Chapter

# Efficacy Evaluation in the Era of Precision Medicine: The Scope for AI

*Dominic Magirr*

## Abstract

Patient stratification and the use of real-world evidence in regulatory decision-making are two key areas where algorithms are having an impact on drug development. The two are linked: increased patient stratification makes it harder to recruit patients into randomized-controlled trials, increasing the pressure on drug developers to find alternative sources of evidence for showing efficacy. In addition to real-world evidence, we are also seeing the emergence of more efficient 'master protocol trials', where multiple targeted agents can be evaluated simultaneously. In this chapter, I will review these developments and investigate the limitations for AI in terms of demonstrating the efficacy of novel targeted agents.

**Keywords:** drug development, precision medicine, statistics

## 1. Introduction

The use of algorithms to find patterns and make predictions from multiple data sources—here referred to as artificial intelligence (AI)—is having an increasingly large impact on clinical drug development.

Algorithms can be applied to combined clinical and genetic data sets to stratify patient populations into subgroups, based on shared characteristics or similar prognostic profiles [1–2]. This would appear to make sense, since the majority of new drugs approved by the US FDA in recent years have been targeted towards specific genetic aberrations [3–4]. If we increase our search, we will find more genetic aberrations, more drug targets, and more potentially efficacious drugs. However, this approach also presents severe challenges in the clinical stages of drug development, as the size, complexity and duration of studies increases.

One way to react to increased cost and duration is to improve the operational efficiency of clinical trials. The last decade has seen the emergence of 'master protocol trials', which allow several substudies to be conducted simultaneously, reducing the rate of screen failures [5]. In addition, there is increasing enthusiasm for augmenting (possibly even replacing) randomized-controlled trials (RCTs) with external and real-world data, where it is claimed that further use of algorithms can protect us from the biases that this approach would otherwise impose [1].

The purpose of this article is three-fold. Firstly, to explain how precision medicine presents challenges to traditional drug development, quantifying the effect of disease stratification on trial recruitment. Secondly, to describe how master

protocol studies have emerged in response to these challenges. Finally, to explore whether it is possible for single-arm studies with 'synthetic control arms' to provide the same standard of evidence as a randomized controlled trial, thus reducing drug development timelines.

## 2. Disease stratification

Consider a patient population that can be stratified according to the value of a diagnostic test. The 'target' population consists of patients who test positive. The 'non-target' population consists of patients who do not test positive. Suppose that a new treatment is expected to be more effective in the target population than in the non-target population. Let $\theta^+$ and let $\theta^-$ denote the treatment effect sizes in target and non-target populations, and $\gamma$ denote the prevalence of the target group. Three things that we would like to demonstrate are:

1. Treatment benefit in the full population, $\gamma\theta^+ + (1-\gamma)\theta^- > 0$.

2. Treatment benefit in the target population, $\theta^+ > 0$.

3. Greater benefit in the target population than in the non-target population, $\theta^+ > \theta^-$.

Which of these is easiest to demonstrate, and which most difficult? To answer this, we compare the standardised statistics, $Z$, that we would use to test the corresponding null hypotheses. For most commonly-used clinical-trial endpoints, the test statistic ends up looking like

$$Z \sim N\left(\theta\sqrt{I}, 1\right), \tag{1}$$

where $\theta$ is the treatment effect size and $I$ is the statistical *information*, which is typically proportional to the sample size [6, 7]. The *power* of a test is the probability that $Z > k$, for a threshold $k$, where $k$ is chosen to ensure a given false-positive rate. The larger the expected value of $Z$, the higher the power. Therefore two trials ('A' and 'B') will have the same power if $\theta_A\sqrt{I_A} = \theta_B\sqrt{I_B}$, or, assuming that information is proportional to sample size, if

$$\theta_A\sqrt{N_A} = \theta_B\sqrt{N_B}. \tag{2}$$

We can use (2) to assess the relative difficulty of our three goals, firstly for the full population versus the interaction (1. versus 3.), and then for the full population versus the target population (1. versus 2.).

### 2.1 Full population versus interaction

It is shown in the appendix that a test of the interaction null hypothesis, $\theta^+ = \theta^-$, with total sample size $N_{\text{int}}$, will have the same power as the test for the full population null hypothesis, $\gamma\theta^+ + (1-\gamma)\theta^- = 0$, with sample size $N$, provided that

$$(\theta^+ - \theta^-)\sqrt{\gamma(1-\gamma)N_{\text{int}}} = \{\gamma\theta^+ + (1-\gamma)\theta^-\}\sqrt{N}. \tag{3}$$

For example, when $\theta^-/\theta^+ = 0.5$, for a prevalence of 50%, the ratio of sample sizes is $N^{\text{int}}/N = 9$. For a prevalence of 5%, $N^{\text{int}}/N \approx 23$. This shows how difficult it is to provide compelling evidence for treatment-biomarker interactions, and why drug development is still focussed on demonstrating average treatment effects. It is also explains why post-hoc data-driven subgroup identification following a clinical trial is often a bad idea. See Gelman [8] for further discussion.

