Guidance for Industry

Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the Federal Register of the notice announcing the availability of the draft guidance. Submit comments to the Division of Dockets Management (HFA-305), Food and Drug Administration, 5630 Fishers Lane, rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the Federal Register.

For questions regarding this draft document contact (CDER) Robert Temple, 301-796-2270, (CBER) Office of Communication, Outreach and Development, 301-827-1800, or (CDRH) Robert L. Becker, Jr., 301-796-6211.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Center for Devices and Radiological Health (CDRH)
December 2012
Clinical Medical
Guidance for Industry

Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products

Additional copies are available from:

Office of Communication
Division of Drug Information, WO51, Room 2201
Center for Drug Evaluation and Research
Food and Drug Administration
10903 New Hampshire Ave.
Silver Spring, MD 20993
Phone: 301-796-3400; Fax: 301-847-8714
druginfo@fda.hhs.gov

or

Office of Communication, Outreach and Development
Center for Biologics Evaluation and Research
Food and Drug Administration
1401 Rockville Pike, suite 200N
Rockville, MD 20852-1448
(Tel) 800-835-4709 or 301-827-1800
ocod@fda.hhs.gov

or

Office of Communication, Education, and Radiation Programs
Division of Small Manufacturers, International, and Consumer Assistance, HFZ-220
Center for Devices and Radiological Health
Food and Drug Administration
1350 Piccard Drive, Rockville, MD 20850-4307
DSMICA E-mail: dsmica@cdrh.fda.gov; DSMICA Fax: 301-443-8818
(Tel) Manufacturers Assistance: 800-638-2041 or 301-443-6597
(Tel) International Staff: 301-827-3993
http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/default.htm

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Center for Devices and Radiological Health (CDRH)

December 2012
Clinical Medical
II. BACKGROUND

Clinical trials are not designed to demonstrate the effectiveness of a treatment in a random sample of the general population. Instead, sponsors use a variety of strategies to select a subset of the general population in which the effect of a drug, if there is one, can more readily be demonstrated. Some of these selection strategies are obvious (e.g., patients are enrolled only if they have the disease that the drug being studied is intended to treat), but there are many more ways in which patients are typically chosen to make detection of a treatment effect more likely. Examples include selecting patients whose disease does not spontaneously disappear or exhibit a large degree of variability, who are likely to comply with treatment, who are likely to have a high rate of disease progression, or who have some

---

This guidance was developed by the Center for Drug Evaluation and Research in coordination with the Center for Biologics Evaluation and Research and the Center for Devices and Radiological Health at the Food and Drug Administration.
characteristic that suggests they can respond to the treatment. All of these selection strategies can be described as *enrichment* of the study population.

For the purposes of this guidance, the term *enrichment* is defined as the prospective use of any patient characteristic to select a study population in which detection of a drug effect (if one is in fact present) is more likely than it would be in an unselected population.

Among others, demographic, pathophysiologic, historical, genetic or proteomic, clinical, and psychological characteristics have been used for enrichment. Enrichment may also refer to the population to be analyzed within a broader population; that is, a study could include patients both with and without the enrichment characteristic, but the primary analysis would be of the subset with the characteristic, an approach that increases the study’s ability to detect a drug effect, but that can also provide some information about patients without the enrichment characteristic. Although this guidance focuses on enrichment directed at improving the ability of a study to detect a drug’s effectiveness, similar strategies can be used in safety assessments.

Enrichment strategies fall into three broad categories:

1. **Strategies to decrease heterogeneity** — These include selecting patients with baseline measurements in a narrow range (decreased inter-patient variability) and excluding patients whose disease or symptoms improve spontaneously or whose measurements are highly variable (decreased intra-patient variability). The decreased variability provided by these strategies increases study power (see section III).

2. **Prognostic enrichment strategies** — choosing patients with a greater likelihood of having a disease-related endpoint event (for event-driven studies) or a substantial worsening in condition (for continuous measurement endpoints) (section IV). These strategies will increase the absolute effect difference between groups but will not alter relative effect.

3. **Predictive enrichment strategies** — choosing patients more likely to respond to the drug treatment than other patients with the condition being treated. Such selection can lead to a larger effect size (both absolute and relative) and permit use of a smaller study population. Selection of patients could be based on a specific aspect of a patient’s physiology or a disease characteristic that is related in some manner to the study drug’s mechanism, or it could be empiric (e.g., the patient has previously appeared to respond to a drug in the same class) (section V).

This guidance describes and illustrates important enrichment strategies within these categories; discusses study design options for different strategies, including advantages and disadvantages of the various designs; and addresses issues of interpretation of the results of enrichment studies.

The enrichment strategies described in this guidance are discussed primarily in the context of randomized controlled trials. In almost all cases, the strategies affect patient selection before randomization (with a few exceptions for adaptive strategies to be noted later). These strategies,
therefore, generally do not compromise the statistical validity of the trials or the meaningfulness of the
conclusions reached with respect to the population actually studied.

The principal concerns with the use of enrichment strategies relate to the generalizability and
applicability of the study results. When considering use of an enrichment design, it is important to
consider whether the enrichment strategy can be used in practice to identify the patients to whom the
drug should be given and whether the drug might be useful in a broader population than will be
studied. The extent to which patients who do not meet the selection criteria for enrichment should be
studied (section VII.B.) is therefore a critical consideration. In addition, the accuracy of the
measurements used to identify the enrichment population and the sensitivity and specificity of the
enrichment criterion in distinguishing responders and non-responders are also critical issues.

III. DECREASING HETEROGENEITY

Approaches to increasing study power – the ability of a clinical trial to demonstrate a treatment effect
if one is present – by decreasing heterogeneity (non-drug related variability) are widely practiced. The
following strategies are useful and generally accepted ways to decrease heterogeneity:

- Defining entry criteria carefully to ensure that entered patients actually have the disease that is
  being studied and training investigators to adhere to protocol-specified entry definitions and
criteria.

- Identifying and selecting patients likely to comply with treatment to decrease variability in
drug exposure. Note: removing poor compliers identified after randomization is generally not
acceptable because such patients are not likely to be a random sample of the study population
and because compliance itself has been linked to outcome, even compliance in taking a
placebo.

- Using placebo-lead in periods prior to randomization to eliminate patients who improve
  spontaneously or have large placebo responses.

- Decreasing intra-patient variability by enrolling only patients who give consistent baseline
  values (e.g., for blood pressure measurements, treadmill exercise tests, pulmonary function
tests, or patient reported outcome (PRO) measures).

- Excluding patients taking drugs that are pharmacologically similar to, or that could interact
  with, the study drug.

- Excluding patients unlikely to tolerate the drug.

- Excluding patients who are likely to drop out for non-medical reasons (e.g., because they have
difficulty getting to the study site).

Other strategies sometimes employed include eliminating patients in outcome trials with concomitant
illness likely to lead to early death or early drop-out and broad exclusions of patients on concomitant
therapies. There are concerns, however, that these strategies can result in studies that provide too little
information about the full range of people who will receive a drug in clinical practice, such as the
elderly, people with multiple illnesses, and those taking multiple drug therapies. It is not clear that
concomitant illnesses that do not affect survival or other endpoint measurements or concomitant drugs unrelated to a test drug really do interfere with assessment of a treatment effect. Therefore, the implications of using these strategies should be carefully considered before they are used. Two of these strategies – encouraging compliance and reducing the numbers of spontaneous improvers or placebo responders – warrant further discussion (see below).

A. Encouraging Compliance

Practices such as encouraging good compliance by making patients aware of the conditions and demands of the trial, avoiding too-rapid titration of drugs that could cause intolerable early side effects, using adherence prompts and alert systems, and counting pills (or using “smart bottles” to monitor drug use) so that non-compliant patients can be encouraged to perform better have become standard. There have also, on occasion, been more specific efforts to identify and enroll good compliers into clinical trials.

1. The VA Cooperative hypertension studies of the late 1960s and early 1970s\textsuperscript{2,3} gave prospective patients placebo tablets containing riboflavin during a single-blind placebo period, then examined patients’ urine for fluorescence and randomized only patients whose urine fluoresced (evidence that the patients had been taking the riboflavin tablets) on two consecutive visits during a 2-4 month observation period.

2. The Physicians’ Health Study\textsuperscript{4} used an 18-week, pre-randomization placebo run-in during which patients (all physicians) self-reported compliance. About one third of the screened patients were not randomized because of self-reported poor compliance. Compliance during the randomized study was reported as a very satisfactory 90% over the 5 years of the study, greatly increasing its power.\textsuperscript{5}

B. Decreasing Placebo Responses and Spontaneous Improvement

In placebo-controlled trials of drugs for symptomatic conditions (e.g., depression, anxiety, angina) or laboratory abnormalities (e.g., dyslipidemia, hypertension), it is relatively common to have a single-blind, placebo lead-in period and to randomize only patients whose signs or symptoms remain above some entry level value. This is in part to identify and not randomize patients who would have had an improvement for a reason other than a response to the test treatment (spontaneous improvement, a placebo response, or effect of expectations or observations) that resolved the patient’s symptoms or signs, making the patient incapable of showing a response to treatment. Also, many signs and symptoms vary spontaneously and, therefore, initial screening values that would support enrollment may represent random highs of the disease course that will be followed by regression to the mean, leaving the patient without the condition to be treated.

IV. PROGNOSTIC ENRICHMENT STRATEGIES—IDENTIFYING HIGH-RISK PATIENTS

A wide variety of prognostic indicators have been used to identify patients with a greater likelihood of having the event (or a large change in a continuous measure) of interest in a trial. These indications include clinical and laboratory measures, medical history, and genomic or proteomic measures.
Selecting such patients allows a treatment effect to be more readily discerned. For example, trials of prevention strategies (reducing the rate of death or other serious event) in cardiovascular (CV) disease are generally more successful if the patients enrolled have a high event rate, which will increase the power of a study to detect any given level of risk reduction. Similarly, identification of patients at high risk of a particular tumor, or at high risk of recurrence or metastatic disease can increase the power of a study to detect an effect of a cancer treatment. Prognostic enrichment strategies are also applicable, or potentially applicable, to the study of drugs intended to delay progression of a variety of diseases, such as Alzheimer’s disease, Parkinson’s disease, rheumatoid arthritis, multiple sclerosis, and other conditions, where patients with more rapid progression could be selected; it is possible, of course, that such patients might be less responsive to treatment (i.e., that rapid progression would be a negative predictor of response), and that would have to be considered.

