Journal of
Medical Statistics and Informatics
ISSN 2053-7662

HOAJ
Herbert Open Access Journals

Research Open Access

Interrelationships among enrichment with diagnostics, probability of success and demonstration of effectiveness in clinical trials

Deepak B. Khatry
Correspondence: khatryd@medimmune.com
MedImmune, Biostatistics/Translational Sciences, One MedImmune Way, Gaithersburg, USA.

Abstract

Background: Prevalence of disease phenotype in clinical practice is often not given adequate importance during formulation, validation, and implementation of diagnostic tests in clinical research and development. After promising biomarkers have been identified as potential screening diagnostics, an important strategic question for optimal decision-making in clinical development of a therapeutic is when to choose an enrichment study design over the traditional all-comer randomized control trial design.

Methods: A hypothetical example of a cholesterol lowering treatment is used to illustrate influences of key statistical criteria and clinical considerations for choosing study designs. Computer simulations demonstrate how results of such analyses can aid in deciding whether or not to choose enrichment study designs.

Results: This study shows how understanding of disease prevalence in practice, predictive values of diagnostic test, and pre-specified establishment of a clinically meaningful minimum effectiveness all need to be integrated to insure clinical trial success and appropriate benefit to targeted patient subgroups. The most important statistical and clinical considerations were the anticipated effect size, phenotype prevalence, predictive values of diagnostics test, study power, and desired clinically meaningful difference.

Conclusions: This study illustrates how successful clinical studies can be designed with careful planning and utilization of computer simulations to increase not only the probability of trial success but also to demonstrate to payers convincing evidence of clinical effectiveness. A six-step checklist is recommended as an evidence-based guideline to assist in decision-making on whether or not to adopt a diagnostics enriched clinical study design.

Keywords: Clinical trial, biomarker, screening diagnostics, clinical effectiveness, personalized healthcare, stratified medicine, computer simulation

Introduction

Many clinical trials fail because a "sufficiently appropriate" group of patients are not studied. Often, inclusion/exclusion criteria in conventional randomized clinical trials (*RCT*) do not optimally match patients to an investigational therapy's specific mechanism of action (*MoA*). Thus, assessment of "efficacy," a measure of a therapy's "average" benefit to patients compared between standard-of-care (*SoC*) and investigational treatment arms, poses two types of risks.

The first risk, designated as "consumer risk," is to individual patients. Because of heterogeneity in the clinical trial population, overall large efficacy signals may be produced by only a small subset of patients matching the therapy's *MoA*. Thus, although a large proportion of study subjects may not be responding to an investigational therapy, the clinical trial, nevertheless, may show statistical significance and result in regulatory approval for marketing. In this latter scenario, there are potential ethical issues in the study itself because many participants with very low probability of benefit from the investigational therapy will have been unnecessarily exposed to potential harm. As the inclusion/exclusion criteria from pivotal trials determine product labels, many patients may later be prescribed the newly-approved therapy causing a large distortion in benefit/harm or benefit/cost ratio in the real world.

The second risk, designated as "producer risk," is to the sponsor. Because of heterogeneity in study sample and inferior concordance with therapy's *MoA*, the ratio of efficacy signal to background noise may be greatly diminished, thereby increasing risk of trial failure. Phase II success rates are lower than at any other phase of development, as evidenced by decline in success rates for new development projects from 28% in 2006-2007 to 18% in 2008-2009 [1]. An estimate of likelihood of a drug successfully progressing through Phase III to launch is 50% [2]. Such high attrition rates in late-stage drug development result in large financial costs to industry from both lost revenue and missed opportunity cost of not pursuing alternate drug candidates or targets.