## 2.2 Full population versus target population

A test for the full population null hypothesis with sample size $N$ will have the same power as a test for the target population null hypothesis, $\theta^+ = 0$, with sample size $N_T$, provided that

$$\{\gamma\theta^+ + (1-\gamma)\theta^-\}\sqrt{N} = \theta^+\sqrt{N_T}, \tag{4}$$

or, equivalently, if

$$\frac{N_T}{N} = \left\{\gamma + (1-\gamma)\frac{\theta^-}{\theta^+}\right\}^2. \tag{5}$$

In (5), we have expressed the relative sample size, $N_T/N$, as a function of the relative efficacy, $\theta^-/\theta^+$ [9]. This relationship is drawn in solid lines in **Figure 1** for two potential prevalences (50% and 5%) when $\theta^-/\theta^+$ is between 0.5 and 1. For a prevalence of 50%, the targeted strategy requires up to 40% fewer patients than the non-targeted strategy. For a prevalence of 5%, a 70% reduction is possible. Note, however, that this is the relative number of patients *enrolled*. What about the

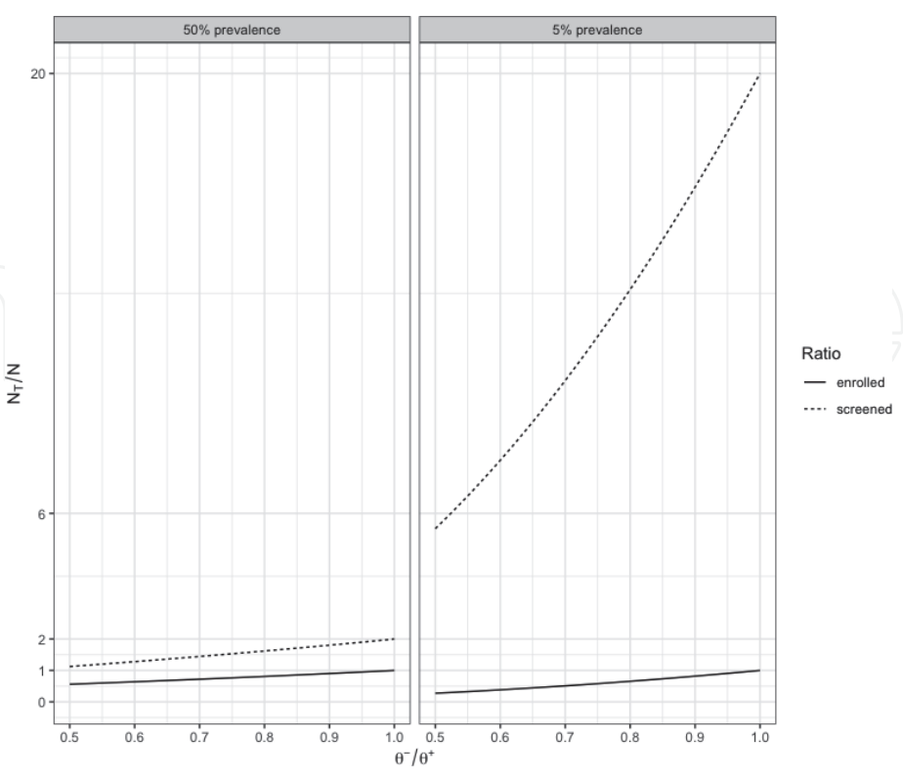



**Figure 1.**
*Relative sample size when testing efficacy in the target population compared to the full population* ($N_T/N$)*, shown in solid lines. The dashed lines show the relative number of patients screened* ($(N_T/\gamma)/N$)*.*

number of patients *screened*? In the full population the minimum number screened is $N$, whereas in the targeted population it is $N_t$ divided by $\gamma$. The ratio, $(N_t/\gamma)/N$, is drawn in dashed lines in **Figure 1**. For the 50% prevalence case, there is a maximum 2-fold increase in the number screened for the targeted compared to the non-targeted trial. But for the 5% prevalence case, there is somewhere between a six-fold and a twenty-fold increase.

