Mahidol University
Wisdom of the Land

Counterfactual Clinical Prediction Models could Help to Infer Individualized Treatment Effects in Randomized Controlled Trials

An Illustration with the International Stroke Trial

Presenter: Sureerat Suwatcharangkoon

RCT

- Average treatment effect is commonly estimated
  - assuming a homogeneous response to the treatment
  - worsening outcomes in a minority of patients
- An individual treatment effect is not directly observed

The need for methods that can provide patient-level evidence about treatment effects

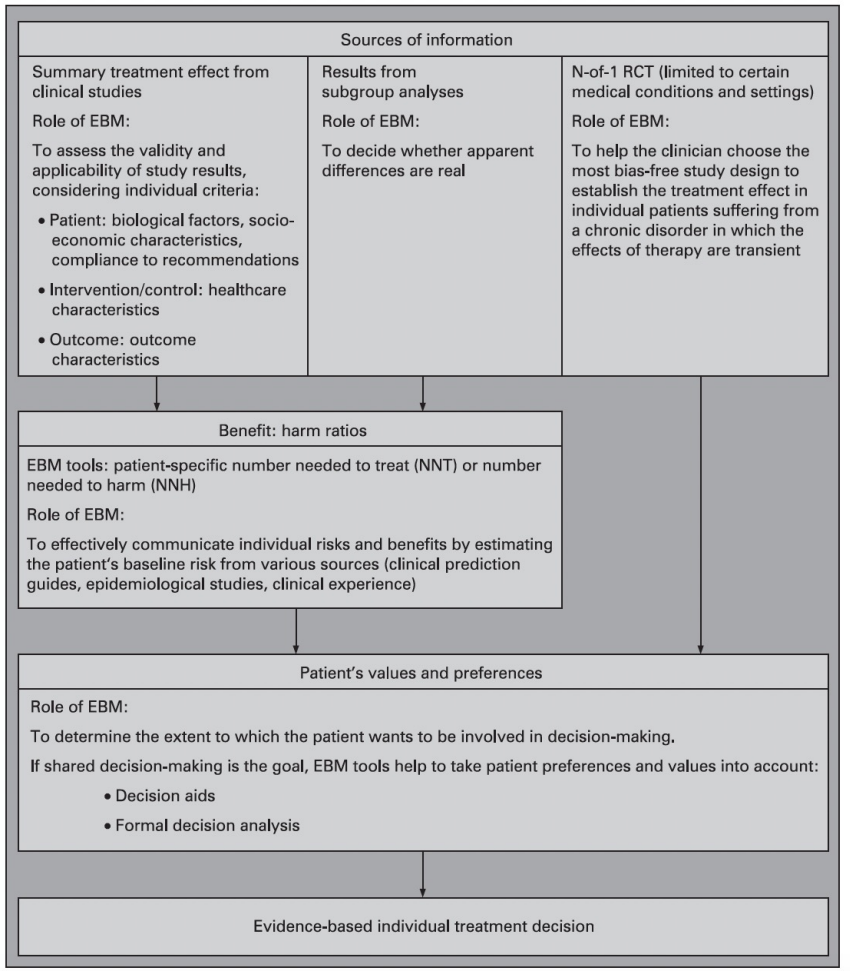
Evidence-based medicine targets the individual patient

Criteria to consider when applying the results of research studies to individual patients

Patient characteristics	Healthcare characteristics	Outcome characteristics
► Biological factors (sex, comorbidities, race, age, severity of pathology) ► Patient compliance with treatment requirements	► Compliance of healthcare providers with treatment requirements ► Resources available for implementation (eg, availability of monitoring) ► Expertise of clinicians	► Did the study measure an outcome of importance to the individual patient?

Evid Based Med 2008;13(4):101-2.

The process of individualised EBM decision making



Evid Based Med 2008;13(5):130-1.

RCT

- Subgroup analyses
  - limited when many underlying characteristics are involved
  - prone to multiple testing -> risk of false-positive findings
- The Predictive Approaches to Treatment effect Heterogeneity (PATH)

THE LANCET

Articles

The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke

*International Stroke Trial Collaborative Group**

P: acute ischemic stroke with onset < 48 h previously

Patients randomised (n=19 435)

Allocations

Aspirin 300 + Heparin 12 500	Aspirin 300 + Heparin 5000	Aspirin 300 + No Heparin	No aspirin + Heparin 12 500	No aspirin + Heparin 5000	No aspirin + No Heparin
2430	2432	4858	2426	2429	4860

No (% randomised) with mortality follow-up

14 days

2430 (100·0%)	2431 (99·99%)	4858 (100·0%)	2426 (100·0%)	2429 (100·0%)	4859 (99·99%)