For any given desired power in an event-based study, the appropriate sample size will depend on effect size and the event rate in the placebo group. Prognostic enrichment does not increase the relative risk reduction (e.g., percent of responders or percent improvement in a symptom), but will increase the absolute effect size, generally allowing for a smaller sample size. For example, reduction of mortality from 10% to 5% in a high-risk population is the same relative effect as a reduction from 1% to 0.5% in a lower risk population, but a smaller sample size would be needed to show a 5% vs. 0.5% change in absolute risk. It is common to choose patients at high risk for events for the initial outcome study of a drug and, if successful, move on to larger studies in lower risk patients.

A. Experience with Prognostic Enrichment Strategies

1. Cardiovascular Studies

In CV disease, the severity of the illness being studied, as well as other factors that can indicate increased risk, such as a history of recent myocardial infarction or stroke; the presence of concomitant illness such as diabetes, hypertension, or hyperlipidemia; and certain blood markers, such as very high LDL cholesterol, low HDL cholesterol and high C-reactive protein (CRP), have been used to identify patients at greater risk for cardiovascular events, considerably reducing the sample sizes needed to show an effect in outcome studies. Outcome studies using ACE inhibitors in congestive heart failure (CHF) and HMG CoA reductase inhibitors in hyperlipidemia (the enalapril and statin trials, respectively) illustrate this approach.

In the enalapril trials, mortality reduction and decreases in morbidity events (such as hospitalization) were first assessed in a very ill CHF population of NYHA Class IV patients (CONSENSUS), then in less ill patients (SOLVD treatment) and eventually in asymptomatic patients (SOLVD prevention). In the later studies, composite endpoints were needed because the number of early deaths was too low to allow an effect to be demonstrated, about 15% at one year on placebo in SOLVD treatment and about 5% in SOLVD prevention, far lower than the 44% placebo 6 month mortality in CONSENSUS. The very high early mortality in CONSENSUS, together with the large effect size (40% reduction), allowed demonstration of a survival effect in just 253 patients, while the studies in less ill patients required studies of 2,000-4,000 patients. The higher risk patients enrolled as a result of prognostic enrichment showed, as expected, a larger absolute effect size, but relative effect size was also greater in the more ill patients, suggesting that severity was also a predictive marker (see section V).
A similar strategy was used in the statin trials. The early CV outcome trials with statins were able to evaluate the effects of the drugs on mortality because they enrolled patients with a history of heart disease and very elevated cholesterol levels, patients whose mortality risk was substantial.\textsuperscript{9,10} As the benefit of statins became established in high-risk patient populations, subsequent CV outcome trials had to enroll patients with less marked LDL cholesterol elevations and without known coronary artery disease – populations that had not yet been shown to benefit from LDL cholesterol lowering and who could still be ethically studied, but who were identified as high risk because of some other illness (e.g., type 2 diabetes mellitus) or risk factor (e.g., low HDL cholesterol, elevated high-sensitivity CRP). As the population’s risk became lower, sample sizes increased considerably but prognostic factors made the studies possible. For example, in the recent JUPITER study\textsuperscript{11} (n=17,802), a statin was shown to have an effect on outcome in patients with normal LDL cholesterol, but who were at higher CV risk based on factors other than LDL cholesterol, including age, one additional CV risk factor, and a high-sensitivity CRP $\geq$ 2 mg/L. As the magnitude of risk declined in these study populations, it often also became necessary to rely on composite endpoints, as the mortality rate was too low to allow a mortality trial of reasonable size.

Choosing patients at relatively high risk of cardiovascular events can also be critical to the success of safety studies intended to rule out a given level of cardiovascular risk. This approach is now recommended for new antidiabetic treatments\textsuperscript{12} and has been a consideration in the design of studies to evaluate the cardiovascular risk of non-steroidal anti-inflammatory drugs.

2. Oncology

It is clear that chemoprevention studies are most likely to be successful in people identified as high risk (e.g., by genetic or other characteristics) for developing a cancer or recurrence of cancer. For example, adjuvant therapy studies of tamoxifen showed that the drug not only delayed development of metastases in patients with breast cancer, but also reduced the risk of contralateral tumors (new primary tumors) in this high-risk group (high risk because they had already had breast cancer). Tamoxifen was then studied in 13,000 high-risk people (calculated using the Gail model) without a prior diagnosis of breast cancer who were followed for 4 years (NSABP P-1).\textsuperscript{13} The study showed a 44% relative reduction in risk of invasive breast cancer, and tamoxifen was labeled for use in reducing the risk of breast cancer in high-risk individuals identified using the Gail model calculator. A study in people at lower risk would have required a substantially larger sample size. For example, a study of people with a risk that was 25% of the risk of the NSABP P-1 study population would have needed about 20,000 people to detect an effect of the size observed in the NSABP P-1 study with 90% power.

B. Potential Strategies for Prognostic Enrichment

There may be additional approaches to identifying high-risk CV patients. A report in 2005\textsuperscript{14} pointed out that a higher resting heart rate, a small increase in exercise heart rate, and delayed recovery of heart rate were all strong predictors of sudden death, suggesting a potential enrichment strategy in studies of drugs to prevent sudden death. More recently, the potential for risk prediction based on genetic factors has been examined,\textsuperscript{15} as has the predictive value of coronary artery calcium score.\textsuperscript{16,17} The actual usefulness of these potential predictors as enrichment tools has yet to be established.
Published reports suggest additional strategies for identifying high-risk breast cancer patients, especially for participation in studies of adjuvant treatment of cancer, where the likelihood of recurrence and survival are critical to demonstrating an effect on these endpoints.

1. Prostate Cancer

It has been reported\textsuperscript{18} that in men with localized prostate cancer following radical prostatectomy, high PS\textsubscript{A} velocity (PS\textsubscript{A} increase $> 2$ ng/mL during prior year) strongly predicted prostate cancer recurrence and mortality over a 10-year period. Prognostic markers such as PS\textsubscript{A} velocity, if validated in future studies, could be used to identify high-risk patients. Studies of adjuvant treatment for prostate cancer would be better able to show an effect on survival if they enrolled patients with a high risk of death.

2. Breast Cancer

Many investigators have reported gene expression profiles that appeared to predict the likelihood of breast cancer recurrence after surgery. Selection of a population with a high rate of recurrence and poor survival is critical if an adjuvant therapy study is to be successful. In a report on the use of five different gene-expression profiling approaches in a non-randomized 285-patient sample treated with local therapy, tamoxifen, tamoxifen plus chemotherapy, or chemotherapy alone, Fan, et al.,\textsuperscript{19} found that four of the five approaches for classifying patients had high concordance and a striking ability to predict large differences in both recurrence and mortality, as illustrated in Figure 1, showing the difference between patients with good and poor 70-gene profiles on relapse-free survival and overall survival. These data need to be verified in actual trials, but gene expression profiling may provide a means of identifying higher risk patients for adjuvant trials.

Figure 1: Relapse Rates and Survival in Breast Cancer Patients Based on 70-Gene Profile

![Figure 1: Relapse Rates and Survival in Breast Cancer Patients Based on 70-Gene Profile](image)

Recently, MammaPrint, an in vitro diagnostic test using the gene expression profile of fresh breast cancer tissue samples to assess a patient’s risk for distant metastasis, was cleared by FDA as a prognostic test for certain breast cancer patients. As noted, use of such a diagnostic test represents a potential enrichment strategy for adjuvant trials to identify a population at higher risk for recurrence.

Women with a deleterious BR\textsubscript{CA} 1 or 2 mutation have a lifetime incidence of breast cancer and ovarian cancer of 60\% and 15-40\%, respectively, compared to a risk of 12\% and 1.4\%, respectively, in
women without a BRCA mutation. Selecting women with such prognostic markers for a primary
prevention trial in breast cancer or ovarian cancer would increase the likelihood of cancer events,
thereby permitting a smaller sample size and a shorter study.

V. PREDICTIVE ENRICHMENT

There are many possible ways to identify patients more likely to respond to a particular intervention,
and these have long been used in clinical trials when selection of patients has been based on a specific
aspect of pathophysiology, past history of response, or a disease characteristic that is related in some
manner to the study drug’s mechanism (e.g., genomic or proteomic factor). For example:

- CHF can result from either systolic or diastolic dysfunction. Presumably, a population with
  systolic dysfunction would be more likely to respond in a study of an inotrope.

- High-renin status predicts a greater anti-hypertensive response to beta-blockers, ACE
  inhibitors, and angiotensin receptor blockers. A population with high renin hypertension
  would be more responsive than a general hypertension population in studies of drugs in these
  classes.

- For some indications, antibacterial drug effects are best analyzed in patients whose organism is
  sensitive to the antibacterial drug. Most commonly, patients are randomized before sensitivity
  is known, but only those patients whose organism is subsequently found to be sensitive to the
  test antibacterial are analyzed for effectiveness.

- An initial screening for response—a biomarker measurement (e.g., radiographic response or
  reduction of ventricular premature beats), early clinical response, or full-fledged clinical
  response—in an open pre-randomization period can be used to identify a responder population
  to be randomized into the controlled study. This approach is of particular value when
  responders constitute only a small fraction of the overall population to be treated.

- A population of non-responders, or people intolerant to another drug, can be randomized to the
  new drug or the original drug. The comparison is enriched because the population is expected
  to have a poor response or a high rate of intolerance to the original drug compared to the test
  drug. These designs cannot be used where effectiveness is critical or the intolerance is
  dangerous.

- Proteomic markers, such as the HER 2/neu marker in breast cancer indicating potential for
  response to trastuzumab, tumor EGFR markers, or genetic markers related to a drug’s
  mechanism of action can be used to identify potential responders, a rapidly growing
  enrichment strategy in oncology.
Identifying a responder population (i.e., a subset of the overall population with a larger than average response to treatment) and studying this population in a clinical trial can provide two major advantages: increased study efficiency or feasibility and an enhanced benefit-risk relationship for the people in the subset compared to the overall population.

**Increased Efficiency or Feasibility**

Identification of a high treatment response population greatly increases the chance that a study of an effective drug will be able to detect a treatment effect and allows a study to succeed with a smaller sample size than a study in an unselected population. 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. When the treatment responder population constitutes only a small fraction of all patients, say 20% (a common situation in oncology settings), enrichment can permit a showing of effectiveness when a study in an overall population may have difficulty showing any effect. Unlike prognostic enrichment, which leads to a larger absolute effect but no change in relative effect, a predictive enrichment population will show both a larger absolute effect and a larger relative effect than the general, unselected population.