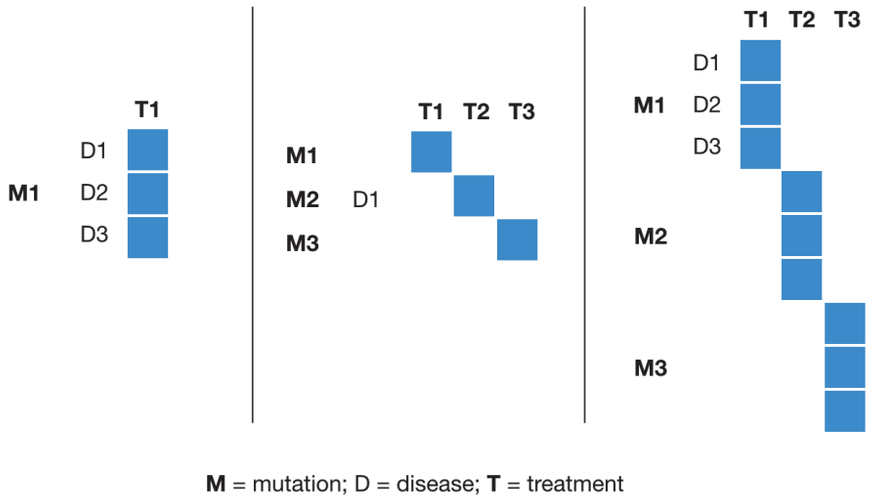
### 2.3 Situations where $\theta^- \ll \theta^+$

The conclusion from **Figure 1** is that population stratification is only likely to be useful if there exists a potential treatment where the treatment effect is considerably higher (e.g. at least two-fold) in the target subgroup than in the rest of the population. Marginal increases in efficacy are not enough in practice. The targeted approach would require a prohibitively large number of patients to be screened, compared to a trial in the full population which would have the same statistical power. Marginal increases are also difficult to establish empirically, as shown in Section 2.1. It follows that successful implementation of precision-medicine drug development is restricted to situations where there is strong biological and preclinical evidence for expecting $\theta^- \ll \theta^+$. Such cases certainly do exist, and the targeted trial is the only sensible approach here. Nevertheless, one still needs to screen a very high number of patients. This is expensive for the sponsor. It is also disheartening for patients who do not meet the eligibility criteria.

## 3. Master protocol trials

The high screen failure rate of precision-medicine trials can be mitigated to some extent by merging multiple sub-studies into a single 'master protocol'. The last decade has seen the emergence of the labels 'basket' and 'umbrella' to describe these complex studies. As a rule of thumb, a basket tends to refer to studies involving the same drug in multiple diseases, whereas umbrella is used when multiple experimental treatments are studied in the same disease. However, as reported by Janiaud and colleagues [10], these terms have not been applied consistently. Their systematic review of master protocol trials in oncology found 30 'basket' trials and 27 'umbrella' trials in a time period of 2006–2018, but with most studies starting after 2015. They explain that some basket trials are mistakenly labeled as umbrella trials, and vice-versa, but there are also trials that contain elements of both and thus become difficult to describe using current language.

Stallard and colleagues [11] propose a refined classification which replaces ambiguous labels with a more precise visual description, as shown in **Figure 2**. In each of the six designs, a small square is representative of a cohort of patients. On the left hand side are the basket-type designs, where there is only one new treatment ($T$) targeting a particular mutation ($M$), but this mutation occurs across diseases ($D_1, D_2, \ldots$). In the middle are the umbrella-type designs, where there are multiple treatments ($T_1, T_2, \ldots$) targeting particular mutations ($M_1, M_2, \ldots$), all within the same overall disease ($D$). The designs on the right hand side combine the features of the basket-type and umbrella-type designs. They allow for multiple disease types within each of the separate treatment-mutation combinations. Note, however, that it is always the mutation that is driving the choice of treatment, rather than the disease type. In all of the designs, for each *T-M-D* sub-study, it is possible to use a single-arm design (**Figure 2a**), or compare with a concurrent control arm (**Figure 2b**).




**M** = mutation; D = disease; **T** = treatment

(a)


**M** = mutation; D = disease; **T** = treatment

(b)

**Figure 2.**
*A classification of master protocol designs by Stallard and colleagues [11]. Each square represents a cohort of patients. (a) Single-arm cohorts. (b) Concurrent control arms.*

### 3.1 Example 1: Vemurafenib in cancers (not melanoma) with BRAF V600 mutations

Hyman and colleagues [12] report the results of a basket-type study with the same structure as the left-hand-side of **Figure 2a**. The treatment (T) was Vemurafenib, the mutation (M) was BRAF V600. There were several cohorts corresponding to different disease types (D):

- Colorectal cancer (CRC)
- Bile duct cancer
- Anaplastic thyroid cancer (ATC)
- Non-small cell lung cancer (NSCLC)
- Erdheim-Chester disease/Langerhans cell histiocytosis (ECD/LCH)

