Med Health Care and Philos (2018) 21:335–345
DOI 10.1007/s11019-017-9807-9

SCIENTIFIC CONTRIBUTION

# Patient-specific devices and population-level evidence: evaluating therapeutic interventions with inherent variation

Mary Jean Walker[1,2]

Published online: 19 October 2017

**Abstract** Designing and manufacturing medical devices for specific patients is becoming increasingly feasible with developments in 3D printing and 3D imaging software. This raises the question of how patient-specific devices can be evaluated, since our 'gold standard' method for evaluation, the randomised controlled trial (RCT), requires that an intervention is standardised across a number of individuals in an experimental group. I distinguish several senses of patient-specific device, and focus the discussion on understanding the problem of variations between instances of an intervention for RCT evaluation. I argue that, despite initial appearances, it is theoretically possible to use RCTs to evaluate some patient-specific medical devices. However, the argument reveals significant difficulties for ensuring the validity of such trials, with implications for how we should think about methods of evidence gathering and regulatory approaches for these technologies.

**Keywords** Medical device · Regulation · Randomised controlled trial · Personalised medicine · 3D printing

Making and using medical devices that are designed for specific patients is becoming increasingly feasible due to developments in 3D imaging software and 3D printing. These technologies enable custom manufacture that is precise, highly detailed, comparatively fast, semi-automated (rather than surgeon-sculpted), and potentially low-cost, prompting some to discuss them as introducing an era of 'mass customisation' (e.g., Wallace et al. 2014).[1]

But how are we to evaluate the safety and performance of such devices? Using patient specific devices is motivated by the idea that it will lead to better outcomes than using standardised devices, and there are some convincing rationales for this. These may be based partly on pathophysiological reasoning. For example, clinical outcomes for knee implant patients are known to be better where the implant achieves a hip-knee-ankle angle of 180 ± 3°. With standardised devices over a quarter of knee replacements miss this optimal outcome (Ng et al. 2012, p. 100). Since the angle achieved is affected by small deviations in the shape or positioning of the implant, we might expect that using implants and surgical guides designed to fit the specific patient will improve accuracy. Similarly, we might reason that replacing a removed heel bone with an implant designed to precisely mirror the shape of the patient's opposite heel bone, being a better fit for the patients' physiology than any other shape, would lead to a better outcome (Imanishi and Choong 2015).

✉ Mary Jean Walker
mary.walker@monash.edu

[1] Ethics Program, ARC Centre of Excellence for Electromaterials Science, Monash University, Clayton, VIC, Australia

[2] Philosophy Department, School of Philosophical, Historical and International Studies, Monash University, Building 11, Clayton, VIC 3800, Australia

[1] Patient-specific medical devices can be considered part of recent trends toward 'personalised medicine', which seeks to provide treatments targeted for specific patients. 'Personalised medicine' most often refers to methods for identifying which pharmacological treatments are likely to work for specific patients by developing drug efficacy data for sub-populations, and methods for identifying which sub-populations specific patients belong to (e.g., genetic tests), such that it relies upon stratified population-level evidence. The personalisation discussed here is somewhat different, referring rather to adapting treatments (surgical or pharmaceutical) to suit the needs of particular patients, or designing treatments for particular patients from scratch (which is not generally feasible for pharmaceutical treatments, but would include bespoke surgeries and prosthetics).

But clinical decision-making should be based on robust evidence, and current paradigms define population-level studies, particularly the randomised controlled trial (RCT), as being the most robust. Approvals of therapeutic goods for sale or use by regulatory authorities rely heavily on this kind of evidence, so it is often required in order to bring new technologies to market, and enable patients to access patient-specific devices on anything more than a small scale.[2] RCTs require giving the same intervention to a group of people, and then derive conclusions based on averaged outcomes (in comparison with a control group). Prima facie, this does not seem possible for patient-specific devices. The lack of fit between current evidence paradigms and patient-specific devices seems to call for a different approach to regulation for these technologies, in order to ensure that their benefits reach patients, while patients remain protected from devices whose safety and performance are not established.[3]

In this paper I examine this lack of fit with an epistemological focus. In investigating whether RCTs could be used to evaluate patient-specific devices, I aim to extend theoretical discussion of the implications of variation between instances of a tested intervention in RCTs.[4] I first distinguish several senses in which devices can be made specific to patients, and will limit the scope of my argument to what I call 'custom-made' devices, in the section on "Kinds of patient-specificity". In the section on "Inherent variation", I overview barriers to RCT assessment of custom-made devices. While most of these barriers can be mitigated or addressed, I argue that variation between instances of a tested intervention poses not only a practical, but a theoretical problem for RCTs of custom-made devices where the variations are believed to benefit patients. Such 'inherent' variation in an intervention is generally recognised to be problematic for RCT assessment, but exactly when and why it is problematic is rarely analysed in any detail. In the section on "Why is variation problematic?", therefore, I develop an argument that while RCTs require that the same intervention is given to individuals in the experimental group in order to have internal validity, what will count as 'the same intervention' depends on identifying the most accurate description of that intervention. I argue that the most accurate description of some patient-specific medical devices, once identified, will be standardisable across a population, and thus RCTs could be used to assess some patient-specific devices.

But the argument also reveals that there will be significant difficulties in doing so. In the section on "Implications for evaluation and regulation of custom-made and other patient-specific devices", I conclude by noting implications of the argument for evaluation and regulation of custom-made devices, and for approaching other kinds of patient-specific device.

## Kinds of patient-specificity

Adapting treatments for particular patients is part of usual practice. Practitioners adjust dosages of pharmacological treatments on the basis of patient needs (e.g., condition severity) or characteristics (e.g., body weight); surgeons vary interventions to take anatomical or other patient differences into account; many medical devices come in standardised 'off-the-shelf' forms, but with variations (such as in size) intended to allow surgeons to select the device most suitable for individuals. Patient-specific medical devices represent a further way in which interventions can be patient-specific, and this may be achieved in several different ways (see Table 1).[5]

Some 'off the shelf' devices can be, or are intended to be, modified or adapted for specific patients. For instance, a breast implant may be modified to achieve a particular size by reducing volume, and hearing aids are typically customised to the shape of a patient's ear as well as calibrated to their hearing loss. Wheelchairs may be modified in size and shape, or have attachments added or removed, for specific patients. These devices can be referred to as 'customisable'.[6] These are often treated similarly to off-the-shelf devices in regulatory terms, and are typically low-risk; they will be left aside for the purpose of the current argument.

