Journal of Clinical Epidemiology 125 (2020) 47–56

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials—An illustration with the International Stroke Trial

Tri-Long Nguyen[a,b,c,d,e,*], Gary S. Collins[b,f], Paul Landais[c], Yannick Le Manach[d]

[a]*Section of Epidemiology, Department of Public Health, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen K, Denmark*
[b]*Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, UK*
[c]*Laboratory of Biostatistics, Epidemiology, Clinical Research and Health Economics, EA2415, Montpellier University, Montpellier, France*
[d]*Departments of Anesthesia & Health Research Methods, Evidence, and Impact, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative Research Group, Population Health Research Institute, Hamilton, Canada*
[e]*Department of Pharmacy, Nimes University Hospital, University of Montpellier, Nimes, France*
[f]*NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK*

Accepted 20 May 2020; Published online 25 May 2020

Abstract

Objective: Causal treatment effects are estimated at the population level in randomized controlled trials, while clinical decision is often to be made at the individual level in practice. We aim to show how clinical prediction models used under a counterfactual framework may help to infer individualized treatment effects.

Study Design and Setting: As an illustrative example, we reanalyze the International Stroke Trial. This large, multicenter trial enrolled 19,435 adult patients with suspected acute ischemic stroke from 36 countries, and reported a modest average benefit of aspirin (vs. no aspirin) on a composite outcome of death or dependency at 6 months. We derive and validate multivariable logistic regression models that predict the patient counterfactual risks of outcome with and without aspirin, conditionally on 23 predictors.

Results: The counterfactual prediction models display good performance in terms of calibration and discrimination (validation *c-statistics*: 0.798 and 0.794). Comparing the counterfactual predicted risks on an absolute difference scale, we show that aspirin—despite an average benefit—may increase the risk of death or dependency at 6 months (compared with the control) in a quarter of stroke patients.

Conclusions: Counterfactual prediction models could help researchers and clinicians (i) infer individualized treatment effects and (ii) better target patients who may benefit from treatments.

Keywords: Counterfactual framework; Causal inference; Clinical prediction models; Heterogeneity of treatment effect; Randomized controlled trial

1. Introduction

The randomized controlled trial (RCT) has long been recognized as the standard experimental method for providing clinical evidence of therapeutic intervention [1]—yet stated at the population level, while clinical decision-making is often made at the individual level [2–5]. As an individual treatment effect is not directly observed in RCTs, an average treatment effect is commonly estimated, thereby assuming a homogeneous response to the treatment, which is unlikely to hold in clinical practice. Given this deficiency of precision, a medicine effective on average can improve outcomes in most patients, although worsening outcomes in a minority of patients—the estimation of these respective proportions being regrettably neglected from RCT analysis. This duality between global evidence-based medicine and personalized decision-making therefore emphasizes the need for methods that can provide patient-level evidence about treatment effects.

To address this issue, subgroup analyses have been used to stratify the treatment effect by subpopulations [6,7].

Conflict of interest: The authors declare that there are no conflict of interest.

Source of funding: G.S.C. is supported by the NIHR Biomedical Research Centre, Oxford, UK, and the Cancer Research UK Programme Grant (C49297/A29084).

* Corresponding author. Section of Epidemiology, Department of Public Health, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen K, Denmark, CSS, Øster Farimagsgade 5, DK-1014 Copenhagen, Denmark, Tel.: +45 26693447.

E-mail address: long@sund.ku.dk (T.-L. Nguyen).

https://doi.org/10.1016/j.jclinepi.2020.05.022
0895-4356/

What is new?

Key findings?

- We illustrate how clinical prediction models used under a counterfactual framework could allow the inference of individualized treatment effects;

What this adds to what was known?

- Counterfactual prediction models return, given a patient, the predicted risks of outcome under different scenarios (e.g. patient risk of outcome under treatment vs. patient risk of outcome under control);

What is the implication and what should change now?

- The comparison of counterfactual predicted risks may help refine clinical therapeutic decision-making at the patient level, as shown in this illustration.

However, these approaches are deemed suboptimal and prone to multiple testing [8,9]. Exploring one variable at a time, based on patient characteristics that are believed to modify the treatment effect, is often limited when many underlying characteristics are involved and may lead to false-positive findings [8,9]. Recent methodological developments have led to considering multivariable predictive approaches to treatment effect heterogeneity [10,11]. In this article, we illustrate how the methodology of clinical prediction models (i.e. non-causal models) used under a counterfactual framework may allow a causal interpretation of individualized treatment effects.

We reanalyze the International Stroke Trial (IST) [12], which evaluated the effect of aspirin in stroke, a disease responsible for 6.7 million deaths in 2012 according to the World Health Organization [13]. With more than 65% of strokes being ischemic [14–16], effective treatments are required, particularly in the large number of patients who cannot receive intravenous thrombolysis. In Western countries, guidelines for the management of acute ischemic stroke recommend the use of aspirin [14,17,18], which demonstrated a clinical, albeit moderate, benefit. Nonetheless, it remains unclear whether this treatment is beneficial to all patients. The original article of the IST reported no clear evidence from the multiple subgroup analyses [12]. In this reanalysis, we derive models that predict the patient counterfactual risks of death or dependency at 6 months after stroke—with and without aspirin. We aim to show how the comparison of counterfactual risks of outcome could refine clinical decision-making on therapeutic strategy, given patient clinical characteristics.