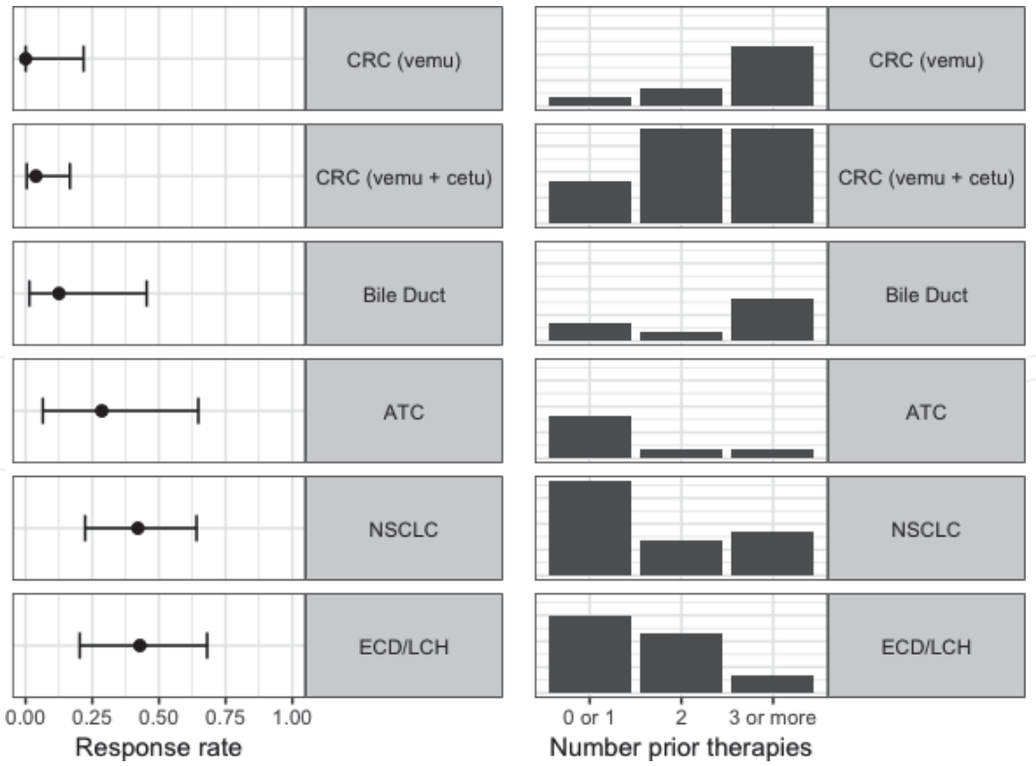



**Figure 3.**
*Results from a basket-type study of Vemuarafenib [12–13].*

A Simon's two-stage design [13] was used for each cohort independently to allow for early futility stopping. Consequently, the cohort sizes ranged from 5 to 27. Hobbs and colleagues [14], did a re-analysis of the data, and their findings are reproduced in **Figure 3**. Looking at the response rate across cohorts, it appears that there is more activity in NSCLC and ECD/LCH than in CRC. However, one can also see a clear inverse relationship between response rate and number of prior therapies, which muddies the water. This example highlights how difficult it can be to interpret uncontrolled studies.

### 3.2 Example 2: FOCUS4

An example of a master protocol trial that does include concurrent control arms is FOCUS4 [15], currently being run by the Medical Research Council Clinical Trials Unit in London. It has an umbrella-type design like **Figure 2b**. The disease setting (D) is advanced colorectal cancer. Mutations (M) include:

- BRAF mutations
- MSI deficient
- PIK3CA mutations
- Wild type

A centralised molecular analysis is performed on each patients tumor. Based on the results, patients are offered entry into an appropriate substudy, where they are randomized to receive either an experimental treatment (T) targeted to their mutation, or a control treatment.

The substudies will be analysed independently, as if they were separate trials. The big advantage over independent studies is the increased efficiency from the centralised molecular analysis, ensuring fewer screen failures. Complications may arise when patients are eligible for more than one substudy, and this has to be planned for in the protocol. Note also the inclusion of the Wild-type cohort in FOCUS4. This maximizes the proportion of patients who undergo screening who are given an option to go on a trial.

## 4. External control arms

Precision medicine is increasing the pressure on drug developers to find innovative ways to demonstrate efficacy without requiring ever larger and lengthier clinical trials. We have seen how operational efficiencies can be found in master protocol trials. A related development is the use of 'big data'—the bringing together of historical RCT data, electronic health records, advanced statistical modeling, and machine-learning—to produce a historical benchmark, or even a so-called 'synthetic control arm', that might allow a single-arm study to take the place of an RCT as a basis for seeking drug approval.