Alternatively, devices may be specifically designed and manufactured for single patients. Among these, some are designed using a standard template which is then adapted for individuals. I call these 'custom-made' devices. Custom-made devices remain standardised in some senses: each has a standard structure or general shape, and materials and production methods are standardised. The procedure for implanting and/or using them may also be standardisable.

[2] Regulatory systems currently allow patient-specific devices to bypass regulatory controls under some conditions (see, e.g., FDA 2014; TGA n.d.). These exemptions typically limit their use to small numbers of people (e.g., less than 5 units per year for the FDA), and more regulatory controls will be appropriate if use is to be scaled up.

[3] Greater use of custom manufacture may also require changes to other kinds of regulatory control, such as different approaches to quality assurance in manufacturing, although I shall not discuss this issue.

[4] Variation in interventions is of course not limited to patient-specific devices, so this argument has relevance for discussions of the use of RCTs in other areas where standardisation is difficult or problematic, such as surgery and complex interventions.

[5] I do not consider diagnostic devices.

[6] This term and the distinction with custom-made devices are borrowed from Australia's Therapeutic Goods Administration (TGA n.d.).

**Table 1** Kinds of patient-specific device

| | Customisable | Custom-made | Bespoke | Tissue-engineered |
|---|---|---|---|---|
| Definition | Available 'off the shelf' but can/are intended to be modified for specific patients | Designed and manufactured for specific patients using a standard template with some variable parameters | Designed and manufactured for specific patients without a template | Created using specific patients' own cells via in vitro tissue culture and/or 3D bioprinting |
| Patient-specific elements | Modifications to size, shape, function | Specific size, shape, function | Size, shape, function | Genetic profile |
| Standardisable elements | Design, manufacture, procedure for customisation, materials used | Template, procedure for design and manufacture, general size and structure, materials used | May be possibilities for standardising (within particular fields of application) of (e.g.) materials used, manufacturing/use procedures, relation of structure of device to patient anatomy | Dependent on how technologies progress, but may include: procedures for manufacture and implantation, non-cell materials, device structure |
| Examples | Hearing aids; some dental implants; wheelchairs | Hip implants; knee implants | 3D printed calcaneus (Imanishi and Choong 2015); 3D printed tracheal splint (Zopf et al. 2013) | Artificial bladders; 3D printed skin, bone, cartilage |

Using custom-made devices on any scale is difficult with traditional manufacturing, due to longer manufacturing times and limitations on pre-operative custom design, so this is an area in which new fabrication methods may have significant impacts. An example is custom-made hip implants. Hip implants have a femoral stem which is inserted into the hollow canal of the patient's femur, an acetabular shell that attaches to the acetabulum, and a liner that sits between these. Custom-made femoral stems have been investigated since the 1980s (when they were manufactured conventionally) (Bargar 1989; Viceconti et al. 2001). There was evidence that using bone cement to fix the implant into the femur could cause complications (such as bone necrosis and osteolysis), but cementless implants require a close fit between the implant stem and the femoral canal (Viceconti et al. 2001). This can be achieved either by cutting the bone to fit an off-the-shelf implant, or by using a custom-made implant, and the latter has often been thought preferable for retaining bone strength (Bargar 1989; Grant et al. 2005). Custom-made implants have also been proposed to have advantages related to stress concentration, load distribution (Colen et al. 2014; Muirhead-Alwood et al. 2010), and primary stability (Viceconti et al. 2001).

A third kind of patient-specific device is, like the above, designed and manufactured for single patients, but without any pre-existing template. I refer to these as 'bespoke' devices.[7] This kind of patient-specificity is one of the most challenging from a regulatory perspective. An example is a heel bone designed for and implanted into a 71-year-old patient in Melbourne, Australia in 2014. The patient's own heel bone was removed due to cancer. This can lead to amputation, and other methods of reconstruction using bone allografts or autografts frequently involve complications (Imanishi and Choong 2015). The prosthesis was 3D printed, with the design based on imaging of the patient's other heel bone. In the absence of a clinical precedent, the surgeons, working with a manufacturing company and Australia's Commonwealth Scientific and Industrial Research Organisation, developed the design by calculating from known properties of various materials in relation to the patient's needs in terms of weight and movement (Imanishi and Choong 2015, p. 84).

A final sense of patient-specificity involves the use of patients' own cells in creating tissue-engineered tissues or organs. Tissue engineering involves growing tissues in vitro for implantation into patients. It can utilise cells' self-organising capacities, and/or 3D bioprinting, where cells can be printed into particular configurations (usually within a hydrogel and utilising a 3D scaffold) (Marro et al.

[7] The custom-made/bespoke distinction and terms are from Chadwick (2014).

2016, p. 8). The creation of entire organs using these techniques is under investigation, and tissue-engineered bladders have been implanted into patients (Atala 2009). These technologies involve a further sense in which the tissue/organ is patient-specific, since it shares their genetic profile.

The argument below will focus on custom-made devices. These are simpler than bespoke devices from an epistemic perspective, since some of their elements are standardisable, suggesting that their safety and performance might be assessable via some of our usual methods, including for regulatory purposes. And they are a simpler starting point for consideration than tissue-engineered products, since the latter are still very experimental, and indeed may not be classified as devices for regulatory purposes in all jurisdictions. Custom-made devices are thus a good starting point for consideration. I discuss some implications of the argument for approaching evaluation of these other kinds of patient-specific device in the section on "Implications for evaluation and regulation of custom-made and other patient-specific devices".