2. Methods

2.1. Data and settings

The IST was a large, multicenter trial assessing the effect of aspirin and heparin on a primary (composite) outcome of death or dependency at 6 months after stroke, using a 2×2 factorial design [12]. It enrolled 19,435 adults with acute ischemic stroke from 36 countries and collected over 99% complete follow-up data. The data set from this RCT was recently released under an open-access license on behalf of the International Stroke Trial Collaborative Group [19]. As we performed a secondary analysis, our study was exempt from patient consent form collection. The trial originally reported a nonsignificant, moderate average risk reduction in the primary outcome at 6 months in the aspirin group (62.2% vs. 63.5%, two-tailed $P = 0.07$) [12]. To estimate the individualized responses to aspirin, we used the methodology of clinical prediction models under a counterfactual framework. This approach aligns with precedents described in the statistical literature, to which we invite readers to refer for further theoretical justification and technical contents [20–27].

2.2. Counterfactual prediction models

Let us define Z the treatment status: $Z=1$ denotes "treated", and $Z=0$ denotes "control". Following Rubin's causal model, let $Y_{(1)}$ and $Y_{(0)}$ denote the potential outcomes (or "counterfactuals") that would be observed if individuals were to receive the treatment or control, respectively [28]. In the IST [12], $Y_{(1)}$ denotes the risk of death or dependency at 6 months under aspirin, while $Y_{(0)}$ denotes that risk under control (i.e. without aspirin). For a particular individual i, the comparison of these two counterfactual outcomes defines the individual treatment effect: $TE_i = Y_{(1)i} - Y_{(0)i}$. (Note, this effect can also be expressed as a ratio.) In an RCT, as an individual can only be either treated or untreated, according to their actual treatment allocation, the individual treatment effect cannot be measured directly (an issue referred to as "the fundamental problem of causal inference". [29]) Clearly, only $Y_{(1)}$ is observed in the aspirin arm, and only $Y_{(0)}$ is observed in the control arm. Denoting the observed outcome by Y, one can write: $Y = ZY_{(1)} + (1-Z)Y_{(0)}$ (which is referred to as "consistency". [30]).

Furthermore, let X denote the baseline covariates. Given complete randomization, Z is assumed independent from X, but also from $Y_{(1)}$ and $Y_{(0)}$. In this sense, treated and control patients are assumed exchangeable: Both potential outcomes and covariates should be similarly distributed across the two groups. Formally, one can write $Z \perp (X, Y_{(1)}, Y_{(0)})$, which entails $Z \perp Y_{(1)}|X$ and $Z \perp Y_{(0)}|X$ (following the rules of conditional independence [31])—which can also be rewritten as $E(Y_{(1)}|X, Z=1) = E(Y_{(1)}|X)$ and $E(Y_{(0)}|X, Z=0) = E(Y_{(0)}|X)$. Thus, under consistency, one can estimate $E(Y_{(1)}|X)$ by fitting a prediction model

to the treated arm, and $E(Y_{(0)}|X)$ by fitting a prediction model to the control arm. For individual i with characteristics x_i, the first model returns $\widehat{E}(Y_{(1)i}|x_i)$ (i.e. patient prediction under aspirin). For the same individual, the second model returns $\widehat{E}(Y_{(0)i}|x_i)$ (i.e. patient prediction without aspirin). The two prediction models (which are noncausal because they do not model the relationship between the treatment and the outcome) can then be used to infer the causal treatment effect: $E(Y_{(1)}|X) - E(Y_{(0)}|X) = E(Y_{(1)} - Y_{(0)}|X)$. This treatment effect is said "individualized" (henceforth, "individualized treatment effect", ITE) because it is conditional on X, the set of patient baseline variables.

This general method can be regarded as similar to the parametric g-formula proposed by Robins [32], with two differences: (i) as we focus on the ITE, we spare the step of treatment effect averaging for a causal interpretation at the population level; (ii) treated and control patients are assumed exchangeable—given randomization—regardless of the set of covariates X included in the prediction models.

We applied this approach to reanalyze the IST.

2.3. Statistical analysis

Before developing the counterfactual prediction models, we split the initial sample of the IST at the hospital center level to generate a derivation sample and a validation sample. By creating two independent sets of patients hospitalized in structures within which practices and measurements were likely to differ, this procedure allowed us to conduct a geographical validation (sometimes called "broad" validation) of the prediction models [33]. We defined a split ratio of 2:1 which ensured that both samples included enough outcomes to avoid overfitting in derivation (>50 events per variable) [34], and to precisely quantify model performance during validation (>200 events) [35].

We fit separate logistic regressions, using 23 predictors (no variable selection), to predict the occurrence of death or dependency at 6 months to each treatment arm (aspirin and control) of the derivation sample. The predictors included in both models were the covariates that the trial investigators had specified for subgroup analyses (i.e. factors originally presumed to be responsible for heterogeneity). The nonlinearity of the continuous variables was handled using restricted cubic splines [36]. Any covariate included in one of the two regressions was also included in the other regression to allow differences in covariate effect across the two models (i.e., effect modification). The large sample size of the IST allowed us to include this large set of covariates. (Note, this procedure should not be conducted without precaution in small samples, in which penalization of regression models might be appropriate for covariate selection. [37]) Given the low rate of missing data, the analysis was performed in complete case data.

We used the two models to predict the probability of the counterfactual outcomes, $\widehat{P}(Y_{(1)} = 1|X)$ and $\widehat{P}(Y_{(0)} = 1|X)$, that would have occurred within 6 months for all individuals who had been treated and not with aspirin, respectively. To evaluate predictive ability of both models, we calculated the discrimination (*c-statistic*) in the derivation and validation samples and calibration (slope and intercept) in the validation sample (intercept and slope will be 0 and 1 by definition for the derivation sample). We also graphically assessed the calibration, using local regression curves [38]. Ninety-five percent confidence intervals (95% CI) were calculated by bootstrapping (500 iterations). We transparently reported our analysis following the TRIPOD statement [33,39].