For this approach to be successful, the key use-case in oncology is a comparison of overall survival (OS). It is typical for inference to focus on the (log) hazard ratio,

$$\theta := \log \frac{\lambda_E(t)}{\lambda_C(t)}, \tag{6}$$

where it is assumed that the hazard of death on the experimental arm, $\lambda_E(t)$, is proportional to the hazard of death on the control arm, $\lambda_C(t)$, for all timepoints $t$. Another way to describe $\lambda(t)$ is that it is your risk of dying on day $t$ given that you were alive at midnight. More stringent than proportional hazards is an assumption of constant hazards, $\lambda_j(t) = \lambda_j$ for all $t$ $(j = E, C)$. Although an over-simplification, this model is often not a bad approximation to reality, and we will use it to compare operating characteristics for a two-arm RCT versus a single-arm trial with an external control arm.

### 4.1 Distribution of treatment effect estimators

The constant-hazards assumption allows us to express the log hazard ratio as the difference between the log-transformed median survival times,

$$\log \frac{\lambda_E}{\lambda_C} = \log m_C - \log m_E. \tag{7}$$

For a two-arm study with equal randomisation and $D$ events, the estimate of the log hazard ratio has the following (approximate) distribution:

$$\hat{\theta} = \log \hat{m}_C - \log \hat{m}_E \sim N(\theta, 4/D). \tag{8}$$

If we were to run a single-arm study instead, but keep the overall sample size the same, i.e. put all patients who would have received the control treatment onto the experimental arm, we could use the test statistic

$$\hat{\theta}^* = \log m_C^* - \log \hat{m}_E \sim N\left(\theta + \log m_C^* - \log m_C, 1/D\right) \tag{9}$$

where $m_C^*$ is our best pre-trial estimate for the median OS on the control arm.

### 4.2 Bias-variance trade-off

We can compare the precision of the two estimates in terms of their mean-squared-errors,

$$\begin{aligned} \mathrm{mse}(\hat{\theta}) &= \mathrm{var}(\hat{\theta}) + \mathrm{bias}(\hat{\theta})^2 \\ &= 4/D + 0 \end{aligned} \tag{10}$$

and

$$\begin{aligned} \mathrm{mse}\left(\hat{\theta}^*\right) &= \mathrm{var}\left(\hat{\theta}^*\right) + \mathrm{bias}\left(\hat{\theta}^*\right)^2 \\ &= 1/D + \left|\log m_C^* - \log m_C\right|^2. \end{aligned} \tag{11}$$

For low values of $D$, variance will be a bigger problem than bias. In this case, $\mathrm{mse}\left(\hat{\theta}^*\right) < \mathrm{mse}(\hat{\theta})$. However, as soon as

$$D > \frac{3}{\left|\log m_C^* - \log m_C\right|^2} \tag{12}$$

the bias will dominate, and the estimate from the two-arm trial will be more precise.

### 4.3 NSCLC example

What is a typical value for $|\log m_C^* - \log m_C|$? This depends on the context. The FDA have published data from 14 large randomized control trials [16] in advanced non-small-cell-lung cancer (NSCLC) conducted between 2003 and 2015. The median survival on the control arm across the studies is shown in **Figure 4**. Three of the studies were targeted towards patients with a particular biomarker. It is immediately obvious that these three data points are different from the rest, and this highlights the dangerous territory we are in. Nevertheless, if we focus on the 11 studies that did not use a targeted approach, the median overall survival ranged from 7 to 13 months. Taking an average value, a sensible choice for $\log m_C^*$ is $\log(9.5)$. We could also think about the 'true' $\log m_C$ for the current study belonging to the same distribution as the 11 other studies, which we might approximate with a normal distribution

$$\log m_C \sim N\left(\log m_C^* = \log(9.5), \sigma_{m_C}^2 = 0.03\right) \tag{13}$$

The expected value of $|\log m_C^* - \log m_C|$ according to (13) is $\sqrt{2/\pi}\sigma_{m_C} \approx 0.14$. Plugging this into (12), the two-arm trial would be more precise than the single-arm trial when $D > 153$.

### 4.4 Reducing the sample size

What if instead of moving patients from the control arm to the experimental arm and keeping total sample size the same, we run a single-arm study with half the number of patients, i.e. we keep the same sample size on the experimental arm and replace the control arm with an historical benchmark? In this case, the