## Inherent variation

Attempts to trial custom-made devices face a number of barriers.[8] The high evidential status of RCTs derives from methodological features that remove potential sources of bias and prevent confounding of results: use of a control group, blinding, and random allocation of participants to groups. In this section I explain these methodological features, show why they will be difficult to achieve in trials of custom-made devices, and identify what I call *inherent* variation as the main theoretical barrier to evaluating custom-made devices using RCTs.

RCTs have (at least) two groups of research subjects, an experimental group who receive the experimental intervention, and a control group who receive either placebo or a standard treatment. Researchers then compare outcomes between these groups to isolate what outcomes are due to the intervention, rather than to other factors introduced by participating in the study (e.g., placebo effects, incidental care received in the course of participating, and so on). Blinding of participants and researchers as to which participants receive the intervention and which do not prevents this knowledge itself from influencing outcomes, or the assessment of outcomes. Random allocation of participants prevents allocation bias (allocating patients more likely to benefit from the intervention to the experimental group) and helps prevent un-blinding, but more importantly, controls for any factors, other than the intervention, that could affect whether different outcomes are obtained in the treatment and control groups. Randomisation is used in recognition that although the same intervention is given to each participant, outcomes will be affected by various individual factors that impact on the effects of the intervention, such as individuals' age, sex, dietary habits, and so on. Results could thus be skewed if there are different rates of such 'confounding' factors in the two groups. Randomisation (along with large enough study groups) ensures that all confounders, both known and unknown, are likely to be evenly distributed in treatment and control groups (for more detail, see Cartwright 1994, p. 62–64, 2010, p. 63–64; Worrall 2011, p. 235–236).

Using these methods, RCTs create a situation where the best explanation for differences in outcome between the two groups of participants is the experimental intervention. RCT results thus provide a measure of the 'efficacy' of an intervention, or how well it works in experimental situations. Measures of it are derived from RCT results by obtaining mean outcome measures in the experimental group, and adjusting for outcomes in the control group (Cartwright 2009, p. 187–188). Since efficacy is an average, it is a population-level measure, and probabilistic: it is consistent with an intervention's being efficacious that it has differing outcomes for particular individuals, some of whom might have no or a negative effect (Cartwright 2009, p. 188). Further, an RCT provides us with information about efficacy without requiring us to know anything about how the intervention achieves its outcomes, what the causal mechanisms at work are (Howick 2011, p. 928). This is a significant advantage of the method given the complexity of biological systems.

These methodological features are difficult to implement for some kinds of intervention, however. This is a well-known problem in relation to surgical procedures and complex (behavioural or community-level) interventions. The literature on barriers to surgical RCTs is thus instructive.[9] First, there are often difficulties having control groups in surgical trials. Placebo surgery is sometimes impractical, and often raises ethical concerns, since it can expose patients to significant risks (such as anaesthesia or incisions) without compensating benefits. Standard treatment comparators can be used if available, though in cases where the standard treatment differs significantly from the experimental intervention, this can prevent blinding. For custom-made devices, standard treatment comparators will often be possible [e.g.,

[8] There are also likely to be various practical barriers, e.g., less available funding [as in surgical trials (Garas et al. 2012)]. In order to focus on the theoretical tension between RCT evidence and custom-made devices, I leave these practical barriers aside.

[9] For further discussion of the barriers discussed in the following three paragraphs, as well as other barriers, see, e.g., Cook (2009), Garas et al. (2012), Lassen et al. (2012), Meshikhes (2015), Stirrat (2004). Some intersections between my argument and literature on complex interventions is noted below.

Grant et al. (2005) and Small et al. (2014) undertook such comparative studies of hip implants].

Next, complete blinding of surgeons and surgical teams is not possible, though they can be blinded up until the point of surgery [as reported in Small et al. (2014, p. 2031)]. If sham surgery is practical and ethically acceptable, or where a standard treatment comparator uses similar procedures, patients can be blinded, although this leaves open the possibility of surgeons (or others involved in follow-up care) unintentionally indicating to patients what procedure they received. Outcomes assessments can also be undertaken by other, blinded researchers. These partial blinding measures can reduce the chance of bias, without entirely removing it.

These barriers can thus be addressed, at least to some extent, in RCTs of custom-made devices. However, a further problem is that surgical procedures are difficult to standardise, such that members of the experimental group receive slightly different interventions. Variations can be due to patient differences, surgeon training and preference, available resources and support, refinements made to a procedure, or the learning curve involved as surgeons learn a procedure new to them (Cook 2009; Meshikhes 2015, p. 161; Pope 2002).

Again, there are methods that can reduce variation, in order to run surgical RCTs: the protocol could require surgeons to use particular methods; restrict how many surgeons are involved; or require surgeons to have reached a level of expertise with a procedure, for example. Of course, using such methods may reduce the generalisability of the RCT results, since it will mean the RCT protocol delivers the intervention in a way that is less reflective of clinical practice. The problem of variation in a tested intervention could thus be taken to cross over with a more general issue, that there is a trade-off between an RCT's internal validity (how well it avoids confounding, and so supports a conclusion of the efficacy of a treatment) and its external validity (how well it supports generalising conclusions of efficacy from the study population to other populations). That is, the methods that support internal validity often reduce external validity, and vice-versa.

This trade-off is reflected in the distinction between 'explanatory' and 'pragmatic' RCTs. The former aim to acquire information on the biological effects of treatments, and require conditions to be as similar as possible between the experimental and control groups (Schwartz and Lellouch 1967, p. 638). But RCTs can also be designed to focus on answering questions of clinical practice, and in such trials it can be appropriate to match conditions in the trial to those of clinical practice. These include variation amongst patients, and lower compliance with treatment protocols, so pragmatic trials usually have fewer exclusion criteria, and less strict protocols.[10] Thus in a pragmatic RCT, more variation between instances of a tested intervention is permissible. While this may lower its capacity to be informative about the biological effects of an intervention, it makes the results more generalisable.