We calculated the $\widehat{ITE}$ as the difference between the two counterfactual prognoses returned by the models, which corresponds to an absolute risk difference.

3. Results

Of the 19,435 included patients, 6,000 patients (62.2%) in the aspirin arm and 6,125 patients (63.5%) in the control arm experienced the primary composite outcome (absolute risk difference = −1.3%, 95% CI: −2.6% to +0.1%, $P = 0.07$; number needed to treat: 77 patients).

After random splitting at the hospital level, the derivation sample included 12,598 patients, while the validation included 6,937. The baseline characteristics at randomization are reported in Table 1. The average treatment effects were equal to −1.3% (95% CI: −3.0% to +0.4%, $P = 0.13$) and −1.1% (95% CI: −3.5% to +1.2%, $P = 0.33$), in the derivation and validation data, respectively.

In the derivation sample, we fit a regression model with 23 covariates to each arm. The aspirin arm model included 59 nonevents per degree of freedom and the control arm model included 58 nonevents per degree of freedom. There were 557 missing values (4.4%) across the primary outcome and the 23 covariates. The two models are presented in Table 2. The predictive performance of both models was good, as measured by discrimination and calibration (Figure 1). Performance was consistent in both the derivation and the validation samples, with no concerns of overfitting. The model predicting the outcome in the presence of aspirin had a c-statistic of 0.815 (95% CI: 0.805 to 0.825) in the derivation sample and 0.798 (95% CI: 0.782 to 0.813) in the validation sample. The calibration slope in the validation sample was 1.009 and the calibration intercept was −0.011. Similarly, the model predicting the outcome in the absence of aspirin had a c-statistic of 0.799 (95% CI: 0.788 to 0.811) in the derivation sample and 0.794 (95% CI: 0.778 to 0.809) in the validation sample, with a calibration slope of 1.026 and an intercept of −0.005 in the validation sample.

We estimated the ITE for each patient as the difference between the counterfactual risks of outcome under aspirin and control, returned by the two prediction models. As depicted in Figure 2, we found that aspirin effect may have been beneficial for certain patients (e.g., reducing the risk

Table 1. Baseline characteristics at randomization and outcomes

	Derivation sample		Validation sample	
Variable	**Aspirin 6,260 (49.7%)**	**Control 6, 338 (50.3%)**	**Aspirin 3, 460 (50.6%)**	**Control 3 377 (49.4%)**
Age (y)	74 (65–80)	74 (65–81)	73 (65–80)	73 (65–80)
Delay (h)	18 (9–28)	19 (9–29)	20 (10–30)	20 (9–30)
Systolic blood pressure (mmHg)	160 (140–180)	160 (140–180)	160 (140–180)	160 (140–180)
Male sex	3,278 (52.4%)	3,358 (53.0%)	1,875 (54.2%)	1,896 (56.1%)
Computerized tomography (CT)	4,175 (66.7%)	4,228 (66.7%)	2,316 (66.9%)	2,305 (68.3%)
Infarct visible at CT	2,036 (32.5%)	2,146 (33.9%)	1,140 (32.9%)	1,093 (32.4%)
Atrial fibrillation	1,092 (17.4%)	1,081 (17.1%)	530 (15.3%)	466 (13.8%)
Missing value	278 (4.4%)	279 (4.4%)	215 (6.2%)	212 (6.3%)
Aspirin within previous 3 d	1,317 (21.0%)	1,340 (21.1%)	644 (18.6%)	639 (18.9%)
Missing value	278 (4.4%)	279 (4.4%)	215 (6.2%)	212 (6.3%)
Face deficit				
Not assessable	89 (1.4%)	84 (1.3%)	34 (1.0%)	40 (1.2%)
No	1,679 (26.8%)	1,658 (26.2%)	888 (25.7%)	864 (25.6%)
Yes	4,492 (71.8%)	4,596 (72.5%)	2,538 (73.3%)	2,473 (73.2%)
Arm/hand deficit				
Not assessable	39 (0.6%)	43 (0.7%)	16 (0.5%)	25 (0.7%)
No	872 (13.9%)	870 (13.7%)	476 (13.7%)	449 (13.3%)
Yes	5,349 (85.5%)	5,425 (85.6%)	2,968 (85.8%)	2,903 (86.0%)
Leg/foot deficit				
Not assessable	94 (1.5%)	77 (1.2%)	39 (1.1%)	45 (1.3%)
No	1,469 (23.5%)	1,473 (23.2%)	803 (23.2%)	757 (22.4%)
Yes	4,697 (75.0%)	4,788 (75.6%)	2,618 (75.7%)	2,575 (76.3%)
Dysphasia				
Not assessable	190 (2.9%)	220 (3.5%)	91 (2.6%)	83 (2.5%)
No	3,250 (53.2%)	3,348 (52.8%)	1,922 (55.6%)	1,822 (53.9%)
Yes	2,820 (43.9%)	2,770 (43.7%)	1,447 (41.8%)	1,472 (43.6%)
Hemianopia				
Not assessable	1,391 (22.2%)	1,375 (21.7%)	596 (17.2%)	583 (17.2%)
No	3,896 (62.2%)	3,949 (62.3%)	2,301 (66.5%)	2,248 (66.6%)
Yes	973 (15.6%)	1,014 (16.0%)	563 (16.3%)	546 (16.2%)
Visuospatial disorder				
Not assessable	1,181 (18.9%)	1,192 (18.8%)	534 (15.4%)	541 (16.0%)
No	4,037 (64.5%)	4,076 (64.3%)	2,379 (68.8%)	2,317 (68.6%)
Yes	1,042 (16.6%)	1,070 (16.9%)	547 (15.8%)	519 (15.4%)
Brainstem/cerebellar signs				
Not assessable	571 (9.1%)	584 (9.2%)	226 (6.5%)	211 (6.3%)
No	4,983 (79.6%)	5,049 (79.7%)	2,865 (82.8%)	2,807 (83.1%)
Yes	706 (11.3%)	705 (11.1%)	369 (10.7%)	359 (10.6%)
Other deficit				
Not assessable	419 (6.7%)	423 (6.7%)	214 (6.2%)	193 (5.7%)
No	5,455 (87.1%)	5,502 (86.8%)	3,026 (87.4%)	2,984 (88.4%)
Yes	386 (6.2%)	413 (6.5%)	220 (6.4%)	200 (5.9%)
Consciousness				
Fully alert	4,742 (75.7%)	4,803 (75.8%)	2,721 (78.7%)	2,655 (78.6%)
Drowsy	1,437 (23.0%)	1,447 (22.8%)	690 (19.9%)	680 (20.1%)
Unconscious	81 (1.3%)	88 (1.4%)	49 (1.4%)	42 (1.3%)
Stroke type				
PACS	2,538 (40.5%)	2,568 (40.5%)	1,382 (39.9%)	1,367 (40.5%)