Physicians, patients, and increasingly payers, who control reimbursement decisions, prefer clinical "effectiveness" over "efficacy" (see reference [3], for a regulatory perspective). Clinical effectiveness measures how well a treatment works in patients in real-world conditions. Ideally, quantification of effectiveness should include proportion of responders and not just average response of a group. Demonstration of effectiveness will generally demonstrate efficacy, but the reverse may not hold true. A study population must exhibit sufficient homogeneity of response to a treatment to demonstrate effectiveness. Thus, a screening diagnostics (*DX*) with an optimum threshold for accurate patient classification may become necessary for assuring higher homogeneity. This paper focuses on key interrelated statistical and clinical considerations for aiding decisions on when to adopt a *DX*-enriched study design. Important among these are the anticipated effect size, prevalence of phenotype, diagnostic test accuracy, study power, and clinically meaningful desired difference. Simulations of a hypothetical example of a cholesterol-lowering drug are used to demonstrate how well-planned computer simulations can aid in deciding when to adopt *DX*-enriched study designs.

Methods

Diagnostic accuracy and probability

Diagnostic tests are undertaken to determine the presence or absence of a phenotype, and this is carried out by making a decision that the condition is or is not present based upon test results [4,5]. In medical decision-making dependent on diagnostic test results, conditional probabilities are conditioned on the outcome rather than the "unknown" truth. Such probabilities are the "inverse probabilities," also known as "Bayesian" probabilities. It is important to understand the direction of a conditional probability as to whether the direction is from truth to outcome or in the reverse direction [4]. Confusion on the directionality of the conditional probability can misinform an understanding of probabilities that affect clinical decision-making. For the purpose of planning and designing a *DX*-enrichment study, positive predictive value (*PPV*) is the most important diagnostic measure of accuracy because it directly impacts probability of trial success (based on statistical significance) as well as the likelihood of demonstrating a pre-specified minimal clinically meaningful difference. The predictive values of a *DX* can be estimated from test sensitivity (*SN*), specificity (*SP*), and pre-test or prior probability (*PP*) using the following equations [6]:

$$PPV = SN \times PP/SN \times PP + (1-SP) \times (1-PP) \text{ ....(1)}$$

$$NPV = SP \times (1-PP)/SP \times (1-PP) + (1-SN) \times PP \text{ ....(2)}$$

Because the inverse probabilities are calculated from the truth-conditional probabilities and the *PP*, it is important to conduct sensitivity analyses using a plausible range of pre-test probabilities when calculating *PPV*. If a plausible range of pre-test probabilities are unknown, but the *SN* and *SP* of a *DX*-test is known with sufficient confidence, the number of test +ves and test -ves from a pilot study or retrospective data can be used with a flat *Beta* prior in a simulated Markov process with a continuous state space to estimate *PP* and Bayesian confidence intervals (see reference [7], for an example with R programming codes). The *PP* can also be approximated using assumed *SN* and *SP* of the diagnostic test and observed proportion of *DX*+ve (*PDXP*) in a pilot or retrospective study using the following equation:

$$PP = (PDXP+SP-1)/(SN+SP-1) \text{ ....(3)}$$

Based on the normal approximation to the binomial distribution, the traditional 95% confidence interval for *PDXP* can be calculated and substituted into Equation 3 to obtain 95% confidence interval for *PP*. However, as cautioned in [7], absurd estimates of prevalence can sometimes result when using Equation 3. The Bayesian method is robust against such absurd estimates and may be preferable, especially considering ready availability of open-access statistical software such as R. Representative values from inside the confidence interval can subsequently be used for sensitivity analyses in downstream simulations of *PPV,* and to examine the effects of *DX* accuracy on clinically meaningful difference.