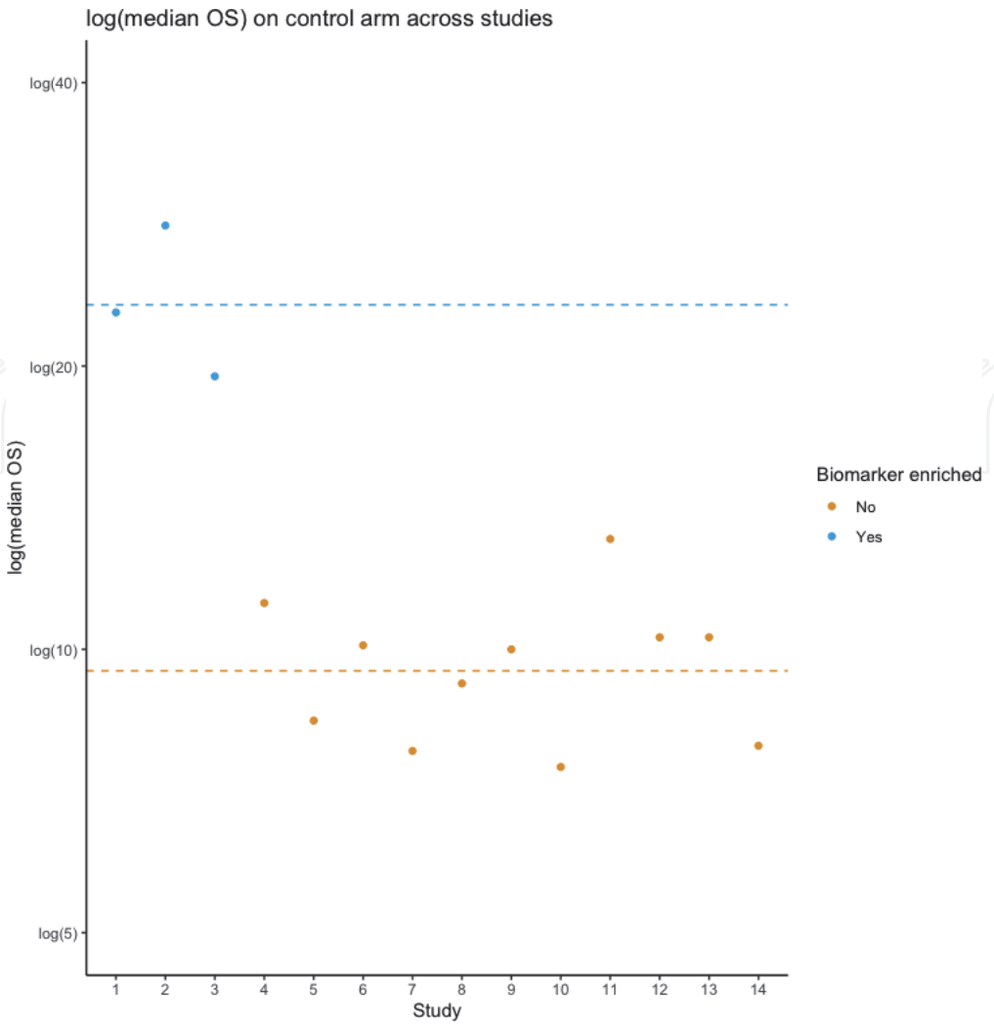



**Figure 4.**
*Between-trial variability in median overall survival time from 14 phase 3 studies in advanced non-small-cell lung cancer trials submitted to the FDA [16].*

mean-squared-error of the estimate from the two-arm trial will be lower than the single-arm equivalent as soon as

$$D > \frac{2}{\left|\log m_C^* - \log m_C\right|^2}, \tag{14}$$

where $D$ is the number of events in the two-arm trial. For our lung cancer example, this would mean as soon as $D > 102$.

### 4.5 More advanced methods

In the previous example we were using the average value from 11 previous studies as a rather crude estimate of $\log m_C^*$. Is it possible to improve the precision using 'big data'—bringing together historical RCT data, electronic health records, advanced statistical modeling, and machine-learning?

We can look to a recent study by Carrigan and colleagues [17]. The group had access to individual patient data from 9 RCTs in advanced NSCLC conducted between 2011 and 2018, as well as electronic health records (EHR) from almost 50,000 patients. They used advanced regression and stratification techniques to estimate treatment effect sizes, and their results are reproduced on the left hand side of **Figure 5**. There is a high correlation (0.86) between the hazard ratio from