The extent to which the sample size needed to adequately power a study can be reduced using a predictive enrichment strategy is a function of the prevalence of the enrichment marker and the relative effectiveness of the drug in the marker-positive and marker-negative populations. Table 2 illustrates how sample size ratios—the ratio of the number of subjects needed in an unselected population versus the number needed if only the marker positive population is studied—change with varying prevalence of marker-positive patients and different magnitudes of treatment effect in marker-negative patients (treatment effect in marker-negative patients of either 0% or 50% of the effect in marker-positives). Table 1 assumes the classification of patients into positive versus negative is 100% accurate.

**Table 1: Sample Size Ratios as a Function of the Prevalence of Marker-Positive Patients**

<table>
<thead>
<tr>
<th>Prevalence of Marker-Positive Patients</th>
<th>Treatment Effect in Marker-Negative Patients (% of Marker-Positive Response)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>100%</td>
<td>1.0</td>
</tr>
<tr>
<td>75%</td>
<td>1.8</td>
</tr>
<tr>
<td>50%</td>
<td>4</td>
</tr>
<tr>
<td>25%</td>
<td>16</td>
</tr>
</tbody>
</table>

In general, the lower the prevalence of marker-positive patients and the smaller the relative effect size in the marker-negative population, the more the sample size can be reduced in a study of marker-positive patients compared to a study in an unselected population. For example, when the prevalence of marker-positives in a population is only 25% and no treatment effect is expected in the 75% of
patients who are marker-negative, the required sample size in a study of an unselected population
would be 16 times the sample size needed for a study that included only marker positive patients. A
more detailed description of these results and the conditions under which the results were obtained has
been presented by Simon and Maitournam.21

**Enhanced Benefit–Risk Relationship**

Identification of a responder population can enhance the benefit–risk relationship of a drug by
avoiding exposure and potential toxicity in people who cannot benefit from the drug. For drugs with
significant toxicity and a low overall response in a general population – factors that could deter further
development – identifying a responder population could make a risk more acceptable and facilitate
continued development and approval. For example, the significant survival advantage (approximately
5 months) seen with trastuzumab in the treatment of metastatic breast cancer in patients with high
HER 2/neu expressing tumors (about 25% of breast cancers) ultimately supported use of the drug in
the marker-selected population despite the significant cardiotoxicity that emerged (see further
discussion in section V.B.3). The much smaller mean effect (less than 2 months) that would have
been observed in an unselected population and the fact that only about one-fourth of patients would
have benefited might have made approval difficult to support in the face of the observed cardiotoxicity
of the drug.

Identifying a more responsive population does not necessarily imply that there is no benefit in the
remaining population. It is therefore generally desirable to have some data in the non-selected (non-
enrichment) population to determine whether they respond less well, or indeed do not respond at all.
These data also can provide an assessment of safety in the non-selected population in the event that
such patients are exposed postapproval. The data need not be obtained in the controlled trials
supporting effectiveness but could be obtained in earlier studies. A qualitative estimate of
effectiveness might also be based on pharmacologic or even pre-clinical data. A strong mechanistic
rationale can make study of the non-enriched population unnecessary (e.g., study of effects in an
infection caused by a resistant organism).

A trial intended to provide evidence of effectiveness to support approval could include a broad range
of patients, but be prospectively designed to have as its primary endpoint the effect in the enriched
population subset. This is a standard (and unavoidable) approach when the baseline characteristic can
only be determined after randomization (e.g., the infectious organism or tumor characteristic), but the
approach (preferably with stratified randomization) is also valuable in other settings to gather some
information on the marker-negative population.

The following discussion considers predictive enrichment strategies in 5 categories:

- empiric strategies
- pathophysiologic strategies
- genomic strategies
- randomized withdrawal studies
- studies in non-responders or patients intolerant to other therapy
A. Empiric Strategies

With an empiric strategy, the selection of likely responders for a study is not based on any understanding of the basis for differences in response between patients, but on observations of a response during screening periods, or prior experience with the drug or related drugs.

1. Open Trial Followed by Randomization

A straightforward enrichment strategy, in cases when a treatment response can be identified shortly after treatment initiation, is to give open-label drug to all patients, identify apparent responders, either on the planned study endpoint or on a biomarker or other short-term response thought to be predictive of clinical response; withdraw the treatment; and then randomize only responders into a placebo-controlled trial. This strategy is particularly useful when there is a low response rate. This strategy has been discussed in the past and was used throughout the 1970s to develop new anti-arrhythmic drugs. Patients were titrated on the investigational agent until they had an acceptable reduction of ventricular premature beats (VPBs). Only the responding patients were then randomized into placebo-controlled trials, often fixed-dose, dose-response studies.

In the mid-1980s, controlled studies to evaluate topical organic nitrate patches randomized only people who were shown immediately before the study to have an exercise angina response, or a blood pressure response, to sublingual nitroglycerin. See also section 2 below (History of Response to a Treatment Class).

The Cardiac Arrhythmia Suppression Trial (CAST) – a study of the mortality effect of suppressing VPBs in patients with a recent acute myocardial infarction (AMI) and at least 6 VPBs per hour – is one of the best known studies conducted in apparent responders. It was known that people with > 10 VPBs per hour after an AMI had a 4-fold increase in the rate of sudden death. Previous failed attempts to show survival benefits with antiarrhythmics had been criticized for the low rate of VPB suppression achieved because many included patients did not respond. The CAST used an open-label screening period to identify responders to two drugs, encaïnide and flecainide, shown to be very effective in suppressing VPBs in a previous study. Only patients who had at least a 70% VPB reduction were randomized. Unfortunately, despite the enrichment effort, these anti-arrhythmic agents did not decrease mortality but instead more than doubled it. This result reflects the inadequacy of the surrogate endpoint of VPB reduction as a predictor of an effect on mortality, not the study design. The enrichment design did, however, yield a study capable of showing an effect of VPB suppression and allowed clear interpretation of the study, which showed, contrary to expectations, that even in VPB responders, the drugs were not helpful and, indeed, were harmful.

Note: Use of an initial open-label phase without a control group does raise some concerns that need to be considered when such a design is used. For example, in the CAST study there were deaths during the screening period (not surprising given the recent infarction) that were difficult to interpret in an open, uncontrolled setting where all patients received active drug. In CAST II (ethmosin vs. placebo), the initial screen for VPB suppression used a randomized comparison of drug to placebo, with the responders then randomized into the placebo-controlled trial. This strategy showed that the drug used in the screen was itself lethal (19 deaths on ethmosin vs. one on placebo) and the study was stopped early.
A similar problem was described in outcome trials of carvedilol in the treatment of chronic CHF.\textsuperscript{27} These studies, unlike CAST, clearly showed a benefit of treatment. In two large studies, some patients were excluded during a run-in period because they could not tolerate carvedilol. Some of those patients died. The drop-out rates during the subsequent controlled trials were undoubtedly decreased by the screening procedure that excluded patients intolerant to beta-blockers, and the results made carvedilol seem better tolerated than it would actually be in patients starting therapy. The randomized comparisons and the benefits demonstrated are fully valid in these trials for the populations studied, but the benefits and risks facing unselected patients, who would be treated in clinical use of the drug, may be different from those observed in the clinical trial, requiring close attention to the screening period results.

There are many other outcome study settings in which it would be possible to select people more likely to benefit from treatment. Patients with a lipid abnormality might be given the planned treatment in a screening period to evaluate their biochemical response. For the randomized trial, only people with a response of a certain size might be randomized, giving a greater mean effect on the lipid level and, presumably, a larger effect on outcome. That approach could be useful in an early outcome trial, but it would also be possible to randomize a broader population stratified by such an initial response with the intent of making the primary study endpoint the result in the high-response subgroup while gaining some information about the less responsive group. Again, the response in such a selected group would not describe the response in an unselected population, an important issue for labeling.

Active, open screening for empiric responders is particularly advantageous when a population is made up of subsets (not identifiable pathophysiologically or genetically prior to treatment) with potentially very different responses to interventions. It is hard to know in advance when this is true, but certain difficult-to-study conditions, such as irritable bowel syndrome or fibromyalgia, might be candidates for this approach.\textsuperscript{22}

The overall strategy (open trial followed by randomization) is a very efficient way to document effectiveness, but it cannot be used prospectively to identify the responder population when the drug is used in clinical practice. In some cases, however, an early response could be used to determine who should stay on the drug, essentially how all symptomatic treatments are used.

2. \textit{An Individual’s History of Response to a Treatment Class}

Information about prior response to a drug in a pharmacologic class, if available, can be used to identify potential responders for a study of a new member of that class. As is the case with an open label trial followed by randomization, use of patient history of response to a drug class can greatly increase the efficiency of a trial in demonstrating effectiveness. In most cases, however, it will not help identify the population to be treated in clinical practice.

A study enriched with prior responders to a pharmacologic class can be useful in demonstrating effectiveness at the proof-of-concept stage. This design may be particularly advantageous for randomized, fixed-dose, dose-response studies (the preferred dose-response study design described in the ICH guidance \textit{E4, Dose-Response Information to Support Drug Registration}\textsuperscript{28}). A responder
population provides a larger overall treatment effect and, therefore, a steeper dose-response curve, which generally allows for easier interpretation of the curve (identifying the steep area and plateau of the curve) and more precise characterization of dose-response, especially for doses providing near-maximum effects. For example, a dose-response study of indapamide in known responders to diuretics demonstrated mean decreases of 29/12 mmHg (systolic/diastolic) for the 2.5 mg dose and 37/15 for the 5 mg dose, an increase in effect with the 5 mg dose considerably larger than that seen in studies of unselected patients, where 2.5 mg and 5 mg gave similar results.

3. Factors Identified in Results from Previous Studies

Analyses of results of previous trials can sometimes point to a substantially greater effect in a specific subset of the overall population and provide a basis for studying that subset in a subsequent study, either as the sole population studied or as the identified primary endpoint subset in a study of a broader population. For example, BiDil, a treatment for severe CHF, was approved on the basis of a placebo-controlled study carried out entirely in self-identified blacks (the African-American Heart Failure Trial (A-HeFT)). The selection of a black population was based on two previous studies (the Vasodilator-Heart Failure Trials (V-HeFT) I and II) of a hydralazine hydrochloride – isosorbide dinitrate combination vs. placebo in a racially mixed population that strongly suggested effectiveness in blacks. In those studies, the combination had not shown an overall benefit, but post-hoc analyses revealed a nominally significant effect in black patients in V-HeFT I and apparent equivalence to enalapril in V-HeFT II. In contrast, there was little or no effect of the combination in whites in V-HeFT I and nominally significant inferiority of the combination to enalapril in whites in V-HeFT II. The replication of the observed effect in blacks was strong, with only a suggestion of a modest effect in whites, perhaps a third of the effect in blacks. A trial to establish this small effect in a white population would have required 16,000 patients. The product was approved for “self-identified blacks” only.