Illustrative example

A hypothetical example of an investigational cholesterol-lowering drug that selectively benefits a targeted sub-group of patients at risk of heart disease is utilized. The example is drawn from an actual Dutch study of cholesterol-lowering therapy as described in [8]. In the Dutch study, familial hypercho lesterolemia (FH) was diagnosed through genetic cascade screening, and the study patients were treated with a cholesterol-lowering drug. After analysis of the study data, it was observed that mean low-density lipoprotein cholesterol (LDL-C) decreased to 124 (± 43) mg/dL, which was statistically significant. However, only 22% of study subjects achieved the LDL-C target level of ≤97 mg/dL recommended in Dutch guidelines. Although questions have been raised about the effectiveness of genetic testing for FH [9,10], it is assumed in this example that a novel predictive *DX* has been developed for selecting likely responders to the investigational lipid lowering treatment. The *DX* will be utilized to enrich the clinical trial population to demonstrate both clinical "effectiveness" and "efficacy" of the new investigational therapy. In this paper, mock simulated studies will be utilized with 10% above the Dutch recommended guideline of ≤97 mg/dL of LDL-C (i.e., ≤107 mg/dL) as the reduced post-treatment target for demonstrating clinical effectiveness. This target level is on the low side of the range for "near ideal" category (100-129 mg/dL) published by the Mayo Clinic [11]. Analyses will incorporate simulation and application of formal statistical tests. First, relationships will be examined between various measures of diagnostic accuracy (*SN*, *SP*, *PPV*, *NPV*, and overall accuracy) (**Table 1**). Subsequently, relationships among predictive values of *DX* tests, effect size, study power, and their contributions to go/no go decisions will be examined (**Figures 1** and **2**). The goal is to not only achieve statistical significance, but also attain

Table 1. Relationship of different measures of diagnostic accuracy with pre-diagnostic test probability (values are means and standard deviations obtained from simulations).

Accuracy (%)	*DX* **Measure**	**Prevalence, ie Pre-Test Probability (%)**				
	--	**30**	**40**	**50**	**60**	**70**
10	***SP***	7.1 (4.7)	8.3 (5.5)	10.0 (6.6)	12.5 (8.3)	16.7 (11.1)
	SN	16.7 (11.1)	12.5 (8.3)	10.0 (6.6)	8.3 (5.5)	7.1 (4.7)
	PPV	6.8 (4.1)	7.9 (4.7)	9.3 (5.5)	11.5 (6.6)	15.1 (8.2)
	NPV	15.1 (8.2)	11.5 (6.6)	9.3 (5.5)	7.9 (4.7)	6.8 (4.1)
20	***SP***	14.3 (8.9)	16.7 (10.3)	20.0 (12.4)	25.0 (15.5)	33.3 (20.7)
	SN	33.3 (20.7)	25.0 (15.5)	20.0 (12.4)	16.7 (10.3)	14.3 (8.9)
	PPV	13.2 (6.8)	15.2 (7.6)	18.0 (8.6)	22.2 (9.8)	29.4 (11.1)
	NPV	29.4 (11.1)	22.2 (9.8)	18.0 (8.6)	15.2 (7.6)	13.2 (6.8)
30	***SP***	21.4 (13.0)	25.0 (15.2)	30.0 (18.2)	37.5 (22.7)	50.0 (30.3)
	SN	50.0 (30.3)	37.5 (22.7)	30.0 (18.2)	25.0 (15.2)	21.4 (13.0)
	PPV	19.3 (8.7)	22.3 (9.5)	26.6 (10.3)	33.4 (10.8)	50.0 (0.0)
	NPV	50.0 (0.0)	33.4 (10.8)	26.6 (10.3)	22.3 (9.5)	19.3 (8.7)
40	***SP***	35.7 (13.0)	33.3 (20.0)	40.0 (24.0)	50.0 (30.0)	50.0 (30.3)
	SN	50.0 (30.3)	50.0 (30.0)	40.0 (24.0)	33.3 (20.0)	35.7 (13.0)
	PPV	22.3 (9.5)	29.6 (10.4)	35.9 (10.4)	50.0 (0.0)	66.6 (10.8)
	NPV	66.6 (10.8)	50.0 (0.0)	35.9 (10.4)	29.6 (10.4)	22.3 (9.5)
50	**SP**	50.0 (13.0)	50.0 (20.0)	50.0 (29.7)	50.0 (30.0)	50.0 (30.3)
	SN	50.0 (30.3)	50.0 (30.0)	50.0 (29.7)	50.0 (20.0)	50.0 (13.0)
	PPV	26.6 (10.3)	35.9 (10.4)	**50.0 (0.0)**	**64.1 (10.4)**	**73.4 (10.3)**
	NPV	73.4 (10.3)	64.1 (10.4)	50.0 (0.0)	35.9 (10.4)	26.6 (10.3)
60	***SP***	64.3 (13.0)	66.7 (20.0)	60.0 (24.0)	50.0 (29.9)	50.0 (30.3)
	SN	50.0 (30.3)	50.0 (29.9)	60.0 (24.0)	66.7 (20.0)	64.3 (13.0)
	PPV	**33.4 (10.8)**	**50.0 (0.0)**	64.1 (10.4)	70.4 (10.4)	77.7 (9.5)
	NPV	*77.7* (9.5)	70.4 (10.4)	64.1 (10.4)	50.0 (0.0)	33.4 (10.8)
70	***SP***	78.6 (13.0)	75.0 (15.2)	70.0 (18.2)	62.5 (22.7)	50.0 (30.3)
	SN	50.0 (30.3)	62.5 (22.7)	70.0 (18.2)	75.0 (15.2)	78.6 (13.0)
	PPV	50.0 (0.0)	66.6 (10.8)	73.4 (10.3)	77.7 (9.5)	80.7 (8.7)
	NPV	80.7 (8.7)	77.7 (9.5)	73.4 (10.3)	66.6 (10.8)	50.0 (0.0)
80	***SP***	85.7 (8.9)	83.3 (10.3)	80.0 (12.4)	75.0 (15.5)	66.7 (20.7)
	SN	66.7 (20.7)	75.0 (15.5)	80.0 (12.4)	83.3 (10.3)	85.7 (8.9)
	PPV	70.6 (11.1)	77.8 (9.8)	82.0 (8.6)	84.8 (7.6)	86.8 (6.8)
	NPV	86.8 (6.8)	84.8 (7.6)	82.0 (8.6)	77.8 (9.8)	70.6 (11.1)
90	***SP***	92.9 (4.7)	91.7 (5.5)	90.0 (6.6)	87.5 (8.3)	83.3 (11.1)
	SN	83.3 (11.1)	87.5 (8.3)	90.0 (6.6)	91.7 (5.5)	92.9 (4.7)
	PPV	84.9 (8.2)	88.5 (6.6)	90.7 (5.5)	92.1 (4.7)	93.2 (4.1)
	NPV	93.2 (4.1)	92.1 (47)	90.7 (5.5)	88.5 (6.6)	84.9 (8.2)