But there are differences between pragmatic RCTs and potential RCTs of custom-made devices which indicate that we should not see the latter as simply one kind of the former. With custom-made devices, the variations to the intervention are introduced, not to reflect clinical practice and increase external validity, but to test a hypothesis that variations which aim to 'match' the device to each individual patient will improve overall outcomes. As such, an RCT of custom-made devices could be designed as either an explanatory or a pragmatic trial,[11] and the kind of variation at stake is different. Notice that most of the sources of variations in an intervention listed above—surgeon training and preference, available resources and support, refinements made to a procedure, and the learning curve—pose a practical problem for RCTs, because they make standardisation difficult to achieve. But many of the variations that are made in response to differences between patients also pose a theoretical problem: they do not just make standardisation *difficult*, they make it *undesirable* (at least, such is the hypothesis being tested). With custom-made devices, as with some other sorts of variations introduced in surgical and complex interventions, variations are introduced because they are considered to be beneficial for particular patients—even though they would not be beneficial for every patient. I will call this *inherent* variation.

Inherent variation, like other variation, could be problematic in an RCT in reducing internal validity. However, the problem of inherent variation differs from that of other variations. The issue is not that of designing a trial along pragmatic or explanatory lines suitably for the trial purpose, but the abstract question of how to assess when variations in a tested intervention are problematic. We know that in general, variation will likely reduce internal validity, but there is no clear or precise way to assess when variation would imply that internal validity is so low as to preclude drawing any conclusions about the intervention [and so undermine trials' external as well as internal validity (Mustafa 2017, p. 187)].

[10] For more detail, and discussion of other differences, see Schwartz and Lellouch (1967); Thorpe et al. (2009); Hey (2015); Nieuwenhuis (2016).

[11] Another way of seeing this is to consider that explanatory trials seek to ensure that the contextual factors surrounding an intervention are the same in the experimental and control groups, and this is possible for RCTs of custom-made devices as long as the variations between these devices are considered part of the intervention, rather than part of the context of the intervention (Schwartz and Lellouch 1967, p. 638). How interventions are defined is discussed further in the section "Variation within the definition of 'the same' intervention".

Further, it seems some variations are permissible; even the strictest explanatory trials allow some variations in an intervention. For example, a surgeon's using a different brand of marker pen to mark the location for an incision would intuitively be irrelevant to outcomes, and insisting on using the same brand—or even the same pen—would seem ridiculous. But while some cases, like this one, seem easy to judge, and there is a range of background knowledge that can help inform judgements about more complex variations, there are reasons to be careful about seemingly obvious judgements of relevance.[12] The same complexity of biological systems which makes it a strength of RCTs that they do not require understanding of pathophysiological processes, also results in a necessary uncertainty in such judgments.

And despite the well-developed literature on pragmatic trials, there is little abstract discussion of exactly why variation between instances of a tested intervention will reduce internal validity (Nieuwenhuis 2016, p. 96). To clarify when variations are innocuous, and when they will influence results, we need a better understanding of why variation is problematic.

## Why is variation problematic?

In this section I examine Cartwright's causal analysis of RCTs (1994, 2009, 2010; Cartwright and Hardie 2012), and show how it helps to illuminate the way that inherent variation could affect internal validity. I then consider two ways of thinking about inherent variation, as a source of confounding (the section "Variation as problematic because it makes a causal difference"), or as contained within the most accurate description of an intervention (the section "Variation within the definition of 'the same' intervention"), and argue that interventions with inherent variation can be evaluated in RCTs.

Cartwright's analysis is developed in relation to the 'problem of external validity', i.e., the problem that a trial with high internal validity may lack external validity, and so not tell us how well the intervention is going to work outside of the controlled conditions of a trial—its 'effectiveness' (Cartwright 2009, p. 187–188; Howick et al. 2013). Inferring effectiveness from efficacy involves inferring that, because the intervention worked, on average, in the treatment population, it will work in other populations. But this will often be an unsafe assumption since, as discussed above, the population and conditions often differ between an (explanatory) trial and clinical practice.

Cartwright argues that deriving conclusions about efficacy from RCT results involves a hidden assumption: that probabilistic dependence (i.e., of an outcome on an intervention) requires a causal explanation. That is, the reason the difference can be attributed to the intervention is that we take the intervention to have played a causal role in the occurrence of the outcomes. This view leads to regarding efficacy as a causal power or capacity, and claims about efficacy as causal claims. That is, when we say that an intervention ($T$) had efficacy for an outcome ($e$) in the study population, we are saying that, in this population, $T$ caused $e$.

Understanding efficacy as a causal capacity means that the causal contribution of an intervention can be considered stable across its different instances (Cartwright and Munro 2010, p. 262).[13] Differing results in different populations are obtained because these represent different causal situations. In some of these causal situations, the causal capacity of the intervention might be less likely to be triggered, or be likely to be triggered but be masked by some countervailing causal factor, or be likely to trigger other causal processes that impact on outcomes. Thus, to predict effectiveness in a new population from efficacy, we need additional reasons to think that an intervention's causal capacity is likely to triggered, and/or not likely to be masked or altered, in the target population (Cartwright and Munro 2010; Cartwright 2010, p. 67–68).

Cartwright's example relates to an RCT of a social intervention in Tamil Nadu. It involved educating mothers about child nutrition, with the aim of improving child nutrition, and showed the intervention to have efficacy. The same intervention was then implemented in Bangladesh, where it did not have the same outcomes. A reason for this, Cartwright suggests, is that in Bangladesh mothers are not in charge of feeding children; others usually have this role. This provides a reason to think that the causal capacity of the intervention, educating mothers about child nutrition, will not be triggered in this causal situation (Cartwright and Hardie 2012, p. 3–4, p. 27–32).

A medical example of how an intervention may have efficacy in some but not other populations is the drug vemurafenib, a treatment for melanoma. This example also shows how Cartwright's framework can help us understand different outcomes within an experimental population, as well as between experimental and other target populations. Vemurafenib has been shown in an RCT to have efficacy

[12] An example is pill-case colour, which has been shown to sometimes influence outcomes in pharmaceutical trials (de Craen et al. 1996). Variations of pill-case colour within a trial might seem innocuous, but could influence results.

