Decision architecture randomisation: extremely efficient clinical trials that preserve clinician and patient choice?

James Flory ,[1] Jessica S Ancker,[2] Scott Yung Ho Kim ,[3] Gilad Kuperman,[1] Andrew Vickers[1]

10.1136/bmjebm-2023-112386

[1]Memorial Sloan Kettering Cancer Center, New York, New York, USA
[2]Vanderbilt University Medical Center, Nashville, Tennessee, USA
[3]Bioethics, National Institutes of Health, Bethesda, Maryland, USA

Correspondence to:
Dr James Flory, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA; floryj@mskcc.org

To cite: Flory J, Ancker JS, Kim SYH, *et al*. *BMJ Evidence-Based Medicine* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjebm-2023-112386

Introduction

Healthcare is full of choices between standard-of-care approaches where one might be better but we do not know which. Examples include 'at what threshold should magnesium be supplemented for critically ill patients?' and 'which insulin formulation should be started in a hospitalised patient with diabetes?'[1–3]

Observational studies of such questions can be conducted relatively easily but are prone to biases, especially selection bias, that prevent them from reliably showing causal relationships between treatments and outcomes.[4] Randomised controlled trials (RCTs) allow stronger causal inference but are major undertakings, typically costing over US$10 000 per patient.[5] Beyond financial cost, traditional RCTs disrupt care, especially by assigning treatment based on random chance rather than clinicians' and patients' preference. Even for patients who merely consider trial participation, weighing benefits and risks and making a decision may create substantial burdens and stress.[6]

Information technology, with most treatment orders now placed through electronic health records (EHRs), has created opportunities to address these challenges by streamlining processes for participant screening, consent, enrolment, randomisation, intervention delivery and outcome ascertainment.[1 6 7] For pragmatic trials that compare two standard-of-care interventions without masking treatment assignment, these innovations have lowered costs.[8] However, these innovations have not addressed the fact that RCTs require patients and clinicians to prioritise a research study's needs over patient care by accepting a randomly chosen treatment even when the patient would have selected another, given the option.

Decision architecture randomisation trials

The options that an EHR offers patients and clinicians when treatment and other orders are placed, and the way in which options are presented, are known as 'decision architecture', and can influence which option is selected. For example, the first on a list of medications might be more likely to be chosen, or a medication might be less likely to be used if it requires a few extra keystrokes or mouse clicks to order. Features of decision architecture like this, that encourage people to select particular options but do not force them to do so, are known as 'nudges'.[9–11] Defaults—the option that obtains if the decision-maker does nothing—are often effective nudges. For example, if a new prescription defaults to 'morning' versus 'bedtime' for the timing of administration, clinicians may be more likely to write prescriptions for the default timing.

We propose a new type of pragmatic clinical trial that leverages the decision architecture of EHRs to introduce useful randomised nudges into otherwise arbitrary prescribing decisions. The approach uses concepts from nudge research, randomised encouragement trials, and A/B testing.[12 13] We call this a decision architecture randomisation trial (DART).

A DART is intended to compare how standard-of-care treatment approaches (A vs B) impact patient outcomes. As a hypothetical example, consider which of two widely used long-acting insulin formulations—detemir (drug A) and glargine (drug B)—should be preferred in hospitalised patients. Although it is unclear which drug is preferable, one of them could plausibly on average lead to better and safer control of blood glucose. A rigorous comparison could benefit tens of thousands of patients every year, but is not a high enough priority to obtain the resources needed for a traditional RCT. This is especially true if the differences are small, requiring a large and hence expensive trial to discern.