B. Pathophysiological Strategies

These strategies involve selection of likely responders based on the patient’s individual physiology or on assessment of disease pathophysiology that suggests that only certain patient subgroups will respond to a particular therapy or that certain subgroups will respond better than others.

1. Metabolism of the Test Drug

For a drug that acts through an active metabolite, as is the case for the antiplatelet drug clopidogrel, patients may differ in their ability to metabolize the prodrug to its active metabolite. Some patients may not form the active metabolite at all and others may not make enough to respond to the dose selected. Including these patients in a trial will dilute the overall drug effect and can also lead to inefficient or inappropriate use of the drug in practice if the two subsets of patients are not identified and treated differently. In some cases, it will be possible to adjust (increase) the dose in the poor metabolizers, but patients who cannot make the active metabolite at all should probably be excluded from the trial or from the planned primary analysis. A closely related approach is the assessment of uptake of the test drug by a tumor. Historically, before treatment of thyroid tumors with I-131, a low dose was given to determine whether the tumor did, in fact, take up iodine and to what extent, so that the needed dose could be estimated.
2. Effect on Tumor Metabolism

It may be possible to select patients for a cancer trial by screening for an effect on a tumor metabolic response, as assessed by a positron emission tomography (PET) scan. For example, response to the tyrosine kinase inhibitors imatinib and sunitinib in patients with gastrointestinal stromal tumors (GIST) has been shown to correlate well with metabolic responses (decreased tumor glucose utilization) assessed by 15F-fluorodeoxyglucose (FDG) PET imaging. The clinical trial could enter only the identified metabolic responders or enter all patients, stratified by metabolic response. In the second case, the primary hypothesis could be the treatment effect in the metabolic responder stratum.

3. Proteomic Markers and Genetic Markers Linked to a Proteomic Marker

Increasingly, cancer treatments are directed at enzymatic, hormonal, or other functions that are tied to tumor surface intracellular receptors. The following examples illustrate use of proteomic markers, or genetic markers that are linked to a proteomic marker, that are known to be essential for the activity of the drug.

- Trastuzumab was developed to bind to the Her-2-neu receptor, which is present on normal and malignant cells but is over-expressed in about 25% of breast cancers. Binding of trastuzumab to the Her-2 neu receptor blocks receptor-mediated growth-stimulating intracellular signaling, decreasing cellular repair after chemotherapy and radiation therapy and also increasing apoptosis. In activity-estimating trials, anti-tumor activity in patients with lower levels of Her-2-neu receptor expression (1+ by immunohistochemical staining) was minimal, so that definitive efficacy trials in patients with metastatic disease were limited to patients with Her-2-neu 2+ or 3+ over-expression. In the treatment of metastatic disease, when added to either of two background regimens, trastuzumab increased survival by a mean of about 5 months, about 3 to 4 times the effect that would have been expected in an unselected population, assuming no response (which a modest amount of testing showed was the case) in the Her-2-neu negative patients. Enrichment thus allowed a modest-sized study to show a striking effect and directed treatment to the population that could benefit. In addition, because the drug was shown to be moderately cardiotoxic in the metastatic breast cancer trials, in designing adjuvant studies it was considered critical to focus on potential responders (i.e., patients with Her-2-neu receptor over-expression).

- Imatinib was developed to treat patients with gastrointestinal stromal tumor (GIST), a tumor not previously responsive to antineoplastic therapy. Imatinib inhibits c-Kit, a receptor tyrosine kinase that is mutated and activated in a large majority of GIST patients, resulting in abnormal proliferation of tumor cells. In a small study (N=147) in patients with a pathologic diagnosis of c-Kit-positive unresectable and/or metastatic malignant GIST, 56 patients responded and 55 of the responding patients demonstrated a durable partial response of 7-38 weeks (median 13 weeks). This study was followed by a placebo-controlled, adjuvant therapy trial using imatinib in GIST patients with c-Kit expression in whom complete gross resection of GIST had been performed. The study showed a substantial increase in recurrence-free survival at a median follow up of 14 months.
Evidence from the metastatic breast cancer setting has demonstrated that the likelihood of response to endocrine therapy is related to the hormone receptor status of the tumor. For example, when treated with tamoxifen, a selective estrogen receptor modulator, patients whose tumors express both estrogen receptors (ER) and progesterone receptors (PR) have a response rate of approximately 70%; patients whose tumors express either ER or PR, but not both, had a 40% response rate; and patients whose tumors are ER and PR negative have a response rate less than 5%. As a consequence, testing of all breast cancer specimens to direct decisions regarding endocrine therapy, in both the early-stage and the advanced setting, has become the standard of care and would be expected in any trials of endocrine therapy.

A more recent illustration is the use of somatic mutations in the gene encoding the serine-threonine protein kinase BRAF to identify potential responders to vemurafenib in melanoma; 40% to 60% of all melanomas carry this activating mutation. In an initial study in 49 patients with melanoma, 11 of 16 patients with BRAFV600E who received vemurafenib had a tumor response, compared to 0 of 5 without the mutation (the remaining 28 patients did not undergo BRAF mutation testing). The phase 3 trial in 675 patients with metastatic or unresectable melanoma who had the BRAFV600E mutation compared vemurafenib to dacarbazine. The trial was stopped after an interim analysis showed a 63% reduction in the risk of death with vemurafenib. The confirmed response rate was 48% for vemurafenib versus 5% for dacarbazine.

The examples of pathophysiologic selection just described reflected, at least initially, tumor receptor variables that could be described as proteomic variables, but that were in many cases later identified as tumor genetic markers (EGFR and BRAF genetics, for example). In such cases the genetic marker defines a pathophysiologic effect.

When proteomic and genetic markers are used in an enrichment strategy, adequate characterization of the test for the marker is critical. An inaccurate assay will undermine an enrichment effort if the study aims to demonstrate superiority or non-inferiority of the test treatment. It is also important to gain as much information as possible about the marker-response relationship (sensitivity and specificity).

C. Genomic Strategies

To date, most genomic enrichment strategies have involved tumor genomics. Although most genomic markers (e.g., for a tumor surface property) have been linked to a pathophysiologic property, this linkage is not essential. Use of a genomic marker could instead be an empiric strategy, identifying responders without providing a pathophysiologic basis for the difference in response.

Any genetic differences that predict response must in the end have some pathophysiologic basis, but enrichment strategies to identify responsive patients could be used before recognition of a mechanism. Studies directed at tumors with specific genomic patterns that appear to predict outcome (e.g., tumors with mutations in a target gene, genome-wide expression profiles, SNP arrays) are becoming increasingly common. One difficulty is that relationships between genetic patterns and outcomes are often found only after study results are known. Markers discovered this way will have credibility problems related to the post-facto nature of the finding and will almost always need confirmation in a
prospectively planned enriched study. Such findings are of particular concern when overall study results show no effect. As a general matter, preservation of study specimens and analysis of results by various genomic markers are increasingly prevalent practices and are of great potential value, whether they lead to an enriched study following a failed trial, or to better targeted therapy after a successful trial.

Freidlin and Simon, however, have proposed a novel, prospective approach that makes use of genomic data collected on all randomized subjects during the conduct of a study:

1. Design the study as usual, but divide it into first and second halves. Prospectively allocate the overall study alpha as 0.04 for the whole population and 0.01 for a patient subset to be identified in the first half of the study.
2. Run the first half of the study and conduct unblinded data analyses, searching for a genetic predictor of response. There would be no limit to the number of such analyses conducted. A single genetic subset appearing to predict response may be identified.
3. Complete the remainder of the study, entering patients according to the original eligibility criteria (both the predicted responders and predicted non-responders) as before.
4. At the conclusion of the study, the effect in the entire study population is tested at an alpha of 0.04, and the genetically identified subset is tested only in the second half at an alpha of 0.01.
5. The study shows evidence of effectiveness if either analysis is positive. When the responder population is a small fraction of the total population, but exhibits a large response, this design can improve the chance of detecting a treatment effect. It also retains good power for the overall study if the drug is more broadly effective.

A similar approach that avoids delay while the genomic marker is being evaluated might also be considered:

1. Run the planned study, again prospectively dividing the available alpha between the overall study population and a subset to be identified later.
2. Whether the overall population analysis is positive or negative, take a random sample of the study population (50%, 33% or another fraction) and search for a subset with a genomic pattern (or, for that matter, any other subset) showing a substantial differential effect. This first analysis would be considered wholly exploratory. It would be important to have genomic assessments on a substantial portion of all subjects.
3. Examine the remaining data, considering only the genomic subset found positive in the first sample. As long as the second part was kept blind until after the first analysis and only the marker-positive subset is tested, study-wide alpha error will be preserved. Any such approach would need scrupulous attention to maintaining the blind, perhaps by using an independent group to do the genomic search.

For anti-viral drugs it can be anticipated that there will be cases in which either host or viral genomics plays a role in determining drug response. For example, genome-wide association studies have identified host IL28B gene variants associated with the likelihood of a viral response to
ribavirin/interferon regimens to treat Hepatitis C virus (HCV). This association has been 
indoctrinally substantiated in multiple studies and can distinguish patients with a low (about 25%) or 
high (about 80%) likelihood of sustained viral response,\textsuperscript{46} to a pegylated interferon plus ribavirin.

It is also well documented that HCV viral genotype determines the needed duration of therapy and 
predicts likelihood of response. Six genotypes have been identified. Genotype 1 is the most prevalent 
genotype in the United States and is more resistant to treatment than genotypes 2 and 3. Patients with 
genotype 1 treated with a pegylated interferon and ribavirin for 48 weeks had a sustained virologic 
response of 40 to 50\%. Patients with genotypes 2 and 3 had sustained responses of 80\% or more with 
only 24 weeks of treatment with pegylated interferon and ribavirin.\textsuperscript{57} Two protease inhibitors, direct-
acting antiviral agents boceprevir and telaprevir, have been shown to provide greatly improved 
sustained viral response and the potential for reduced duration of total treatment in patients with 
genotype 1 virus, when added to ribavirin/interferon.