[13] This is a controversial claim amongst philosophers concerned with the nature of causation, but I shall not seek to deal with this here. Cartwright's analysis can be regarded as useful in understanding the significance of inherent variation even if we resist its metaphysical commitments.

for producing tumour regressions and promoting survival, for patients whose melanoma has a V600E BRAF mutation (Chapman et al. 2011). Consistently with the point that efficacy is a population-level, probabilistic measure, not all patients with a melanoma with this mutation have these outcomes. Further research suggests that vemurafenib will not work for patients who: have cancer cells that over-express the protein PDGFRB, have a mutated NRAS (Nazarian et al. 2010), or have stromal cells that secrete hepatocyte growth factor (Wilson et al. 2012). In Cartwright's terminology, we can say that in these cases, the causal capacity of vemurafenib, to bring about tumour regression and increased survival time, remains the same as it is in cases where these outcomes do occur. But in these non-responsive cases, the causal capacity is either blocked by these other factors, so they do not occur at all, or it is triggered but is masked by the presence of other causal mechanisms, or vemurafenib triggers other mechanisms that alter the outcomes.[14]

Cartwright's analysis provides a way of thinking about therapeutic interventions that can be used to give content to what it means for an intervention to be 'the same' across its various instantiations, while also explaining why these different instantiations may have different outcomes. Though Cartwright's discussion is focused on the problem of external validity, the point about differing background causal situations also applies within an experimental group: different outcomes in the experimental group occur because each individual represents a different causal background situation into which the intervention is introduced. While the tested intervention makes a stable causal contribution, this contribution may be blocked, masked, or altered in some causal situations (Cartwright and Hardie 2012, p. 26–36). With these concepts in hand, two ways of thinking about when variation between instances of an intervention is problematic are investigated below.

## Variation as problematic because it makes a causal difference

We might consider variations to the intervention to be potential threats to the intervention's causal capacity. A variation might alter the intervention such that it no longer makes the same causal contribution, such that the intervention's causal contribution is no longer triggered, or is masked or altered by other causal processes. On this understanding, variations would be problematic because they would act in similar ways to confounders. They would undermine internal validity because their presence would mean that different causal interactions would be introduced to different members of the experimental group.

On this view, variations are permissible when they are causally innocuous. This would explain many of our intuitive judgements about permissible variations, such as the marker pen brand example above, and is perhaps what underlies such judgements: we assume they are causally innocuous. This view may be useful where we are able to judge causal innocuousness with some confidence. However, as discussed above, there are reasons to be cautious about intuitive judgements of relevance, and in some cases we will not be able to make this judgement.[15]

More problematically for my purposes, with custom-made devices the rationale for the variation itself implies that the variation is not causally innocuous, since inherent variation is motivated by its potential benefits for patients. If variation is problematic whenever it makes a causal difference, inherent variation will therefore undermine internal validity. This would also apply to RCTs of other interventions that involve inherent variation, such as surgeries.

## Variation within the definition of 'the same' intervention

But there is another way to understand inherent variation, which allows that at least some non-causally-innocuous variations do not reduce internal validity, suggesting that RCT assessments are possible. Cartwright's view implies that different instances of an intervention count as 'the same' when the causal contribution they make is stable across those instances. This implies that we can consider 'the intervention' to be defined as whatever it is that has the stable causal capacity that is intended.[16]

[14] Though further research is needed, current research suggests that for patients with over-expressed PDGFRB or mutated NRAS, vemurafenib's causal capacity (which is thought to involve blocking how the cancer grows and maintains itself) is triggered, but this effect is negated by other mechanisms (as over-expressed PDGRFB or mutated NRAS mean the cancer has other routes by which to grow and maintains itself). In the case of patients whose stromal cells secrete hepatocyte growth factor, it would appear that vemurafenib's capacity to block cancer growth and maintenance is itself prevented from occurring (see Nazarian et al. 2010; Straussman et al. 2012; Wilson et al. 2012).

[15] This view would imply that in explanatory trials, no variations should be acceptable without strong reasons to think they are causally innocuous. In pragmatic trials, some variations would still be acceptable since this view is consistent with thinking that the causal contribution of a varied intervention could remain similar across its various instances, albeit not the same. The causal capacity of, for instance, taking the same drug every day at the same time might be *similar* to that of taking the same drug most days at slightly different times. The point at which variations made internal validity so low that the trial results tell us nothing at all would, again, often be very difficult to assess.

[16] A similar view is found in the work of Schwartz and Lellouch (1967), as discussed below. Something like this view is also implicated in discussions of intervention 'fidelity' or 'integrity', that is, the extent to which an intervention is carried our according to a defined

Though Cartwright does not (to my knowledge) consider this specific issue, she suggests that interventions can be described in different ways, and some of the descriptions are more accurate than others. Consider the child nutrition example: if we describe the intervention as 'educating mothers about child nutrition', it has efficacy in the test population in Tamil Nadu, but not in Bangladesh. But we could alternatively describe the intervention as 'educating those in charge of feeding children about child nutrition'. It seems possible that under this description, the intervention may well have worked in Bangladesh. Cartwright thus argues that "the right description is the one that plays the same causal role in the target as in the study" (Cartwright and Hardie 2012, p. 46). It is more accurate to describe the child nutrition intervention as 'educating those in charge of feeding children about child nutrition' because this is the description that isolates the relevant causal contribution.

While this is phrased in relation to different populations, again, we may say the same thing about the different individuals within an experimental group: that the most accurate description of 'the intervention' is the one that captures whatever it is that plays the same causal role, among the various instances of the intervention given to this group.

A similar implication can be drawn from Schwartz and Lellouch's (1967) discussion of treatment definitions. In explanatory trials, we try to make elements of the background causal situation (which Schwartz and Lellouch refer to as 'contextual factors') in the experimental and control groups the same. In a pragmatic trial, contextual factors are instead determined by the purposes of clinical practice. Schwartz and Lellouch claim this does not undermine comparing the two groups because in a pragmatic trial, contextual factors "become themselves part of the therapies to be compared and are thus distinguished from non-contextual factors for which comparability must be assumed … the treatments … 'absorb' into themselves the contexts in which they are administered" (1967, p. 638). That is, the definition of the intervention (in either kind of trial) includes all elements which have a causal impact on differences in outcomes between the treatment and control groups.