cholesterol reduction to a hypothetical mean target threshold of ≤107 mg/dL LDL-C in the treatment arm. Thus, the interest is in assessing what level of *PPV* will be necessary to achieve this target, and what will be the associated study power for demonstrating statistical significance to help decide if a *DX*-enriched clinical trial should be undertaken.

Simulation

Key features of the simulations and the statistical methods/tests used in this study are shown inside the text box. Simulation was carried out to generate artificial data representing prevalence of phenotype in 10% increments. SYSTAT (v. 11.0) and the open access statistical software R (v. 2.10.1) were

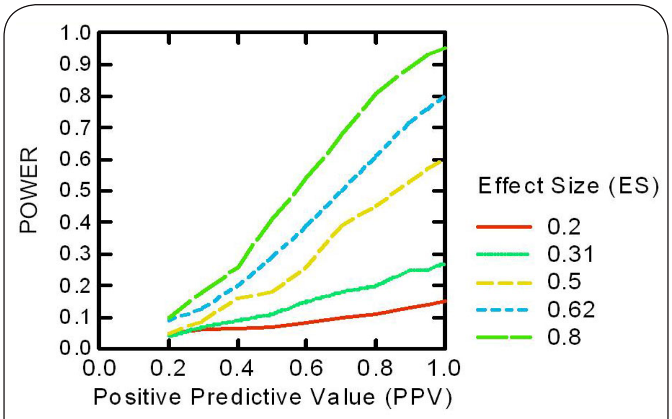

Figure 1. Relationship among study power, *PPV*, and effect size [n=82, power=80%, α=0.05, and reference effect size=0.62 (corresponding to treatment mean=97 mg/dL, SoC mean=124 mg/dL, pooled standard deviation=43 mg/dL)].