**D. Randomized Withdrawal Studies**

In a randomized withdrawal study, patients who have an apparent response to treatment in an open-
label period or in the treatment arm of a randomized trial are randomized to continued drug treatment 
or placebo. As such trials generally involve only patients who appear to have responded, this is a 
study enriched with apparent responders, an empiric strategy. The study evaluation can be based on 
signs or symptoms during a specified interval (e.g., blood pressure or angina rate), on recurrence of a 
condition that had been absent (e.g., depression), or on the fraction of patients developing a rate or 
severity of symptoms that exceeds some specified limit (a failure criterion).

The randomized withdrawal design was proposed as a way to establish long-term effectiveness of 
drugs in settings in which long-term use of a placebo would not be acceptable (e.g., most psychiatric 
and antihypertensive drug treatments).\textsuperscript{46} Even in settings in which long-term placebo use may be 
acceptable, however, it is generally difficult to recruit patients, and drop-out rates are often high, 
posing difficult analytic problems. A randomized withdrawal design in which the study population is 
on treatment for an extended duration followed by blinded, randomized withdrawal of treatment for a 
short duration could provide evidence of prolonged effectiveness with only brief exposure to placebo. 
The design allows a patient to be removed from the study (for having reached an endpoint) when the 
condition returns at specified severity, avoiding long-term exposure to an ineffective treatment.\textsuperscript{48}

The randomized withdrawal design can also be used as an initial trial to show effectiveness when there 
is an existing population of patients in an open-label treatment setting (e.g., under an IND or as an off-
label use of an approved drug), as illustrated by the cases of nifedipine and gamma-hydroxybutyrate.

The approval of nifedipine for vasospastic angina (the first drug approved for this condition) illustrates 
the utility of this design. An open-label, historically controlled trial\textsuperscript{47} was considered inadequate to 
support approval because the natural history of vasospastic angina was not well-established. A 
randomized withdrawal design (see Figure 2) was conducted in patients already receiving the drug, 
with a primary endpoint of recurrence of severe vasospastic angina leading to study withdrawal. A 
total of 28 patients participated in the study. One-third of the patients randomized to placebo 
withdrew early, as compared to no withdrawals in patients randomized to nifedipine (see Table 2).
Figure 2: Nifedipine Randomized Withdrawal Trial in Vasospastic Angina

![Diagram of Nifedipine Randomized Withdrawal Trial]

### Table 2: Results of Nifedipine Randomized Withdrawal Study

<table>
<thead>
<tr>
<th></th>
<th>Nifedipine</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>Early withdrawal</td>
<td>0</td>
<td>5*</td>
</tr>
<tr>
<td>Early withdrawal or AMI</td>
<td>0</td>
<td>6*</td>
</tr>
</tbody>
</table>

* Statistically significant at p ≤ 0.05

Another example in which patients already using a drug were studied was gamma-hydroxybutyrate (GBH, sodium oxybate), which was approved for treatment of cataplexy on the basis of a single placebo-controlled study of conventional design and a second, small, randomized withdrawal study in 55 long-term (7 to 44 months) users randomized to 2 weeks of continued treatment with GBH or placebo. The second study produced a clinically and statistically impressive result, as shown in Table 3, and needed little time for recruitment.

### Table 3: Randomized Withdrawal Study of GHB in Cataplexy

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>Median Attacks/2 weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
</tr>
<tr>
<td>Placebo (n=29)</td>
<td>4.0</td>
</tr>
<tr>
<td>GHB (n=26)</td>
<td>1.9</td>
</tr>
</tbody>
</table>

P < 0.001

By randomizing patients to different doses, the randomized withdrawal design can also be used to obtain long-term dose-response data. For example, this design was used to demonstrate effectiveness of a single weekly dose of fluoxetine in preventing recurrence of depression; patients on 20 mg/day were randomized to placebo, fluoxetine 70 mg/week as a single dose, or continued 20 mg per day. Both fluoxetine groups were superior to placebo in reducing the rate of recurrence.
A randomized withdrawal design may also be useful in establishing the duration of benefit of long-term treatments, an issue of great interest for drugs with long-term toxicity concerns and a potential for long, post-treatment benefit. For example, the original node negative adjuvant tamoxifen breast cancer trial (NSABP B-14)\textsuperscript{48} randomized 2,892 women with node-negative, hormone receptor-positive breast cancer to receive 5 years of tamoxifen versus placebo. Tamoxifen reduced the risk of recurrence by 41% and the risk of death by 33% and was rapidly adopted as the standard of care. A second randomization was then performed in which the 1,166 patients who had received tamoxifen were assigned to receive either an additional 5 years of tamoxifen or placebo. The trial was halted by the data monitoring committee after an interim analysis revealed that continuing the study could not demonstrate an advantage for continued tamoxifen use and suggested that tamoxifen was harmful. There was a non-statistically significant trend in both disease free survival (DFS) and overall survival (OS) favoring the placebo group and an increase in new malignancies in the tamoxifen group. Similar designs have been used to examine long-term effects of bisphosphonates.

E. Studies in Non-responders or Patients Intolerant to Other Therapy

A study can be enriched by selection of patients who failed to respond to an existing drug, or who were intolerant of that drug. Although non-responders or treatment intolerants are not more likely than an unselected population to respond to or tolerate the new drug, they would generally be less likely to respond to or tolerate the existing drug, giving the test drug an enrichment advantage. Because patients in a trial sometimes respond to a drug to which they had previously failed to respond, in most cases studies in non-responders are informative for the between-drug comparison of effectiveness only if patients are randomized to both the new and failed drug (i.e., not simply placed onto the new drug in a single arm study or randomized to new drug versus placebo). This approach can also provide important information to practitioners; it is critical to know whether another member of a pharmacologic class or a member of a different class can be useful in patients who fail on a previous treatment. The approach may be useful in two settings:

- To demonstrate the effect in non-responders to previous therapy of a drug that may not be more effective overall than existing therapy in an unselected population, but that has a different responder population. A drug that acts through a mechanism different from that of existing treatments might be effective in non-responders to the existing drug.

- To efficiently demonstrate the treatment effect of a new drug that is actually moderately superior to the existing drug, but where a very large study would be needed to show superiority if the study included unselected patients, many of whom would respond to the less effective drug. For example, if the new drug response rate is 90% and the existing drug response rate is 80%, a study with 90% power to detect that 10% difference would require about 600 patients. In contrast, if only non-responders to the existing drug were randomized (20% of the patients treated with the existing drug), few would respond to the existing drug, and at least half of the patients would respond to the new drug, a difference detectable with fewer than 40 of the non-responders.

Note: In neither case would showing an advantage for the new drug in non-responders to previous therapy establish overall superiority of the new drug.
1. Studies in Non-Responders

a. Captopril

To support a claim in severe hypertension unresponsive to other agents, a study was designed to evaluate patients who had not responded to standard triple therapy (propranolol 320 mg, hydrochlorothiazide 100 mg, and hydralazine 200 mg) (there were many escape pathways for non-responding patients so that they would not be untreated). Patients who had failed triple therapy and had diastolic pressure that was severely elevated were observed for 1 to 2 weeks on the same regimen (triple therapy lead-in) and, if their diastolic pressure did not exceed a defined limit, randomized to the same standard triple therapy they had failed on, or to captopril, with a 2:1 captopril to triple therapy randomization ratio. The number of responders (diastolic pressure less than 90 mmHg or fall of at least 10 mmHg, but not to below 90 mmHg diastolic) clearly favored captopril in this difficult-to-treat population (Table 4).

<table>
<thead>
<tr>
<th>Table 4: Results of the Captopril Severe Hypertension Trial for the Group Randomized to Captopril or Triple Therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total number analyzed</strong></td>
</tr>
<tr>
<td>----------------------------</td>
</tr>
<tr>
<td>Week 4</td>
</tr>
<tr>
<td>Normalized DBP ≤ 90</td>
</tr>
<tr>
<td>Reduction in DBP ≥ 10</td>
</tr>
<tr>
<td>Week 8</td>
</tr>
<tr>
<td>Normalized (DBP ≤ 90 mmHg)</td>
</tr>
<tr>
<td>Reduction in DBP &gt; 10 mmHg</td>
</tr>
</tbody>
</table>

Note: approximately 25% of the triple therapy non-responders did respond to the previously failed therapy in the new trial. This finding reinforces the need for randomization to the new and reportedly failed therapy in a study in non-responders. A study that had merely switched patients from the failed therapy to the new one and found 25% responders might have been interpreted as showing an effect of the new drug in the non-responders to prior therapy when, in fact, it would not have demonstrated that.

b. Clozapine

Clozapine is an antipsychotic agent associated with serious toxicity, a greater than 1% rate of potentially fatal agranulocytosis. For clozapine to be approved, it was crucial to show that it offered a clear advantage over safer alternatives. To show this, a study was conducted in hospitalized schizophrenic patients with a history of poor response to neuroleptics who, in addition, had failed to respond to 6 weeks of treatment with haloperidol. These patients were randomized to 4 weeks of treatment with clozapine or chlorpromazine plus benztropine. The results showed a striking advantage for clozapine on Clinical Global Impression (CGI) and British Psychological Rating Scale (BPRS) standard measures in antipsychotic drug trials (see Table 5). Despite its serious risk, clozapine was approved for use in patients not responding to other anti-psychotic agents.
Table 5: Results of Clozapine Study in Non-responders to Standard Psychotropic Agents

<table>
<thead>
<tr>
<th>Measure</th>
<th>Clozapine</th>
<th>Chlorpromazine</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGI (decrease &gt; 1)</td>
<td>71</td>
<td>37</td>
</tr>
<tr>
<td>BPRS Items (decrease &gt; 1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concept disorganization</td>
<td>60</td>
<td>39*</td>
</tr>
<tr>
<td>Suspiciousness</td>
<td>64</td>
<td>42*</td>
</tr>
<tr>
<td>Hallucinations</td>
<td>59</td>
<td>51</td>
</tr>
<tr>
<td>Thought content</td>
<td>15</td>
<td>2*</td>
</tr>
<tr>
<td>CGI and BPRS</td>
<td>15</td>
<td>2*</td>
</tr>
</tbody>
</table>

* p < 0.05

CGI = Clinical Global Impression
BPRS = British Psychological Rating Scale

c. Rofecoxib

It is widely believed that individual patients respond differently to different NSAIDs. To examine this belief, a controlled trial was conducted in which osteoarthritis patients identified as non-responders to celecoxib were randomized to celecoxib or rofecoxib. In fact, there was no observed difference (Figure 3).