On this kind of view, interventions may count as 'the same' even though they involve variations. This would be the case where variations are made in order to support the intervention's making the same causal contribution in its different instances. And arguably, this is the intention with which patient-specific variations are introduced. For example, studies of custom-made femoral stems describe these devices in terms of their similarities. In the "Kinds of patient-specificity" section I noted similarities of using a common template, material, and design process. But another similarity between the differing devices is the parameters and aims (e.g. for particular clinical outcomes) that drive the design of the variations. For example, Bargar (1989) states that the custom-made implants in his study were designed such as to achieve four aims: better initial stability, better contact over ingrowth surfaces, uniformity of stress transfer, and restoration of the femoral head position. Colen et al. (2014) indicate that their custom prostheses aimed to achieve better fit and fill of the femoral canal while retaining more bone for better endosteal loading and stress shielding. Grant et al. (2005) consider optimal load transfer, close fit to the femur, and mechanical stability. That is, the variations themselves are driven by their enabling the devices to meet parameters that are the same across different patients, and so, by the aim of facilitating the same causal contribution from the intervention.[17]

Thus for custom-made devices, the intervention can be described not in terms of the differences between the devices, but in terms of the similarities that drive the variations. This is a more accurate description of the intervention, because it is in virtue of meeting these parameters that the implants have the causal capacity that they do (if in fact custom-made implants do improve outcomes). This allows recognition that for custom-made devices, the variations between devices are proposed because they are part of what is supposed to make the device work as it does. As such, RCTs could be used to evaluate the efficacy of custom-made devices, in cases where variations between particular devices support the 'sameness' of each particular instance of the intervention, at a different level of description.

Footnote 16 (continued)

protocol, primarily discussed in relation to complex interventions. Though discussions of fidelity are partly concerned with how to ensure a treatment known to be effective is implemented as planned, some of this literature recognises that resolving issues related to variation will involve identifying a clear definition of an intervention, and identifying what aspects of an intervention are causally implicated in the desired outcomes (see, e.g., Gearing et al. 2011, p. 82; Medical Research Council 2008; Moncher and Prinz 1991, p. 250). However, exactly why variations are tied to these matters is not made explicit in this literature.

[17] Hawe et al. (2004) offer a view of complex interventions which is consistent with this account. As they put it, "[r]ather than defining the components of the intervention as standard—for example, the information kit, the counselling intervention, the workshops—what should be defined as standard are the steps in the change process that the elements are purporting to facilitate or the key functions that they are meant to have" (2004, p. 1562). What Hawe and colleagues refer to as the 'change process' or 'function' of the intervention, I take it, maps onto what I have called the causal contribution of the intervention, though these authors do not elaborate on the conceptual basis for their view.

## Implications for evaluation and regulation of custom-made and other patient-specific devices

This argument shows that we can, in principle, use RCTs to evaluate custom-made devices. It also reveals that doing so will be very difficult. To conclude, I draw out some implications for approaching evaluation of custom-devices for regulatory purposes, and comment on implications of the argument for evaluating other patient-specific devices.

### Kinds of evidence and evaluating custom-made devices

In order to assess whether variations between devices will impact on internal validity, we need to find the description of the intervention that correctly isolates the causal contribution that remains the same for each device. While such descriptions might be posited—consistently with whatever rationale is motivating the attempt to make the intervention specific for patients—this will rely on having reasonably detailed knowledge about the causal mechanisms that connect the intervention to its outcomes. In many cases, such knowledge will be incomplete or unavailable, unclear, and/or contentious. For example, the case of custom-made hip implants showed that a number of parameters and aims were referred to in designing variations into the implants. There does not appear to be a consensus on which of these (or other proposed) parameters are necessary or most important, how each is related to the desired clinical outcomes, how variations to the devices alter them, or how they might interact with each other in relation to possible outcome measures.

The argument in the section "Why is variation problematic?" thus supports the view that in evaluating custom-made devices, other kinds of evidence are needed. RCTs on their own do not provide knowledge about causal mechanisms, though they can test hypotheses derived from it, and contribute to theory development (Hey 2015). Relevant knowledge in the hip implant case is developed in, for instance, engineering studies of the properties of implants; cadaveric bone studies of the biomechanical interaction of bone with implant; computer modelling studies of how implant design will impact on load distribution; and so on. These kinds of studies can contribute causal knowledge to help develop consensus on what parameters and aims are appropriate, and how these will relate to particular clinical outcomes.

While evidence hierarchies usually place evidence about causal mechanisms very low in the hierarchy, or omit it altogether (Clarke et al. 2014, p. 341), recent arguments have questioned this stance, pointing out that although causal reasoning can introduce bias, it is necessary for designing trials, and using their results (Clarke et al. 2014; Howick 2011). This discussion of custom-made devices adds to these arguments the point that evidence about causal mechanisms is in some cases necessary for identifying the correct description of an intervention, which will allow us to determine what counts as 'the same intervention', correctly define it, and understand when variations to it are problematic. This is not to say that evidence about causal mechanisms should be pursued in place of RCT evidence. On the contrary, it shows that one reason for improving our background causal understanding related to custom-made devices like hip implants is to enable us to devise adequate RCT protocols. It would also be appropriate to collect longer-term outcomes data in observational studies or registries to help build this background knowledge. As both Clarke et al. (2014) and Hey (2015, p. 1323–1324) show in some detail, RCTs contribute to theory development as well as providing information about efficacy (or lack thereof) of specific interventions for specific outcomes, and thus can be used in combination with other sorts of evidence in developing this knowledge.

Further, whether RCTs of custom-made devices are designed to answer questions of clinical practice (such as whether to use a custom-made over a standardised device), or to build the surrounding knowledge base (such as whether custom-made devices result in better initial stability), it should be borne in mind that the robustness of RCTs of interventions with inherent variations will depend on how well understood the causal properties of the intervention are. Running RCTs before this knowledge is available is unlikely to be informative (and may waste resources and expose patients to risks for no purpose[18]). Similar points can be made about RCTs of other interventions likely to involve inherent variation.