(*Continued*)

Table 1. Continued

Variable	Derivation sample		Validation sample	
	Aspirin 6,260 (49.7%)	Control 6, 338 (50.3%)	Aspirin 3, 460 (50.6%)	Control 3 377 (49.4%)
TACS	1,546 (24.7%)	1,539 (24.3%)	781 (22.6%)	772 (22.8%)
LACS	1,428 (22.8%)	1,474 (23.3%)	898 (26.0%)	857 (25.4%)
POCS	733 (11.7%)	735 (11.6%)	388 (11.2%)	372 (11.0%)
Other	15 (0.2%)	22 (0.3%)	11 (0.3%)	9 (0.3%)
Region				
Europe	5,243 (83.8%)	5,309 (83.8%)	2,876 (86.0%)	2,804 (86.0%)
North America	96 (1.5%)	94 (1.5%)	28 (0.8%)	30 (0.9%)
South America	205 (3.3%)	213 (3.4%)	142 (4.3%)	133 (4.1%)
Africa	33 (0.5%)	32 (0.5%)	2 (0.1%)	2 (0.1%)
Middle East	107 (1.7%)	107 (1.7%)	93 (2.8%)	93 (2.8%)
North Asia	44 (0.7%)	45 (0.7%)	18 (0.5%)	17 (0.5%)
South Asia	112 (1.8%)	117 (1.8%)	81 (2.4%)	79 (2.4%)
Oceania	420 (6.7%)	421 (6.6%)	105 (3.1)	104 (3.2%)
Death/dependency at 6 mo	3,896 (62.2%)	4,027 (63.5%)	2,104 (60.8%)	2,098 (62.1%)
Missing value	43 (0.7%)	42 (0.07%)	38 (1.1%)	27 (0.8%)

Medians (interquartile ranges) and counts (proportions) are reported for continuous and binary or categorical variables, respectively.

of death/dependency by more than 20%), but harmful for others (increasing the risk by 20% or more). Following the suggestion of an anonymous reviewer, we reported the calibration and discrimination performance of these predicted ITEs in Appendix.

We then stratified the trial with regard to the predicted ITE (i.e., stratum with expected benefit, $\widehat{ITE} < 0$; stratum without expected benefit, $\widehat{ITE} \geq 0$). These two strata represented 74.0% and 26.0% of the overall trial, respectively. In the beneficial stratum, the average aspirin effect was more than two-fold greater than the one originally reported in the trial with, in the derivation sample, an absolute risk difference equal to −3.4% (95% CI: −5.5% to −1.4%, $P < 0.001$; number needed to treat: 29 patients), which was confirmed in the validation sample: −3.3% (95% CI: −6.1% to −0.4%, $P = 0.025$; number needed to treat: 30 patients). In the stratum without expected benefit, the average effect of aspirin was equal to +3.3% (95% CI: 0.3% to 6.3%, $P = 0.031$; number needed to harm: 30 patients) in the derivation sample and to +1.6% (95% CI: −2.9 to +6.1, $P = 0.49$; number needed to harm: 63 patients) in the validation sample.

4. Discussion

We have illustrated how the methodology of clinical prediction models may be used under a counterfactual framework to predict individualized treatment responses. By reanalyzing the IST, we show that using counterfactual risk prediction models may help clinicians determine which patients with suspected ischemic stroke may benefit from aspirin.

The fairly good predictive performances of the two prediction models suggest consistent predictions that allow ITEs to be estimated. Our analyses in the derivation sample concur with those conducted in the validation sample, in finding that aspirin had a heterogeneous effect across the population, with a benefit in three-quarters of patients. This finding raises a concern of the naïve analysis of RCTs: If a significant average effect (as previously demonstrated for aspirin) shows a glass half full, an analysis of individualized effects can show it half empty. Meta-analyses have established a benefit of aspirin on average [40,41], yet a quarter of patients may instead experience harmful effects under aspirin.