Figure 3: Study Comparing Rofecoxib to Celecoxib in Celecoxib Non-Responders

Note: there was considerable and prompt improvement in pain in both groups. A baseline-controlled, single-arm trial of rofecoxib would have led to a clearly erroneous conclusion, and even a placebo-controlled trial of rofecoxib in this population might have shown an effect that would have been incorrectly interpreted as an effect in celecoxib non-responders.
2. Study in Intolerants: Angiotensin Receptor Blockers (ARBs) in People Who Cough on Lisinopril

Studies of the tolerability of a new drug in people who do not tolerate a previous treatment are also informative and efficient. Comparative studies in an unselected population could provide some information on relative tolerability, but a very large study would be needed to show small differences. For example, if the true rates of cough for an angiotensin converting enzyme inhibitor (ACEI) and an angiotensin II receptor antagonist (ARB) were 5% and 1%, respectively, a study with 90% power to show a difference in an unselected population would need about 800 patients. In contrast, a study in patients known to cough on ACEIs would need fewer than 20 patients, if, for example, the cough rate were > 90% in the ACEI arm and 20% in the ARB arm.

This approach was used in a study of 84 elderly hypertensive patients with a history of coughing on an ACEI. They were then withdrawn from their ACEI and given 8 weeks of the ACEI lisinopril, which had to cause at least moderate coughing for patients to continue on the study. Lisinopril was then withdrawn for 4 weeks and coughing had to disappear. The patients were then randomized to losartan 50 mg, lisinopril 10 mg, or metolazone (diuretic active control that does not induce coughing) for 10 weeks. The study achieved a very persuasive result with this small population (see Table 6).

Table 6: Comparison of Coughing Rates with ARB, ACEI, and an Additional Active (Non-Cough-Inducing) Control

<table>
<thead>
<tr>
<th></th>
<th>Lisinopril</th>
<th>Losartan</th>
<th>Metolazone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any Cough</td>
<td>N = 28</td>
<td>N = 28</td>
<td>N = 28</td>
</tr>
<tr>
<td></td>
<td>97%</td>
<td>18%</td>
<td>21%</td>
</tr>
<tr>
<td>P</td>
<td>&lt; 0.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

VI. ENRICHMENT STUDY DESIGN AND OTHER CONSIDERATIONS

A. General Considerations

In general, enrichment studies should incorporate the established principles of well-controlled studies, controlling bias (randomization and blinding), and studywise type I error (see section VI.C).

An enrichment design should be explicitly described in the protocol and study report and should fully detail the enrichment maneuvers and their impact on interpretation of results. For example, if only half of patients screened meet the selection criterion, the implications of this finding on expected response rate in the overall population and generalizability of the results should be evaluated. Such descriptions are particularly important for trials in which high-risk patients (prognostic enrichment) and probable responders (predictive enrichment) have been selected, when the description is critical to knowing to which patients the results apply.
1. Performance Characteristics of a Screening Strategy for Selecting Patients

Some prognostic and predictive enrichment strategies depend heavily on a screening measurement for selecting the enriched population. For less well-established measures, such as use of proteomic and genomic markers to identify responder populations, there are generally more performance uncertainties than for phenotypically well-known prognostic factors such as blood pressure or cholesterol. It is critical to understand, to the extent possible, the accuracy of the test used for enrichment, as well as the performance characteristics (sensitivity, specificity) of the measurement used. If a selection criterion does not accurately distinguish patients, the effect of enrichment is weakened. If a classifier is overly inclusive (lacks specificity), the estimated difference between the effect in the enriched and non-enriched populations will be attenuated, defeating the goal of the enrichment strategy. If a classifier is overly exclusive (i.e., lacks sensitivity), patients who could benefit from the drug will not be studied, and study subjects will be needlessly difficult to find. In a non-inferiority study, an inaccurate classifier would bias the study toward a finding of no difference.

Apart from the adverse impact of a poor classifier on a study, the performance characteristics of the classifier remain important if the classifier is to be used in selection of patients after approval of the drug. In some cases, the performance characteristics of a classifier will be well-documented in earlier studies, but the enriched trial may itself constitute an important source of information about the sensitivity and specificity of the classifier. For example, by including patients above and below a presumed classifier cut-off and examining results by classifier status, it is possible to assess the quantitative relation of the enrichment factor to response (e.g., the relationship of a receptor level such as HER2/neu to response). This relationship can be explored even if the primary analysis defines the marker-positive population of interest on the basis of a pre-specified cut-off. The classifier is pertinent not only to selection of patients for the study, but to selection of patients for treatment once the drug is marketed. Planned interim evaluations could be used to narrow selection criteria, especially if there were a measure of effect that could be assessed early (e.g., short-term pharmacologic effect).

2. When Should a Classifier Be Developed and Characterized?

During a development program using an enrichment strategy, it is not always clear when the performance of a classifier should be characterized, nor how well the classifier and assay methodologies must perform if they do in fact define a population with increased response. Answers to these questions depend in part on when during development it becomes clear that an enrichment strategy should be used and how critical enrichment will be to successful drug development and approval. When there is not strong pre-existing information defining a responder subset, early exploratory studies that include patients over a broad range of values of the enrichment factor can be used to develop criteria for classifying patients for subsequent enrichment studies. If the need for classifier-based patient selection does not become apparent until late in development, it is also possible to first characterize a classifier in a phase 3 trial. When the classifier is not based on an intrinsically binary measurement, the trial could explore a range of threshold values for a classifier that would be used to identify patients for the primary analysis, but this must be planned in advance to control the Type I error rate of the study. With few exceptions, the enrichment characteristics used in confirmatory studies should be measured at baseline, and patients who are classified as having, or not having, the predictive marker should be stratified and randomly assigned to treatments if both subgroups of patients are to be included. If the classifier is known only after randomization, but is a
baseline characteristic, randomization should be effective even without stratification, as long as the sample sizes in the treatment groups within each marker-defined subgroup are large enough to balance important prognostic baseline factors.\textsuperscript{43}

**B. Which Populations to Study**

As will be described in more detail below, trials can be designed (1) to include only patients with the enrichment factor or (2) to include patients with and without the enrichment factor, but with an intent to analyze only the patients with the enrichment factor as one of the primary study hypotheses. Studies including both populations need not include a wholly unselected population of patients with the disease to be treated, but can designate separate sample sizes for patients with and without the enrichment characteristic to collect sufficient information to demonstrate effectiveness in the enriched subgroup and also to allow a reasonable estimate of effect in the non-enriched group. Many design alternatives have been discussed in the literature.\textsuperscript{50,51}

A critical question in all settings in which enrichment is used is therefore the extent to which the enrichment marker-negative population should be studied, an issue that may bear importantly on how a drug would be labeled. In some cases, study of the general population (one including the marker-negative population) would not be expected. For example, if prognostic enrichment is used to ensure that there are sufficient events to make a trial feasible, even if it is thought that the treatment effect would also be present in the lower-risk population without the marker (but at a lower absolute effect size), it may not be possible to design a trial that includes a significant fraction of the marker-negative population without greatly increasing the sample size needed – a strategy that may make the trial impractical and defeat a major purpose of prognostic enrichment. The Clinical Studies section of labeling and sometimes the Indications section would identify the population studied and comment, as appropriate, on use in other populations. Advice regarding use in the untested marker-negative population would depend on the particular circumstances. For example, the presence of significant toxicity could lead to doubts about the advisability of using the drug in the lower-risk population. The heterogeneity-reducing factors discussed in Section III would not ordinarily call for study of the population lacking the enrichment factor (e.g., poor compliers).

It is principally in the area of predictive enrichment, especially predictive enrichment using a pathophysiological or genomic marker, that the question of studying the population without the enrichment factor is most germane. Experience suggests that the selected enrichment factors often do not prove to precisely dichotomize patients into subpopulations that will and will not respond, so that it is usually desirable to obtain some information on the marker-negative population to assess performance of the factor. However, even an imperfectly characterized predictive marker can greatly increase the power and likelihood of success of a study. Moreover, in treating serious and life-threatening illnesses, especially when there are alternative treatments, using the test treatment in patients thought unlikely to respond raises critical ethical issues.

Efforts to use predictive enrichment thus offer a number of design choices. The study designs illustrated below are fixed sample size designs that can be used with predictive enrichment strategies (also see section VI.D below concerning adaptive enrichment and non-fixed sample size). The examples all describe trials intended to show superiority of the test treatment to a control (placebo, standard of care), but non-inferiority studies would present similar issues.
1. **Studying Marker-Positive Patients Only**

A study randomizing only marker-positive patients is shown in Figure 4. Because a study that uses a marker-positive only population will provide no direct information about the marker-negative population, its use should generally be limited to situations in which information about the marker-negative population is not needed or is not feasible given the objectives of the study. For example, if it appears clear, based on mechanistic, pre-clinical, or early clinical data, that the marker-negative patients will have no or minimal response or would be exposed to unreasonable risk, inclusion of the marker-negative patients would, in most cases, not be justified.

**Figure 4**

Prospective, Screened - no possible effect in the marker-negative group

---

The study shown in figure 4 would support an effectiveness claim for the enriched population, but it would overstate the actual effectiveness for an unselected population, so that the fraction of patients with the marker would be important information. The study would provide no new clinical evidence with respect to the marker negative population and would not further characterize the predictiveness of the marker because there would be no ability to compare effectiveness. Because there will be no effectiveness or safety information on the enrichment-marker-negative patients, it is implicit in this approach that the selection process would be fully described in the labeling and that in clinical practice, in most cases, all patients would be tested for the enrichment marker before exposure. Moreover, because assessment of marker status is critically important to effectiveness in patients in clinical practice, it would generally be expected that the enrichment marker would be measured after
approval using an established, FDA approved, laboratory test explicitly labeled for this purpose as a
companion diagnostic.\(^5\)

2. Studying Both Marker Positive and Negative Patients

We encourage inclusion of some predictive marker-negative patients in most trials intended to provide
primary effectiveness support, unless it has been well established in earlier studies that the marker-
negative patients do not respond, or there is a strong mechanistic rationale that makes it clear that they
will not respond. In general, the greater the uncertainty about the marker cut-off and responsiveness
of marker-negative patients, the more important it is to include a reasonable sample of marker-
negative patients. When substantial incentive exists to use the drug in the marker-negative population
(e.g., for serious diseases with few alternative therapies), characterization of the response in the
marker-negative population is more important, especially if the drug has important safety concerns.

