Averaged versus Individualized: Pragmatic N-of-1 Design as a Method to Investigate Individual Treatment Response

European Journal for Philosophy of Science
https://doi.org/10.1007/s13194-023-00559-0

Davide Serpico
Department of Economics and Management, University of Trento
Interdisciplinary Centre for Ethics & Institute of Philosophy, Jagiellonian University

Mariusz Maziarz
Doctoral School in the Humanities, Faculty of Philosophy, Jagiellonian University
Interdisciplinary Centre for Ethics & Institute of Philosophy, Jagiellonian University
Email: mariusz.maziarz@uj.edu.pl

Abstract: Heterogeneous treatment effects represent a major issue for medicine as they undermine reliable inference and clinical decision-making. To overcome the issue, the current vision of precision and personalized medicine acknowledges the need to control individual variability in response to treatment. In this paper, we argue that gene-treatment-environment interactions (G×T×E) undermine inferences about individual treatment effects from the results of both genomics-based methodologies - such as genome-wide association studies (GWAS) and genome-wide interaction studies (GWIS) - and randomized controlled trials (RCTs). Then, we argue that N-of-1 trials can be a solution to overcome difficulties in handling individual variability in treatment response. Although this type of trial has been suggested as a promising strategy to assess individual treatment effects, it nonetheless has limitations that limit its use in everyday clinical practice. We analyze the existing variability within the designs of N-of-1 trials in terms of a continuum where each design prioritizes epistemic and pragmatic considerations. We then support wider use of the designs located at the pragmatic end of the explanatory-pragmatic continuum.

Keywords: Gene-environment interactions; Conflicting results; Randomized controlled trials (RCTs); P-medicine; Genome-wide association studies (GWAS); N-of-1 trials.

Introduction

Heterogeneous treatment effects are widely considered a major issue for medicine as they undermine reliable inference and clinical decision-making. Under the strain of empirical literature reporting conflicting results across clinical studies, researchers have pointed to the need for a more precise or personalized approach in medicine (generally named *P-Medicine*) to account for the variation among patients and thus improve diagnosis and treatment. However, the question of whether P-medicine has the conceptual and methodological resources to deliver on its promises remains open (Gamma 2016; Lemoine 2017; Plutynski 2020).

In this paper, we approach the problem of individual treatment effect heterogeneity and argue that gene-treatment-environment interactions (G×T×E) undermine the results of both randomized controlled trials (RCTs) and the repertoire of genomics-based P-medicine — genome-wide association studies (GWAS) and genome-wide interaction studies (GWIS). We then support the use of N-of-1 trials as a source of evidence for predicting individual treatment responses and informing therapeutic decisions. Below is a detailed structure of the article.

In Section 1, we explain that the evidence-based medicine (EBM) movement focuses on average causal effects in developing its evidence appraisal tools and argue that such averages are not representative of individual treatment effects, which may differ significantly from the average. The heterogeneity of individual treatment effects makes clinical decisions based on average treatment effects (ATEs) likely ineffective in cases where individual outcomes differ from the population-wide average.

In Section 2, we argue that part of the heterogeneity in individual treatment effects depends on the genetic variability of populations and variability in environmental exposures that interact with treatments. In other words, individual response to treatment is generated not just by interactions between the absence/presence of specific genetic variants and drugs (G×T), as it is often assumed in pharmacogenomics studies, but also by further interactions with environmental exposures that are difficult to operationalize and control for (G×T×E). We will review pharmacogenetics studies on asthma to highlight major limitations in the systematic and reliable identification of G×T×E. Findings on asthma represent an interesting case study because asthma, compared with more complex phenotypes (e.g., major depression), is a relatively simple trait related to well-known physiological mechanisms. Thus, shortcomings in the study of asthma unlikely depend on the operationalization of the trait but rather on more general issues relating to the control of population stratification in genetic and environmental variability.

In Section 3, we argue that N-of-1 trials can be a solution to overcome difficulties in handling individual variability in treatment response. This type of trial has been suggested as a promising strategy to assess individual treatment effects, but major limitations affect this approach, too, including their low feasibility in everyday clinical practice. By drawing an analogy with the explanatory/pragmatic RTCs distinction, we thus discuss the plurality of existing single-patient designs in terms of a continuum ranging from explanatory and pragmatic aims: on this view, different N-of-1 designs put different emphasis on methodological rigor (at the expense of lower feasibility) or pragmatic considerations (at the expanse of lower internal validity). Finally, we outline the main features of N-of-1 designs that are closer to the pragmatic end of the continuum as potential ways in which this type of trial can be simplified to make it more feasible without negatively impacting the results' integrity.

1. The Omission of the Individual Patient by the Evidence-based Medicine Approach

The standard approach to assess evidence in medicine and inform clinical decisions has been developed by the evidence-based medicine (EBM) movement. This approach to the appraisal of evidence for treatment effectiveness and safety (or, more broadly, causal generalizations, i.e., type-level causality) is based on assessing the risk of bias or confounding of each study type (Borgerson 2009; La Caze 2009). Accordingly, factors unaccounted for in a study (e.g., genetic differences, environmental exposures, or the researchers' expectations) make a difference between treatment and control groups and undermine an accurate assessment of treatment effectiveness.

According to the EBM approach, randomized-controlled trials (RCTs) are prioritized over non-randomized interventional studies and other observational designs (e.g., cohort and case-control studies) (OCEBM Levels of Evidence Working Group 2009; National Institute for Health and Care Excellence 2014). RCTs allow for estimating the average treatment effect and measuring the dispersion of individual treatment responses. However, as we will argue, they provide little information regarding what confounders mediate treatment effectiveness and about individual treatment effects $TE(n)$. This problem is further aggravated when the evidence produced by RCTs is aggregated with meta-analyses: sample sizes, in such analyses, are much larger than in individual studies and hence deliver more precise estimates for the average treatment effects (ATE). However, obtaining ATE estimates with narrower confidence intervals does not change the distribution of individual treatment effects; hence, the empirical rule cannot be applied to estimate the dispersion of outcomes (Maziarz 2022).[1]

RCTs deliver the most trustworthy evidence for average treatment effects but are unable to inform regarding individual treatment effects. As Borgerson put it,

> RCTs produce data that is averaged over the patients in the trial. Physicians and practitioners encounter individual patients. The gap between the average patient (after inclusion and exclusion criteria) and the individual patient [despite equaling zero in expectancy] is a significant one, and is the first thing critics of RCTs mention when listing the problems with the RCT (2008, p. 190).

This feature of RCTs inspired the view that the EBM movement oversimplifies the complexity of clinical decision-making because it ignores heterogeneity in treatment responses. For instance, Aron (2020) observes that treatment response is a function of not only intervention but also of the context in which it is delivered (constituted by an organism and its peculiar characteristics). Additionally, individual treatment responses are further shaped by environmental exposures and disease severity. As Feinstein observed,

> [p]harmaceutical companies, regulatory agencies, and public policymakers may be satisfied to receive those average results, but practicing clinicians and patients are not. The clinicians and patients want to know the results in subgroups having a pertinent ‘clinical resemblance’ to the current patient (1995, p. 73)

The Potential Outcomes Approach (POA) seems to be the predominant position underlying inferences from RCTs despite often being considered too restrictive about the notion of *cause* (Vandenbroucke et al. 2016). The POA defines treatment effect in terms of the difference between the outcome observed by the patient receiving the intervention under investigation and the outcome observed when the n-th patient is treated with the comparator drug:

$$TE(n) = Y_T(n) - Y_C(n)$$

Where:
$TE(n)$ — n-th patient treatment effect
$Y_T(n)$ — the outcome of n-th patient receiving treatment
$Y_C(n)$ — the outcome of n-th patient receiving control

[1] The empirical rule states that 99.7% of observations of a normally distributed variable fall within three standard deviations from the average.

The impossibility of observing, at the same time, both the outcome of the treatment with intervention and control of the same patient constitutes the fundamental problem of causal inference (Rubin 1974; Rubin 2005). The solution to this problem endorsed by the proponents of the EBM movement is to focus on the population-wide average treatment effects that can be estimated by comparing (calculating the difference in means between) the average outcomes observed in the treatment group and in the control group (Hernan & Robins 2018):

$$A\hat{T}E = \frac{1}{N}$$

This solution relies on randomization, which, in the long run, balances the overall impact of confounders between the treatment and the control group, so that the only explanation for the observed difference in means is the intervention under test (La Caze 2013). However, the estimate of the average treatment effect ($A\hat{T}E$) does not inform the dispersion of individual treatment effects $TE(n)$ in the population of patients. Indeed, the variance in outcomes (Y) is generated by individual differences in the values of confounding variables. To illustrate how confounders impact individual outcomes, consider the following situation analyzed by Greenland:

> [s]uppose I wish to study whether lidocaine prophylaxis prevents death within the 72 hours following hospital admission for acute myocardial infarction. I will enroll two patients for this study, two successive admissions to a hospital emergency room. When the first patient is admitted, I will toss a fair coin: If heads, the first patient will receive lidocaine and the second will not; if tails, the second admission will receive lidocaine and the first will not. Suppose now that the first admission is massively compromised and is certain to die within 72 hours of admission, whereas the second is a mild case and is certain to survive, whether or not either of them receives lidocaine therapy (1990, p. 421).

To obtain warranted conclusions regarding the ATE, researchers need to recruit a sample of a size sufficient to ensure that the impact of confounders on an outcome of interest will average out (e.g., both the treatment and control groups will include similar numbers of mild and severe cases). This sample size is determined at the research design stage, given a chosen power (β) and a threshold of statistical significance (α). The exact number of patients that need to be recruited depends on the absolute effect size ($|\hat{\mu}_T - \hat{\mu}_C|$) and the dispersion of outcomes (measured by their variance σ^2) (see Cook & DeMets 2008, pp. 115-139; Chow et al. 2018, pp. 47-49).

