Comparative Effectiveness Study: New Novel Treatment vs. Oral Antipsychotics in Schizophrenia
There are multiple challenges in the design of comparative effectiveness studies including selecting endpoints, comparator(s), and study population(s). These studies may be larger and longer than placebo-controlled Phase III trials, requiring substantial investment for the study sponsor. Additionally, limited comparative information may be available at the time of protocol development on which to base the decision on endpoints, comparator(s) and study population(s). The ISCTM Adaptive Design Working Group will discuss the issues associated with these decisions and explore the use of multiple adaptations within a single Phase IIIB/IV comparative effectiveness study in schizophrenia.
Presentation Agenda

• Study Rationale

• EMR Clinical Trial
  – Concept
  – Candidate Health System

• EMR Prospective Study Design

• Simulations

• Discussion
Study Rationale
# Product Development for Two Worlds...........FDA & Payers

*Meeting Requirements for Regulatory Approval and Patient Access*

<table>
<thead>
<tr>
<th>Purpose</th>
<th>FDA</th>
<th>Reimbursement by Payers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product Approval Processes</td>
<td>Patient Access</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidentiary Requirements</th>
<th>FDA</th>
<th>Reimbursement by Payers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized Controlled Trials</td>
<td></td>
<td>Real World Evidence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study Endpoints</th>
<th>FDA</th>
<th>Reimbursement by Payers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defined Regulatory Endpoints to Prove Efficacy and Safety of a Product</td>
<td></td>
<td>Generalizable Outcome Measures i.e. quality measures, cost</td>
</tr>
</tbody>
</table>
NEW VALUE PROPOSITION

- Volume based care → value based care (quality over cost)
- These programs focus on new methods of paying physicians and health systems for the care of patients and are moving from “volume” to “value” payments.
- It will be imperative to Understand the new Value Proposition in the evolving healthcare environment.
- Significant number of generic agents in schizophrenia. New novel treatment may need to establish clear value difference against the oral agents.

Organizations that align their products and services to contribute to improvement in performance metrics will better support providers and health systems in these new programs.
EMR Clinical Trial
A Point-of-Care (or Electronic Medical Record) Clinical Trial

• **Definition** – A randomized clinical trial comparing two treatments or interventions in equipoise during normal routine clinical care in the context of an electronic medical records system.

• **Application of EMR-CT Concept** – Pilot study EMR-CT trial comparing insulin regimens in Boston VA health system (*Clinical Trials* 2011; 8: 183-195)
Electronic Medical Record (EMR) Potential Advantages

• Clinical research performed within normal clinical workflow of the health system
• Comparable patient characteristics between treatment groups (i.e., randomization)
• Outcome data captured passively
• Results directly relevant to healthcare system
• Ability to assess long-term clinical and functional outcomes (lower cost)
• Facilitation of economic analysis
Endpoint Selection

- Frequency of psychiatric-related hospitalization

- Frequency of utilization of services including
  - ER visit (psychiatric, all-causes)
  - Hospitalization for reasons other than psychiatric

Key determinants for inclusion/exclusion of an endpoint are:

- Relatively high risk for occurrence of the particular type of event

- Plausible argument that endpoint will meaningfully differentiate as a function of new novel treatment versus oral antipsychotic therapy

- High cost associated with the event
EMR Research Question

How does new novel treatment compare to psychiatrist choice of oral antipsychotic in preventing psychiatric hospitalizations (or alternative quality measure-related endpoint) in patients with schizophrenia?

Are the findings generalizable to overall patient population?
EMR Study Design

- Prospective, open-label, randomized
- Treatment groups
  - New Novel Treatment — as in normal practice, using label as guide
  - Oral antipsychotic (physician’s choice) — as in normal practice, using label as guide
- Study duration — 1 year
- Primary outcome measure (EMR retrospective database analysis)
  - Frequency of psychiatric hospitalization
  - Frequency of utilization of services
- Inpatients or outpatients with unstable schizophrenia
  - Identify and utilize clinical risk factor(s) for psychiatric hospitalization, while preserving generalizability
- Adults aged 18 years or older
EMR Study Design