6 months

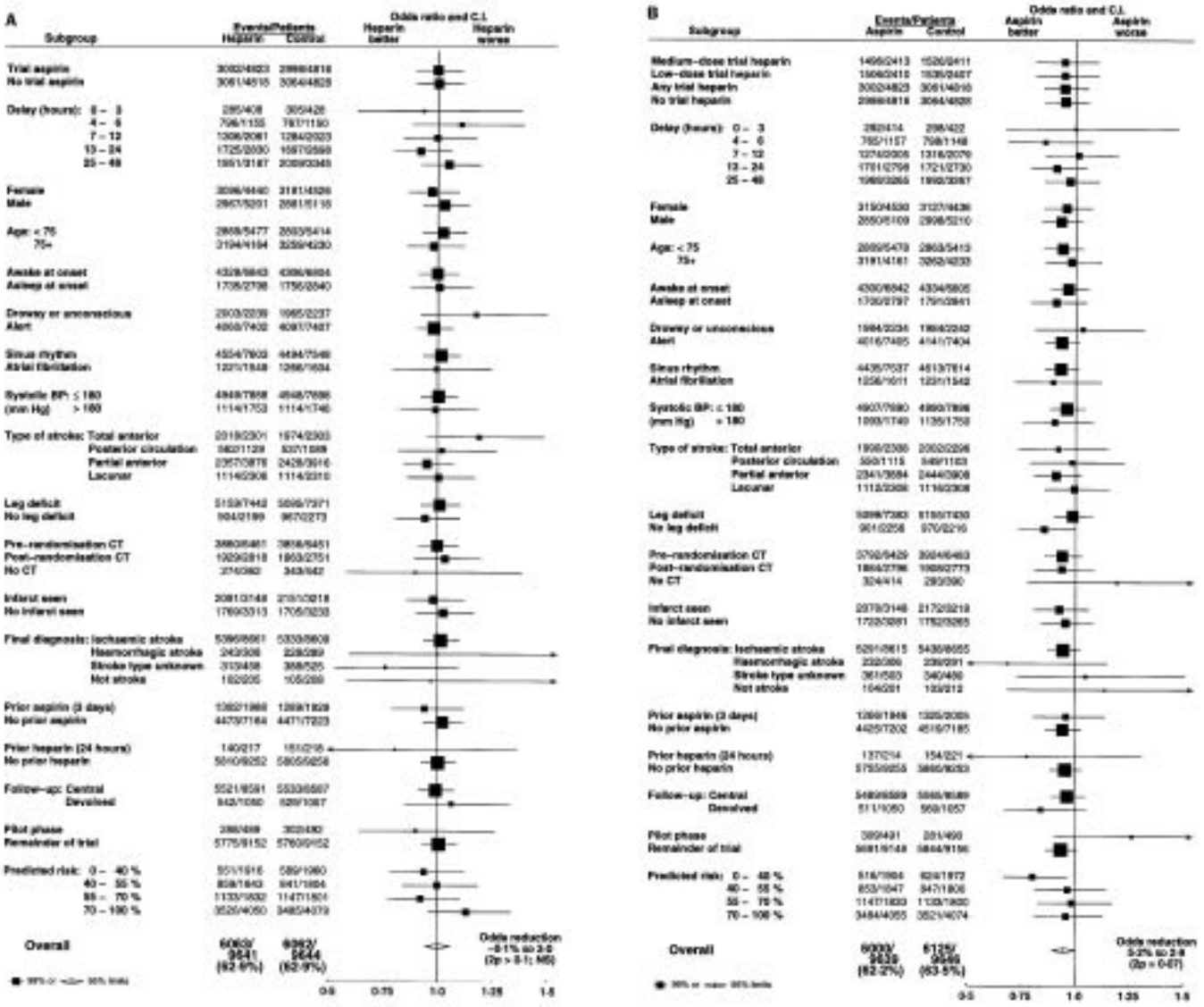
2413 (99·3%)	2410 (99·1%)	4816 (99·1%)	2411 (99·4%)	2407 (99·1%)	4828 (99·3%)

O: death within 14 days and death or dependency at 6 months

Lancet. 1997 May 31;349(9065):1569-81.

1º outcome at 6 months in the aspirin group (62.2% vs. 63.5%, P = 0.07)

IST Subgroup analyses



Counterfactual prediction models

Theory

1. Definitions

Rubin's causal model

$$Z_i \in \{0; 1\}$$

$$TE_i = Y_{(1)i} - Y_{(0)i} \quad (1)$$

Z; the treatment status: $Z_i = 1$; *'treated'*, and $Z_i = 0$; *'control'*

$Y_{(1)i}$ and $Y_{(0)i}$; the *potential outcomes (or 'counterfactuals')*

i; a particular individual

Theory

1. Definitions

Consistency

$$Y_i = Z_i Y_{(1)i} + (1 - Z_i) Y_{(0)i} \quad (2)$$

Y_i ; the *observed outcome*

$Z_i = 1$; *'treated'*, and $Z_i = 0$; *'control'*

$Y_{(1)i}$ and $Y_{(0)i}$; the *potential outcomes (or 'counterfactuals')*

$$ATE = E(Y_{(1)} - Y_{(0)}) = E(Y_{(1)}) - E(Y_{(0)}) \quad (3)$$

Theory

1. Definitions

$$Xi \in \mathcal{X}$$

$$ITE = E\left(Y_{(1)} - Y_{(0)}|X\right) = E\left(Y_{(1)}|X\right) - E\left(Y_{(0)}|X\right) \quad (4)$$