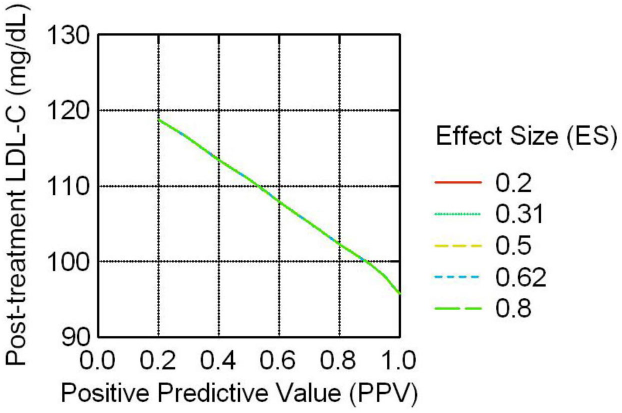

Figure 2. Relationship of clinical effectiveness, different effect sizes (each with correctly specified sample size), and *PPV*.

used for data analyses.

- Control arm: mean LDL-C = 124 (± 43) mg/dL (n=5000).
- Treatment arm: mean LDL-C = 97 (± 43) mg/dL (n=5000).
- Effect size (the difference between control and treatment arms divided by the pooled standard deviation): 0.2, 0.31, 0.5, 0.62 (reference, from the Dutch Study), and 0.8.
- *PPV*: 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.95, 1.0 (1.0 = perfect *DX* accuracy),
- Statistical test and other parameters: two-sample t-test, α=0.05, β=0.2, n=2000 bootstrap samples.

Results

Simulation results and observed relationships between the different measures of diagnostic accuracy and pre-diagnostic probabilities are summarized in (**Table 1**). The shaded cells with bold numbers in the table indicate thresholds of *PPVs* below which choosing a *DX*-enriched study design will not make sense because the posterior probabilities offer no advantage over the pre-test probabilities.

(**Table 1**) shows that higher overall accuracy of a *DX* is need ed for lower prevalence in order to cross the potential minimal utility threshold of *PPV* (a *PPV* of 50% is equivalent to tossing an unbiased coin when pre-test probability is 50%). Thus, for investment in a *DX*-enriched study design, an acceptable target *PPV* will need to be predetermined taking into consideration the assumed pre-test probability and level of willingness of the sponsor to risk trial failure. As an example, a *PPV* threshold of 70% may constitute such a target deemed to be acceptable. Thus, in order to cross a 70% *PPV* threshold, an overall *DX* accuracy of 80, 80, 70, 60, and 50% are required for prevalence of 30, 40, 50, 60, and 70%, respectively (**Table 1**).

(**Figure 1**) demonstrates relationships among study power, *PPV*, and different effect sizes for a fixed sample size (n=82). As effect size decreases from the reference effect size, study power obtained from gains in diagnostic accuracy (*PPV*) declines rapidly, exhibiting progressively lower relative impact of *PPV* on the study power (bottom three lines in **Figure 1**). However, for an underestimated or a correctly specified effect size, there is rapid gain in study power as *PPV* increases (top two lines in **Figure 1**). Thus, as per expectation, a study's power is affected by both the sample size and the *PPV* of a *DX*.

However, having sufficient study power by correct specification of sample size does not guarantee demonstration of clinical effectiveness. This point was previously emphasized in reference [12], in which the authors caution and illustrate that even large changes in statistical significance levels can correspond to small, non-significant changes in the underlying quantities of practical interest. Once a sample size has been adequately specified for an unknown true effect size (i.e., the influence of variance has been sufficiently accounted for) to achieve the desired study power, clinical effectiveness will then depend upon the *PPV* of a *DX*-test as shown by the steep negative slope of the superimposed lines representing the different effect sizes in (**Figure 2**) (the lines for the different effect sizes lie on top of one another because sample sizes have been adjusted for differences in variances). Thus, a *PPV* of 70% (0.7 on the *X*-axis) would insure that a *DX* +ve study should result in expected mean LDL-C level of 105 mg/dL (which satisfies the clinical effectiveness target of less than 107 mg/dL on the *Y*-axis), irrespective of the effect size. Note that even though the clinical effectiveness target is expected to be met, the study power would only be ~ 60% (**Figure 1**, for the reference effect size of 0.62), not 80%. This difference in study power arises because the sample size for the 80% power was calculated assuming a *PPV* of 100%, not 70%. A *PPV* of 20% on (**Figure 2**). would be expected to reduce LDL-C level to only an expected mean of 119 mg/dL, a number significantly above the desired target LDL-C level set for demonstrating clinical effectiveness, even when statistical significance is attained. Of course, a *PPV* of 20% will likely fail to attain statistical significance because of low study power, but a consumer risk does exist, nevertheless.