- Psychiatrist/mental health care provider/investigator schedules patient visits as usual; no specific study visits are scheduled

- EMR-CT provision of randomized treatment
  - Oral treatment arm – Oral antipsychotic from predefined list can be prescribed; Rx coupon will be provided to the patient
  - New novel treatment arm – Physician will prescribe new novel treatment as in normal clinical practice; Rx coupon will be provided to the patient

- Physicians may treat patients as in normal clinical practice, even if it means switching randomized medication
Potential EMR Retrospective Data Analysis

• Rationale:
  – To characterize EMR patients diagnosed with schizophrenia disorder
  – Findings from this analysis will be used to inform the prospective EMR clinical trial
    • Identify an appropriate patient population
    • Study design parameters (duration, sample size, etc.)
    • Appropriate endpoints

• End goal:
  – Is a prospective EMR feasible?
  – Do we have the right patients to answer our research questions in this EMR?
  – Predictors of re-hospitalization?
    • Duration of Illness, Substance Abuse, Prior Hospitalization Rate
  – Do the endpoints support the new novel treatment value proposition?
EMR Data Collection Sketch

General Data Collection

• Note: Data variables will be one of two forms, raw or derived
• Demographic characteristics
• Clinical characteristics
  – Diagnosis codes (eg ICD-9 codes)
  – Psychiatric history
  – Psychotropic treatment history (prior 6-9 months)
  – Concomitant medication therapy
  – Comorbid conditions
  – Medical history
  – Adverse events
EMR Data Collection Sketch

General Data Collection
- Health outcomes, resource utilization and cost data (inpatient/outpatient)
  - Hospitalization
  - ER Visits
  - Adherence to post-discharge outpatient follow-up visits
  - Medication utilization
  - Clinical/Functional Scales
  - Mental health care service utilization
    - Type of treatment setting/care (inpatient, intensive outpatient, outpatient in Psychosocial Rehabilitation and Recovery Center (PRRC), regular outpatient, residential, supported work settings)
    - Psychotherapeutic management (cognitive behavioral therapy, acceptance and commitment therapy (ACT), interpersonal therapy (IPT))
    - Substance abuse management (motivational enhancement therapy, cognitive behavioral therapy, Opioid Treatment Programs (OTPs), residential, work therapies)
    - Suicide prevention
    - Mental Health Intensive Case Management (MHICM)/RANGE
    - Work therapies (transitional work experience, supported employment, incentive therapy)
EMR Trial Set-Up

Value Proposition
- Frequency of psychiatric hospitalization
- Frequency of utilization of services

Candidate Health System was identified
- Management bought the concept of EMR

Study Design Parameters
- Extensive computer simulations need to be performed to determine study design parameters and operating characteristics
- Optimization of study resources is a potential benefit of methodology
- Simulations will be based on following operating characteristics
  - Overall Type I error
  - Overall Power
  - Study stopping rules to be determined \textit{a priori}
    - For stopping early for success or for futility
    - May allow us to randomize patients to subgroups of interest
    - May allow us to change the primary endpoint
Looking at the trial design

• WARNING – this is work in progress
  – For a number of reasons getting a ‘good’ design has proved harder than previous case studies

• We will look at but not have definite conclusions on some, possibly, interesting aspects of this trial …
  – Two endpoints
  – Subgroups
  – What advantages, if any, can adaptation bring

• Imagine, this is a mid point review of a trial design process, the design team is going to present where its got to.

• What are your suggestions? What would you prioritise?
Starting point

• Design Requirements:
  – 12 month recruitment period, 12 month follow up.
  – Younger in disease: 12%, Substance use: 40%
  – Difference in hospitalisation may not be enough, can we use secondary endpoint too?