Although the concept of evidence-based medicine has been widely implemented in clinical practice, evidence obtained from RCTs is stated at the population level, while clinical decisions are often made at the patient level [4,5]. This contrast warrants the need for methods to estimate treatment effects at individualized and subpopulation levels [6,7]. In traditional subgroup analysis, a population is subdivided on one variable at a time according to what researchers and clinicians consider to be potentially modifying treatment effects; thereby limiting how many characteristics can simultaneously explain heterogeneity in therapeutic response [8,9]. Recent approaches propose stratifying populations using disease risk scores [10,11,42–44], that is, prediction models without treatment (i.e., models that return $\widehat{P}(Y_{(0)} = 1|X)$). In contrast to one-variable-at-a-time analyses, these methods rely on multivariable models, which enable to create subpopulations that differ by many covariates (“multivariable” subgroup analysis). Nonetheless, the way thresholds are defined to create such strata may be arbitrary, particularly in cases of

Table 2. Models with and without aspirin predicting death or dependency at 6 months

	With aspirin		Without aspirin	
Variable	**Odds ratio (95% CI)**	***P***	**Odds ratio (95% CI)**	***P***
Intercept	0.08 (0.03–0.20)		0.12 (0.05–0.32)	
Age (y)	1.03 (1.02–1.04)	<0.001	1.03 (1.02–1.04)	<0.001
(Age)'	1.03 (1.01–1.04)		1.03 (1.01–1.04)	
Delay (h)	1.00 (1.00–1.01)	0.061	1.00 (1.00–1.01)	0.001
Systolic blood pressure (mmHg)	1.00 (0.99–1.00)	0.001	1.00 (0.99–1.00)	0.003
(Systolic blood pressure)'	1.00 (0.99–1.01)		1.00 (1.00–1.01)	
Male sex	0.76 (0.67–0.86)	<0.001	0.79 (0.70–0.90)	<0.001
Computerized tomography (CT)	0.55 (0.47–0.64)	<0.001	0.55 (0.47–0.64)	<0.001
Infarct visible at CT	1.47 (1.26–1.73)	<0.001	1.51 (1.30–1.76)	<0.001
Atrial fibrillation	1.19 (0.99–1.43)	0.046	1.28 (1.06–1.54)	0.005
Aspirin within previous 3 d	1.20 (1.03–1.40)	0.094	1.28 (1.10–1.48)	0.005
Face deficit (reference: No)		<0.001		<0.001
Not assessable	1.13 (0.55–2.32)		0.87 (0.43–1.78)	
Yes	1.24 (1.07–1.44)		1.18 (1.02–1.36)	
Arm/hand deficit (reference: No)		<0.001		<0.001
Not assessable	0.57 (0.20–1.58)		1.03 (0.32–3.32)	
Yes	1.42 (1.13–1.79)		1.41 (1.12–1.76)	
Leg/foot deficit (reference: No)		<0.001		<0.001
Not assessable	1.93 (0.96–3.86)		2.10 (0.89–4.98)	
Yes	2.21 (1.84–2.64)		1.97 (1.65–2.35)	
Dysphasia (reference: No)		0.002		0.397
Not assessable	2.36 (1.12–4.97)		1.18 (0.72–1.94)	
Yes	1.14 (0.96–1.35)		1.20 (1.02–1.43)	
Hemianopia (reference: No)		<0.001		<0.001
Not assessable	1.53 (1.16–2.01)		1.41 (1.08–1.85)	
Yes	1.70 (1.30–2.22)		1.66 (1.27–2.15)	
Visuospatial disorder (reference: No)		<0.001		<0.001
Not assessable	1.57 (1.22–2.03)		1.69 (1.31–2.18)	
Yes	1.59 (1.28–1.99)		1.79 (1.44–2.23)	
Brainstem/cerebellar signs (reference: No)		0.019		0.414
Not assessable	1.20 (0.85–1.69)		1.10 (0.80–1.50)	
Yes	2.87 (0.89–9.26)		1.98 (0.78–5.07)	
Other deficit (reference: No)		0.001		0.233
Not assessable	1.57 (1.05–2.34)		0.75 (0.53–1.06)	
Yes	1.60 (1.21–2.13)		1.07 (0.82–1.40)	
Consciousness (reference: Fully alert)		<0.001		<0.001
Drowsy	2.84 (2.31–3.49)		2.73 (2.22–3.36)	
Unconscious	8.98 (2.05–39.39)		11.57 (3.38–39.67)	
Stroke type (reference: PACS)		<0.001		<0.001
TACS	1.14 (0.86–1.50)		1.08 (0.82–1.42)	
LACS	0.93 (0.76–1.14)		0.88 (0.72–1.08)	
POCS	0.32 (0.10–1.03)		0.45 (0.18–1.13)	
Other	0.81 (0.21–3.20)		0.92 (0.33–2.57)	
Region (reference: Europe)		<0.001		<0.001
North America	0.38 (0.23–0.65)		0.81 (0.49–1.32)	
South America	0.52 (0.37–0.72)		0.62 (0.45–0.85)	
Africa	0.27 (0.11–0.67)		0.46 (0.20–1.08)	

(*Continued*)

Table 2. Continued

Variable	With aspirin		Without aspirin	
	Odds ratio (95% CI)	*P*	Odds ratio (95% CI)	*P*
Middle East	0.83 (0.52–1.34)		0.70 (0.44–1.10)	
North Asia	0.50 (0.24–1.06)		0.53 (0.26–1.07)	
South Asia	0.93 (0.59–1.47)		0.65 (0.42–1.02)	
Oceania	0.66 (0.51–0.84)		0.58 (0.46–0.74)	

A restricted cubic spline with three knots was used to describe the effects of age (knots at 56, 74 and 85 years) and systolic blood pressure (knots at 130, 160 and 200 mmHg).

Abbreviations: PACS, partial anterior circulation syndrome; TACS, total anterior circulation syndrome; LACS, lacunar syndrome; POCS, posterior circulation syndrome

nonmonotonic relationships between the prognosis without treatment and the treatment effect itself. In this regard, we highlight a practical issue of disease risk–stratified analysis: because the analyst might be unable to properly define thresholds, they may be inclined to repeatedly create strata and conduct statistical analyses until finding results that satisfy their hypothesis (i.e., multiple analyses increase the risk of false-positive findings). Using counterfactual prediction models, there is no need to define thresholds: the ITE is directly inferred from comparing the counterfactual risks of outcome (i.e., $\widehat{P}(Y_{(1)}=1|X)$ and $\widehat{P}(Y_{(0)}=1|X)$). As treatment effects are estimated at individualized levels (these estimates can then be averaged at the (sub) population level), this refers to a "bottom-up" approach as opposed to previous methods estimating effects from the population to the patient ("top-down").