### Implications for evaluating other patient-specific devices

While the argument has focused on custom-made devices as the simpler case, it has some implications for further consideration of bespoke and tissue-engineered devices. Of course, bespoke devices will not typically be assessable on the population level just because there are not sufficient numbers of patients with sufficiently similar treatment needs. If patient-specific devices come to be used more, it is possible that some devices currently made as 'bespoke' could come to be standardised enough to move across to the 'custom-made' category. In this case, the argument above would apply, indicating that the barrier for using RCTs to assess bespoke devices is practical rather than theoretical. However, as long as this practical barrier remains—as it presumably will for at least some cases where bespoke devices could be utilised—other approaches to evaluation and regulation will be needed. Again here, evidence relating to causal mechanisms might play a role: a more sophisticated understanding of how

[18] Thank you to an anonymous reviewer for pointing this out.

the properties of bespoke devices (including characteristics of the materials used, but also more generally what causal contributions the devices need to make to particular bodily functions) could help to improve clinical outcomes for bespoke devices. As long as bespoke devices continue to reach patients through custom device exemptions (or under regulations for research), data on the mechanistic assumptions used in their design, and on outcomes, could also be collected in a registry, in order to develop theoretical knowledge and guide research into causal mechanisms.

What the argument implies regarding tissue-engineered senses of patient-specificity is less clear, and may depend on how those technologies develop, as well as on developing empirical knowledge about what difference using patients' own cells makes. One rationale for using patients' cells is that it will remove the chance of rejection and the need for immune suppression (Atala 2009, p. 575). If correct, this would indicate that this variation is one that supports a common clinical outcome, and thus can be regarded as supporting the stability of the organ's causal contribution. However, it is not yet clear if the use of patients' own cells might alter the causal contributions relevant to various kinds of interventions in other ways (Gilbert et al. 2017).

In examining the possibility of assessing custom-made devices in RCTs, this paper has developed an account of the reasons variations in instances of a tested intervention undermine internal validity, giving some guidance as to when variations are problematic. The argument suggests that in order to understand when interventions with inherent variation can be evaluated in RCTs, we need to identify when variations support rather than undermine the stability of the causal contribution made by the intervention. In relation to custom-made devices, this may require improving our overall understanding of how the intervention works. This might be developed through a range of kinds of study, including but not limited to basic science research.

**Acknowledgements** My thanks to Robert Sparrow, Wendy Rogers, David Wotton, Tajanka Mladenovska, and the anonymous referees for this journal, for their comments on previous drafts. The funding was provided by Australian Research Council (Grant No. CE140100012)

## References

Atala, A. 2009. Engineering organs. *Current Opinion in Biotechnology* 20: 575–592.

Bargar, W. L. 1989. Shape the implant to the patient: A rationale for the use of custom-fit cementless total implants. *Clinical Orthopedics and Related Research* 249: 73–78.

Cartwright, N. 1994. *Nature's capacities and their measurement*. Oxford: Oxford University Press.

Cartwright, N. 2009. What is this thing called efficacy? In *Philosophy of the Social Sciences*, ed. C. Mantzavinos, 185–206. Cambridge: Cambridge University Press.

Cartwright, N. 2010. What are randomised controlled trials good for? *Philosophical Studies* 147: 59–70.

Cartwright, N., and J. Hardie. 2012. *Evidence-based Policy: A practical guide to doing it better*. Oxford: Oxford University Press.

Cartwright, N., and E. Munro. 2010. The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice* 16: 260–266.

Chadwick, R. 2014. The ethics of personalized medicine: A philosopher's perspective. *Personalized Medicine* 11 (1): 5.

Chapman, P. B., et al. 2011. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *New England Journal of Medicine* 364 (26): 2507–2516.

Clarke, B., D. Gillies, P. Illari, F. Russo, and J. Williamson. 2014. Mechanisms and the evidence hierarchy. *Topoi* 33: 339–360.

Colen, S., A. Dalemans, A. Schouwenaars, and M. Mulier. 2014. Outcome of custom-made IMP femoral components of total hip arthroplasty. A follow-up of 15–22 years. *Journal of Arthroplasty* 29: 397–400.

Cook, J. A. 2009. The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. *Trials* 10: 9.

de Craen, A. J., P. J. Roos, A. L. de Vries, and J. Kleijnen. 1996. Effect of colour of drugs: Systematic review of perceived effect of drugs and of their effectiveness. *BMJ (Clinical Research Ed.)* 313 (7072): 1624–1626.

Food and Drug Administration (FDA). 2014. *Custom device exemption: Guidance for industry and Food and Drug Administration staff*. Rockville, MD: US Department of Health and Human Services.

Garas, G., A. Ibrahim, H. Ashrafian, K. Ahmed, V. Patel, K. Okabayashi, P. Skapinakis, A. Darzi, and T. Athanasiou. 2012. Evidence-based surgery: Barriers, solutions and the role of evidence synthesis. *World Journal of Surgery* 36: 1723–1731.

Gearing, R. E., N. El-Bassel, A. Ghesquiere, S. Baldwin, J. Gillies, and E. Ngeow. 2011. Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review* 31: 79–88.

Gilbert, F., C. D. O'Connell, T. Mladenovska, and S. Dodds. 2017. Print me and organ? Ethical and regulatory issues emerging from 3D bioprinting in medicine. *Science and Engineering Ethics* 1–19.

Grant, P., A. Aamodt, J. A. Falch, and L. Nordsletten. 2005. Differences in stability and bone remodeling between a customized uncemented hydroxyapatite coated and a standard cemented femoral stem. A randomized study with use of radiostereometry and bone densitometry. *Journal of Orthopedic Research* 23 (6): 1280–1285.

Hawe, P., A. Shiell, and T. Riley. 2004. Complex interventions: How "out of control" can a randomised controlled trial be? *BMJ (Clinical Research Ed.)* 328 (7455): 1561–1563.

