E-values for early decisions in clinical trials

Pilot/feasibility trials and single-arm phase II designs

UNIVERSITY OF TWENTE.

Rianne de Heide | University of Twente and Centrum Wiskunde & Informatica, Amsterdam

Phase II e-trial project: joint work with Nynke Luijten (University of Twente) and Vincent van der Noort (Netherlands Cancer Institute)

CWI Centrum Wiskunde & Informatica

The building block of any-time valid designs: the e-process

For a null hypothesis H_0, an **e-process** $(E_t)_{t\geq 0}$ is a nonnegative process satisfying

$$\sup_{P\in H_0} \mathbb{E}_P[E_\tau] \leq 1 \qquad \text{for every stopping time } \tau.$$

A stopping time is any rule for deciding whether to stop that uses only the information available so far. Thus, even if the decision of when to stop is data-dependent, an e-process cannot have expected value exceeding one when the null hypothesis is true.
Ville's inequality gives the any-time valid guarantee

$$\sup_{P\in H_0} \Pr_P\Big(\sup_{t\geq 0} E_t \geq \frac{1}{\alpha}\Big) \leq \alpha.$$

One may inspect evidence, continue collecting data, or stop at the first crossing of $1/\alpha$, without inflating the Type-I error.

E-value-based designs combine all patient observations into one any-time valid analysis with strong statistical guarantees, supporting timely efficacy and null-acceptance decisions, while accommodating prespecified feasibility or futility rules.

In the pilot/feasibility–definitive setting, this retains early observations and allows the total sample size to be driven by the accumulating evidence; in the single-arm phase-II setting, the e-value design can substantially reduce expected sample size relative to Simon's two-stage designs.

Project 1: pilot/feasibility trials

Goal. We develop a combined pilot/feasibility–definitive framework in which early observations are retained in one any-time valid inferential procedure, rather than being discarded after an external pilot study.

Core idea. Recruit participants one by one or in small batches and work with an e-process.

Example. We use the *Turner e-variable* to test the equality null for two Bernoulli streams observed in blocks of one observation per arm [1]. For completed block j, let $(A_j, B_j) \in \{0,1\}^2$ be the two outcomes and let $\mathcal{F}_{j-1}$ denote the past. Before observing block j, form predictable alternative forecasts $q_{A,j}$ and $q_{B,j}$; here these are Beta–Bernoulli posterior predictive probabilities. Under the equality null, both arms use the pooled prediction

$$q_{0,j} = \frac{q_{A,j} + q_{B,j}}{2}.$$

The blockwise e-variable is

$$L_j = \frac{q_{A,j}^{A_j}(1-q_{A,j})^{1-A_j}\, q_{B,j}^{B_j}(1-q_{B,j})^{1-B_j}}{q_{0,j}^{A_j+B_j}(1-q_{0,j})^{2-A_j-B_j}}, \qquad E_m = \prod_{j=1}^{m} L_j.$$

The pooled construction ensures

$$\mathbb{E}_{H_0}[L_j \mid \mathcal{F}_{j-1}] \leq 1,$$

so $(E_m)_{m\geq 0}$ is an e-process. Thus, rejecting when $E_m \geq 1/\alpha$ is any-time valid.

Conventional trial. In a conventional external-pilot approach, these first observations would be used only to assess feasibility or estimate an event rate; they would not contribute to the later confirmatory treatment-effect analysis. For binary-outcome pilots, a review of registered UK studies found a median target sample size of 50 participants per arm [4], while simulation-based guidance recommends at least 60 per arm to estimate an event rate with adequate precision [5]. A separate definitive trial would still be required.

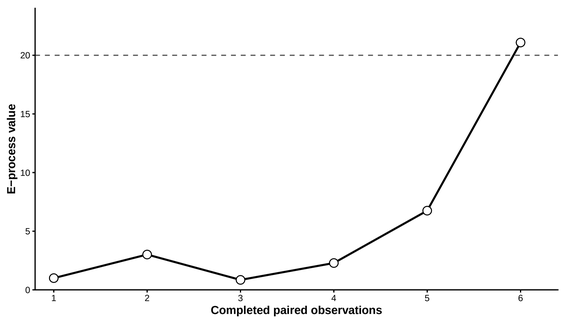
Improvement.