Counterfactual prediction modeling may address issues faced by existing approaches. Those seeking to assess heterogeneity are often limited by the unobserved distribution of the ITEs [45,46], while existing methods for predicting ITEs do not consider heterogeneity at all [47,48]. For example, Dorresteijn et al. suggest calculating the ITE by assuming a homogenous treatment effect and multiplying the pretreatment risk of outcome (obtained from an existing model) by the average treatment effect [47]. Other approaches, such as that proposed in the study by van Kruijsdijk et al. [49], or Yeh et al. [50], require appropriate interaction terms to be included in a modeling step to handle treatment effect heterogeneity [10,11]. However, precedents have shown that modeling strategies that omit interactions may result in misleading estimates of ITE [51,52]. Counterfactual prediction modeling uses a different paradigm: where testing interactions can only suggest statistically significant differences in effects between subpopulations, estimating separate models allow differences that are informative at the individual level to be captured. In fact, this corresponds to a model including all (two-way) interactions possible with the treatment variable. This flexible approach can still be completed by including additional high-order interactions. The use of

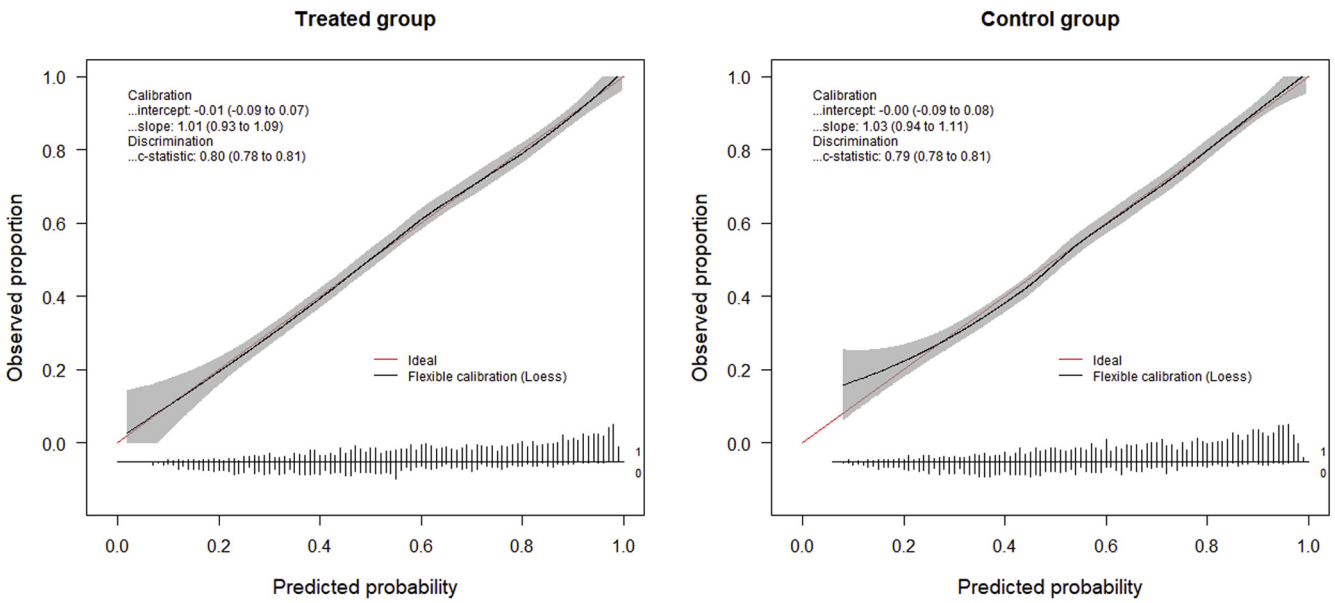

Fig. 1. Calibration curves of the counterfactual prediction models within each treatment group of the validation sample. The red dotted lines refer to ideal calibration. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