Random differences in the distribution of confounders are not an obstacle to sound inferences, as the hypothesis of treatment effectiveness is tested statistically. Usually, the null hypothesis of no difference is chosen (H_0: $\mu_T = \mu_C$) versus the alternative (H_1: $\mu_T \neq \mu_C$), although the more warranted choice would be to test if the difference between trial arms is larger than the minimal clinically important difference (MCID) (McGlothlin & Lewis 2014; Lawler & Zimmermann 2021). As we mentioned, the randomization procedure is expected to assert that the confounders are distributed equally between the trial arms and their impact averages out (Deaton & Cartwright 2018). While this is a demanded feature of RCTs if one is interested in the population-wide average treatment effects, the loss of the individual characteristics that determine treatment outcomes is detrimental to predicting individual treatment responses. Indeed, only a small number of patients will experience treatment outcomes similar to the population-wide average.

For simplicity, let us take a trial testing a treatment against a placebo and no placebo effects. In that case, the difference in mean outcome for the treatment and control groups ($\hat{\mu}_T - \hat{\mu}_C$) measures the effect size of the intervention (instead of an average difference in the effectiveness of two alternative therapies). The differentiation of individual treatment responses in the population of all patients fulfilling the inclusion and exclusion criteria is measured by the variance (σ^2) of the primary outcome. In particular, the empirical rule (see above) allows for calculating the range including about 95% of individual treatment responses, which is given by the formula $\hat{Y}_T - 2\sigma; \hat{Y}_T + 2\sigma >$ (Freund & Wilson 2010, p. 27).

Kent et al. (2016) re-analyzed data from 32 large (phase III) trials and observed that “the absolute risk reduction between the extreme risk quartiles ranged from -3.2 to 28.3%” (p. 2075) despite the phase III trials “are often characterized as enrolling relatively homogenous populations” (p. 2084). The surprising level of treatment effect heterogeneity made Kent and colleagues conclude that

> clinically important differences in effect across predicted risk are likely to be common in trials with statistically significant average treatment effects. However, even when these factors are taken into account, considerable variation remains unexplained and could potentially be attributable to genetic differences between patients (2016, p. 2085).

However, the empirical rule is only valid for inferences concerned with outcomes distributed normally. A growing body of evidence suggests that there are non-linear effects of substantial size in cases when treatments interact with moderators that produce non-normal distributions of individual treatment responses. For example, patients who inherited a thiopurine S-methyl transferase deficiency are more than ten times more sensitive to the effects of a leukemia drug on marrow suppression (Coulthard et al. 2002). Another example of treatment outcome heterogeneity that does not follow the Gaussian distribution is the outcome distribution of glioblastoma patients, which effectiveness is determined by the presence/absence of one single genetic variant (Blunt 2019).[2]

Another problem related to applying ATE estimates to individuals is that clinical trials are usually characterized by relatively strict inclusion criteria (e.g., excluding polypharmacy patients or those with comorbidities) resulting in samples being not representative of the general population of patients (Stegenga 2018), which creates the problem of extrapolation: even if an individual patient sufficiently resembles the average of all patients in the clinic, the $A\hat{T}E$ reported by a clinical trial may be different from the average treatment effect of the population of patients in the clinic. But this is only one side of the problem of extrapolation, as strict inclusion and exclusion criteria narrow down the estimates of variance ($\hat{\sigma}_Y$) of the primary outcome and hence the variability in treatment responses observed in the clinic may be larger than the variance measured in a clinical trial. What follows, more than about 5% of patients will experience treatment effects deviating from the average by more than the interval described by the empirical rule.

Notably, the farther away from the average an individual treatment effect is, the less accurate the clinical decision concerning that patient based on average treatment effect estimates stemming from large RCTs or meta-analyses. This inaccuracy of applying population-wide averages to individuals is related to the following two problems: first, uncertainty about the outcomes of untreated disease and, second, uncertainty about the individual response to treatment.

For illustration, suppose that a patient suffers from a condition for which only one treatment is available. The patient may either be a moderate case or be unsusceptible to that drug and experience only limited benefits from the treatment while being exposed to the risk of adverse events (leaving the disease untreated). Or the patient may either be a severe case whose benefits outweigh potential risks and harm or a moderate case that tolerates the treatment well and still benefits from treatment. As Kravitz et al. put it,

> misapplying averages can cause harm, by either giving patients treatments that do not help or denying patients treatments that would help them (2004, p. 662).

To sum up, RCTs are designed to control for individual-level confounders by averaging the effects across individual patients in treatment and control groups. However, this strategy risks overlooking important aspects of individual variability (Deaton &

[2] As we argue in Section 2, however, genetic heterogeneity is not the only confounding factor that moderates treatment response: other sources of uncontrolled heterogeneity are environmental exposures and interactions between them and genetic differences.

Cartwright 2018; Greenhalgh et al. 2014; Kent et al. 2010). While this criticism to the EBM evidence hierarchies is not new, in the next section we show that it also applies to typical evidentiary sources for genomics-based P-Medicine.

2. P-Medicine and Individual Variability

One of the major aims of P-medicine is to deliver evidence for therapeutic decisions concerned with individual patients and overcome the problem of applying population-wide averages in the clinic. Knowledge about individual differences in heredity, environmental exposures, lifestyle, and epigenetic profiles would help understand variability in treatment response, prescribe more effective drugs, and avoid prescribing drugs with negative side effects. In this sense, P-medicine differs from the standard 'one-size-fits-all approach' where medical treatments are designed for the 'average patient'.[3]

Generally speaking, the presence of gene-environment interactions (G×E) implies that the effect of an environmental factor (E) on the phenotype is mediated by genetic factors (G). For instance, the effects of environmental exposure can depend on the presence/absence of a certain allele and thus have a different impact on different individuals. This type of interaction has been observed in several complex traits and diseases, including cancer, psychopathologies, obesity, and general intelligence (see e.g., Caspi et al. 2003; Hyde et al. 2011; Serpico & Borghini 2021; Turkheimer et al. 2003).

Current trends in pharmacogenomic use GWAS to identify statistical associations between genetic variation (G) and response to treatment (T).[4] An increasing number of studies identifies G×T as a major source of treatment effect heterogeneity, suggesting that part of the variability in response to treatment can depend on how the drug interacts with the relevant genes. Here, a treatment is taken as the environmental factor (T=E). The major strength of pharmacogenomics studies is that drugs are relatively simple compared to other environmental factors that may interact with genetic variability: as Ritz and colleagues (2017) argue, drugs are often associated with a specific outcome phenotype (e.g., lowering blood pressure) and their mechanism of action and metabolic pathways are well understood.

[3] P-medicine is a heterogeneous field involving a variety of evidentiary sources and methodologies, ranging from genomics to proteomics, metabolomics, and many others (Snyderman 2012). In our analysis, we mostly focus on pharmacogenomics studies on the role of genetic differences in response to drugs targeting multifactorial diseases, although we acknowledge that personalized health care can involve much more than this. The analysis of individual genetic profiles plays such a key role in the emerging vision of P-medicine that it is *genetic* P-medicine that is usually practiced (Abettan & Welie 2020; Gamma 2016).

[4] GWAS are a hypothesis-free methodology that scan hundreds of thousands of single-nucleotide polymorphisms (SNPs), the most common type of genetic variants in the human genome.

A context in which pharmacogenomics studies have been performed widely is the study of asthma, a complex condition characterized by chronic airway inflammation (Global Initiative for Asthma, GINA 2019). Here, interactions have been identified between dozens of genetic variants and a variety of treatments, including short-acting beta$_2$-agonists (SABAs), long-acting beta$_2$-agonists (LABAs), inhaled corticosteroids (ICS), and leukotriene modifiers (LTMs) (Farzan et al. 2018; Kersten & Koppelman 2018; Lima et al. 2006; Turner 2009; Wang & Tantisira 2016).

Unfortunately, evidence on interactions between genetic variability and asthma treatments (G×T) is unsystematic and usually inconsistent across studies: reported results are often not replicated and associations between genetic variants and treatment response do not reach the threshold of statistical significance, with the consequence that much variability remains unexplained.[5] As Farzan and colleagues conclude (2018, p. 3), these genomics markers are currently not ready for clinical application. And, indeed, while GINA (2019) acknowledges different treatment responses to standard therapies (e.g., inhaled corticosteroids, p. 52), its recommendations still adhere to the one-size-fits-all approach: patients with poorly controlled symptoms are advised to receive a next-step treatment based on the results of clinical trials.

To clarify, the case of asthma is not isolated: similarly unclear are the findings obtained through genomics techniques on other complex conditions, such as major depression and obesity (Chang et al. 2015; Giacomelli et al. 2021; Keers & Uher 2012; Pedersen 2017). Asthma represents to us an interesting case for two main reasons: first, it is a widely investigated condition; second, it is related to well-understood physiological mechanisms and symptoms and is thus a relatively ‘simple’ phenotype compared with more complex traits like psychiatric ones. For these reasons, methodological issues in the identification of G×T in asthma treatment are unlikely dependent on limited data or conceptual imprecision in the definition of asthma. For instance, in studies on conditions such as major depression, questions may arise about how the trait is operationalized and the severity of symptoms assessed through psychometric methods — including questions on whether we should consider fine-grained phenotypes (e.g., serotonin dysregulation) rather than major depression itself. So, considerations that are often made about the genetics of human behavior can be made for simpler traits, too: the literature on asthma suggests that conflicting results regarding individual response to treatment do not depend on the lack of data or due to mere phenotypic complexity, but rather emerge because of difficulties with controlling for population stratification and genotype-treatment interactions (we focus on such difficulties in the next section).

[5] On ICS and LTMs, see Farzan et al. (2018). On conflicting outcomes for SABAs and LABAs, see Kersten & Koppelman (2018).

Contradictory findings are usually explained in terms of methodological limitations or biases.[6] However, we suggest that some of these inconsistencies could be read in a different light: if we consider the variability that characterizes any human population, such results are unsurprising and can rather tell us something interesting about how treatments happen to interact with genetic and environmental factors that differ across individuals. As we explain below, the problem is that our current ability to detect and control for individual variability in G×T may be severely impaired by the complexity at stake.