Discussion

As drug costs are escalating with an increase in development costs nearing $2 billion for each marketed drug, success rate has declined from ~12% to ~7% [13]. "Fail fast, fail cheap," "shots on goal," the use of biomarkers, and changing governance and organizational models have been some of the strategies adopted by industry since 2001 [14]. As the science of individualized medicine matures, empiricism and the probabilistic underpinnings of medical practice are increasingly replaced by specific targeted diagnosis and treatments with mechanism-based deterministic precision [13]. Drug developers must now provide evidence of differentiation and clinical value in order to convince major payers to offer reimbursement for new medicines at a fair price [15,16]. Both public and private payers in rich and emerging economies are becoming increasingly interested in using evidence to inform health-care resource allocation decisions and for preferential coverage in health plans [17,18]. One way to achieve and demonstrate such evidence is through practical enrichment, i.e., seeking to reduce noise (variability of measurement) and minimizing heterogeneity of patients in clinical trials [19]. A sufficiently accurate screening diagnostics, as illustrated in this paper, can be an invaluable tool for such a purpose. Ideally, relevant diagnostics should be evaluated during drug development and be available for use in efficacy trials [20]. So what should such an evaluation encompass? Sensitivity and specificity of a diagnostic test estimate the probability of a positive or negative test result when the gold standard (truth-surrogate outcome) is known. These two commonly reported diagnostic measures are useful in selecting a test from among different competing tests. Positive and negative predictive values, on the other hand, measure the probability of making a correct choice from a test result. These latter two diagnostic measures are useful in clinical decision making and in patient screening for enrolment in clinical trials. Sensitivity infers the probability of a +ve *DX* test, given that the patient has the disease or phenotype of interest. With a test result in hand, however, the clinician wants to know the probability that the patient has the specified disease or phenotype given a +ve or –ve *DX* test. As predictive values are a function of both the sensitivity and specificity of the *DX* test and the pre-test probability, clinical decision making or patient selection for clinical trials should pay significant attention to making sure the pre-test probabilities are not seriously over- or under-estimated. Test sensitivities and specificities can only be correctly interpreted for decision-making when the unknown true pre-test probabilities have been estimated reasonably accurately. At low pre-test probabilities, diagnostic utility is often limited by inadequate accuracy of tests. As pre-diagnostic probability decreases further away from 50%, the *DX*-test will need to be increasingly more accurate in sensitivity while possessing increasingly higher specificity. At high pre-test probabilities, diagnostic tests may not be necessary for designing clinical trials because the study population will already be relatively homogenous. Any marginal gain in enrichment by using a diagnostics in such a situation may be inefficient from a time and cost perspective. Honig [21] expresses the lack of importance given to disease/trait prevalence in determining the predictive value, clinical utility, cost-effectiveness, and generalizability of screening, testing, or enrichment in published papers that discuss diagnostic measures, particularly those overtly emphasizing only sensitivity and specificity. Diagnostic tests will have highest utility in clinical trial designs when pre-test probabilities are in the mid-range [21,22].