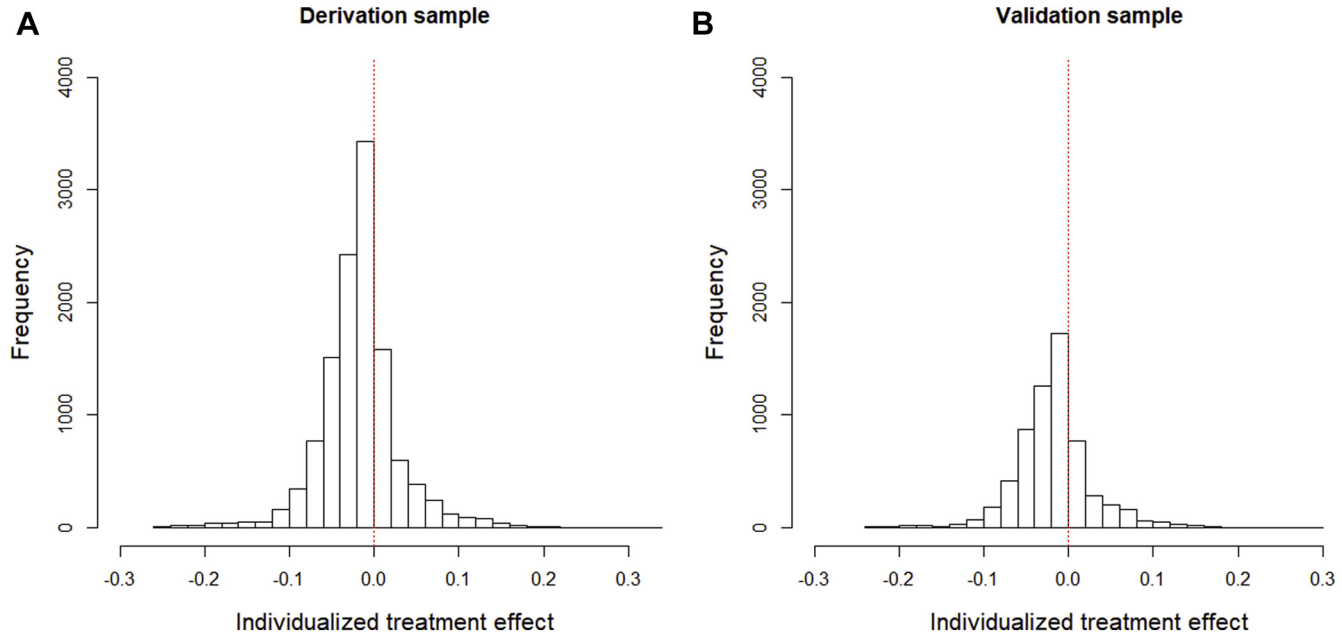

Fig. 2. Distribution of the individualized effect of aspirin (absolute risk difference). Negative values correspond to an outcome risk reduction under aspirin (beneficial effect), whereas positive values denote an increase of risk under aspirin (harmful effect).

separate counterfactual models complies with recent proposed approaches for identifying and targeting beneficial subpopulations [20–27]. Because counterfactual prediction modeling allows a causal interpretation of ITE based on prediction models, it may combine advantages and solve concerns of the disease risk model approach and the effect-interaction model approach—both described in a recent statement on predictive approaches to treatment effect heterogeneity [10,11].

Diverse limitations have to be considered to counterfactual prediction models. This methodology should not be applied without precaution. As with any clinical prediction model, three key steps should be undertaken: model development, external validation, and impact analysis [53], with models being transparently reported as stated for diagnostic and prognostic research [33,39]. In addition to assessing the predictive performances of the counterfactual prediction models, further methods are needed for calibrating the estimated ITEs. Ideally, counterfactual prediction models approach should be applied to identify responders in RCTs which have demonstrated a significant benefit; failing that, they may be useful to refine inclusion criteria for secondary trials. Appropriate confirmatory studies must nonetheless be conducted to prove the benefits revealed by such a reanalysis. Optimally, with regard to our reanalysis of the IST, further studies on external trials should be conducted to confirm our results; we intended to provide an illustrative example rather than results ready to be applied in clinical practice. From an analytic perspective, counterfactual prediction models combines two regression models (or more, in the case of multiple treatment arms), which might require more meticulous practices than usual. Further studies are needed to explore the robustness of this approach against model misspecification. The applicability of this method for reanalyzing RCTs may be limited by the need for large RCTs because samples including sufficient outcomes within each treatment arm are required to avoid overfitting [34]. Our illustration takes advantage of the considerable sample size of the IST, which may not be found in most RCTs. In the (likely) case of smaller trials, penalization of regression models might be required [37]. Finally, it is worth noting that the "individualized treatment effect", which is defined on a limited set of covariates, is to be distinguished from the individual—"indivisible", etymologically—treatment effect, which is nonidentifiable [10,11]. In this sense, our approach is to be understood as a support for clinical decision-making based on evidence inferred in (fine) groups of patients sharing similar characteristics. Epistemic uncertainty is therefore to be acknowledged in this decision-making: uncertainty about the evidence drawn from the groups, and uncertainty due to the gap between groups and individuals.

In conclusion, we have illustrated how using the methodology of clinical prediction models under a counterfactual framework may potentially help infer individualized therapeutic responses.

CRediT authorship contribution statement

Tri-Long Nguyen: Conceptualization, Methodology, Formal analysis, Writing - original draft. **Gary S. Collins:** Conceptualization, Methodology, Validation, Writing - review & editing. **Paul Landais:** Conceptualization, Methodology, Validation, Writing - review & editing. **Yannick Le Manach:** Conceptualization, Methodology, Validation, Writing - original draft.

Acknowledgments

The authors thank Sarah Kabani for her valuable editing help, Prof Jean-Pierre Daurès and Dr Truong Minh Nguyen for their helpful suggestions, and Prof Peter A.G. Sandercock for releasing the International Stroke Trial data on behalf of the investigators.

Authors' contributions: T.L.N. contributed to conceptualization, methodology, analysis, and original draft preparation. G.S.C. contributed to conceptualization, methodology, validation, and editing. P.L. contributed to conceptualization, methodology, validation, review, and editing. Y.L.M. contributed to conceptualization, methodology, validation, and original draft preparation.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.05.022.