2.1 Genetic and Environmental Heterogeneity

There are major limitations affecting methodologies investigating G×T in large populations. A recent set of methods to test systematically for interaction effects between each single-nucleotide polymorphism (SNP) and a specific environmental factor (like a drug) is provided by GWIS. Like GWAS, GWIS is a hypothesis-free approach, and for this reason, it is also affected by the methodological issues usually imputed to GWAS.[7] Here, we would like to focus on limitations relating more specifically to population heterogeneity at the genetic and environmental levels — how it can bias the results of GWAS and GWIS, how heterogeneity is usually handled, and why such strategies are often ineffective.

The first issue regards *population stratification*, i.e., undetected heterogeneity in allele frequencies due to non-random mating and geographical isolation (Hellwege et al. 2018; Lawson et al. 2020). In any population, there are arguably different *sets of individuals* that differ systematically in both the genetic ancestry and the phenotype under investigation. If the effects of stratification are not properly corrected, spurious associations can arise due to differences in ancestry, especially in large meta-analyses (Uffelmann et al. 2021).[8]

The stratification problem is intertwined with other sources of heterogeneity, particularly *variability in disease etiology and mechanisms* (Ogino et al. 2013a, 2013b), including their genetic basis (Gravel et al. 2011; Fuller 2021). For instance, a sample may comprise subgroups of individuals with similar phenotypes (e.g., asthma typical

[6] On G×E, see Dick et al. (2015). On behavioral genetics, see Chabris et al. (2012, 2013); Hewitt (2012). More generally on clinical trials, see Ioannidis (2005).

[7] On the difficulty of making causal claims based on genome-wide methods, see Craver et al. (2020); Kaplan & Turkheimer (2021); Oftedal (2022). On statistical biomarkers more generally, see Tabb & Lemoine (2021).

[8] Note that genetic studies can be affected by stratification biases even in relatively homogenous populations with common geographic origins (Sarmanova et al. 2020).

symptoms and immunological biomarkers), but such phenotypes may be due to different mechanisms associated with different genetic variants. At the same time, even in single-gene diseases, carrying a given genetic variant can bring about different phenotypic effects in different individual (see Chen et al. 2016; Cooper et al. 2013; Katsanis 2016; Lynch 2021). In all such cases, statistical associations would likely be spurious, and the results reported by different studies in conflict.

The ideal strategy to avoid stratification biases would be ensuring that the sample is homogenous at the genetic level (Rivadeneira et al. 2021). The trouble is that statistical associations are investigated through hypothesis-free methods like GWAS precisely when we know little about the genetic composition of a population and the genetic basis of a given disease. There exist other ways to correct for stratification, but current methods (e.g., principal component analysis and linear mixed models) come with important shortcomings (Lawson et al. 2020). Particularly worrying is the fact that they are based on common variants, but the genetic basis of complex diseases involves a variety of *types* of genetic variants beyond SNPs that are difficult to capture through GWAS, such as rare genetic variants (frequency <1%), copy-number variants, and structural variants (Baverstock 2019; Burt 2022; Fries 2020; Génin 2020; McClellan & King 2010; Uffelmann et al. 2021; Zaidi & Mathieson 2020).

Environmental factors enter this already very complex picture by multiplying exponentially the number of moderators of individual treatment effects. Indeed, stratification biases regard not only genetic factors but also environmental factors and thus epigenetic markers (i.e., subgroups of individuals in a large sample can be exposed to different environmental influences). Moreover, inconsistency is to be expected when the interactions involve rare genetic variants that are difficult to capture through genome-wide methods.

This leads us to a second major issue, which depends on the difficulty of assessing environmental exposures and thus controlling for individual variability in such factors. As explained above, pharmacogenomics studies usually focus on interactions between genetic variants and treatments (G×T). However, we argue that the causal network generating an individual's response to treatment can involve not just a given treatment (T) and the relevant gene (G) but also undetected environmental factors that can interact with both G and T, generating multiplicative interactions that we will call *G×T×E*. In the case of asthma, such environmental variables can involve air pollution and allergens, for example.[9]

[9] These triadic interactions have been investigated, for instance, in major depression. Chang et al. (2015) pointed out that interactions between corticotropin-releasing hormone (CRH) polymorphisms and antidepressants is mediated by stressful life events. Unfortunately, assessing environmental variables like

Over the past two decades, scholars have repeatedly called for an increase in sample sizes as the solution to the many limitations of genome-wide studies: ideally, bigger numbers would come with more statistical power, and confounding factors of any sort are more likely to average out in larger trials. However, there is disagreement as to whether this strategy alone could bring substantial benefits. In fact, with the development of better techniques, geneticists have become able to test thousands of individuals, but inconsistencies and low replicability have never fully disappeared. This led many to believe that genetic effect sizes are even smaller and more elusive than initially expected, rather than questioning the very reliability of genome-wide methods.[10]

Something very similar is going on in the study of gene-environment interactions, including studies where the investigated environmental factor is just one single drug. Even by considering only interactions between genetic variants and one environmental factor, interactions have turned out to be extremely elusive: indeed, an environmental effect on phenotypic variance can be weak in the general population but extremely relevant in a subgroup of individuals that carry a relevant allele. To scan effectively for G×E through GWIS, the required numbers are thus much greater than in standard GWAS (Dai et al. 2018, p. 470).

The power of GWIS to detect sources of heterogeneity in response to treatment is likely to decrease further if we consider what we said above about G×T×E: if we take treatment as the only environmental factor at stake, we might be unable to account for the actual network of relevant interactions, which arguably includes not just the genotype and the drug, but also uncontrolled (and often poorly understood) environmental variance.

How far should we go with an increase in sample sizes before considering a different approach? There is clearly a tradeoff between analyzing increasingly larger populations and focusing on smaller groups: although larger sample sizes may allow for more generalizable results, this will also bring more genetic and environmental heterogeneity into the analysis, making it even harder to get biologically significant or interpretable results (note that larger studies are more acutely affected by stratification biases, see Hellwege et al. 2018; Marchini et al. 2004). By contrast, smaller and more homogeneous samples allow for finer phenotyping and better control of the extensive

stressful life events can be difficult due to the lack of standardized measures, which limited the integrity of data collected by Chang and colleagues (for a review including other studies, see Keers & Uher, 2012). Moreover, due to the small sample size (193 and 149 individuals in the control and case groups, respectively), Cheng et al. (2015) could not stratify the populations according to *types of antidepressants* with different mechanisms of action. This is a further source of heterogeneous treatment effects that is beyond the aims of this paper.

[10] See long-standing debates on the missing heritability problem (Downes & Matthews 2019; Maher 2008; Manolio et al. 2009; Matthews & Turkheimer 2022; Turkheimer 2011).

genetic and environmental heterogeneity involved in treatment response (for similar considerations, see Giangrande et al. 2022).

The literature on G×T×E suggests that individual-level variation is not 'an exception' or a factor to 'average out' from clinical studies: accounting for individual variability is rather necessary given the aims of P-medicine. As we have shown, heterogeneity in individual treatment response is, however, an obstacle that neither RCTs nor GWAS seems to be able to handle easily: when major G×T×E are present, running effective studies ideally needs subtyping the population in such a way to track down actual biological differences; however, this requires (often missing) prior knowledge from GWAS, candidate-gene studies, and environmental epigenetics on what specific G×T×E can affect treatment response.

3. N-of-1 Trials as a Potential Solution

In the previous sections, we explained that individual treatment responses are determined by individual-level genetic and environmental characteristics and their interactions with an intervention. We also argued that existing methods to assess such interactions have major limitations and that incremental improvements in such techniques may be unable to overcome the issue. If so, pharmacogenomics will not solve the problem of predicting individual treatment response as it would require screening every relevant factor to which a given patient is exposed and understanding their role in shaping phenotypic outcomes. This might turn out to be an unachievable ideal due to the difficulty of controlling for G×T×E in systematic and unbiased ways.

However, one research design already used in some areas of medicine allows for estimating individual treatment responses even when the interactions among the treatment, environment, and genes remain unknown: in N-of-1 studies, single patients undergo cycles of a treatment under test followed by the appropriate control conditions. For instance, Nikles suggested that

> [u]ntil pharmacogenetics […] becomes further developed and widely available, N-of-1 trials remain the best method of identifying patients who respond to certain drugs (2015, p. 13).

Measuring outcomes repeatedly allows for averaging out random environmental exposures or spontaneous deteriorations and improvements and measuring the immediate treatment effects (as opposed to long-term effects). Together with randomization, such features have various advantages, e.g., they help ensure the integrity of results and offer a solution to the problem of extrapolation that we discussed in Section 1.

N-of-1 trials have mostly fallen outside the range of topics studied by philosophers of medicine (Jukola 2019). However, a few voices speak for their potential. Guyatt et al. (1990) concluded that N-of-1 studies are feasible and useful in clinical practice. N-of-1 trials have been observed to be a promising source of evidence regarding individual treatment response, especially in chronic conditions (Duan et al. 2013), and have a track record of informing clinical decisions that allowed the reduction of pharmaceutical treatments (e.g., the number of prescribed drugs) or the prescription of more effective drugs for individual patients. For instance, N-of-1 trials of methylphenidate (Nikles et al. 2015) proved that some individuals benefit from the treatment while others suffer from its side effects, which makes the ATE estimate close to zero. Scuffham et al. (2010) observed that fewer treatments were prescribed after N-of-1 trials aimed at finding treatments most effective for those individuals. Both patients and physicians questioned by Kronish et al. (2017) in New York Presbyterian Hospital perceived N-of-1 studies as useful for individualizing treatments (see also Moise et al. 2018). Recently, Vogt (2022, p. 66) voiced his belief that N-of-1 "studies do present one promising way forward for precision medicine in aligning with the tenets of evidence-based medicine." Finally, the Oxford Centre for Evidence-Based Medicine (OCEBM) guidelines elevated this research design to the highest level of evidence quality for evaluating treatment effectiveness (Bradbury et al. 2020).