- e-trial: 12 observations in total
- conventional: ≥ 120 observations (recommended for estimating a binary event rate), plus a separate definitive trial

Design aim: learn early, use all data, and preserve an explicit error guarantee.

E-trial

Six alternating-night pairs compared two reusable-nappy boosters. The observed sequence led to $E_6 = 21.09$, exceeding the pre-specified threshold $1/\alpha = 20$ for $\alpha = .05$.



Project 2: single-arm phase II trials

Simon's two-stage design. Let $X_i \sim \text{Bernoulli}(p)$ denote the binary response of patient i. Simon's widely used design (e.g. https://drupstudy.nl) [2] tests

$$H_0 : p \leq p_0 \qquad \text{versus} \qquad H_1 : p \geq p_1.$$

First, enrol n_1 patients. If at most r_1 respond, stop for futility; otherwise enrol $n - n_1$ additional patients. The treatment is declared promising only if more than r responses are observed among all n patients. The four integers (n_1, r_1, n, r) are fixed in advance to control Type-I error at α and attain power $\geq 1-\beta$ at $p = p_1$.

E-design. Fix a design alternative $q > p_0$ (usually $q \geq p_1$). The efficacy e-process is the likelihood ratio for q against p_0. For null-acceptance, use the reverse likelihood ratio, which is an e-process under $p \geq q$. We stop and draw the corresponding conclusion if

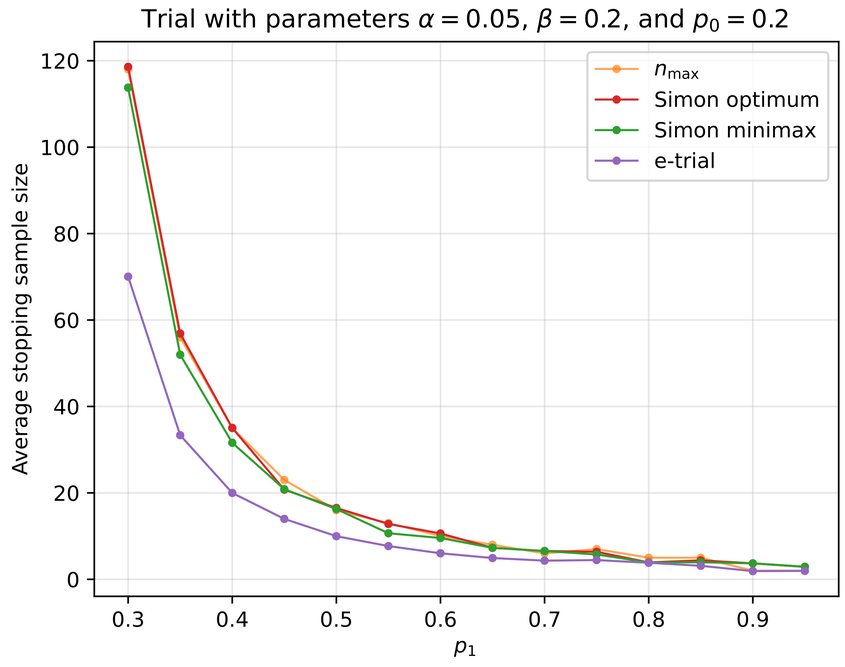
$$E_n^{\text{eff}} \geq \frac{1}{\alpha} \quad \text{or} \quad E_n^{\text{null}} \geq \frac{1}{\beta} \quad \text{or we stop inconclusively at } n_{\max}.$$

Indeed, E_n^{eff} is a nonnegative test supermartingale for every $p \leq p_0$, while E_n^{null} is one for every $p \geq q$. Hence Ville's inequality gives

$$\Pr_{p\leq p_0}\big(\exists n:\ E_n^{\text{eff}} \geq 1/\alpha\big) \leq \alpha, \qquad \Pr_{p\geq q}\big(\exists n:\ E_n^{\text{null}} \geq 1/\beta\big) \leq \beta.$$

No multiplicity correction is required since the two hypotheses tested here are disjoint. No error guarantee is claimed in $p_0 < p < q$.