We can introduce randomisation into this decision by randomly presenting half of providers with an order set that has drug A prechecked, and the other half with drug B prechecked. The precheck serves as the default: if the provider accepts the order without altering it, the prechecked drug will be ordered. Even though the default selection is easily over-ridden, the default 'nudges' providers, making them more likely to use the selected agent. The strength of such nudges can be as weak as changing a few percent of decisions or as strong as changing 50% or more.[13 14]

Some EHRs include established processes for randomly assigning providers to different versions of order sets.[15] Such randomisation can be based either on individual patient or provider identity. In situations where individual-level randomisation is infeasible and cluster randomisation is appropriate, different versions of the

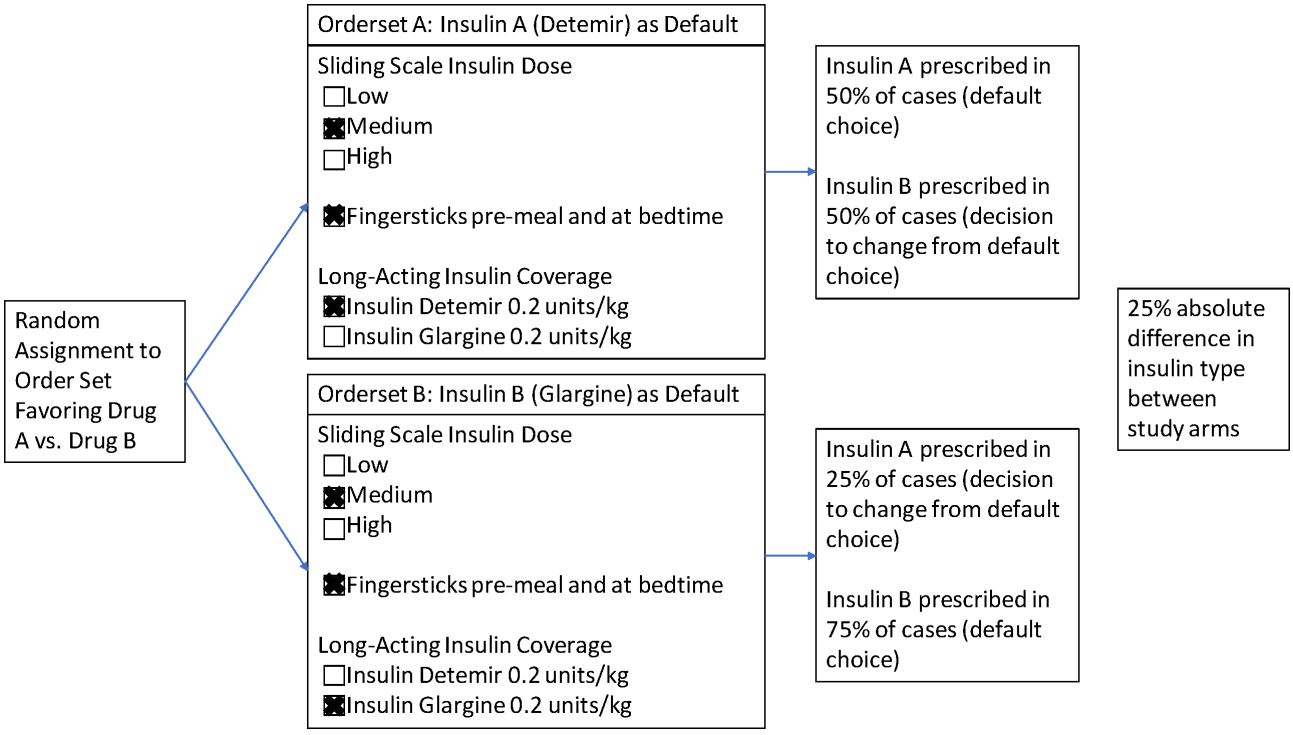

Figure 1 Example of a possible decision architecture randomisation trial.

order set could be randomly assigned to geographic clusters such as hospital floor.