$Xi \in \mathcal{X}$ denote the *baseline covariates*

ITE; *individualised treatment effect*

Mahidol University
Wisdom of the Land

Theory

2. Identification

Conditional independence

$$Z \perp (X, Y_{(1)}, Y_{(0)}) \quad (5)$$

Z; the treatment status; X; the baseline covariates, Y; the potential outcome

$$E(Y_{(1)}|X, Z = 1) = E(Y_{(1)}|X) \quad (6)$$

$$E(Y_{(0)}|X, Z = 0) = E(Y_{(0)}|X) \quad (7)$$

Mahidol University
Wisdom of the Land

Theory

2. Identification

Conditional independence

$$E\left(Y_{(1)}|X,Z=1\right)=E\left(Y_{(1)}|X\right) \quad (6)$$

$$E\left(Y_{(0)}|X,Z=0\right)=E\left(Y_{(0)}|X\right) \quad (7)$$

Consistency

$$Y_i = Z_i Y_{(1)i} + (1-Z_i)Y_{(0)i} \quad (2)$$

$$E(Y|X,Z=1)=E\left(Y_{(1)}|X,Z=1\right)=E\left(Y_{(1)}|X\right) \quad (8)$$

$$E(Y|X,Z=0)=E\left(Y_{(0)}|X,Z=0\right)=E\left(Y_{(0)}|X\right) \quad (9)$$

Mahidol University
Wisdom of the Land

Theory

2. Identification

$$E(Y|X,Z=1) = E(Y_{(1)}|X,Z=1) = E(Y_{(1)}|X) \quad (8)$$

$$E(Y|X,Z=0) = E(Y_{(0)}|X,Z=0) = E(Y_{(0)}|X) \quad (9)$$

ITE

$$E(Y_{(1)}|X) - E(Y_{(0)}|X) = E(Y_{(1)} - Y_{(0)}|X)$$

f; *function*

Mahidol University
Wisdom of the Land

Theory

2. Identification

$$E(Y|X,Z=1) = E\left(Y_{(1)}|X,Z=1\right) = E\left(Y_{(1)}|X\right) \quad (8)$$

$$E(Y|X,Z=0) = E\left(Y_{(0)}|X,Z=0\right) = E\left(Y_{(0)}|X\right) \quad (9)$$

Estimator of ITE

$$\hat{E}(Y|X,Z=1,\alpha) = f(X,\alpha) \quad (10)$$

$$\hat{E}(Y|X,Z=0,\beta) = f(X,\beta) \quad (11)$$

f; *function*

Mahidol University
Wisdom of the Land

Calibration and Discrimination Performance of Predicted ITE

- ITE (individualised treatment effects) are never observed– but estimated
- Consistence between prediction and observation ✗

  , but rather between different estimates

Discrimination Performance of Predicted ITE

- Use “c-index”

Calibration Performance of Predicted ITE

- Agreement between the predicted ITE and the corresponding "observed" ITE.

1. Stratify the validation sample according to quintiles of predicted ITE
2. Compute the marginal average difference in observed outcome across the treatment groups

"observed" ITE

Statistical analysis

1. **Split the initial sample** to generate a derivation sample and a validation sample (2:1)
   - enough outcomes to avoid overfitting in derivation (> 50 events/variable)
   - precisely quantify model performance during validation (> 200 events)

Statistical analysis

2. **Fit separate logistic regressions**, using 23 predictors (no variable selection), to predict the outcome to each treatment arm of the derivation sample
   - effect modification

Mahidol University
Wisdom of the Land

Statistical analysis

3. Predict the probability of the counterfactual outcomes

$$\hat{P}\left(Y_{(\mathbf{1})} = 1|X\right)$$
$$\hat{P}\left(Y_{(\mathbf{0})} = 1|X\right)$$

Mahidol University
Wisdom of the Land

Statistical analysis

4. **The discrimination:** calculate the discrimination (c-statistic) in the derivation and validation samples

5. **The calibration:**
   - calculating the calibration (slope and intercept) in the validation sample
   - using local regression curves
   - 95% CI were calculated by bootstrapping (500 iterations)

Mahidol University
Wisdom of the Land

Statistical analysis

6. **Calculated the $\widehat{ITE}$** (difference between the 2 counterfactual prognoses returned by the models)

