Why simply increasing sample size for obtaining greater trial power often does not work:

caveats you should know

Stephen Brannan, M.D.
Chief Medical Officer
KARUNA PHARMACEUTICALS
Basic thoughts on maximizing power for your CNS study

• A quick google search shows that the basic elements needed for a power calculation involves sample size, anticipated Rx difference (effect size), and variance

• Quick examination shows that if you hold the last 2 elements as unchanged, that increasing the sample size will increase your power

• So why won’t this work every time one runs a trial?
Assumptions: better check them closely

• The problem is that very seldom will the anticipated Rx effect and even more so, the variance remain equal as one increases ones sample size to any significant degree

• In a recent antipsychotic program “placebo response” increased from 7.4 to 10.3 to 15.1 in successive trials

• There is also evidence in CNS that drug effects also may decrease over time, which is thought to possibly be related to a large % of a population being “successfully treated” (such as with SSRIs or atypical antipsychotics)

• Though this phenomena may be worse in CNS, it is also seen in part in other therapeutic areas as well
  • “A higher-than-expected response in the control group of the cholesterol study pushed down shares, despite a positive outcome”
  • Arif Khan has reported on the increase in the magnitude of placebo response over time in studies of epilepsy, hypertension, and diabetes
Examples of when the simplistic approach of just increasing size did not work?

• Does anyone remember the US Reboxetine trials that failed?

• Biogen AD trial- "We did see more variability on the primary endpoint than assumed when we did the original sample size estimations," explained Sandrock to ..."So we've decided to increase the sample size to maintain 90% power."

• Recent programs in “Cognition” and “Mood Disorder” in the following slides
Phase 2 and Phase 3 trials from a Cognitive Enhancer

Phase 3 trials represent average trial scores for both studies
Phase 2 and Phase 3 trials from a Mood Disorder program

Phase 3 trials represent average trial scores for both phases
Caveats worth considering

How they can lead to difficulties
Some caveats

• Non adherence
  • Adds variability and increases with the size of a study

• Increasing the number of sites (and the regions they are located in)
  • Often shows regional variations
  • There is a limit to the number of “high quality sites”
    • And of high quality raters/ratings
    • Higher sample size was associated with a higher placebo response
    • Role of Professional subjects

• Unintended consequences from “minor” changes in protocols

• Problems related to recruiting a large number of subjects in a “timely” fashion
  • Subject characteristics are not stable
    • the 4th quarter of subjects enrolled show especially high placebo response*

*McCann DJ et al JCP 2015
Trial subject compliance by PK showed a number of patterns (including about 5% with apparently zero compliance)

*Plasma concentrations during Phase 3 Studies; examples of individual subjects patterns*

- 74 (17%) and 89 (18%) of active-treatment subjects in Studies -015 and -16, respectively, had a post-Day 1 concentration of “0” at some time in the study
- 19 (4%) and 26 (5%) of active-treatment subjects in Studies -015 and -016, respectively, had a concentration of “0” at every sampled time in the study
Nonadherence to study medication increases with study size