Once randomisation is achieved, the defaults randomly nudge providers towards drug A or drug B (figure 1). The association between the assignment and the actual drug used may be weak, because the provider could easily order the alternative. But, even associations as weak as 10%–20% can be used as an instrumental variable (IV) to estimate an unbiased effect of treatment.[16]

The use of a weak association between randomisation and treatment as an IV is perhaps the least novel aspect of a DART. Statistically, DART is just a 'randomised encouragement' trial, in which randomisation encourages, but does not mandate, an intervention.[12] The IV approach to analysing such trials has been well developed. Roughly speaking, the IV estimates the true intervention effect size by dividing the observed effect size by the absolute difference in treatment rates. For example, a strong nudge (ie, one that is not often over-ridden by clinician or patient preference) might make the use of drug A 50% higher in arm A than in arm B. This means any difference in outcomes between arms is diluted by half, so in IV analysis the observed effect size would be doubled to get the true effect. If the nudge were weaker (ie, more often over-ridden) and resulted in a 20% difference in treatment assignment, the observed difference would be multiplied by 5. The strength of the nudge as an IV is a consequence of how often the nudge is over-ridden and can be assessed by calculating how often treatment selection aligns with the direction of the nudge.

The difference between DART and typical randomised encouragement trials is that most randomised encouragement is explicit. Classic examples include mailed pamphlets or text messages; in the EHR setting, examples include explicit pop-up messages disclosing that a trial is underway and requesting the provider consider changing treatment.[1 12 17] In contrast, DART relies on unobtrusive nudges that do not call for extra deliberation or intentional deviation from routine practice and can be scaled without disrupting care. These qualities make DART resemble A/B testing, a type of rapid and unobtrusive RCT used in business to determine whether a design change, especially on a digital platform, changes behaviour.

An example of A/B testing is showing two different versions of a news headline and measuring which one results in a higher rate of 'clicks' on the article.[18] In healthcare, an example is comparing different EHR nudges to encourage providers to vaccinate patients for influenza.[13] DART differs from such A/B testing in that, instead of studying how effective the nudge is, DART uses the nudge to generate a difference in rates of use of some clinical intervention, then uses IV analysis to extract the effect of the intervention on a clinically important outcome. For example, a study that used nudges to create groups with different rates of influenza vaccination and then measured the impact of vaccination on mortality would be a DART.

Another attribute DART shares with A/B testing is automated data collection. Baseline characteristics and outcomes of participants should be derived from routinely collected data in the EHR.[1 7] The resulting trial should place no burden on patients or providers, allowing it to be run at any length of time or scale with no additional cost.

Risk and consent for DART

Because DART enrolment and randomisation occur when a nudge is delivered and the nudges are likely imperceptible, study-specific informed consent is likely infeasible ('impracticable', in regulatory terminology) for many DARTs, making it imperative to explore under what circumstances individual patient consent could ethically be waived. US regulations specify three more criteria, in addition to impracticability, that must be met for such a waiver. Of these, the requirement that the proposed study not exceed minimal research risk has received the most scrutiny.[19] Minimising risk to participants to achieve this standard is a strength of DART.

The way that DART minimises risk is that randomisation only delivers a nudge, which does not preclude choice. It is important

that nudges be easily over-riden—ideally with a single click—so even a weak treatment preference can take precedence over the nudge. It is crucial that the design leaves patient and provider autonomy intact, preserving clinically motivated decision-making. Since both treatments will already be in clinical use (by design), and since the postnudge treatment assignments are consistent with patient and provider preference, one can argue that risks from DART are routine clinical risks, with negligible research risk.[20]

Sample size

The price of the weakened association seen with an IV is loss of statistical precision. The weaker the nudge, the larger the sample size needed. If a standard RCT would require 100 patients in each arm, a DART with a 50% difference in treatment rates between arms would require approximately 400 patients per arm and a DART with a 25% difference would require 1600 patients per arm.[16]

The fact that a DART using a weak nudge would require a much larger sample size than a traditional RCT appears to be a disadvantage, until one considers that a traditional RCT may have to contact dozens of patients for each that is actually enrolled.[21] In contrast, DART enrols and randomises every eligible patient automatically when their provider encounters the nudge. In contrast to alternative designs that use explicit prompts, DART does not create any new cognitive burden or 'alert fatigue'. With no cost or burdens from enrolment, screening, randomisation, treatment or outcome ascertainment, there is no marginal cost to add additional patients. For many clinical questions these efficiency gains should more than make up for increased sample size requirements. Because DART trials should accrue rapidly, an increased sample size should not importantly delay publication of trial results.