Mahidol University
Wisdom of the Land

Results

Articles

The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke

*International Stroke Trial Collaborative Group**

Outcome	Heparin *vs* no heparin			Aspirin *vs* no aspirin		
	Heparin	No heparin	Events prevented per 1000 (SD)	Aspirin	No aspirin	Events prevented per 1000 (SD)
No randomised	9717	9718		9720	9715	
No with 6 month data	9641 (99·2%)	9644 (99·2%)		9639 (99·2%)	9646 (99·3%)	
Fully recovered, independent	1655 (17·2%)	1641 (17·0%)	−2 (5)	1694 (17·6%)	1602 (16·6%)	−10 (5)
Not recovered, but independent	1923 (19·9%)	1941 (20·1%)	2 (6)	1945 (20·2%)	1919 (19·9%)	−3 (6)
Dependent	3898 (40·4%)	3986 (41·3%)	9 (7)	3927 (40·7%)	3957 (41·0%)	3 (7)
Dead from any cause	2165 (22·5%)	2076 (21·5%)	−9 (6)	2073 (21·5%)	2168(22·5%)	10 (6)
Dead or dependent	6063 (62·9%)	6062 (62·9%)	0 (7)†	6000 (61·2%)	6125(63·5%)	13 (7)‡

†After adjustment for prognosis predicted at baseline, the benefit from heparin was 0 (SD 6), NS. ‡After adjustment for baseline stroke severity, the benefit from aspirin was 14 (SD 6), ($2p$=0·03). Negative numbers; same conventions as in table 2.
*$2p$<0·05, **$2p$<0·01, ***$2p$<0·001, ****$2p$<0·00001.

Table 3: **Outcome at 6 months**

Lancet. 1997 May 31;349(9065):1569-81.

Mahidol University
Wisdom of the Land

Table 1. Baseline characteristics at randomization and outcomes

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%)
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%)
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.

D = 12,598

V = 6,937

Mahidol University
Wisdom of the Land

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

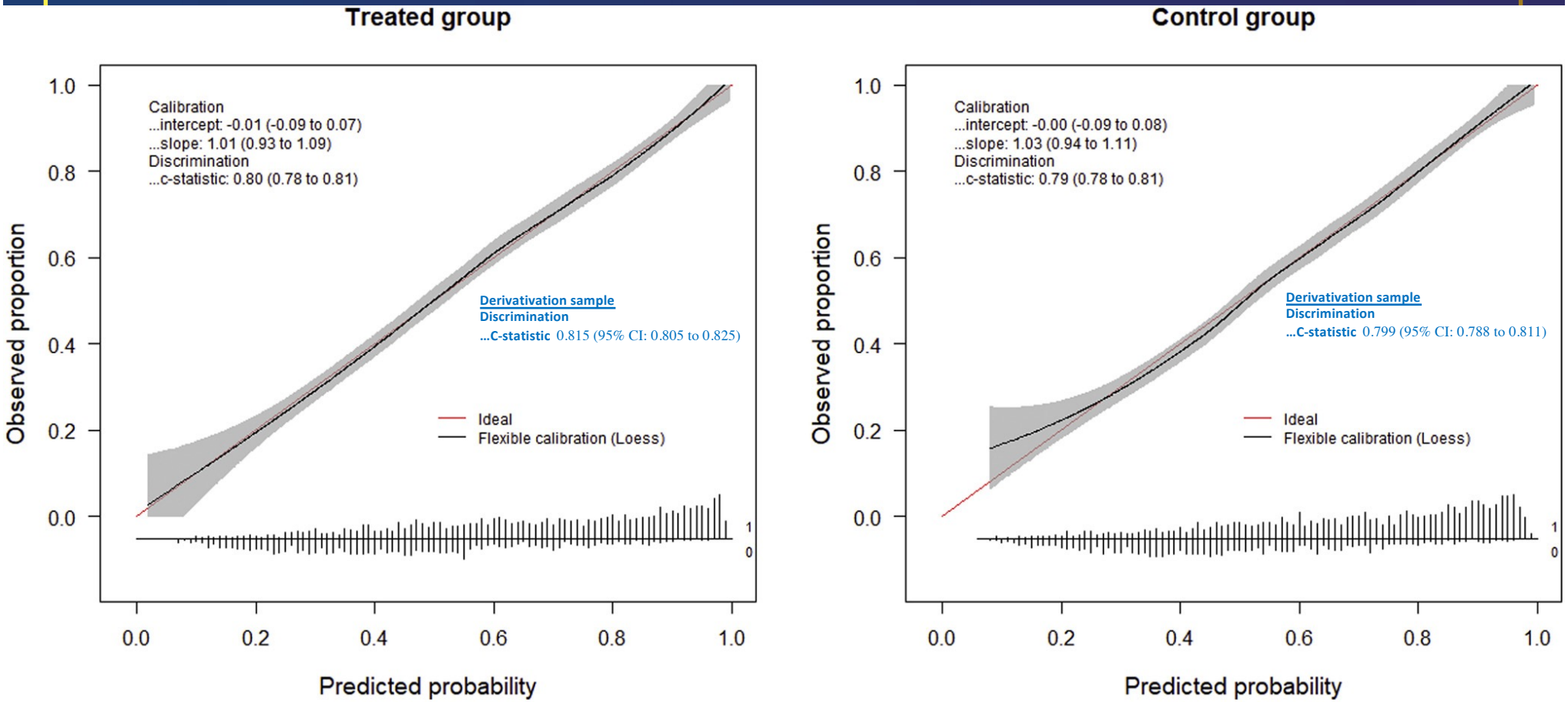
Variable	With aspirin		Without aspirin	
	Odds ratio (95% CI)	P	Odds ratio (95% CI)	P
Intercept	0.08 (0.03–0.20)		0.12 (0.05–0.32)	
Age (y)	1.03 (1.02–1.04)	<0.001	1.03 (1.02–1.04)	<0.001
(Age)'	1.03 (1.01–1.04)		1.03 (1.01–1.04)	
Delay (h)	1.00 (1.00–1.01)	0.061	1.00 (1.00–1.01)	0.001
Systolic blood 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)	
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