There are two cases to consider in studies that include marker-positive and marker negative patients:
(1) when the marker can be assessed before randomization and (2) when the marker can be assessed
only after randomization. Figures 5A and 5B provide sample study designs for these two cases.\(^5\)

**Figure 5A**

Patients Prospectively Screened and Stratified - where there is possible effect in the marker-negative group

In the first case, marker status is determined for all patients, and randomization is stratified by marker
status. The primary study objective would usually be a statistically rigorous demonstration of the
treatment effect in the marker-positive patients, and the study would therefore be powered for the
effect in that group. The size of the marker-negative group would be determined separately (i.e., it
would not be necessary to randomize all marker-negative patients). Because the treatment effect
would be expected to be much smaller (if there were any effect) in the marker-negative population, the
size of the marker-negative population would usually be too small to give a definitive answer on the
effect in that population; however, it would provide at least some estimate of the effect in that
population. The design could also provide an overall risk–benefit assessment for the drug in a general
population, which would be advantageous if some exposure in marker-negatives is anticipated in
clinical practice (e.g., because the test is not widely available). When there is substantial uncertainty about whether a marker is predictive, (i.e., can select a population in which treatment is effective), the primary endpoint could be the effect in the overall population, or study alpha could be divided between the two endpoints (overall population and marker-positive population). \(^{52}\)

**Figure 5B**

All patients randomized because marker cannot be assessed prior to randomization

The second case (Fig 5B) is one in which a drug that is expected to be effective only in the marker-positive subset (e.g., only in patients with a sensitive organism) must nonetheless be given to all patients because the marker result is not available at randomization. It would still be appropriate to have the primary study outcome be the effect in the marker-positive subset, but the risk–benefit assessment would reflect results in the entire population (i.e., the population that would be exposed to treatment). In these cases, when marker test results for patients using the drug will not be known before drug administration and if no patient management decisions will be made on the basis of the test result (e.g., decisions to discontinue treatment in marker-negative patients), FDA approval of the test prior to approval of the drug is not needed.

3. *Studies in Alternative Therapy Non-Responders to Support Claim of Effectiveness in Non-Responders*

Knowing that a drug is effective in patients who have not responded to, or who have responded inadequately, to other therapy is an important clinical finding. The lack of apparent response to prior treatment, however, is not a reliable predictor of how the patient would respond in a new setting (i.e., in a new randomized trial). It is therefore usually essential, in order to demonstrate effectiveness of the new drug in non-responders, to randomize patients to both the failed drug and the test drug (Figure 6) because in most cases it cannot be assumed that the non-responders to the other therapy outside the trial will not respond to it in the setting of the new trial, as the examples in Section V illustrate. There are cases in which previous non-response will almost certainly be repeated (e.g., non-response to an oncologic treatment) so that re-randomization to the failed treatment may not be necessary. Great care should be taken in reaching this conclusion, however, as the altered conditions in a new study might affect the response to an apparently failed prior treatment (e.g., by improving compliance). It would, therefore, be important to ensure that the failed therapy was properly administered. Finally, it should
be appreciated that re-randomization to a failed treatment could pose significant ethical problems in a serious or life-threatening condition.

Figure 6

Studies in Non-Responders

standard drug

standard drug

non-responder

new drug

Figure 7

Studies in Intolerants

Adverse Drug Reaction, drug 1

drug 1

drug 2

4. Studies in Patients Intolerant of a Prior Treatment

As in the case of studies of non-responders, knowing that patients who have experienced important adverse effects on an available therapy do not have such effects on a new agent is a clinically important finding. In this case, too (as in studies of non-responders), it is critical to randomize to the poorly tolerated previous therapy and the new drug (Figure 7) to reach a conclusion that the new drug has superior safety, because the adverse effects do not always reappear when a treatment is repeated. For the same reason, it would also be advisable to include a placebo group to be certain that the side effect was indeed reproduced in the previous treatment group. This study design is not feasible if the adverse effect was dangerous to the patient. In that case, showing that a new drug lacks the side effect is probably feasible only in a trial in a previously untreated population, unless it is possible to know very confidently that the adverse effect would have recurred in the previously treated population that did not tolerate the initial treatment (i.e., use of a historically controlled design).
C. Type I Error Rate Control for Enriched Study Subpopulations

Generally, even if patients both with and without an enrichment characteristic are studied, the primary endpoint is expected to be driven by the result in the enriched subgroup. In some enrichment designs that enroll patients both with and without the enrichment characteristic, the type-I error rate for the study can be shared between a test conducted using only the enriched subpopulation and a test conducted using the entire population. This alpha allocation scheme allows for assessment of the treatment effect in the entire entered population when there may be some effect in the patients without the enrichment characteristic while also allowing the assessment in the enriched subgroup. Determining the required sample size that will provide reasonable power to test the different hypotheses while controlling type-I error (usually including a pre-specified order of testing or a multiple testing procedure allowing testing of both hypotheses) is challenging.

D. Adaptive Enrichment

Although an enrichment characteristic should almost always be specified before a study begins, certain adaptive designs can use enrichment strategies that identify predictive markers during the course of the study. Specifically, entry criteria or sample size can be modified for later stages of a trial if factors can be identified that increase event rate or treatment response (e.g., discovery that the enrichment factor has a greater impact on response than anticipated or that the patients without the enrichment factor have a very low response or safety concern). Such changes will need appropriate type-I error rate control to account for interim, unblinded analyses of the accumulating data as well as type-I error allocation if there were analyses of multiple subgroups, but adjustment may not be needed for the enrichment if all randomized patients are included. However, the issue of whether the statistical testing results obtained by such an adaptive enrichment strategy are reproducible needs to be addressed. For a full discussion of adaptive designs see FDA’s draft Adaptive Design Guidance.

During a trial, information may be obtained from accruing patients that could inform a pre-planned opportunity for modification of the trial’s size or design (e.g., to be certain that the enriched population is of adequate size). In a trial that enrolls patients with and without enrichment factors, the relevant sample size would generally be based on the number of subjects entered with the enrichment characteristic, so that sample size adjustment based on a blinded analysis of the proportion of marker-positive patients would usually be recommended. It is possible for enrichment to occur to a greater degree in the later stages of a trial. For example, a broader population might be studied initially with altered entry criteria implemented in later stages of the trial based on interim results. Such a procedure should be described in the study protocol, and the analysis should adequately account for the changes. A number of examples of such trials have been reported in the literature, but there is little experience as yet with their use in drug development.

Although there has been little practical experience with enriched study designs whose sample size changes after the start of the study, or where other changes in the design are pre-planned to be based on accrued information during a trial, a number of adaptive designs seem potentially applicable:

1. In a study that includes both marker-positive and negative patients, an interim look could reveal, either on an early endpoint (e.g., imaging or PD biomarker or tumor response rate) or
later endpoint (e.g., progression-free survival) that the marker-negative population has a much lower response than the marker-positive group. Additional enrollment of marker-negative patients could be reduced or stopped entirely.

2. The design could be used to obtain more precise information on the performance characteristics of a classifier, for example, by examining an early endpoint using several different classifier cut-offs to determine the optimal cut-off value. If a cut-off proved too low (i.e., selected a population with few responses), it could be raised. Such plans would need to be carefully specified in advance as part of the phase 3 study, but could also be examined in exploratory studies.

3. When the optimal marker cutpoint is not well known prior to the study, use of several different cut-offs creates multiplicity issues; cut-offs should be pre-specified or corrected for the multiple choices. Jiang, Friedlin, and Simon have suggested a study design that tests the treatment effect in the overall population and in a subset defined by a marker cut-off point when the cut-off point is determined after the study is over using a statistical procedure that controls for multiplicity.55

4. Interim analyses could suggest changing entry criteria to emphasize a better-responding subgroup; if all randomized patients were included in the final analysis, such a step would not appear to need alpha adjustment, although describing results could be challenging.53 It should be appreciated, however, that unplanned adjustments of entry criteria based on early data can have unexpected effects that may not enhance the ability of the study to show an effect.

5. Sample size planning in these designs can be difficult, because such designs are generally used when there is uncertainty about the prevalence of a marker, its predictiveness, and what sample size or entry criteria adjustments are contemplated. In some more complex situations, use of statistical simulations may help calculate study power and the impact of design choices.

E. Cautions in Interpretation

Any use of an enrichment design should be explicit in the protocol and study report and should fully detail the rationale, specific enrichment maneuvers, and their impact on interpretation of results. For example, if only half of the patients screened passed the entry test, that should be noted, and the impact of this selection in terms of the expected response rate in the overall population and on the generalizability of the results should be evaluated. The importance of such descriptions is obvious for trials in which high-risk patients (prognostic enrichment) and probable responders (predictive enrichment) have been selected, where the description is critical to knowing to which patients the results apply, but such descriptions are important for all types of enrichment studies. Given the potentially complex interpretation of studies using enrichment designs, we strongly recommend early discussions with the Agency on plans to use them.

When enrichment depends on a proteomic or genetic test, particularly if the test is intended for use in practice to identify patients to be treated, the analytical validity of the test is critical. In addition to assay validity, for any marker used to select patients, even a familiar one, its sensitivity and specificity and positive and negative predictive values should be well characterized. To the extent an enrichment strategy successfully identifies patients with high event rates or high response rates and leads to a successful study, study results could be said to speak for themselves (i.e., the randomized trial did
show an effect; the event rate was high enough) and certainly support the effectiveness of the drug in the population studied. Again, however, the enrichment strategies should be clearly described to indicate how the drug is to be used and to whom the results might apply (groups of patients that do and do not benefit).

Selection of the optimal predictive enrichment study design, specifically, whether to include both marker-positive and negative patients, and whether to introduce adaptive elements can be difficult to determine in the face of uncertainty about the properties of the enrichment marker. Many publications have addressed these issues. One conclusion is that the greater the uncertainty regarding the marker cut-off and responsiveness of marker-negative patients, the more sense it makes to include a reasonable sample of marker-negative patients, perhaps using an adaptive design to exclude such patients if they are seen not to respond. In general, especially when marker prevalence in the population studied is relatively low, it would generally be sensible to stratify by marker status.