Despite their virtues, the popularity of N-of-1 trials is limited so far. At first sight, this is surprising if we consider two aspects. First, using them to make clinical decisions could allow for the reduction of overtreatment and overall healthcare costs compared to standard care (Scuffham 2010). Second, there is an increasing prevalence of chronic diseases and elderly patients suffering from multiple comorbidities. This corresponds to an increase in the number of patients that would benefit from a careful assessment of their individual treatment effects, which provides the perfect environment for wider use of N-of-1 trials, given that the N-of-1 trials can mainly be used to study chronic conditions that are stable in time. Indeed, such studies are suitable for patients with chronic conditions, multiple comorbidities, polypharmacy, and rare diseases (Vohra et al. 2015) and less adequate for studying individual treatment responses in acute or progressive conditions (Duan et al. 2013).

The limited use of N-of-1 trials in standard clinical practice seems to result from the low feasibility of such studies and the burden imposed both on the physicians willing to use them and on patients whose treatment response is to be assessed. This, for instance, is the explanation provided by Kravitz et al. (2008) based on a literature review and in-depth interviews with proponents of N-of-1 trials, who pointed at the physicians' lack of interest in reducing uncertainty about individual treatment response. As Mirza and colleagues put it:

> The obstacles to conducting N-of-1 trials as an element of routine clinical practice have been too great. For many pharmacists, preparing identical drug and placebo combinations proved too labour-intensive. For clinicians, N-of-1 trials take too much time, even with easy-to-use guidance: preparing questionnaires, instructing patients, and examining the results all require clinician commitment (2017, p. 334).

Furthermore, Selker et al. (2022) observed recently that the stakeholders have not sufficiently recognized the benefits of using N-of-1 studies in clinical practice and listed the requirements for N-of-1 studies to be adopted more broadly: (1) clear articulation of the reasons for patients to participate in the N-of-1 studies; (2) definition of needs and costs of N-of-1 studies; (3) understanding the inter-patient heterogeneity; (4) specification of the criteria for covering participation in N-of-1 studies; (5) understanding how N-of-1 studies help patients and healthcare systems; (6) specification of the types of evidence stemming from N-of-1 trials required by regulatory agencies for drug approval.

It is beyond our aim to consider all such facets of this complex issue. Below, we will focus on various versions of N-of-1 designs involving different methodological choices. Our aim is to assess the epistemic and pragmatic trade-offs of such trials and encourage wider use of the N-of-1 trials that rely on more pragmatic choices. In this view, the N-of-1 design can be simplified to achieve higher feasibility without significantly impacting study integrity.

Let us emphasize that, as in other aspects of scientific research, balancing epistemic (e.g., methodological choices) and non-epistemic aspects (e.g., feasibility) involved in clinical trials is crucial: indeed, such trials do not represent a *value-free* epistemic enterprise but are rather entangled with crucial pragmatic considerations regarding their very applicability. If the aim of a clinical study is to impact medical practice (e.g., by helping us select the best treatment option for a given patient), we do need not only strict methodological requirements but also agile and feasible practices that can be applied in real-world scenarios by clinicians. In other words, N-of-1 trials are susceptible to an adequacy-for-purpose evaluation exactly like scientific models, which demands considering how epistemic factors promote or facilitate their practical aims.[11]

[11] For recent discussions on this type of evaluation in scientific models, see Luck & Elliott (2022); Parker (2020).

3.1 Towards a Greater Feasibility of N-of-1 Trials

Some attempts have already been made to increase the feasibility of N-of-1 trials in day-to-day clinical practice. One way to limit the burden for physicians and patients is to use new technologies for measuring outcomes and reporting. For example, the mobile health app *Trialist* allows for designing and conducting personalized N-of-1 studies. The app was studied in an RCT, where patients suffering from chronic pain were assigned to either the Trialist app or standard care (Barr et al. 2015). Despite patients' positive opinions about the app, no statistically significant difference in pain management was observed (Kravitz et al. 2008). Another attempt that relies on technological developments is described by Mande et al. (2022) pilot study on the iMTracker app involving the N-of-1 design to self-manage chronic conditions such as chronic pain, headaches, anxiety, and depression.

An alternative way to make the N-of-1 trials more feasible would be to simplify their design. This possibility was considered by Kravitz and colleagues, who pointed out that

> in a single-patient head-to-head trial of (generic) omeprazole versus Nexium® for acid reflux, considerable information might be gleaned by simply alternating the two medications (without blinding) every fortnight for a total of eight to twelve weeks and asking patients to keep detailed symptom diaries. Research is needed to determine whether the reduction in costs and burden and the gain in acceptability from such diluted designs would be worth the reduction in scientific rigor (2008, p. 548-549).

In what follows, we will consider a *continuum* of alternative N-of-1 designs, each of which involves different methodological choices and comes with different degrees of feasibility. Before considering such a continuum, though, we need to introduce a distinction between the standard vs. pragmatic dimensions of trials. This distinction draws, by analogy, on a distinction made by Schwartz and Lellouch (1967) between *standard RCTs* (also known as *explanatory*) and *pragmatic RCTs*.

The main purpose of standard RCTs is to assess treatment effectiveness (i.e., assert internal validity). To narrow down the variability of outcomes and make valid conclusions about effectiveness, such RCTs have inclusion and exclusion criteria that make trial participants differ systematically from the population of patients suffering from the condition targeted by the tested treatment. Some standard RCTs thus exclude certain subpopulations (e.g., patients with comorbidities) that experience outcomes systematically different from the population-wide averages. By contrast, pragmatic RCTs aim to test the effects of treatment on the population of patients suffering from the condition targeted by the tested treatment, i.e., establishing that treatment benefits the *actual* population of patients (pragmatic RCTs have become increasingly popular, see Patsopoulos 2011). For this purpose, they have broader inclusion criteria and fewer exclusion criteria, pose only a limited burden related to participation, and rely on outcome measures that are relevant to study participants and patients (Loudon et al. 2015). Moreover, some pragmatic RCTs do not use blinding to mask patient assignment, but substantial heterogeneity exists in the design of pragmatic trials (Dal-Ré et al. 2018).

It is worth noting that pragmatic RCTs do not represent a solution to the problems we analyzed in Sections 1 and 2. In fact, by analyzing a broader population, the outcomes observed in a pragmatic RCT may be more heterogeneous than in an explanatory RCT; thus, such trials may require larger sample sizes to achieve the same statistical power under the assumption of the same effect size. But even if pragmatic RCTs deliver effect size estimates that are closer to the actual average benefit of the population of patients suffering from a condition, obtaining outcome measures that are closer to the true average effect size of the target population does not address the problem of individual treatment effect heterogeneity. As La Caze argued:

> The main selling point for large pragmatic trials is that by allowing considerably more variability in the patients recruited and in the non-experimental treatments that they receive, the trial provides more insight into the likely effects of the treatment in routine clinical care. This is true to an extent. A well-conducted *successful* large pragmatic trial provides good evidence that the *average* effects of giving the treatment are positive. However, in extending the results of such a trial to a given specific population or individual, the critical assumption is that the positive average effects are consistent across the many subpopulations included in the trial. Sometimes this seems to be a reasonable assumption, but often it is an assumption that is difficult to justify (2016, p. 204-205).

Thorpe et al. (2010) suggested that the distinction between pragmatic and explanatory trials is not to be considered dichotomous but in terms of a continuum (see also Patsopoulos 2011). As we mentioned, the same can be said about single-patient designs and the alternations of N-of-1 design and the traditional trial-and-error approach to choosing a therapy. On this view, like the distinction between alternative RCTs, various types of N-of-1 trials can be put on a continuum that takes into account epistemic and pragmatic aspects (see Figure 1).

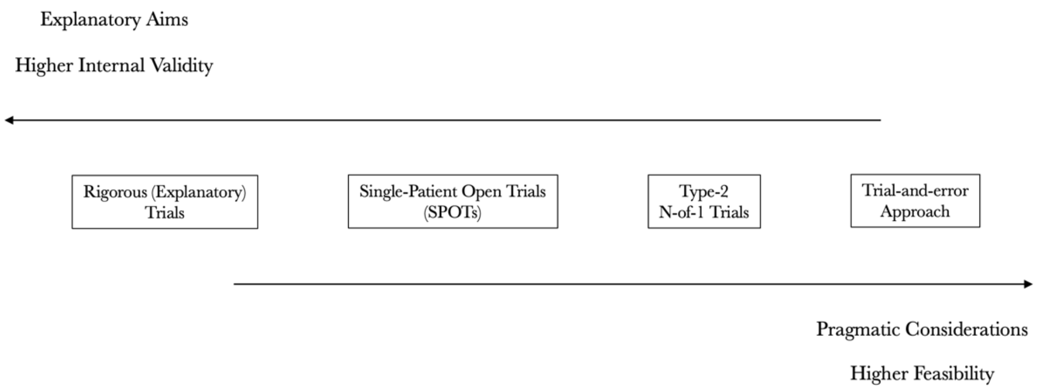

Figure 1: A continuum of single-patient designs prioritizing validity and feasibility to different degrees.

The left end of the continuum is populated by *standard (explanatory) trials*, which aim to higher results' integrity but pose a burden on participants and physicians. These can be understood as such trials that include all possible measures to assert integrity, such as blinding, randomized assignment, and washout periods. On the right end of the continuum, we have instead *trial-and-error approaches* to select therapies. Such approaches, which are the default strategy to making therapeutic decisions when a patient does not respond to the treatment of first choice, rely on an informal assessment of treatment response and the prescription of alternatives when the patient is not content with their treatment outcome due to poor symptoms control or adverse effects (Kravitz et al. 2014, ch. 1).

Although such two extremes capture idealized versions of existing trials, introducing this distinction might be useful for clinicians and patients willing to make decisions adjusted to a single patient. Moreover, the scheme in Figure 1 could be used (and further refined) to frame various designs within a comprehensive framework where single-patient designs can be seen as somewhat explanatory and somewhat pragmatic depending on the emphasize a design puts on methodological rigor (at the expense of lower feasibility) or pragmatic considerations (at the expanse of lower internal validity).