Simulation: average stopping sample size and planned $n_{\max}$



Simon's design vs. the e-trial

Simon	One planned interim futility decision; a fixed final analysis.
E-trial	Evidence is updated after each patient or batch; efficacy and null-acceptance can trigger at any time.

Why are the observed gains so large? Simon's design can stop only at two fixed analyses. The e-process is any-time valid and may stop as soon as the accumulated evidence crosses an efficacy or null-acceptance boundary.

GROwth of the evidence. Let $Y_i = \log\{f_q(X_i)/f_{p_0}(X_i)\}$ be the efficacy log-likelihood-ratio increment. The efficacy likelihood-ratio e-process is q-GRO for the point alternative $p = q$ (equivalently, W-GRO for $W = \delta_q$) [3]. Under a true response rate p and i.i.d. Bernoulli(p) outcomes, define

$$\mu_{\text{eff}}(p) := \mathbb{E}_p[Y_i] = p\log\left(\frac{q}{p_0}\right) + (1-p)\log\left(\frac{1-q}{1-p_0}\right).$$

This is the *per-patient e-power* of the efficacy e-process:

$$\mathbb{E}_p\big[\log E_n^{\text{eff}}\big] = n\mu_{\text{eff}}(p).$$

In particular, $\mu_{\text{eff}}(q) = D(\text{Bern}(q)\,\|\,\text{Bern}(p_0))$. Since the null-acceptance e-process is the reciprocal likelihood ratio, its per-patient e-power equals $-\mu_{\text{eff}}(p)$. There is a unique $p^\star \in (p_0, q)$ with $\mu_{\text{eff}}(p^\star) = 0$. For $p > p^\star$, the efficacy e-process has positive e-power; for $p < p^\star$, the null-acceptance e-process has positive e-power. Thus, away from $p^\star$, evidence accumulates faster and boundary-crossing times are typically shorter. For the clinically relevant region of the parameter space, this leads to large gains.

Future directions

- **Finite-horizon optimal e-processes.** Characterise e-processes that optimise a finite-horizon objective, such as expected sample size or power at $n_{\max}$, rather than only asymptotic log-growth.
- **Composite $\mathbf{H_1}$.** Mixture or plug-in e-processes tailored to the relevant response-rate ranges.
- **Principled decision framework for efficacy, null-acceptance, futility, and inconclusive stopping.**
- **Realistic trial settings.** Delayed outcomes, time-to-event (surrogate) endpoints, covariate adjustment, unequal batches.
- **Pilot/feasibility trials: stable estimands under adaptation.** Feasibility learning may change recruitment, treatment delivery, outcome measurement, or the target population: develop e-processes that remain conditionally valid under adaptations.

References and funding

1. Turner, R. J., Ly, A., and Grünwald, P. D. (2024). Generic e-variables for exact sequential k-sample tests that allow for optional stopping. *Journal of Statistical Planning and Inference*, 230, 106116.
2. Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10(1), 1–10.
3. Grünwald, P., de Heide, R., and Koolen, W. (2024). Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(5), 1091–1128.
4. Totton, N., et al. (2023). A review of sample sizes for UK pilot and feasibility studies on the ISRCTN registry from 2013 to 2020. *Pilot and feasibility studies*, 9(1), 188.
5. Teare, M. D., et al. (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials*, 15(1), 264.

Both projects are ongoing; preprints are planned later in 2026. Rianne de Heide is supported by NWO Veni grant VI.Veni.222.018.