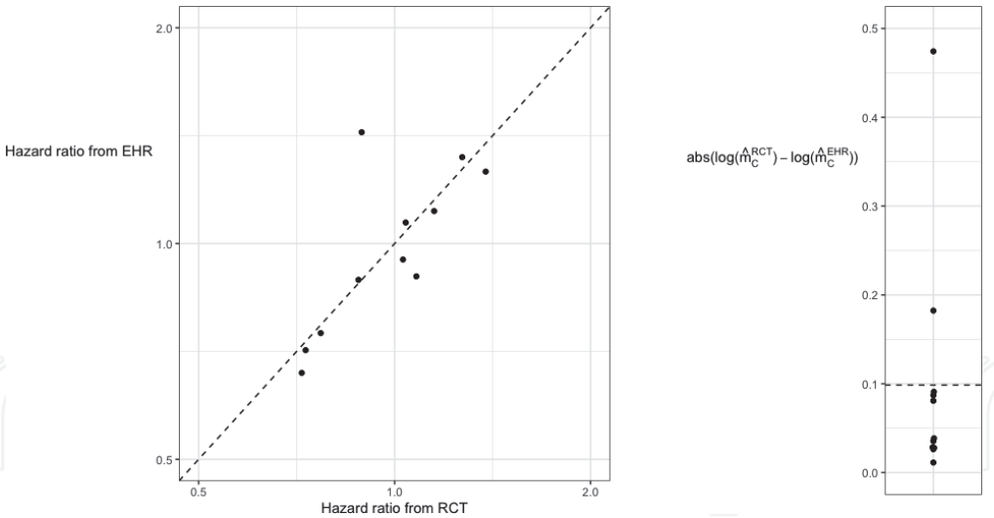



**Figure 5.**
*Correlation between RCT-derived and EHR-derived hazard ratios from nine studies in advanced non-small-cell lung cancer [17]. On the right-hand-side, the results have been converted into an approximation of the bias when estimating the median survival time on the control arm using EHR data.*

the RCTs and the hazard ratio that would have been observed had the control arm been replaced with electronic health record data. On the right hand side, the data points have been transformed into an estimate of the bias $|\log m_C^* - \log m_C|$, assuming constant hazards. The mean value is 0.1 and according to (12) this means that a two-arm trial would be more precise than a single-arm trial of the same total sample size whenever $D > 300$. Similarly, using (14), a two-arm trial will be more precise than a single-arm trial with half the sample size when $D > 200$.

To put these findings in some context, for a study with one-sided type-1 error of $\alpha = 0.025$, 300 events would give 90% power when HR $= 0.69$. Likewise, 100 events would give 90% power when HR $= 0.52$.

## 5. Conclusions

Advances in pattern-recognition and prediction algorithms have the potential to improve health outcomes, as well as making the drug development process more efficient. Nevertheless, it is important to have a strong grasp of some limiting factors, to avoiding spending time on futile endeavors.

The stratification of patient populations into ever finer subgroups is only likely to prove useful when there exist potential treatments with very large differential treatment effects. Marginal is not enough—it needs to be 100% more efficacious in the target subgroup than in the non-target subgroup. Otherwise, a clinical trial in the full population would have the same statistical power with far fewer patients screened. This means that we need strong biological rationale and robust pre-clinical evidence. In addition, it is essential that the diagnostic test has high sensitivity and specificity. Otherwise, a large treatment effect in the *true* biomarker-positive population would become diluted in the *observed* biomarker-positive population.

In cases where there is a strong rationale for a targeted approach, recruitment will be challenging. Master protocol trials can be an excellent option. They are an efficient way to test novel agents, and they increase the chance that a patient entering screening will be able to join a clinical trial.

Improvements in the quality of electronic health records, as well as better algorithms to interrogate this data, are a positive development that can enhance our

understanding of health outcomes, and help enormously with clinical trial design and interpretation. Nevertheless, we should not forget the fundamental benefits of concurrent control [18], and should remain realistic about the ability of synthetic control arms to replace the real thing. We have seen that under favorable circumstances (highly prevalent disease, patient-level data from numerous high-quality large RCTs, tens of thousands of electronic health records, well-defined and accurately-measured primary endpoint, careful analysis), a single-arm study can provide similar precision to a two-arm randomized comparison with sample size in the low hundreds [17]. It is plausible, therefore, that for a new drug in this space with a very large treatment effect, a single-arm study may provide convincing evidence of efficacy. But one should expect this to be the exception, not the norm.

## Conflict of interest

Dominic Magirr is an employee of Novartis Pharma AG.

## Abbreviations

| | |
|---|---|
| AI | artificial intelligence |
| FDA | Food & Drug Administration |
| RCT | randomized controlled trial |
| NSCLC | non-small-cell lung cancer |
| OS | overall survival |
| EHR | electronic health record |

## Appendix

Based on the test statistics (1) for the target and non-target populations,

$$Z^+ \sim N\left(\theta^+ \sqrt{\gamma I_{\text{int}}}, 1\right)$$

and

$$Z^- \sim N\left(\theta^- \sqrt{(1-\gamma) I_{\text{int}}}, 1\right),$$

we can define an interaction test statistic

$$Z^{\text{int}} := \sqrt{1-\gamma} Z^+ - \sqrt{\gamma} Z^- \sim N\left((\theta^+ - \theta^-)\sqrt{\gamma(1-\gamma) I_{\text{int}}}, 1\right).$$

By (2), this test will have the same power as the full population test with sample size $N$ if

$$(\theta^+ - \theta^-)\sqrt{\gamma(1-\gamma) N_{\text{int}}} = \{\gamma\theta^+ + (1-\gamma)\theta^-\}\sqrt{N}.$$