Hey, S. P. 2015. What theories are tested in clinical trials? *Philosophy of Science* 82: 1318–1329.

Howick, J. 2011. Exposing the vanities – and a qualified defense – of mechanistic reasoning in health care decision making. *Philosophy of Science* 78 (5): 926–940.

Howick, J., P. Glasziou, and J. K. Aronson. 2013. Can understanding mechanisms solve the problem of extrapolating from study to target populations (the problem of 'external validity')? *Journal of the Royal Society of Medicine* 106: 81–86.

Imanishi, J., and P. F. M. Choong. 2015. Three-dimensional printed calcaneal prosthesis following total calcanectomy. *International Journal of Surgery Case Reports* 10: 83–87.

Lassen, K., A. Høye, and T. Myrmel. 2012. Randomised trials in surgery: The burden of evidence. *Reviews on Recent Clinical Trials* 7: 244–248.

Marro, A., T. Bandukwala, and W. Mak. 2016. Three-dimensional printing and medical imaging: A review of the methods and applications. *Current Problems in Diagnostic Radiology* 45: 2–9.

Medical Research Council (2008) *Developing and evaluating complex interventions: New guidance*, http://www.mrc.ac.uk/complexinterventionsguidance.

Meshikhes, A. N. 2015. Evidence-based surgery: The obstacles and solutions. *International Journal of Surgery* 18: 159–162.

Moncher, F. J., and R. J. Prinz. 1991. Treatment fidelity in outcome studies. *Clinical Psychology Review* 11: 247–266.

Muirhead-Allwood, S. K., N. Sandiford, J. A. Skinner, J. Hua, C. Kabir, and P. S. Walker. 2010. Uncemented custom computer-assisted design and manufacture of hydroxyapatite-coated femoral components. *The Journal of Bone and Joint Surgery* 92-B: 1079–1084.

Mustafa, F. A. 2017. Notes on the use of randomised controlled trials to evaluate complex interventions: Community treatment orders as an illustrative case. *Journal of Evaluation in Clinical Practice* 23(1): 185–192.

Nazarian, R., H. Shi, Q. Wang, X. Kong, R. C. Koya, H. Lee, Z. Chen, M. K. Lee, N. Attar, H. Sazegar, T. Chodon, S. F. Nelson, G. McArthur, J. A. Sosman, A. Ribas, and R. S. Lo. 2010. Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* 468: 973–977.

Ng, V. Y., J. H. DeClaire, K. R. Berend, B. C. Gulick, and A. V. Lombardi. 2012. Improved accuracy of alignment with patient-specific positioning guides compared with manual instrumentation in TKA. *Clinical Orthopaedics and Related Research* 470: 99–107.

Nieuwenhuis, J. B., E. Irving, K. O. Rengerink, E. Lloyd, I. Goetz, D. E. Grobbee, P. Stolk, R. H. H. Groenwold, and M. Zuidgeest. 2016. Pragmatic trial design elements showed a different impact on trial interpretation and feasibility than explanatory elements. *Journal of Clinical Epidemiology* 77: 95–100.

Pope, C. 2002. Contingency in everyday surgical work. *Sociology of Health and Illness* 24 (4): 369–384.

Schwartz, D., and J. Lellouch. 1967. Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Diseases* 20: 637–648.

Small, T., V. Krebs, R. Molloy, J. Bryan, A. K. Klika, and W. K. Barsoum. 2014. Comparison of acetabular shell position using patient specific instruments vs. standard surgical instruments: A randomized clinical trial. *Journal of Arthroplasty* 29: 1030–1037.

Stirrat, G. M. 2004. Ethics and evidence based surgery. *Journal of Medical Ethics* 30: 160–165.

Straussman, R., R. Morikawa, K. Shee, M. Barzily-Rokni, Z. R. Qian, J. Du, A. Davis, M. M. Mongare, J. Gould, D. T. Frederick, Z. A. Cooper, P. B. Chapman, D. B. Solit, A. Ribas, R. S. Lo, K. T. Flaherty, S. Ogino, J. A. Wargo, and T. Golub. 2012. Tumour micro-environment elicits innate resistance to RAF inhibitors through HGF secretion. *Nature* 487: 500–504.

Therapeutic Goods Administration (TGA). No date. *Custom-made medical devices* (*fact sheet*). Available at: https://www.tga.gov.au/custom-made-medical-devices.

Thorpe, K. E., M. Zwarenstein, A. D. Oxman, S. Treweek, C. D. Furberg, D. Altman, S. Tunis, E. Bergel, I. Harvey, D. Magid, and K. Chalkidou. 2009. A pragmatic-explanatory continuum indicator summary (PRECIS): A tool to help trial designers. *Journal of Clinical Epidemiology* 62: 464–475.

Viceconti, M., D. Testi, R. Gori, C. Zannoni, A. Cappello, and A. De Lollis. 2001. HIDE: A new hybrid environment for the design of custom-made hip prosthesis. *Computer Methods and Programs in Biomedicine* 64: 137–144.

Wallace, G. G., R. Cornock, C. O'Connell, S. Beirne, S. Dodds, and F. Gilbert. 2014. *3D bioprinting: Printing parts for bodies*. Wollongong: ARC Centre of Excellence for Electromaterials Science. https://3dbioprint.creatavist.com/3dbioprinting.

Wilson, T. R., J. Fridlyand, Y. Yan, E. Penuel, L. Burton, E. Chan, J. Peng, E. Lin, Y. Wang, J. Sosman, A. Ribas, J. Li, J. Moffat, D. P. Sutherlin, H. Koeppen, M. Merchant, R. Neve, and J. Settleman. 2012. Widespread potential for growth-factor-driven resistance to anticancer kinase inhibitors. *Nature* 487: 505–509.

Worral, J. 2011. Causality in medicine: Getting back to the Hill top. *Preventive Medicine* 53: 235–238.

Zopf, D. A., S. J. Hollister, M. E. Nelson, R. G. Ohye, and G. E. Green. 2013. bioresorbable airway splint created with a three-dimensional printer. *New England Journal of Medicine* 368: 2043–2045.