Thus, if the success of a clinical trial depends on sufficient enrichment, then both diagnostic accuracy and pre-test probability of a disease or phenotype are important criteria. If the true prevalence of a disease or genetic trait of interest is low, the *PPV* of the test/screen also will be correspondingly low and, even with high sensitivity and specificity, the majority of positive tests will tend to be falsely positive [21,22]. A sponsor will want to invest in a clinical trial only if there is sufficient confidence in the trial's probability of success. Although pre-diagnostic probability can be biased toward trial success by altering a study's inclusion/exclusion criteria, such a manipulation often comes with a trade-off cost in higher rate of false negatives (i.e., higher screen failure rate and correspondingly smaller potential market size). Therefore, from a purely *DX*-enrichment viewpoint, positive predictive value must be recognized as the "primary" diagnostic measure of interest because it directly impacts the probability of success of a clinical trial. Effect size affects the power of a study if a study's signal/noise (*S/N*) ratio differs significantly from the initially assumed *S/N* used in calculating the sample size. Clinical trials designed with prognostic or predictive biomarkers as screening diagnostics can greatly increase the efficiency of trials because enrichment positively affects the *S/N* ratio and, consequently, leads to smaller sample size requirements for demonstrating both clinical efficacy and effectiveness. According to a recently published guidance document of the Food and Drug Administration (FDA) on clinical trial enrichment strategies, "the strategy can be particularly useful for early effectiveness studies because it can provide clinical *proof of concept* and contribute to selection of appropriate doses for later studies" [23]. The Agency further states, "The decision to use an enrichment design is largely left to the sponsor of the investigation, but like the entire research and clinical communities, FDA is very interested in targeting treatments to the people who can benefit from them (i.e., individualization)."

Conclusion

In conclusion, the following six-step checklist is recommended as a generic evidence-based guideline to aid decision-making on whether or not to adopt a *DX*-enriched clinical study design: (1) be reasonably confident that assumption of prevalence is not erroneous, (2) insure that *DX* yields a sufficiently high *PPV*, (3) insure that the expected *PPV* will likely result in a pre-specified minimum clinically meaningful average value for the study's primary endpoint, (4) seek assurance that assumption of effect size is not faulty in order to insure sufficient power from the planned study sample size, (5) conduct simulations to assess effects of plausible under-or over-estimations of the critical assumptions (i.e., conduct sensitivity analyses), and lastly (6) make go/no go decisions based upon careful evaluation of assumptions and synthesis of simulation results from all preceding five steps in conjunction with risk tolerance for clinical trial failure deemed to be acceptable by the sponsor. As only statistical significance is affected by a study's sample size, it is especially important to pre-specify and satisfy a minimum clinically-meaningful average treatment difference in both *DX*-enriched and *RCT* clinical studies. Adherence to the above guidelines will increase the likelihood of not only successful demonstration of efficacy, safety and benefit-risk management to obtain regulatory approval, but also the achievement of what Honig [24] has characterized in a recent editorial as the "fourth hurdle" to successful commercialization of biopharmaceutical products--increased likelihood of reimbursement.

Competing interests

The author declares that he has no competing interests.

Publication history

Received: 20-Sep-2013 Revised: 01-Nov-2013
Re-Revised: 06-Nov-2013 Accepted: 13-Nov-2013
Published: 21-Nov-2013