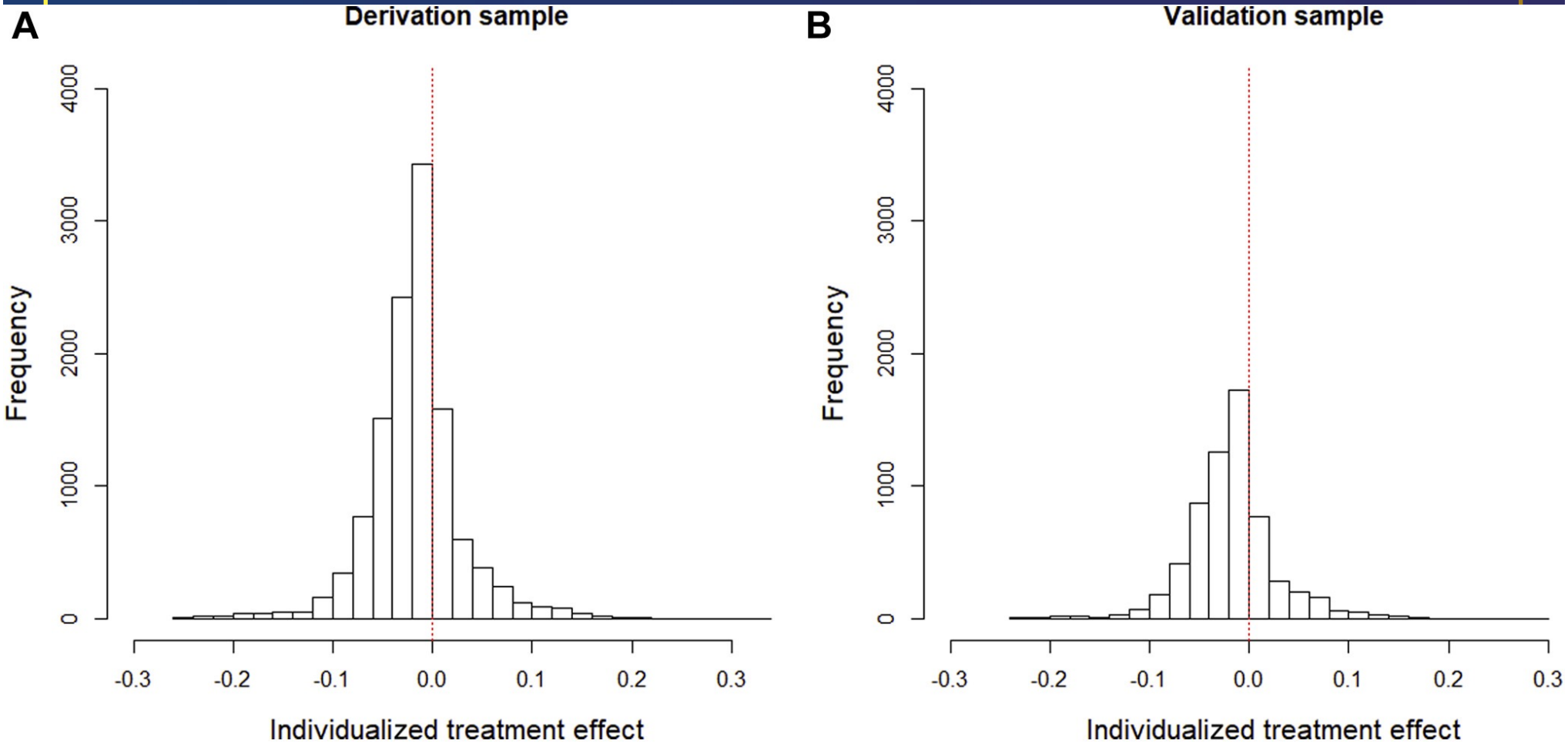
Mahidol University
Wisdom of the Land

Fig 1. Calibration curves of the counterfactual prediction models within each treatment group of the validation sample.