References

[1] Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ, et al. What is "quality of evidence" and why is it important to clinicians? BMJ 2008;336:995−8.
[2] Bassler D, Busse JW, Karanicolas PJ, Guyatt GH. Evidence-based medicine targets the individual patient, part 1: how clinicians can use study results to determine optimal individual care. Evid Based Med 2008;13(4):101−2.
[3] Bassler D, Busse JW, Karanicolas PJ, Guyatt GH. Evidence-based medicine targets the individual patient, part 2: guides and tools for individual decision-making. Evid Based Med 2008;13(5):130−1.
[4] Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet 1995;345:1616−9.
[5] Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;365:82−93.
[6] Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA 2007;298:1209−12.
[7] Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176−86.
[8] Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004;57:229−36.
[9] Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess 2001;5:1−56.
[10] Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med 2019;172:35−45.
[11] Kent DM, van Klaveren D, Paulus JK, D'Agostino R, Goodman S, Hayward R, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement: explanation and elaboration. Ann Intern Med 2019;172:W1−25.
[12] International Stroke Trial Collaborative Group. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. International Stroke Trial Collaborative Group. Lancet 1997;349:1569−81.
[13] World Health Organisation. Global status report on noncommunicable diseases. 2014. Available at https://www.who.int/nmh/publications/ncd-status-report-2014/en/. Accessed June 15, 2020.
[14] Jauch EC, Saver JL, Adams HP Jr, Bruno A, Connors JJB, Demaerschalk BM, et al. Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2013;44(3):870−947.
[15] Tsai CF, Thomas B, Sudlow CL. Epidemiology of stroke and its subtypes in Chinese vs white populations: a systematic review. Neurology 2013;81:264−72.
[16] Warlow C. Stroke, transient ischaemic attacks, and intracranial venous thrombosis. Donaghy, M Brain's Diseases of the Nervous System. 11th ed. Oxford Univeristy Press; 2001:775−896.
[17] European Stroke Organisation Executive Committee, ESO Writing Committee. Guidelines for management of ischaemic stroke and transient ischaemic attack 2008. Cerebrovasc Dis 2008;25(5):457−507.
[18] Royal College of Physicians. National Clinical Guidelines for Stroke. 2016. Available at https://www.rcplondon.ac.uk/guidelines-policy/stroke-guidelines. Accessed June 15, 2020.
[19] Sandercock PA, Niewada M, Czlonkowska A. International stroke trial collaborative G. The international stroke trial database. Trials 2011;12:101.
[20] Lamont A, Lyons MD, Jaki T, Stuart EA, Feaster DJ, Tharmaratnam K, et al. Identification of predicted individual treatment effects in randomized clinical trials. Stat Methods Med Res 2016;27:142−57.
[21] Li J, Zhao L, Tian L, Cai T, Claggett B, Callegaro A, et al. A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative controlled clinical studies. Biometrics 2016;72:877−87.
[22] Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics 2011;12(2):270−82.
[23] Kang C, Janes H, Huang Y. Combining biomarkers to optimize patient treatment recommendations. Biometrics 2014;70:695−707.
[24] Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. Stat Med 2011;30:2867−80.
[25] Zhao L, Tian L, Cai T, Claggett B, Wei LJ. Effectively selecting a target population for a future comparative study. J Am Stat Assoc 2013;108:527−39.
[26] Porcher R, Jacot J, Wunder JS, Biau DJ. Identifying treatment responders using counterfactual modeling and potential outcomes. Stat Methods Med Res 2018;28(10−11):3346−62.
[27] Huang Y, Gilbert PB, Janes H. Assessing treatment-selection markers using a potential outcomes framework. Biometrics 2012;68:687−96.
[28] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;6:688−701.
[29] Holland PW. Statistics and causal inference. J Am Stat Assoc 1986;81:945−60.
[30] Pearl J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? Epidemiology 2010;21:872−5.
[31] Dawid AP. Conditional independence in statistical theory. J R Stat Soc Ser B Methodol 1979;41(1):1−31.
[32] Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Model 1986;7(9−12):1393−512.
[33] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1−73.
[34] Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-

sample validity of logistic regression models. Stat Methods Med Res 2017;26(2):796−808.
[35] Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med 2016;35:214−26.
[36] Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med 2016;35:4124−35.
[37] van Klaveren D, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. J Clin Epidemiol 2019;114:72−83.
[38] Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med 2014;33:517−35.
[39] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594.
[40] Chen ZM, Sandercock P, Pan HC, Counsell C, Collins R, Liu LS, et al. Indications for early aspirin use in acute ischemic stroke : a combined analysis of 40 000 randomized patients from the Chinese acute stroke trial and the international stroke trial. On behalf of the CAST and IST collaborative groups. Stroke 2000;31(6):1240−9.
[41] Sandercock PA, Counsell C, Tseng MC, Cecconi E. Oral antiplatelet therapy for acute ischaemic stroke. Cochrane Database Syst Rev 2014CD000029.
[42] Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. Circ Cardiovasc Qual Outcomes 2014;7(1):163−9.
[43] Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. Int J Epidemiol 2016;45:2075−88.
[44] Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010;11:85.
[45] Schlattmann P. Medical Applications of Finite Mixture Models. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg; 2009.
[46] Adams C. Estimating Heterogeneous Treatment Effects in Randomized Control Trials. 2014. Available at http://dx.doi.org/10.2139/ssrn.2433064. Accessed June 15, 2020.
[47] Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ 2011;343:d5888.
[48] van der Leeuw J, Ridker PM, van der Graaf Y, Visseren FL. Personalized cardiovascular disease prevention by applying individualized prediction of treatment effects. Eur Heart J 2014;35:837−43.
[49] van Kruijsdijk RC, Visseren FL, Ridker PM, Dorresteijn JA, Buring JE, van der Graaf Y, et al. Individualised prediction of alternate-day aspirin treatment effects on the combined risk of cancer, cardiovascular disease and gastrointestinal bleeding in healthy women. Heart 2015;101:369−76.
[50] Yeh RW, Secemsky EA, Kereiakes DJ, Normand SL, Gershlick AH, Cohen DJ, et al. Development and validation of a prediction rule for benefit and harm of dual antiplatelet therapy beyond 1 Year after percutaneous coronary intervention. JAMA 2016;315:1735−49.
[51] van Klaveren D, Vergouwe Y, Farooq V, Serruys PW, Steyerberg EW. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. J Clin Epidemiol 2015;68:1366−74.
[52] Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TPA, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. J Clin Epidemiol 2016;78:90−100.
[53] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10(2):e1001381.