There is, in fact, much heterogeneity in how N-of-1 trials can be designed. But as Kravitz et al. (2014, ch. 1) pointed out, the defining feature of N-of-1 trials is the use of multiple crossovers conducted on a single patient. What are, then, the peculiarities of different designs? What kind of methodological choices do they typically make?

As we mentioned, on the explanatory end of the continuum, we find designs that leverage blinding, randomized assignment, and washout periods to achieve higher rigor.

As regards *randomization* and *blinding*, they are used in the majority but not all N-of-1 studies (Punja et al. 2016). Considering, however, that the decision-makers in clinical practice are usually concerned with the overall effect size of treatments rather than the net effect of the therapy in comparison to the placebo, blinding may not be necessary (Kravitz et al. 2014, ch. 1). N-of-1 trials dispensing of blinding can report biased results in cases when patients are too optimistic about one of the treatments or exaggerate a treatment's harms (Howard & Rajasundaram 2022), but the nocebo effect seems to have only a limited impact on patients' decisions after the N-of-1 trial concludes (Tudor et al. 2022). In contrast, if a patient is expected to have positive views about only the tested treatment and not the control (e.g., in a N-of-1 study that tests the effects of an expensive drug and its generic version), then blinding might be necessary as the placebo effect confounds results.

Some N-of-1 studies use *washout periods*, whose application depends on a pharmacokinetic understanding of a drug's metabolism and effects duration. In a paper discussing the use of cross-over trials in drug development, Senn (2001) admitted that the trialists should determine the length of the treatment period or washout based on knowledge concerning carry-over effects. Since some treatments are less likely to have carryover effects, washout periods can sometimes be omitted without much impact on the risk of biases. However, N-of-1 studies differ regarding design even when testing similar drugs' effectiveness. For example, Kronish et al. 2018) reviewed five N-of-1 studies assessing the effectiveness of depression drugs: three out of five psychiatric trials reported using a washout period shorter than or equal to one day, and two other studies set its duration at one or six weeks.[12]

In some cases, methodological rigor imposes a burden that exceeds the patient's and physician's resources, which undermines the use of N-of-1 trials altogether. Resigning from some methodological aspects, potentially reducing rigor, may nonetheless make N-of-1 trials more popular and benefit the patients participating in them. In fact, each of the methodological decisions above results from pragmatic considerations (a study's feasibility, the worry that having washout periods will lead to patient deterioration), also bearing in mind the context (a patient's values, diseases-specific characteristics, available resources, etc.). In other words, differences among N-of-1 designs are not shaped exclusively by epistemic reasons, but also by non-epistemic factors.

[12] Another aspect that could reduce the burden on physicians, thus increasing the feasibility of single-patient trials, regards using *outcome measures* that are easier to self-report. While some N-of-1 studies use objective outcome measures, Gabler et al. (2011, p. 764) reported that 82% of trials employed "patient-reported outcome measurement such as a patient diary (46%), visual analog scale (27%), or a Likert scale (12%)".

Closer to the pragmatic end of the continuum (depicted in Figure 1) are thus attempts to simplify the N-of-1 design.

For example, Smith, Yelland & Del Mar supported the use of *Single Patient Open Trials* (SPOTs), which "lie somewhere in between formal N-of-1 trials and totally informal trials of treatment in terms of rigor" (2015, p. 195). SPOTs employ at least one crossover with in-between washout, rely on patient-centered outcome measures, and do not require physicians to arrange the study in a way that asserts blinding, randomized assignment, or statistical analysis of results. The rationale for using SPOTs instead of the standard N-of-1 design is that they are less demanding to arrange than the latter.

Still, SPOTs are more demanding than the trial-and-error approach but promise higher validity of results. The reason is that repeated crossovers make confounding effects less likely, washout periods prevent carryover effects, and predefining outcome measures assert that neither patients nor doctors choose outcome assessment post-factum based on non-epistemic values. However, the higher feasibility of SPOTs is nevertheless related to the higher risk of biases. For example, using patient-centered outcome measures without conducting statistical analysis poses a risk of interpreting random differences between treatment regimes as resulting from drugs' actions.

A more radical alternative is the *type-2 N-of-1* design. Selker et al. (2022) recently argued that in some cases (e.g., when testing treatments for severe, rare diseases), having only one cycle of candidate treatment alternation is sufficient. But such studies can also be used to study the effects of interventions targeting common chronic diseases in cases when the expected effect size of the intervention significantly exceeds the potential impact of all other confounders. In a sense, type 2 N-of-1 studies can be considered a more pragmatic version of SPOTs. However, these studies can be seen as close to a trial-and-error approach that uses formal outcome assessment defined prior to trying a new therapy: if the treatment effect size is expected to vastly exceed the summary impact of confounders (such as expected deterioration during the trial duration), then this type of design offers a promising way of testing treatment candidates.

Overall, there exists a menu of alternations in the N-of-1 designs aimed at choosing the best treatment options that differ with respect to the use of assignment procedures, blinding, outcome assessment, the number of crossovers, and washout periods. Decisions concerned with each of those characteristics of N-of-1 trials can arguably be made separately depending on the patient's values, the resources available to the physician, and treatment- and disease-specific characteristics. This implies that the question of which design would do better is highly contextual and will depend on the explanatory and pragmatic aims at stake.

3.2 Potential Caveats

In the previous section, we introduced a distinction between explanatory (standard) and pragmatic trials. Two observations are in place here.

First, the distinction between explanatory and pragmatic RCTs, on which our distinction draws by analogy, has received criticism in the literature. For example, Karanicolas et al. (2009a) criticized the notion of pragmatic trials on the grounds that there are varying perspectives in clinical decision-making, and hence the results of such trials are not directly applicable to each decision problem at hand. Kent & Kitsios (2009) argued that extrapolating the results of pragmatic RCTs to individual patients may be as problematic as the extrapolation of outcomes reported by explanatory trials and warned that diminishing the problem of extrapolation in such cases may lead to introducing harmful policies. Pawson (2019) pointed out that the problem of extrapolation is simplified in the literature concerned with the pragmatic-explanatory trials distinction, and regardless of where a particular trial is located on this continuum, no single result can be generalized without a mechanistic understanding of how an intervention works in a particular context. Recently, Tresker (2022) analyzed the relationship between the pragmatic/explanatory distinction with generalizability, internal validity, external validity, efficacy, and effectiveness, and argued that the distinction is conceptually problematic. However, despite being aware of some drawbacks of the distinction, other authors support the use of pragmatic trials in medicine (e.g., Patsopoulos 2022; Casey et al. 2022). We think that the distinction is useful as it allows one to focus on the trade-off between feasibility and epistemic rigor. This is particularly relevant given that the variety of pragmatic trials has grown in the last ten years (Palakshappa 2022), which speaks of their growing importance, even if some conceptual issues remain to be resolved in future research.

Second, applying the explanatory/pragmatic distinction to N-of-1 trials fruitfully or coherently may be difficult for the reason that N-of-1 trials “enable us to compare two treatments under the conditions in which they would be applied in practice” (Schwartz and Lellouch 1967, p. 638) and deliver evidence “aimed at *decision*" (p. 647). While we fully agree that all N-of-1 trials are aimed at assessing treatment effectiveness for the patient participating in them (and some constitute evidence amalgamated with other N-of-1 trials), for patients other than trial participants, some N-of-1 trials create an artificial context, as some features of N-of-1 studies are unlikely to be used in clinical practice (e.g., in the traditional trial-and-error approach to choosing therapy). For instance, wash-out periods, which are used in some N-of-1 studies (those located towards the explanatory end of the continuum) are unlikely to be used when treatments are changed in the clinic because they pose a risk of deterioration for the patient not receiving any treatment for their condition.

Other distinctions have been introduced to replace the explanatory/pragmatic divide. For instance, Karanicolas et al. (2009b) distinguished between mechanistic trials that assess a biological relationship and 'practical' studies that deliver evidence for decision-makers in the clinic. Our distinction could be read in the latter sense, in terms of 'feasibility', so that N-of-1 designs that are closer to the explanatory end of the continuum are less feasible, i.e., more difficult to execute in everyday clinical practice, while those closer to the pragmatic end of the continuum are more feasible. However, such a simplification would omit the matter of fact that the trials that are easier to implement in clinical practice are epistemically inferior to those that are less feasible. For this reason, again, we still think that it can be useful for providing a workable taxonomy of the many existing trials.

Furthermore, the literature does include some suggestions about applying the explanatory/pragmatic distinction to N-of-1 studies. For example, in a recent article criticizing the distinction, Tresker (2022) considered whether pragmatic trials are better in terms of representativeness of the population of patients and observed that

> [r]epresentativeness can certainly be important in certain contexts, though it is inadequate as a unifying conceptual approach for indicating a trial's potential for informing valid treatment effectiveness claims. Possibly only in N-of-1 trials is the "population" the same, although even here the "population" is different at different time points, which complicates simple inferences of effectiveness because of carryover effects and other issues (pp. 315-316).

This suggests that some types of N-of-1 trials (e.g., those including washout periods) might have a higher degree of verisimilitude to the counterfactual situation of a patient being treated in a clinic.

Before concluding, it is worth asking how one can evaluate the success of different types of N-of-1 designs. At present, there is no definitive evidence of the effects of participating in standard versus pragmatic versions of N-of-1 trials. The lack of such evidence can depend on the mixed results of existing RCTs comparing the use of standard N-of-1 trials in clinical practice to standard care (Samuel et al. 2022).

The ideal way to compare alternative designs would be to run an RCT where patients are randomly assigned to either a treatment group involved in a pragmatic version of an N-of-1 study or a control group employing the standard N-of-1 design. If the two trial arms tested the same treatment and were sufficiently powered, the observed difference in outcomes (if any), could be ascribed to how the two types of N-of-1 trials are designed.