What we still need to know

Many assumptions about the DART's feasibility are supported by evidence. For instance, the literature supports that it is feasible to modify EHRs to deliver reasonably strong nudges reaching large enough populations for relevant clinical questions.[11 13 14] Practical experience implementing DART may be the best way to uncover unexpected challenges to its underlying assumptions amidst the complexities of real-world information technology implementation and clinical practice. Conversely, such experience might show that DART could be deployed very frequently and at large scales in a learning health system.[22]

If DART in practice is very easy to implement, risk of publication bias may be high if large numbers of negative DARTs are not published. Requiring advance registration of DARTs could mitigate this issue.

Other questions pertaining to DART's safety and acceptability deserve direct methodological study before DART becomes widely used. One possible problem would be if providers assume defaults in the EHR never change and sign off on orders without inspecting them. This might be especially likely with inexperienced or overworked providers, or for order sets that are standardised to the point that providers are discouraged from reviewing and personalising the orders. DART designs should be discussed with the providers affected to anticipate and avoid such risks.

Past uses of A/B testing for research outside healthcare have attracted criticism for inadequate oversight, and DART should be done with empirically informed ethical guidelines.[23] Preliminary engagement with patients and caregivers suggests that they find DART acceptable but would like at least a broad up-front notification that DART studies are ongoing. Engagement with providers and information technology specialists has also highlighted the importance of engaging and notifying those stakeholders when a DART is planned. Our group is conducting a series of democratic deliberative sessions with patients, providers and institutional officials to more rigorously elicit stakeholder views.[24]

Conclusions

DART uses unobtrusive nudges to produce randomisation and compare standard-of-care treatments without compromising provider–patient decision-making, thus respecting their relationship and the patient's interests. DART may have ethical and practical advantages over traditional trial designs, but the first obligation of researchers is to make sure randomisation of decision architecture is acceptable to patients and providers.

Contributors JF is a clinical epidemiologist and endocrinologist and is the writer and guarantor of this manuscript. JSA edited the manuscript as an expert in the optimal use of information and information technology to improve decision-making. SYHK edited the manuscript as an expert in bioethics. GK edited the manuscript as an expert in information technology and healthcare. AV edited the manuscript as an expert in clinical trial design.

Funding Funding from the Patient Centered Outcomes Research Institute (ME-2022C1-26378) is gratefully acknowledged.

Competing interests None declared.

Patient and public involvement statement This proposal was discussed by a group of approximately 15 patients and caregivers at a meeting at Memorial Sloan Kettering's Patient and Family Advisory Council for Quality. This input is gratefully acknowledged and briefly summarised in the main text.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
James Flory http://orcid.org/0000-0002-0259-969X
Scott Yung Ho Kim http://orcid.org/0000-0002-9444-4627

