	130 (43%)
Male	186 (48%)	278 (52%)	65 (49%)	142 (47%)
Female	200 (52%)	255 (48%)	67 (51%)	157 (53%)
Race/ethnic group				
Asian	25 (6%)	34 (6%)	8 (6%)	17 (6%)
White or Caucasian	318 (82%)	442 (83%)	110 (83%)	253 (85%)
Other	43 (11%)	57 (11%)	14 (11%)	29 (10%)
ECOG score				
0	54 (14%)	63 (12%)	27 (20%)	65 (22%)
1	215 (56%)	311 (58%)	74 (56%)	162 (54%)
1.5	11 (3%)	16 (3%)	7 (5%)	12 (4%)
2	68 (18%)	90 (17%)	15 (11%)	40 (13%)
2.5	9 (2%)	13 (16%)	1 (1%)	3 (1%)
3	21 (5%)	29 (5%)	8 (6%)	11 (4%)
3.5	1 (0%)	1 (0%)	0 (0%)	2 (61%)
4	7 (2%)	10 (2%)	0 (0%)	4 (1%)
Smoking status				
Current/Former	235 (61%)	340 (64%)	107 (81%)	240 (80%)
Never	151 (39%)	193 (36%)	25 (19%)	59 (20%)
Tumor hist. features				
Squamous	92 (24%)	133 (25%)	29 (22%)	80 (27%)
Non-squamous	294 (76%)	400 (75%)	103 (78%)	219 (73%)
Disease status				
(I)	24 (6%)	33 (6%)	10 (8%)	20 (7%)
(II)	16 (4%)	22 (4%)	9 (7%)	20 (7%)
Locally advanced (III)	67 (17%)	92 (17%)	29 (22%)	66 (22%)
Metastatic (IV)	225 (58%)	308 (58%)	69 (52%)	134 (45%)

Table 6: Baseline characteristics of *expanded* RWE cohort with the PD-L1 TPS low and not-negative strata.

	Pembrolizumab		**Chemotherapy**	
PD-L1 TPS	Low (n=83)	Not Neg (n=762)	Low (n=115)	Not Neg (n=1414)
Age (years)	72.9 ± 10.7	72.0 ± 10.4	66.0 ± 10.4	67.2 ± 9.7
<65	18 (22%)	210 (28%)	52 (45%)	583 (41%)
Male	52 (62%)	376 (49%)	55 (58%)	689 (49%)
Female	31 (37%)	386 (51%)	60 (52%)	725 (51%)
Race/ethnic group				
Asian	4 (5%)	50 (7%)	6 (5%)	52 (4%)
White or Caucasian	75 (90%)	623 (82%)	98 (85%)	1223 (86%)
Other	4 (5%)	89 (12%)	11 (10%)	139 (10%)
ECOG score				
0	6 (7%)	79 (10%)	28 (24%)	196 (14%)
1	51 (61%)	486 (64%)	59 (51%)	941 (67%)
1.5	4 (5%)	18 (2%)	3 (3%)	30 (2%)
2	14 (17%)	111 (15%)	18 (16%)	161 (11%)
2.5	1 (1%)	13 (2%)	2 (2%)	13 (1%)
3	6 (7%)	41 (5%)	1 (1%)	56 (4%)
3.5	0 (0%)	1 (0%)	2 (2%)	3 (0%)
4	1 (1%)	13 (2%)	2 (2%)	14 (1%)
Smoking status				
Current/Former	53 (64%)	460 (60%)	93 (81%)	1093 (77%)
Never	30 (36%)	302 (40%)	22 (20%)	321 (23%)
Tumor hist. features				
Squamous	27 (33%)	184 (24%)	35 (30%)	398 (28%)
Non-squamous	56 (67%)	578 (76%)	80 (70%)	1016 (72%)
Disease status				
(I)	7 (8%)	57 (7%)	8 (7%)	109 (8%)
(II)	5 (6%)	44 (6%)	6 (5%)	110 (8%)
Locally advanced (III)	17 (21%)	126 (17%)	29 (25%)	338 (24%)
Metastatic (IV)	45 (79%)	416 (55%)	46 (50%)	710 (50%)