This methodology has been applied to compare the effects of participating in (standard) N-of-1 study versus standard care for patients suffering from irreversible chronic airflow limitation (Mahon et al. 1996; Mahon 1999). Such research shows that using N-of-1 trials to assess individual response to theophylline allows for reducing drug use without adverse effects. The results of 39% of 57 N-of-1 trials conducted at McMaster Hospital convinced physicians to change their treatment advice before patients participated in the trial (*ibid.*).

Samuel et al. (2022) reviewed the literature comparing the outcomes of N-of-1 trials to standard care using parallel arm design. Only one out of 12 studies showed the superiority of the N-of-1 arm in the primary outcome, and five studies reported statistically significant and positive differences in at least one secondary outcome. However, all those studies suffered from methodological drawbacks such as the lack of blinding patients and outcome assessors, and non-randomized assignment. As we mentioned in Section 3, other studies reported positive effects experienced by patients participating in N-of-1 studies (e.g., Duan et al. 2013; Guyatt et al. 1990; Nikles et al. 2021; Scuffham et al. 2010).

Although the success of different N-of-1 designs is yet to be assessed in randomized trials, we believe that the implications of simplifying the standard N-of-1 design can be predicted based on an empirically informed methodological analysis of the decisions involved in planning and executing N-of-1 trials in clinical practice. Designing N-of-1 trials in a more pragmatic way would allow practitioners to choose a therapy more suitable for a given patient instead of using the recommendations for the average patient or applying the trial-and-error approach. Even if the epistemic gain from using such a pragmatic N-of-1 trial is lower compared with the application of standard designs, pragmatic trials are more feasible and hence more likely to become part of the standard clinical practice.

Conclusions

In this paper, we argued that neither RCTs (a key research design in EBM) nor GWAS/GWIS (the main tools of genomics-based P-medicine) can easily elucidate and predict individual treatment responses. A convincing solution to handle individual variability lies in N-of-1 trials. Unfortunately, their use in everyday clinical practice is limited at present. We have analyzed a continuum of single-patient designs that range from restrictively designed N-of-1 trials that mimic explanatory RCTs to the trial-and-error approach. We have argued that the N-of-1 trials that are closer to the pragmatic end of the continuum are more suited for everyday clinical practice while their epistemic tradeoffs are limited.

More specifically, in Sections 1 and 2, we argued that both standard RCTs and GWAS struggle with the characterization and control of inter-individual heterogeneity at various levels of analysis: first, we showed that gene-treatment interactions (G×T) and gene-treatment-environment interactions (G×T×E) are an ineliminable source of individual differences in response to treatment that undermine using the results of RCTs to inform therapeutic decisions concerned with a single patient; second, systematic attempts to investigate such interactions through genomics methods come with major limitations. This may suggest the need for larger populations with the hope that genetic and environmental variability would 'average out.'

However, here we considered a different strategy: identifying principled methods to *capitalize* on individual variability rather than trying to exclude it from the picture. This basic idea is consistent with recent trends toward P-medicine. In classical clinical trials aimed at establishing universally applicable treatments, the variability in populations is often perceived as an impediment and a 'threat' to the reliability of the results. But, in P-medicine, such variability is arguably the main source of information: understanding where it comes from and using such knowledge for a patient's good, are key epistemic goals. Heterogeneity in treatment response is thus precisely the kind of factor that P-medicine should aim to include into medical models and clinical decisions.

In Section 3, we considered N-of-1 trials as one potential methodology that would help handle individual variability effectively. However, standard N-of-1 trials pose a significant burden on practitioners and patients, and their complication is likely a factor that limits their use in clinical practice despite the growing prevalence of chronic diseases and comorbidities. We thus applied the distinction between explanatory and pragmatic RCTs to analyze the differences among the menu of single-patient trials and argued for the use of N-of-1 studies that are designed in a more pragmatic way.

The main selling point of pragmatic N-of-1 trials is that they would solve the problem of extrapolation and uncertainty about individual-level gene-environment-treatment interactions: indeed, the evidence informing therapeutic decisions about a given patient stems from the outcomes of *that* patient. For this reason, applying N-of-1 trials in everyday clinical practice would lead to more precise therapeutic decisions. So far, the N-of-1 trials are rarely used in everyday clinical practice; due to the lower burden for both the practitioner and the patient, pragmatic N-of-1 trials represent a more suitable choice for everyday clinical practice than the standard design. If compared with standard N-of-

1 trials, pragmatic designs such as SPOTs, type-2 N-of-1 studies, and other alternations to the single-patient designs involve methodological choices such as the use of a non-randomized assignment procedure and pragmatic outcome measures as well as the possibility of resigning from blinding and washout periods. Although such designs might be more susceptible to biases than the standard one, they would outperform both informal trials of therapy with the trial-and-error approach and decisions based on population-wide averages.

Although more pragmatic alternations to the N-of-1 design would bring about substantial benefits in terms of both simplicity and feasibility — maximizing the overall value of N-of-1 trials — we need to point out a few limitations. First, like any type of single-patient trials, pragmatic N-of-1 trials have a specific area of application: they can only be used to inform therapeutic decisions regarding patients suffering from a stable, chronic condition and treatments that tend to alleviate the symptoms but do not cure them (see Nikles & Mitchell 2015, pp. 51 et seq.). Second, N-of-1 trials designed pragmatically should not, due to their epistemic shortcomings compared to the standard N-of-1 trials, be understood as a method to gather evidence for new treatments, but rather to inform therapeutic decisions concerned with a single patient when two or more alternative therapies have been approved by a drug agency. However, standard N-of-1 studies have been applied in the field of precision oncology to develop fine-tuned treatments (Gouda et al. 2023) and supported as a cost-effective strategy for drug development in other fields (Mirza et al. 2017). Pragmatic N-of-1 trials will prove useful in deciding about treatments whose mechanism of action is poorly understood, including details about individual-level G×T×E. This design can also be applied to studying harms in cases when two or more alternative therapies are effective but cause negative side effects — for instance, Herret et al. (2021) conducted a series of N-of-1 trials to assess the relationship between muscle symptoms and the use of different types of statins.

In other words, the type of evidence that pragmatic N-of-1 trials would help gather is about a single patient's response to treatment. It should be noted, however, that this evidence could in principle have a 'second use' to inform a new hypothesis on G×T×E to be further assessed. For instance, if a patient suffering from asthma reacts positively to ICS and not to montelukast, data can be collected about the patient's systematic environmental exposures to allergens or air pollutants. This way, pragmatic trials have the potential to provide evidence on relevant therapy-environment interactions and advise further, more systematic investigation of such interactions. Furthermore, evidence stemming from N-of-1 trials testing the same compounds might also be useful for patients not participating in them when amalgamated in aggregate N-of-1 trials. Further research is needed to assess the impact of changes in N-of-1 designs on the reliability of such amalgamated treatment effect estimates.

Acknowledgements: Author contributions Both authors contributed equally to the conceptual development and the writing of this manuscript. Funding The first draft of this work and those changes that resulted from the first round of revisions have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 805498).

References

Abettan, C., & Welie, J. V. (2020). The impact of twenty-first century personalized medicine versus twenty-first century medicine's impact on personalization. *Philosophy, Ethics, and Humanities in Medicine*, *15*(1), 1-8.

Anjum, R. L., Copeland, S., & Rocca, E. (2020). *Rethinking Causality, Complexity and Evidence for the Unique Patient: A CauseHealth Resource for Healthcare Professionals and the Clinical Encounter*. Springer.

Aron, D. C. (2020). Managing Patients: Evidence-Based Medicine Meets Human Complexity. In *Complex Systems in Medicine* (pp. 63-74). Springer, Cham.

Barr, C., Marois, M., Sim, I., Schmid, C. H., et al. (2015). The PREEMPT study-evaluating smartphone-assisted n-of-1 trials in patients with chronic pain: study protocol for a randomized controlled trial. *Trials*, *16*(1), 1-11.

Baverstock, K. (2019). Polygenic scores: Are they a public health hazard? *Progress in biophysics and molecular biology*, *149*, 4-8.

Blunt, C. J. (2019). The Dismal Disease: Temozolomide and the Interaction of Evidence. Available at SSRN 3444926.

Borgerson, K. (2008). Valuing and Evaluating Evidence in Medicine. https://tspace.library.utoronto.ca/bitstream/1807/11182/1/Borgerson_Kirstin_200806_PhD_Thesis.pdf (Accessed July 9, 2023).

Borgerson, K. (2009). Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. *Perspectives in Biology and Medicine*, *52*(2), 218-233.

Bradbury, J., Avila, C., & Grace, S. (2020, January). Practice-based research in complementary medicine: could N-of-1 trials become the new gold standard? In *Healthcare* 8(1), 15.

Burt, C. H. (2022). Challenging the Utility of Polygenic Scores for Social Science: Environmental Confounding, Downward Causation, and Unknown Biology. *Behavioral and Brain Sciences*, 1-36.

Casey, J. D., Beskow, L. M., Brown, J., Brown, S. M., Gayat, É., Gong, M. N., ... & Collins, S. P. (2022). Use of pragmatic and explanatory trial designs in acute care research: lessons from COVID-19. *The Lancet Respiratory Medicine*, *10*(7), 700-714.

Caspi, A., Sugden, K., Moffitt, T. E., et al. (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science*, 301(5631), 386-389.

Chabris, C. F., Hebert, B. M., Benjamin, D. J., et al. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological science*, *23*(11), 1314-1323.

Chabris, C. F., Lee, J. J., Benjamin, D. J., et al. (2013). Why it is hard to find genes associated with social science traits: Theoretical and empirical considerations. *American journal of public health*, *103*(S1), S152-S166.

Chang, H. S., Won, E., Lee, H. Y., Ham, B. J., & Lee, M. S. (2015). Association analysis for corticotropin releasing hormone polymorphisms with the risk of major depressive disorder and the response to antidepressants. *Behavioural Brain Research, 292*, 116-124.

Chen, R., Shi, L., Hakenberg, J., et al. (2016). Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nature biotechnology*, *34*(5), 531-538.