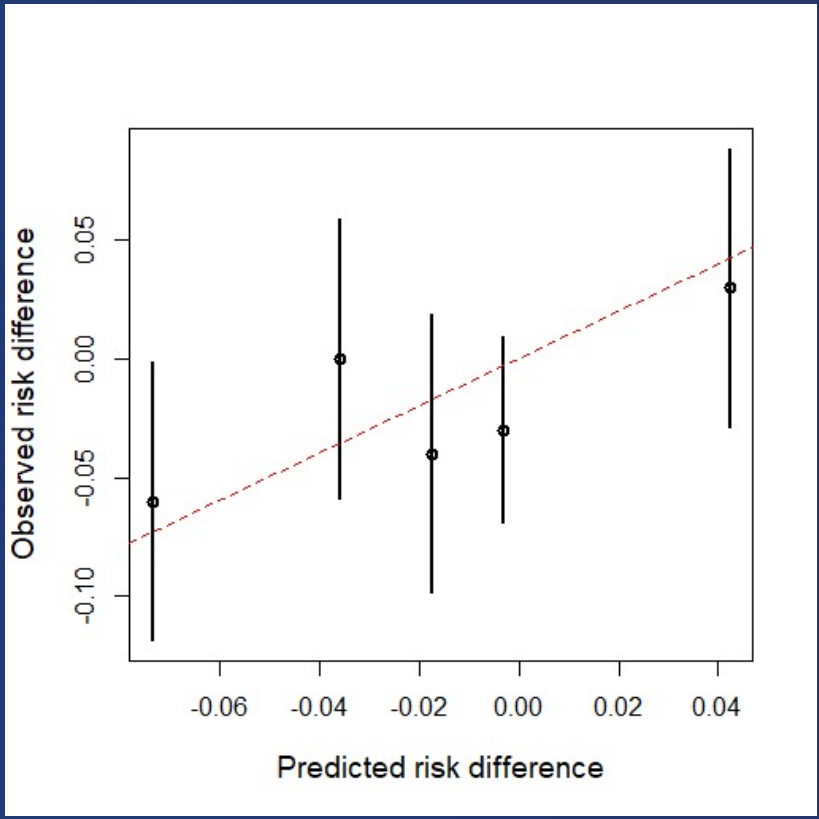


Mahidol University
Wisdom of the Land

Fig 2. Distribution of the individualized effect of aspirin (absolute risk difference).



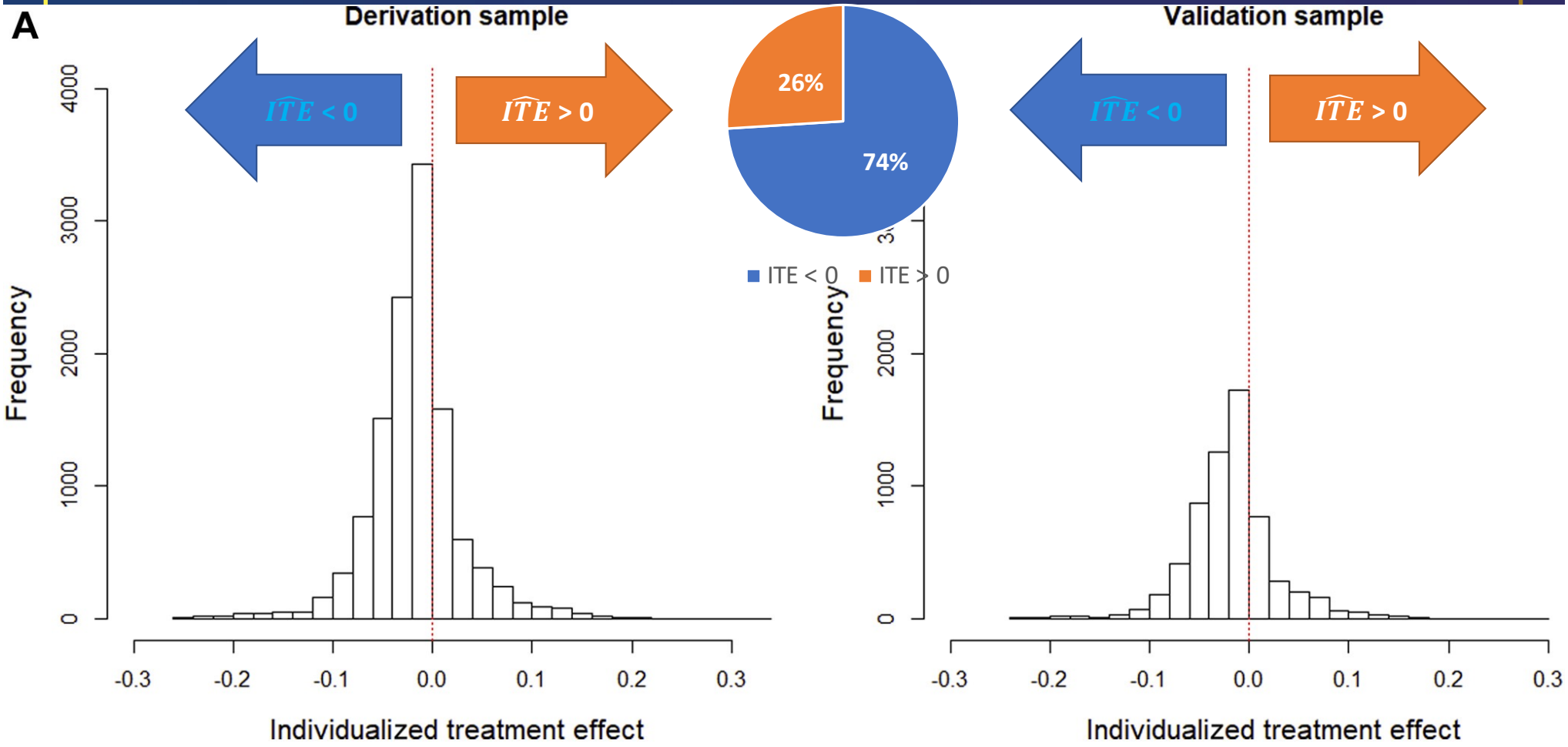
Calibration of the predicted ITE on an absolute risk difference scale