VII. ENRICHMENT – REGULATORY ISSUES

A. Summary – The Decision to Use an Enrichment Strategy

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). FDA’s interests also include the adequacy of the study (Will it successfully assess effectiveness in a defined population and, in so doing, support marketing approval?) as well as the degree to which study findings can be described in drug labeling.

As discussed above, there are many reasons to use such designs, including an enhanced benefit–risk relationship if a population with an increased likelihood of response can be identified, and efficiency in drug development, as smaller studies can often be used to demonstrate effectiveness. There are, however, two critical considerations when contemplating the use of enrichment designs.

1. Can the Enrichment Strategy Be Used to Identify the Patients to Whom the Drug Should Be Given?

When patients with an increased likelihood of response can be defined before treatment by a predictive marker (a pathophysiologic or genomic characteristic, a short-term screen such as response to a test dose), a straightforward method is available for selecting patients for treatment. In contrast, some empiric strategies that provide predictive enrichment (e.g., studying known responders in a conventional study or in a randomized withdrawal study) can efficiently establish the effectiveness of a drug in a subset of the population, but provide no way for prescribers to prospectively identify patients with a greater likelihood of response, or predict the magnitude of response in an unselected patient. Although this type of untargeted treatment may seem troubling (treatment of many to attain a response in only some), the reality is that this is generally the case with treatments that are approved on the basis of conventional studies in a non-enriched population, where there is typically a wide range of responses, including no effect at all, or even harm in some cases. However, it needs to be understood and made clear that the magnitude and/or likelihood of a treatment response for an unselected patient could be substantially less than the mean response observed in a clinical trial that
employed an empiric enrichment strategy. When the prescriber is reasonably able to gauge the
effectiveness of a drug in an individual patient (e.g., pain is relieved, cholesterol is reduced), the pre-
treatment ability to predict the likelihood of a drug response with accuracy may not be as critical.

In some cases, enrichment cannot be used to select patients for study because the enrichment factor is
not known until after the treatment is initiated, but is used to identify the subset of the studied
population to be analyzed – for example, the patients with a sensitive organism in studies of
antimicrobial drugs. Again, the subset analysis documents effectiveness, but the treated population, at
least initially, will be the unselected patients (i.e., a larger group than the population of potential
responders). Such situations are unavoidable, however, if treatment is urgent and must be initiated
before the enrichment measure is available.

Finally, the quality (sensitivity and specificity) of the enrichment strategy may not be critical to
knowing that a drug has an effect in the study subjects (if the selected study population shows an
effect, the drug was in fact effective), but it is important to therapeutic use, as it is plainly undesirable
to inadvertently include non-responders and exclude potential responders because the test has poor
precision or because cut-off points were poorly selected.

The above problems noted, however, if the enrichment strategy allows a drug of value to be developed
and shown to be effective when disease and response variability would make non-enriched studies
unable or unlikely to succeed, there is clearly an important gain from such strategies. Labeling will
reflect limitations and concerns, but it seems clear that a drug shown effective in an enriched study
should be available even if the responder population is not identified as precisely as would be
desirable.

2. Might the Drug Be Useful in a Broader Population Than Was Studied?

The data that should be obtained for the marker-negative patients will be considered below, but it can
be anticipated that less information will be available about them and there will be greater uncertainty
as to their response to the treatment. Studies in unselected patients (i.e., a non-enriched population),
the typical basis for drug approval, simply ignore the question of identifying responders and lead to
treatment of many patients who will not benefit. There would thus seem to be a gain from a process
that seeks to establish the characteristics that predict a drug response, rather than ignoring the varied
responses and overcoming them by simply increasing sample sizes.

In general, then, FDA is prepared to approve drugs studied primarily or even solely in enriched
populations and will seek to ensure truthful labeling that does not overstate either the likelihood of a
response or the predictiveness of the enrichment factor. But the extent of data that should be available
on the non-enriched subgroup should always be considered. Postmarket commitments or requirements
may be requested to better define the full extent of a drug’s effect (including efficacy and safety
studies and trials in a broader population).

B. Data That Should Be Obtained for the Marker-Negative Patients

Well-controlled enrichment studies, if successful, provide clear evidence of effectiveness in the
population studied. In many cases, however, questions will remain as to how to identify the patients to
which the data apply and the magnitude of effect in the marker-negative patients. In general, the heterogeneity-reducing efforts raise few problems of this kind, but the prognostic and, especially, the predictive enrichment strategies do raise them, and the remedies before or even after approval are not always clear and may be circumstance-specific.

For studies of serious or irreversible endpoints, showing an effect in a high-risk population (e.g., high blood pressure, high LDL with a history of MI, severe CHF) has, historically, been a prelude to later demonstrations of effects in lower-risk patients, but the effect size has sometimes been smaller (CHF), the endpoints have sometimes changed (mortality versus composite in the lower risk patients), and consequently, benefit–risk considerations may change. FDA has generally accepted the results from prognostically enriched studies, approved a claim based on the observed effect, and described the study, including the patient population, in the Clinical Studies section of labeling, with any enrichment selection criteria noted. The specific patient population studied has sometimes, but certainly not always, been given as the indicated population in the Indications section of labeling in addition to its description in the Clinical Studies section.

The most challenging situation is a finding of benefit in a clinically, demographically, pathophysiologically, proteomically, or genetically selected population used for predictive enrichment, because in that case there is inevitably the question of how sure we are that other patients could not benefit, even if the benefit were smaller. Ideally, therefore, there will be at least some data on the marker-negative population; the study designs in Figure 5 illustrate this approach, but it must be appreciated that a study, if sized to show an effect in the enriched population, will have relatively little capacity to detect or rule out the anticipated substantially smaller effect in the marker-negative population. Nevertheless, the design does provide some information on that population. It could, for example, show that the estimated treatment effect is not likely to be as large in the marker-negative population, but could show a trend suggesting a smaller effect.

In deciding how much information should be available on the marker-negative population, both before and after approval, risk–benefit assessments of several kinds should be considered. When the treatment is a critical advance for the enriched group, it would generally be unreasonable to delay approval for the enriched group, even if few data on the group without the enrichment factor were available and even if some off-label use were anticipated despite appropriate labeling (although how much off-label use would be unacceptable would depend on the expected risk of the treatment). For less important benefits, factors considered would include the expected risks to the marker-negative population from the use of the drug (related to its observed toxicity), the relative size of the marker-positive and marker-negative populations, and how convincing it is that there is no useful treatment effect in the marker-negative population, so that labeling and other information might make off-label use unlikely. If the risks of the drug are substantial, FDA will want greater confidence that patients who will not benefit from the drug will not be treated in the course of patient care. Conversely, if the risks appear low, less assurance regarding the marker-negative population would be sufficient. Ironically, a very important medical benefit (e.g., survival or prevention of significant disability), one that realistically and ethically cannot be delayed, could raise the greatest concerns because there would be great desire to use the drug in the marker-negative population. During development it is important to discuss with FDA review staff how much information should be available before drug approval.
Note: The information on the marker-negative group could come from all of the studies in the clinical program, including both studies in an unselected population and studies in the marker-negative population. The information that will describe the effect in the marker-negative group includes direct empirical evidence of clinical results in patients, pathophysiologic information, or combinations of various kinds of information.

A number of considerations would support collection of less (sometimes even no) information on the non-enrichment-factor population (again, this should always be discussed with FDA review staff):

- A clear pathophysiologic basis for concluding that the non-enriched population will not respond (e.g., because they lack the molecular target of the drug or because they cannot convert a pro-drug to its active metabolite); this could be supported by pre-clinical or clinical pharmacologic and biomarker studies
- Early clinical studies that show very marked difference in response between the enrichment and non-enrichment populations.

Important toxicity such that use in a less responsive population, even if a small effect could be present, would not be attractive.

### D. Labeling

The use of enrichment designs will often have implications for labeling, especially the Indications and Usage, Dosage and Administration, and Clinical Studies sections. As noted above, prognostic enrichment will be described in Clinical Studies and has sometimes led to a description of the studied population in Indications. Use of predictive enrichment will usually lead to an indication directed at the predictive enrichment population, often with recommended testing, and a description of the selection in clinical studies. If no marker-negative patients are studied, it will be difficult to describe the effect of enrichment fully. In some cases, however, earlier data may show clearly that the marker-negative patients cannot respond. In general, significant toxicity would lead to stronger direction toward the marker-positive patients and strong directions to avoid the marker-negative patients. When an enrichment design results in the need for an approved or cleared companion diagnostic device, coordinated labeling for the therapeutic product and the diagnostic device should be provided.\textsuperscript{55,56}
APPENDIX

Additional Guidance Related to Enrichment

A number of Agency guidances have been issued that provide additional and related information about clinical trial designs (including enrichment designs) and demonstrating effectiveness. See especially the following draft and final guidances. Once finalized the draft guidances will represent FDA’s thinking on their respective topics.


- FDA’s draft guidance *Clinical Pharmacogenomics: Premarketing Evaluation in Early Phase Clinical Studies* focuses particularly on use and evaluation of genomic strategies in early drug development, and highlights identification of enrichment options for later trials.

- FDA’s draft guidance for industry and FDA staff on *In Vitro Companion Diagnostic Devices*\(^5^6\) defines IVD (in vitro diagnostic) companion diagnostic devices that are essential for the safe and effective use of their corresponding therapeutic products. The draft guidance describes the Agency’s policies for approval and clearance and for labeling companion diagnostics, contemporaneously with approval and labeling of the therapeutic product.

- FDA’s draft guidance *Adaptive Design Clinical Trials for Drugs and Biologics*\(^5^2\) considers the case of enrichment approaches introduced only after randomization and based on interim evaluations. Such a retrospective finding would have to be carefully implemented and highly compelling to be accepted without further study.

- FDA’s guidance *Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products*\(^5^7\) describes the amount and type of evidence needed to demonstrate effectiveness, and is applicable to studies using enrichment designs.
REFERENCES


2. Veterans Administration Cooperative Study Group on Antihypertensive Agents. Effects of Treatment on Morbidity in Hypertension: Results in Patients with Diastolic Blood Pressures Averaging 115 through 129 mm Hg. JAMA 1967; 202: 1028-1034.


37. US Food and Drug Administration Drug Approval Summaries: Imatinib Mesylate, Mesna Tablets, and Zoledronic Acid; The Oncologist 2002; 7: 393-400.


Contains Nonbinding Recommendations
Draft – Not for Implementation


56. FDA draft guidance for industry: In Vitro Companion Diagnostic Devices.

57. FDA draft guidance for industry: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products.