## Author details

Dominic Magirr
Novartis Pharma AG, Basel, Switzerland

*Address all correspondence to: dominic.magirr@novartis.com

IntechOpen

## References

[1] Gao F, Wang W, Tan M, Zhu L, Zhang Y, Fessler E, et al. DeepCC: A novel deep learning-based framework for cancer molecular subtype classification. Oncogene. 2019;**8**(9):1-2. DOI: 10.1038/s41389-019-0157-8

[2] Bailey P, Chang DK, Nones K, Johns AL, Patch AM, Gingras MC, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016;**531**(7592):47. DOI: 10.1038/nature16965

[3] Shah P, Kendall F, Khozin S, Goosen R, Hu J, Laramie J, et al. Artificial intelligence and machine learning in clinical development: A translational perspective. npj Digital Medicine. 2019;**2**(1):69. DOI: 10.1038/s41746-019-0148-3

[4] FDA. Novel Drug Approvals for 2018. Available from: https://www.fda.gov/drugs/new-drugs-fda-cders-new-molecular-entities-and-new-therapeutic-biological-products/novel-drug-approvals-2018. 2018. [Accessed: 20 August 2019]

[5] Woodcock J, LaVange LM. Master protocols to study multiple therapies, multiple diseases, or both. New England Journal of Medicine. 2017;**377**(1):62-70. DOI: 10.1056/NEJMra1510062

[6] Whitehead J. The Design and Analysis of Sequential Clinical Trials. Chichester: John Wiley & Sons; 1997

[7] Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. Boca Raton: Chapman and Hall/CRC; 1999. DOI: 10.1201/9780367805326

[8] Gelman A. You need 16 times the sample size to estimate an interaction than to estimate a main effect. In: Statistical Modeling, Causal Inference, and Social Science. 2018. Available from: https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/ [Accessed: 14 August 2019]

[9] Simon R. The use of genomics in clinical trial design. Clinical Cancer Research. 2008;**14**(19):5984-5993. DOI: 10.1158/1078-0432.CCR-07-4531

[10] Janiaud P, Serghiou S, Ioannidis JP. New clinical trial designs in the era of precision medicine: An overview of definitions, strengths, weaknesses and current use in oncology. Cancer Treatment Reviews. 2019;**73**:20-30. DOI: 10.1016/j.ctrv.2018.12.003

[11] Stallard N, Todd S, Parashar D, Kimani PK, Renfro LA. On the need to adjust for multiplicity in confirmatory clinical trials with master protocols. Annals of Oncology. 2019;**30**(4):506. DOI: 10.1093/annonc/mdz038

[12] Hyman DM, Puzanov I, Subbiah V, Faris JE, Chau I, Blay JY, et al. Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. New England Journal of Medicine. 2015;**373**(8):726-736. DOI: 10.1056/NEJMoa1502309

[13] Simon R. Optimal two-stage designs for phase II clinical trials. Controlled Clinical Trials. 1989;**10**(1):1-0. DOI: 10.1016/0197-2456(89)90015-9

[14] Hobbs BP, Kane MJ, Hong DS, Landin R. Statistical challenges posed by uncontrolled master protocols: Sensitivity analysis of the vemurafenib study. Annals of Oncology. 2018;**29**(12):2296-2301. DOI: 10.1093/annonc/mdy457

[15] Kaplan R, Maughan T, Crook A, Fisher D, Wilson R, Brown L, et al. Evaluating many treatments and biomarkers in oncology: A new design. Journal of Clinical Oncology: Official

Journal of the American Society of Clinical Oncology. 2013;**31**(36):4562

[16] Blumenthal GM, Karuri SW, Zhang H, Zhang L, Khozin S, Kazandjian D, et al. Overall response rate, progression-free survival, and overall survival with targeted and standard therapies in advanced nonsmall-cell lung cancer: US Food and Drug Administration trial-level and patient-level analyses. Journal of Clinical Oncology. 2015;**33**(9):1008. DOI: 10.1200/JCO.2014.59.0489

[17] Carrigan G, Whipple S, Capra WB, Taylor MD, Brown JS, Lu M, et al. Using electronic health records to derive control arms for early phase SingleArm lung Cancer trials: Proof of concept in randomized controlled trials. Clinical Pharmacology & Therapeutics. 2020; **107**(2):369-377. DOI: 10.1002/cpt.1586

[18] Senn S. Control in clinical trials. In: Data and Context in Statistics Education: Towards an Evidence-Based Society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8 2010 July). 2010. Available from: https://pdfs.semanticscholar.org/d36e/873d830932dd17c9ddf14e34dc542d14b63c.pdf [Accessed: 21 August 2019]