References

1. Arrowsmith J. **Trial watch: Phase II failures: 2008-2010**. *Nat Rev Drug Discov.* 2011; **10**:328-9. | Article | PubMed
2. Arrowsmith J. **Trial watch: phase III and submission failures: 2007-2010**. *Nat Rev Drug Discov.* 2011; **10**:87. | Article | PubMed
3. Eichler HG, Abadie E, Breckenridge A, Flamion B, Gustafsson LL, Leufkens H, Rowland M, Schneider CK and Bloechl-Daum B. **Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response**. *Nat Rev Drug Discov.* 2011; **10**:495-506. | Article | PubMed
4. Swets J.A., Dawes R.M., and Monahan J. **Psychological Science Can Improve Diagnostic Decisions**. *Psychological Science in the Public Interest*. 2000; **1**:1-26. | Pdf
5. Hubert L and Wainer H. **A Statistical Guide for the Ethically Perplexed**. *CRC Press, Boca Raton, FL.* 2013; 565. | Article
6. Kelly H, Bull A, Russo P and McBryde ES. **Estimating sensitivity and specificity from positive predictive value, negative predictive value and prevalence: application to surveillance systems for hospital-acquired infections**. *J Hosp Infect.* 2008; **69**:164-8. | Article | PubMed
7. Suess E.A. and Trumbo B.E. **Introduction to Probability Simulation and Gibbs Sampling with R**. *Springer.* 2010; 307. | Pdf
8. Huijgen R, Kindt I, Verhoeven SB, Sijbrands EJ, Vissers MN, Kastelein JJ and Hutten BA. **Two years after molecular diagnosis of familial hypercholesterolemia: majority on cholesterol-lowering treatment but a minority reaches treatment goal**. *PLoS One.* 2010; **5**:e9220. | Article | PubMed Abstract | PubMed Full Text
9. Stein E.A. and Raal F.J. **Polygenic familial hypercholesterolaemia: does it matter?** *Lancet.* 2013; **381**:1255-1257. | Article
10. Talmud PJ, Shah S, Whittall R, Futema M, Howard P, Cooper JA, Harrison SC, Li K, Drenos F and Karpe F et al. **Use of low-density lipoprotein cholesterol gene score to distinguish patients with polygenic and monogenic familial hypercholesterolaemia: a case-control study**. *Lancet.* 2013; **381**:1293-301. | Article | PubMed
11. Mayo Clinic Staff. **Cholesterol levels: What numbers should you aim for?** 2013. | Website
12. Gelman A. and Stern H. **The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant**. *The American Statistician*. 2006; **60**:328-331. | Article
13. Waldman SA and Terzic A. **Pharmacoeconomics in the era of individualized medicine**. *Clin Pharmacol Ther.* 2008; **84**:179-82. | Article | PubMed
14. Cioffe C. **Portfolio selection and management in pharmaceutical research and development: issues and challenges**. *Clin Pharmacol Ther.* 2011; **89**:300-3. | Article | PubMed
15. Tunis SR, Stryer DB and Clancy CM. **Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy**. *JAMA.* 2003; **290**:1624-32. | Article | PubMed
16. Chalkidou K. **The (possible) impact of comparative effectiveness research on pharmaceutical industry decision making**. *Clin Pharmacol Ther.* 2010; **87**:264-6. | Article | PubMed
17. Lalonde RL and Willke RJ. **Comparative efficacy and effectiveness: an opportunity for clinical pharmacology**. *Clin Pharmacol Ther.* 2011; **90**:761-3. | Article | PubMed
18. Schneeweiss S, Gagne JJ, Glynn RJ, Ruhl M and Rassen JA. **Assessing the comparative effectiveness of newly marketed medications: methodological challenges and implications for drug development**. *Clin Pharmacol Ther.* 2011; **90**:777-90. | Article | PubMed
19. Temple R. **Enrichment of clinical study populations**. *Clin Pharmacol Ther.* 2010; **88**:774-8. | Article | PubMed
20. Woodcock J. **Assessing the clinical utility of diagnostics used in drug therapy**. *Clin Pharmacol Ther.* 2010; **88**:765-73. | Article | PubMed
21. Honig P.K. **Prevalence and Clinical Utility**. *Clinical Pharmacology & Therapeutics.* 2011; **89**: 488-489. | Article
22. Khatry D.B. **When Is a Biomarker-based Study Design in Drug Development Likely to Succeed? Proceedings of the Joint Statistical Meetings, Biometrics Section: Alexandria, VA: American Statistical Association**. 2011; 1982-1990. | Website
23. FDA. **Guidance for Industry: Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products (Draft Guidance), CDER/CBER/CDRH.** *Clinical Medical.* 2012; 41. | Pdf
24. Honig PK. **Comparative effectiveness: the fourth hurdle in drug development and a role for clinical pharmacology**. *Clin Pharmacol Ther.* 2011; **89**:151-6. | Article | PubMed

Citation:
Khatry DB. **Interrelationships among enrichment with diagnostics, probability of success and demonstration of effectiveness in clinical trials.** *J Med Stat Inform.* 2013; **1**:4.
http://dx.doi.org/10.7243/2053-7662-1-4