References

1 **Wilson MG**, Asselbergs FW, Miguel R, *et al*. Embedded point of care randomisation for evaluating comparative effectiveness questions: prospector-critical care feasibility study protocol. *BMJ Open* 2022;12:e059995.
2 **Hochberg I**. Insulin detemir use is associated with higher occurrence of hypoglycemia in hospitalized patients with hypoalbuminemia. *Diabetes Care* 2018;41:e44–6.
3 **Danforth KN**, Patnode CD, Kapka TJ, *et al*. Comparative effectiveness topics from a large, integrated delivery system. *Perm J* 2013;17:4–13.
4 **Freemantle N**, Marston L, Walters K, *et al*. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 2013;347:6409.
5 **Speich B**, von Niederhäusern B, Schur N, *et al*. Systematic review on costs and resource use of randomized clinical trials shows a lack of transparent and comprehensive data. *J Clin Epidemiol* 2018;96:1–11.
6 **Lee SJC**, Murphy CC, Geiger AM, *et al*. Conceptual model for accrual to cancer clinical trials. *J Clin Oncol* 2019;37:1993–6.
7 **Fiore LD**, Brophy M, Ferguson RE, *et al*. A point-of-care clinical trial comparing insulin administered using a sliding scale versus a weight-based regimen. *Clin Trials* 2011;8:183–95.
8 ADAPTABLE results. n.d. Available: https://www.pcori.org/research-results/2015/comparing-two-aspirin-doses-prevent-heart-attacks-and-strokes-people-living-heart-disease-adaptable-study

9 **Johnson EJ**. *The elements of choice: why the way we decide matters*. New York: Riverhead Books, an imprint of Penguin Random House LLC, 2021: 390.

10 **Thaler RH**, Sunstein CR. *Nudge: the final edition*. New Haven: Yale University Press, 2021: 366.

11 **Raban MZ**, Gates PJ, Gamboa S, *et al*. Effectiveness of non-interruptive nudge interventions in electronic health records to improve the delivery of care in hospitals: a systematic review. *J Am Med Inform Assoc* 2023;30:1313–22.

12 **Green EP**, Augustine A, Naanyu V, *et al*. Developing a digital marketplace for family planning: pilot randomized encouragement trial. *J Med Internet Res* 2018;20:e10756.

13 **Horwitz LI**, Kuznetsova M, Jones SA. Creating a learning health system through rapid-cycle, randomized testing. *N Engl J Med* 2019;381:1175–9.

14 **Malhotra S**, Cheriff AD, Gossey JT, *et al*. Effects of an e-prescribing interface redesign on rates of generic drug prescribing: exploiting default options. *J Am Med Inform Assoc* 2016;23:891–8.

15 **Mutter M**, Martin M, Yamamoto Y, *et al*. Electronic alerts for acute kidney injury amelioration (ELAIA-1): a completely electronic, Multicentre, randomised controlled trial: design and rationale. *BMJ Open* 2019;9:e025117.

16 **Walker VM**, Davies NM, Windmeijer F, *et al*. Power calculator for instrumental variable analysis in pharmacoepidemiology. *Int J Epidemiol* 2017;46:1627–32.

17 **Flory J**, Karlawish J. The prompted optional randomization trial: a new design for comparative effectiveness research. *Am J Public Health* 2012;102:e8–10.

18 **Bulik M**. Which headlines attract most readers. *New York Times* 2016.

19 **Kim SY**. Comparative effectiveness research, learning health systems, and pragmatic randomized controlled trials. In: **MD IA**, ed. *The oxford handbook of research ethics*. Oxford University Press, 2020.

20 **Chen SC**, Kim SY. A framework for analysis of research risks and benefits to participants in standard of care pragmatic clinical trials. *Clin Trials* 2016;13:605–11.

21 **McCarney R**, Fisher P, van Haselen R. Accruing large numbers of patients in primary care trials by retrospective recruitment methods. *Complement Ther Med* 2002;10:63–8.

22 **Wouters RHP**, van der Graaf R, Voest EE, *et al*. Learning health care systems: highly needed but challenging. *Learn Health Syst* 2020;4:e10211.

23 **Verma IM**. Editorial expression of concern: experimental evidence of massive-scale emotional contagion through social networks. *Proc Natl Acad Sci U S A* 2014;111:10779.

24 Available: https://www.pcori.org/research-results/2022/can-ab-testing-be-adapted-ethical-and-useful-approach-patient-centered-outcomes-research