### TABLE 1. Medication Nonadherence in AstraZeneca Psychiatry Studies, 2001 to 2011

<table>
<thead>
<tr>
<th>Indication</th>
<th>No. Subjects Receiving Active Treatment</th>
<th>Name of Drug Under Study</th>
<th>ClinicalTrials.gov Identifier (NCT no)</th>
<th>Subjects With Any PK Sample BLQ (%)</th>
<th>Subjects With &gt; Half of PK Samples BLQ (%)</th>
<th>Subjects With all PK Samples BLQ (%)</th>
<th>Nonadherence Calculated From Pill Counts (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDD</td>
<td>39</td>
<td>AZD2066*</td>
<td>NCT01145755</td>
<td>12.8</td>
<td>12.8</td>
<td>2.6</td>
<td>NC</td>
</tr>
<tr>
<td>MDD</td>
<td>91</td>
<td>AZD7268†</td>
<td>NCT01020799</td>
<td>28.6</td>
<td>16.5</td>
<td>12.1</td>
<td>2.9</td>
</tr>
<tr>
<td>MDD</td>
<td>100</td>
<td>AZD5077† (quetiapine)</td>
<td>NCT00326144</td>
<td>26.0‡</td>
<td>26.0‡</td>
<td>26.0‡</td>
<td>2.2</td>
</tr>
<tr>
<td>GAD</td>
<td>169</td>
<td>AZD7325‡</td>
<td>NCT00807937</td>
<td>33.0</td>
<td>22.3</td>
<td>16.3</td>
<td>2.7</td>
</tr>
<tr>
<td>GAD</td>
<td>309</td>
<td>AZD7325†</td>
<td>NCT00808249</td>
<td>33.7</td>
<td>21.7</td>
<td>13.6</td>
<td>5.1</td>
</tr>
<tr>
<td>CIAS</td>
<td>313</td>
<td>AZD3480§</td>
<td>NCT00528905</td>
<td>34.8</td>
<td>20.1</td>
<td>15.0</td>
<td>4.6</td>
</tr>
<tr>
<td>MDD</td>
<td>331</td>
<td>AZD5077† (quetiapine)</td>
<td>NCT00320268</td>
<td>23.3‡</td>
<td>23.3‡</td>
<td>23.3‡</td>
<td>0.0</td>
</tr>
<tr>
<td>GAD</td>
<td>413</td>
<td>AZD5077† (quetiapine)</td>
<td>NCT00329264</td>
<td>39.2‡</td>
<td>39.2‡</td>
<td>39.2‡</td>
<td>NC</td>
</tr>
</tbody>
</table>

* LQ = 1.00 nmol/L.  
† LQ = 0.5 ng/mL.  
‡ LQ = 0.05 ng/mL.  
§ LQ = 0.04 nmol/L.  
‖ Only 1 PK sample was obtained in the study.

MDD indicates major depressive disorder; GAD, generalized anxiety disorder; CIAS, cognitive impairment associated with schizophrenia; NC, not calculated; LQ, limit of quantification.
In the overall ITT population for the NCC, European sites average a gain of ~4.5 points in the placebo arms.

Least squares means (+/- SE) of change from baseline in neurocognitive MCCB score over time by treatment group - Europe.
In the overall ITT population for the NCC, non-European sites average a gain of 2-2.5 points in the placebo arms.

Least Squares Means (+/-SE) of change from baseline in the NCC score over time by treatment group - non-Europe
Sites that Recruited less than 6 subjects
Show almost no differences between Treatment Groups
Importance of monitoring early-
“Dual” scoring variance relative to BPRS interview length (n=392)

- 23% of BPRS interviews were conducted in less than 15 min
- Shorter interviews had greater variability and discordance
- In many shorter interviews, the rater asked fewer questions

Targum et al. Eu Neuropsychopharm 2015
Randomization of a trial Looking at a cognitive outcome
To accelerate enrollment, subjects could be enrolled if their “Stable” regimen included 2 Antipsychotics.

At the end of the trial, the breakdown of how many medications subjects were on was:
- Stable on One,  N= 1343(90.55%)
- Stable on Two,  N= 140 (9.44%)

Despite the request from the sites to allow this, only a little less than 10% of subjects enrolled were actually stable on 2 medications; though there was no reason to believe this group would differ from those stable on 1 Medication, post hoc analysis showed this not to be the case.
Surprisingly, the 10% of study subjects stable on 2 antipsychotics showed a very different placebo response than those on only one antipsychotic.

One Medication

>=Two Medications
A similar pattern is seen in the functional co-primary with the placebo arm doing markedly better than the drug arms.

One Medication

>= Two Medications
The balance between Speed, cost & quality in clinical trials

AN EXAMPLE OF WHAT CAN GO WRONG
Notable differences, particularly in the placebo arm, in the time windows

Rando – 8/2013 (N: 24, 23, 25)

8/2013 - 8/2014 (N: 249, 244, 235)


After 2/2015 (N: 121, 114, 117)
Similar findings are also seen in the functional co-primary
Summary

• Simply increasing the size of a trial without paying attention to any number of issues that are unintended, but nonetheless predictable, does not increase power, it leads to failed trials
  • There may be ways to mitigate some of these issues

• If not pushed past a certain size (which likely varies depending on the indication), some of these issues will be less problematic