The red dashed line refers to the ideal calibration.

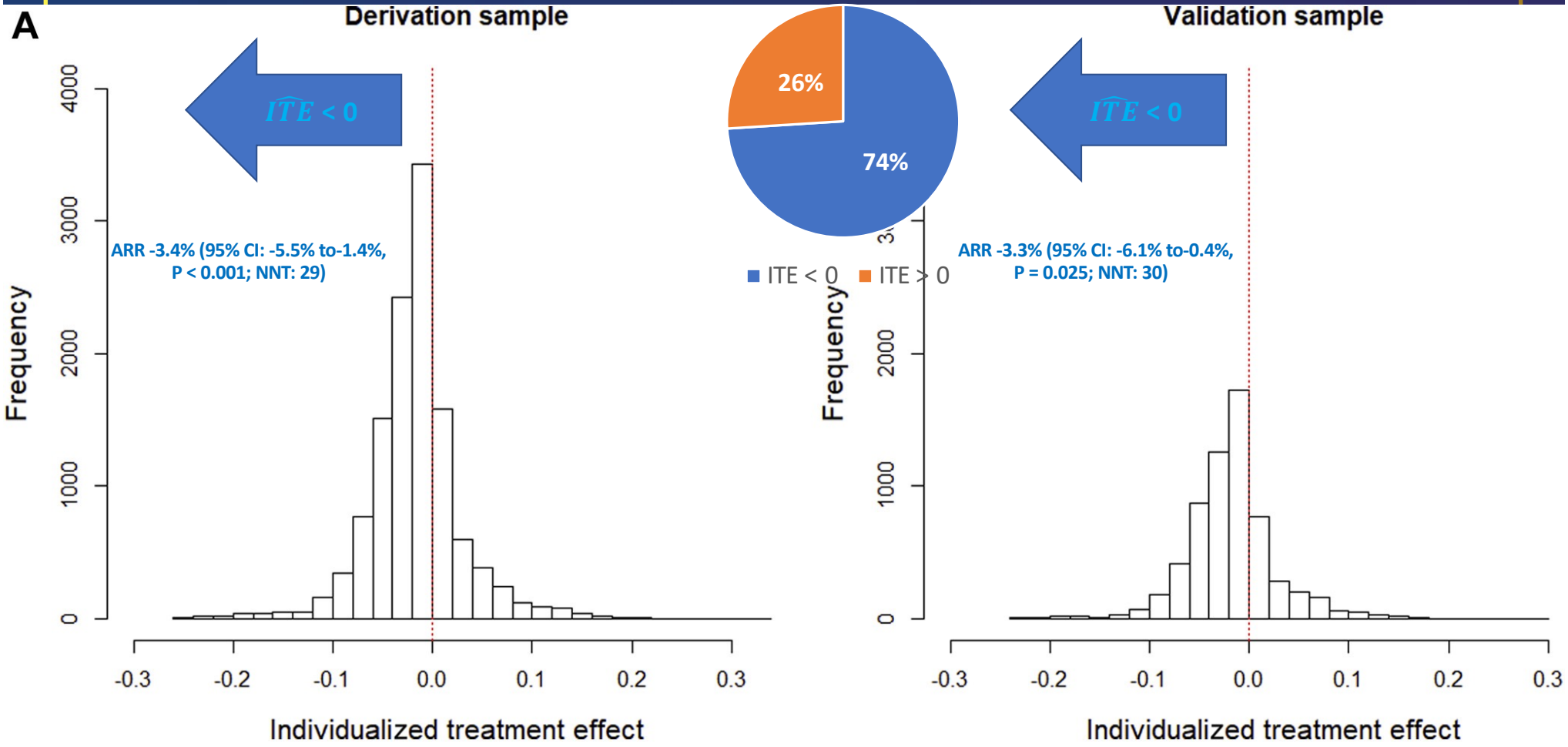
Mahidol University
Wisdom of the Land

Fig 2. Distribution of the individualized effect of aspirin (absolute risk difference).