Chow, S. C. (Ed.). (2018). Encyclopedia of biopharmaceutical statistics-four volume set. CRC Press.

Cook, T. D., & DeMets, D. L. (2007). Introduction to statistical methods for clinical trials. CRC

Cooper, D. N., Krawczak, M., Polychronakos, C., et al. (2013). Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human genetics*, *132*(10), 1077-1130.

Coulthard, S. A., Hogarth, L. A., Little, M., Matheson, E. C., Redfern, C. P., Minto, L., & Hall, A. G. (2002). The effect of thiopurine methyltransferase expression on sensitivity to thiopurine drugs. Molecular Pharmacology, 62(1), 102–109.

Craver, C. F., Dozmorov, M., Reimers, M., & Kendler, K. S. (2020). Gloomy prospects and roller coasters: finding coherence in genome-wide association studies. *Philosophy of Science*, *87*(5), 1084-1095.

Dai, J. Y., Hsu, L., & Kooperberg, C. (2018). Two-stage testing for genome-wide gene-environment interactions. In *Handbook of Statistical Methods for Case-Control Studies* (pp. 459-474). Chapman and Hall/CRC.

Dal-Ré, R., Janiaud, P., & Ioannidis, J. P. A. (2018). Real-world evidence: How pragmatic are randomized controlled trials labeled as pragmatic? *BMC Medicine, 16*(1), 49.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, *210*, 2-21.

Dick, D. M., Agrawal, A., Keller, M. C., et al. (2015). Candidate gene–environment interaction research: Reflections and recommendations. *Perspectives on Psychological Science, 10*(1), 37-59.

Downes, S. M., & Matthews, L. (2019). Heritability. In N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2020/entries/heredity (Accessed July 9, 2023).

Duan, N., Kravitz, R. L., & Schmid, C. H. (2013). Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of Clinical Epidemiology, 66*(8), S21-S28.

Farzan, N., Vijverberg, S. J., Kabesch, M., et al. (2018). The use of pharmacogenomics, epigenomics, and transcriptomics to improve childhood asthma management: where do we stand? *Pediatric pulmonology*, *53*(6), 836-845.

Feinstein, A. R. (1995). Meta-analysis: statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, *48*(1), 71-79.

Freund, R. J., & Wilson, W. J. (2003). *Statistical Methods*. Elsevier.

Fries, G. R. (2020). Polygenic risk scores and their potential clinical use in psychiatry: are we there yet? *Brazilian Journal of Psychiatry*, *42*, 459-460.

Fuller, J. (2021). The myth and fallacy of simple extrapolation in medicine. *Synthese, 198*(4), 2919-2939.

Gabler, N. B., Duan, N., Vohra, S., & Kravitz, R. L. (2011). N-of-1 trials in the medical literature: a systematic review. *Medical Care, 49*(8), 761-768.

Gamma, A. (2016). Personalized and Precision Medicine. In *The Routledge Companion to Philosophy of Medicine* (pp. 411-421). Routledge.

Génin, E. (2020). Missing heritability of complex diseases: case solved? *Human Genetics*, *139*(1), 103-113.

Giacomelli, R., Afeltra, A., Bartoloni, E., et al. (2021). The growing role of precision medicine for the treatment of autoimmune diseases; results of a systematic review of literature and Experts' Consensus. *Autoimmunity Reviews*, *20*(2), 102738.

Giangrande, E. J., Weber, R. S., & Turkheimer, E. (2022). What Do We Know About the Genetic Architecture of Psychopathology? *Annual Review of Clinical Psychology, 18,* 19-42.

Glass, G. V. (2000). Education in Two Worlds: Meta-Analysis at 25: A Personal History. https://nepc.colorado.edu/blog/meta-analysis (Accessed July 9, 2023).

Global Initiative for Asthma GINA (2019). Global Strategy for Asthma Management and Prevention. https://ginasthma.org/wp-content/uploads/2019/06/GINA-2019-main-report-June-2019-wms.pdf. Updated version: https://ginasthma.org/wp-content/uploads/2022/07/GINA- Main-Report-2022-FINAL-22-07-01-WMS.pdf (Accessed July 9, 2023).

Gouda, M. A., Buschhorn, L., Schneeweiss, A., Wahida, A., & Subbiah, V. (2023). N-of-1 Trials in Cancer Drug Development. *Cancer Discovery*, *13*(6), 1301-1309.

Gravel, S., Henn, B. M., Gutenkunst, R. N., et al. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, *108*(29), 11983-11988.

Greenhalgh, T., Howick, J., & Maskrey, N. (2014). Evidence based medicine: a movement in crisis? *The British Medical Journal*, *348*.

Greenland, S. (1990). Randomization, statistics, and causal inference. *Epidemiology*, 421-429.

Guyatt, G. H. & Rennie, D. (eds.) (2002). *Users' guide to the medical literature: Essentials of evidence-based clinical practice*. Chicago: American Medical Association Press.

Guyatt, G. H., Keller, J. L., Jaeschke, R., et al. (1990). The n-of-1 randomized controlled trial: clinical usefulness: our three-year experience. *Annals of Internal Medicine*, *112*(4), 293-299.

Hellwege, J. N., Keaton, J. M., Giri, A., et al. (2017). Population stratification in genetic association studies. *Current Protocols in Human Genetics*, *95*(1), 1-22.

Hernan, M. A., & Robins, J. M. (2018). *Causal Inference: What If.* New York: CRC Press.

Herrett, E., Williamson, E., Brack, K., et al. (2021). Statin treatment and muscle symptoms: series of randomised, placebo controlled n-of-1 trials. *The British Medical Journal*, *372*.

Hewitt, J. K. (2012). Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behavior Genetics*, *42*(1), 1.

Howard, J., & Rajasundaram, S. (2022). Role of Blinding in N-of-1 Trials. *Circulation: Cardiovascular Quality and Outcomes*, *15*(6), e008914.

Hyde, L. W., Bogdan, R., & Hariri, A. R. (2011). Understanding risk for psychopathology through imaging gene-environment interactions. *Trends in Cognitive Sciences*, *15*(9), 417-427.

Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Jama*, *294*(2), 218-228.

Jukola, S. (2019). Casuistic Reasoning, Standards of Evidence, and Expertise on Elite Athletes' Nutrition. *Philosophies*, *4*(2), 19.

Kaplan, J. M., & Turkheimer, E. (2021). Galton's Quincunx: Probabilistic causation in developmental behavior genetics. *Studies in History and Philosophy of Science*, *88*, 60-69.

Karanicolas, P. J., Montori, V. M., Devereaux, P. J., Schünemann, H., & Guyatt, G. H. (2009b). A new'Mechanistic-Practical" Framework for designing and interpreting randomized trials. *Journal of clinical epidemiology*, *62*(5), 479-484.

Karanicolas, P. J., Montori, V. M., Schünemann, H., & Guyatt, G. H. (2009a). "Pragmatic" clinical trials: from whose perspective?. *BMJ Evidence-Based Medicine*, *14*(5), 130-131.

Katsanis, N. (2016). The continuum of causality in human genetic disorders. *Genome Biology*, *17*(1), 1-5.

Keers, R., & Uher, R. (2012). Gene–environment interaction in major depression and antidepressant treatment response. *Current Psychiatry Reports*, *14*(2), 129-137.

Kent, D. M., & Kitsios, G. (2009). Against pragmatism: on efficacy, effectiveness and the real world. *Trials*, *10*(1), 1-3.

Kent, D. M., Nelson, J., Dahabreh, I. J., et al. (2016). Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. *International Journal of Epidemiology*, *45*(6), 2075-2088.

Kent, D. M., Rothwell, P. M., Ioannidis, J., et al. (2010). Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*, *11*(1), 1-11.

Kersten, E. T., & Koppelman, G. H. (2017). Pharmacogenetics of asthma: toward precision medicine. *Current Opinion in Pulmonary Medicine, 23*(1), 12-20.

Kravitz, R. L., Duan, N., & Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, *82*(4), 661-687.

Kravitz, R. L., Duan, N., Niedzinski, E. J., et al. (2008). What Ever Happened to N-of-1 Trials? Insiders' Perspectives and a Look to the Future. *The Milbank Quarterly*, *86*(4), 533-555.

Kravitz, R. L., Duan, N., Vohra, S., Li, J., & DEcIDE Methods Center N-of-1 Guidance Panel. (2014). Introduction to N-of-1 trials: indications and barriers. Design and implementation of N-of-1 trials: A user's guide. AHRQ Publication No. 13(14)-EHC122-EF.

Kronish, I. M., Alcántara, C., Duer-Hefele, J., et al. (2017). Patients and primary care providers identify opportunities for personalized (N-of-1) trials in the mobile health era. *Journal of clinical Epidemiology*, *89*, 236-237.

Kronish, I. M., Hampsey, M., Falzon, L., et al. (2018). Personalized (N-of-1) trials for depression: a systematic review. *Journal of Clinical Psychopharmacology*, *38*(3), 218.

La Caze, A. (2009). Evidence-based medicine must be…. *Journal of Medicine and Philosophy*, *34*(5), 509-527.

La Caze, A. (2013). Why randomized interventional studies. *Journal of Medicine and Philosophy*, *38*(4), 352-368.

La Caze, A. (2016). The randomized controlled trial: internal and external validity. In Solomon et al. (eds.) *The Routledge Companion to Philosophy of Medicine* (pp. 209-222). Routledge.

Lawler, I., & Zimmermann, G. (2021). Misalignment between research hypotheses and statistical hypotheses: A threat to evidence-based medicine? *Topoi*, *40*(2), 307-318.

Lawson, D. J., Davies, N. M., Haworth, S., et al. (2020). Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Human Genetics*, *139*(1), 23-41.

Lemoine, M. (2017). Neither from words, nor from visions: understanding p-medicine from innovative treatments. *Lato Sensu: Revue de la Société de Philosophie des Sciences*, 4(2).

Lima, J. J., Zhang, S., Grant, A., et al. (2006). Influence of leukotriene pathway polymorphisms on response to montelukast in asthma. *American Journal of Respiratory and Critical Care Medicine*, *173*(4), 379-385.