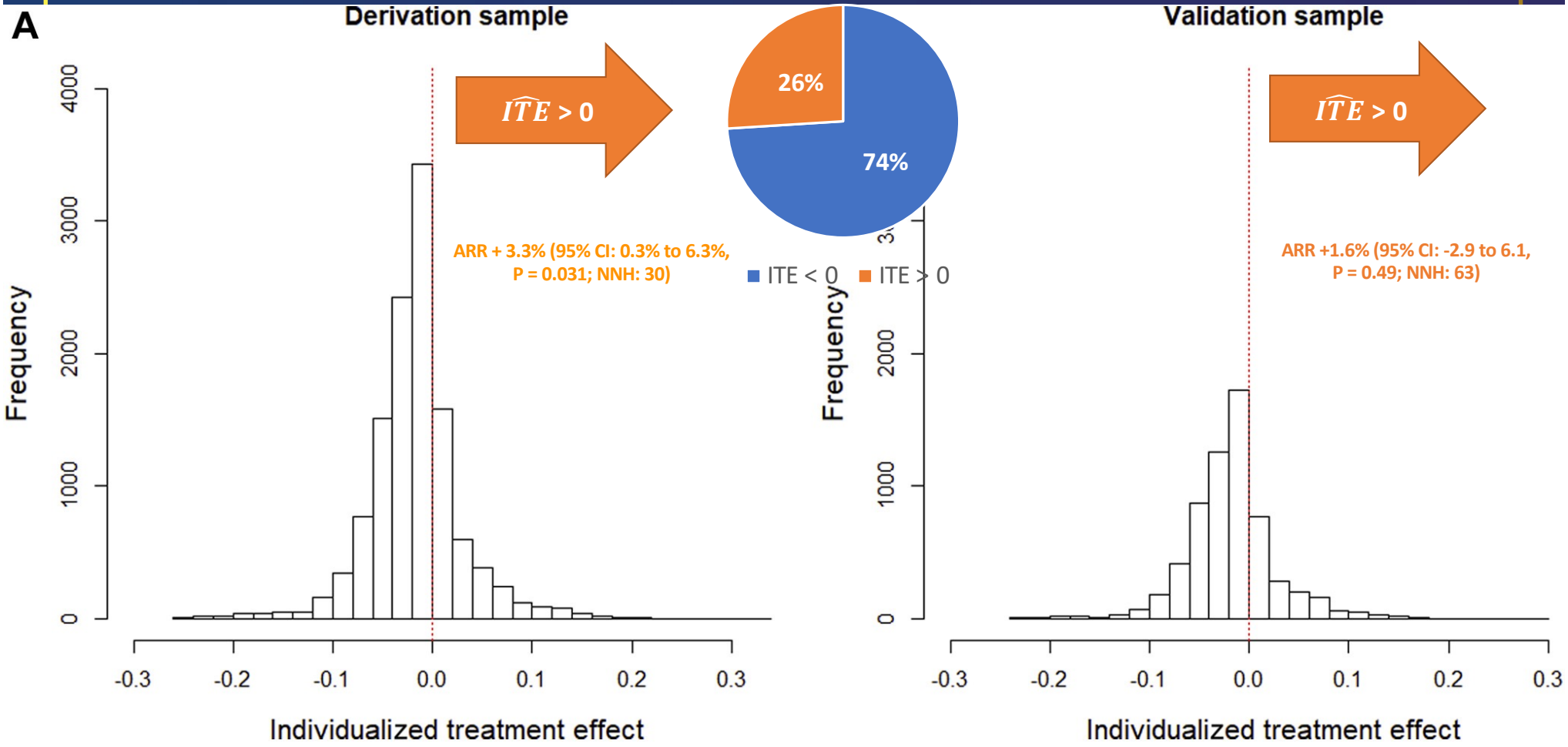
Mahidol University
Wisdom of the Land

Fig 2. Distribution of the individualized effect of aspirin (absolute risk difference).



Mahidol University
Wisdom of the Land

Fig 2. Distribution of the individualized effect of aspirin (absolute risk difference).



Mahidol University
Wisdom of the Land

Discussion

Subgroup analysis	• one variable- at-a-time analyses • may lead to false-positive findings
‘‘multivariable’’ subgroup analysis	• rely on multivariable models using disease risk scores • unable to properly define thresholds • increase the risk of false-positive findings
Counterfactual prediction models	• no need to define thresholds • ITE is directly inferred from comparing the counterfactual risks of outcome

Limitations

1. Model development
   - Models are transparently reported as stated for diagnostic and prognostic research
2. External validation
   - Refine inclusion criteria for secondary trials
3. Impact analysis
   - Require more meticulous practices than usual

Limitations

- Need for large RCTs
- Further studies are needed to explore the robustness of this approach against model misspecification
- ITE
  - evidence inferred in (fine) groups of patients sharing similar characteristics
  - uncertainty due to the gap between groups and individuals

Mahidol University
Wisdom of the Land

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 is new?

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)

Mahidol University
Wisdom of the Land

What is new?

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.