Loudon, K., Treweek, S., Sullivan, et al. (2015). The PRECIS-2 tool: designing trials that are fit for purpose. *The British Medical Journal*, *350*.

Lusk, G., & Elliott, K. C. (2022). Non-epistemic values and scientific assessment: an adequacy-for-purpose view. *European Journal for Philosophy of Science*, *12*(2), 1-22.

Lynch, K. E. (2021). The meaning of "cause" in genetics. *Cold Spring Harbor Perspectives in Medicine*, *11*(9), a040519.

Mahon, J. L., Laupacis, A., Hodder, R. V., et al. (1999). Theophylline for irreversible chronic airflow limitation: a randomized study comparing n of 1 trials to standard practice. *Chest*, *115*(1), 38-48.

Mahon, J., Laupacis, A., Donner, A., & Wood, T. (1996). Randomised study of n of 1 trials versus standard practice. *The British Medical Journal*, *312*(7038), 1069-1074.

Mande, A., Moore, S. L., Banaei-Kashani, F., et al. (2022). Assessment of a Mobile Health iPhone App for Semiautomated Self-management of Chronic Recurrent Medical Conditions Using an N-of-1 Trial Framework: Feasibility Pilot Study. *JMIR Formative Research*, *6*(4), e34827.

Manolio, T. A., Collins, F. S., Cox, N. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747-753.

Marchini, J., Cardon, L. R., Phillips, M. S., & Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*, *36*(5), 512-517.

Martinez, M., & Teira, D. (2020). Why Experimental Balance is Still a Reason to Randomize. http://e-spacio.uned.es/fez/eserv/bibliuned:501130/MartinezTeiraBJPSRandomizationBalance.pdf (Accessed July 9, 2023).

Matthews, L. J., & Turkheimer, E. (2022). Three legs of the missing heritability problem. *Studies in History and Philosophy of Science*, *93*, 183-191.

Maziarz, M. (2022). Is meta-analysis of RCTs assessing the efficacy of interventions a reliable source of evidence for therapeutic decisions? Studies in History and Philosophy of Science, 91, 159–167.

McClellan, J., & King, M. C. (2010). Genetic heterogeneity in human disease. *Cell*, *141*(2), 210-217.

McGlothlin, A. E., & Lewis, R. J. (2014). Minimal clinically important difference: defining what really matters to patients. *Jama, 312*(13), 1342-1343.

Mirza, R. D., Punja, S., Vohra, S., & Guyatt, G. (2017). The history and development of N-of-1 trials. *Journal of the Royal Society of Medicine*, *110*(8), 330-340.

Moise, N., Wood, D., Cheung, Y. K. K., et al. (2018). Patient preferences for personalized (N-of-1) trials: a conjoint analysis. *Journal of Clinical Epidemiology*, *102*, 12-22.

National Institute for Health and Care Excellence. (2014). Developing NICE guidelines: the manual. London: National Institute for Health and Care Excellence. https://www.nice.org.uk/media/default/about/what-we-do/our-programmes/developing-nice-guidelines-the-manual.pdf (Accessed July 9, 2023).

Nikles, J., & Mitchell, G. (Eds.). (2015). *The essential guide to N-of-1 trials in health* (pp. 1-7). New York: Springer.

Nikles, J., Daza, E. J., McDonald, S., et al. (2021). Creating evidence from real world patient digital data. *Frontiers in Computer Science*, 61.

Nikles, J., Daza, E. J., McDonald, S., et al. (2021). Creating evidence from real world patient digital data. Frontiers in Computer Science, 61.

OCEBM Levels of Evidence Working Group. (2009). *The Oxford 2009 Levels of Evidence*. Oxford Centre for Evidence-Based Medicine. https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence (Accessed July 9, 2023).

Oftedal, G. (2022). Proportionality of single nucleotide causation. *Studies in History and Philosophy of Science*, *93*, 215-222.

Ogino, S., Fuchs, C. S., & Giovannucci, E. (2013a). How many molecular subtypes? Implications of the unique tumor principle in personalized medicine. *Expert Review of Molecular Diagnostics*, *12*(6), 621-628.

Ogino, S., Lochhead, P., Chan, A. T., et al. (2013b). Molecular pathological epidemiology of epigenetics: emerging integrative science to analyze environment, host, and disease. *Modern Pathology*, *26*(4), 465-484.

Palakshappa, J. A., Gibbs, K. W., Lannan, M. T., Cranford, A. R., & Taylor, S. P. (2022). Systematic Review of the "Pragmatism" of Pragmatic Critical Care Trials. *Critical Care Explorations*, *4*(7).

Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, *87*(3), 457-477.

Patsopoulos, N. A. (2011). A pragmatic view on pragmatic trials. *Dialogues in Clinical Neuroscience, 13(2),* 217-224.

Patsopoulos, N. A. (2022). A pragmatic view on pragmatic trials. *Dialogues in clinical neuroscience*. 13*(2)*, 217-224.

Pawson, R. (2019). The shrinking scope of pragmatic trials: a methodological reflection on their domain of applicability. *Journal of Clinical Epidemiology*, *107*, 71-76.

Pedersen, L. H. (2017). The risks associated with prenatal antidepressant exposure: time for a precision medicine approach. *Expert Opinion on Drug Safety*, *16*(8), 915-921.

Plutynski, A. (2020). Why precision oncology is not very precise (and why this should not surprise us). *Philosophical Issues in Precision Medicine*. Springer.

Press.

Punja, S., Bukutu, C., Shamseer, L., et al. (2016). N-of-1 trials are a tapestry of heterogeneity. *Journal of Clinical Epidemiology*, *76*, 47-56.

Ritz, B. R., Chatterjee, N., Garcia-Closas, M., et al. (2017). Lessons learned from past gene-environment interaction successes. *American Journal of Epidemiology*, *186*(7), 778-786.

Rivadeneira, F., & Uitterlinden, A. G. (2021). Genetics of osteoporosis. In *Marcus and Feldman's Osteoporosis* (pp. 405-451). Academic Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology, 66*(5), 688.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322-331.

Samuel, J. P., Wootton, S. H., Holder, T., & Molony, D. (2022). A scoping review of randomized trials assessing the impact of n-of-1 trials on clinical outcomes. *Plos One*, *17*(6), e0269387.

Sarmanova, A., Morris, T., & Lawson, D. J. (2020). Population stratification in GWAS meta-analysis should be standardized to the best available reference datasets. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.09.03.281568v1.full (Accessed July 9, 2023).

Schwartz, D., & Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Chronic Diseases, 20*(8), 637-648.

Scuffham, P. A., Nikles, J., Mitchell, G. K., et al. (2010). Using N-of-1 trials to improve patient management and save costs. *Journal of General Internal Medicine*, *25*(9), 906-913.

Selker, H. P., Cohen, T., D'Agostino, R. B., Dere, W. H., Ghaemi, S. N., Honig, P. K., ... & Eichler, H. G. (2022). A Useful and Sustainable Role for N-of-1 Trials in the Healthcare Ecosystem. *Clinical Pharmacology & Therapeutics*, *112*(2), 224-232.

Senn, S. (2001). Cross-over trials in drug development: theory and practice. *Journal of Statistica Planning and Inference*, *96*(1), 29-40.

Serpico, D. & Borghini, A. (2021). From obesity to energy metabolism: Ontological perspectives on the metrics of human bodies. *Topoi*, 40(3), 577-586.

Smith, J., Yelland, M., & Del Mar, C. (2015). Single patient open trials (SPOTs). *The Essential Guide to N-of-1 trials in Health*, 195-209.

Snyderman, R. (2012). Personalized health care: from theory to practice. *Biotechnology Journal*, *7*(8), 973-979.

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 465-472.

Stegenga, J. (2018). *Medical Nihilism*. Oxford University Press.

Tabb, K., & Lemoine, M. (2021). The prospects of precision psychiatry. *Theoretical Medicine and Bioethics*, *42*(5), 193-210.

Thorpe, K. E., Oxman, A. D., Treweek, S., & Furberg, C. D. (2010). Pragmatic trials are randomized and may use a placebo. *Journal of Clinical Epidemiology*, *63*(6), 694-695.

Tresker, S. (2022). Treatment effectiveness, generalizability, and the explanatory/pragmatic-trial distinction. *Synthese*, *200*(4), 316.

Tudor, K., Brooks, J., Howick, J., Fox, R., & Aveyard, P. (2022). Unblinded and blinded N-of-1 trials versus usual care: a randomized controlled trial to increase statin uptake in primary care. *Circulation: Cardiovascular Quality and Outcomes*, *15*(6), e007793.

Turkheimer, E. (2011). Still missing. *Research in Human Development*, *8*(3-4), 227-241.

Turkheimer, E., Haley, A., Waldron, M., et al. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science*, 14(6), 623-628.

Turner, S. W. (2009). Genetic predictors of response to therapy in childhood asthma. *Molecular Diagnosis & Therapy*, *13*(2), 127-135.

Uffelmann, E., Huang, Q. Q., Munung, N. S., et al. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 1-21.

Vandenbroucke, J. P., Broadbent, A., & Pearce, N. (2016). Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology, 45*(6), 1776-1786.

Vogt, H. (2022). The precision paradox: How personalized medicine increases uncertainty. In Barilan, Y.M., Brusa, M. & Ciechanover, A. (eds.) *Can Precision Medicine be Personal; Can Personalized Medicine be Precise?* Oxford University Press (pp. 61-74).

Vohra, S., Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Tate, R., Nikles, J., Zucker, D. R., Kravitz, R., Guyatt, G., Altman, D. G., & Moher, D. (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. BMJ, 350, h1738.

Wang, A. L., & Tantisira, K. G. (2016). Personalized management of asthma exacerbations: lessons from genetic studies. *Expert Review of Precision Medicine and Drug Development*, *1*(6), 487-495.

Zaidi, A. A., & Mathieson, I. (2020). Demographic history mediates the effect of stratification on polygenic scores. *Elife*, *9*, e61548.