<table>
<thead>
<tr>
<th></th>
<th>YU (4.8%)</th>
<th>YN (7.2%)</th>
<th>OU (35.2%)</th>
<th>ON (52.8%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AC</td>
<td>Trt</td>
<td>AC</td>
<td>Trt</td>
</tr>
<tr>
<td>HR</td>
<td>40%</td>
<td>25%</td>
<td>25%</td>
<td>18%</td>
</tr>
<tr>
<td>TFR</td>
<td>70%</td>
<td>50%</td>
<td>35%</td>
<td>24%</td>
</tr>
</tbody>
</table>

• Treatment failure includes hospitalisation – so we have subtracted hospitalisation rates to create non-Hospitalisation Treatment Failure as the second endpoint
1 A tale of two endpoints

- How to analyse both endpoints?
- Two possible schemes
  - Sequential testing: take the one sided alpha limit (0.025) and divide it, testing first for significance in the HR endpoint claiming success if that is successful, otherwise then testing for significance in the TFR endpoint. If the HR is twice as important (and harder to test) as TFR we could test the HR at 0.025 * 3/4 and the TFR at 0.025 * 1/4.
  - Utility function: assign scores to our estimates of the treatment effect on the HR and nHTHF to reflect relative importance. (e.g. so a difference in rates of 5% in HR scores ‘1’ whilst a difference in rates of 10% in TFR scores ‘1’) and add them together, to see if probability that overall utility is > 1 exceeds some threshold.
• HR utility function based on difference in rate relative to active control:
  – Utility 0 at a treatment diff of 0 or less
  – Linearly rising to utility of 1 at a treatment diff of 0.05

• TFR utility function:
  – Utility 0 at a treatment diff of 0 or less
  – Linearly rising to utility of 1 at a treatment diff of 0.1

• Overall utility ... the sum of the two
Compare endpoint analyses

- Sequential test performs best when strong effect is seen on one endpoint. Utility function performs best when there is some effect on both endpoints.
- $Pr(\text{success})$ assessed by simulation (10,000 of each case)
- Looked at Null case (0% of predicted treatment effect on either endpoint) and combinations of 100%, 75% and 50%.
- Looked at test in whole population using the given rates combined weighted by expected population proportion
- $HR: 0.178 \text{ vs } 0.115$, $nHTFR: 0.154 \text{ vs } 0.094$

<table>
<thead>
<tr>
<th>HR</th>
<th>TFR</th>
<th>YU (4.8%)</th>
<th>YN (7.2%)</th>
<th>OU (35.2%)</th>
<th>ON (52.8%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AC</td>
<td>Trt</td>
<td>AC</td>
<td>Trt</td>
</tr>
<tr>
<td>40%</td>
<td>25%</td>
<td>25%</td>
<td>18%</td>
<td>25%</td>
<td>15%</td>
</tr>
<tr>
<td>70%</td>
<td>50%</td>
<td>35%</td>
<td>24%</td>
<td>40%</td>
<td>25%</td>
</tr>
</tbody>
</table>
Comparing endpoint analyses

• Comparing using the whole population:

<table>
<thead>
<tr>
<th>HR</th>
<th>nHTFR</th>
<th>Seq Tst</th>
<th>Util Fn</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0%</td>
<td>2.55%</td>
<td>2.46%</td>
</tr>
<tr>
<td>100%</td>
<td>100%</td>
<td>92.0%</td>
<td>93.1%</td>
</tr>
<tr>
<td>75%</td>
<td>100%</td>
<td>82.8%</td>
<td>80.2%</td>
</tr>
<tr>
<td>75%</td>
<td>75%</td>
<td>68.2%</td>
<td>71.7%</td>
</tr>
<tr>
<td>75%</td>
<td>50%</td>
<td>57.2%</td>
<td>61.6%</td>
</tr>
<tr>
<td>50%</td>
<td>100%</td>
<td>73.3%</td>
<td>59.2%</td>
</tr>
<tr>
<td>50%</td>
<td>75%</td>
<td>50.7%</td>
<td>47.5%</td>
</tr>
<tr>
<td>50%</td>
<td>50%</td>
<td>33.8%</td>
<td>37%</td>
</tr>
</tbody>
</table>

• Less difference than expected. I’ve used utility going forward (without harm) but not with a strong feeling of confidence either
2 Divide and Conquer?

- Ppn successful instead of all comers the trial is restricted to a different sub-population (and corresponding sample size?)

<table>
<thead>
<tr>
<th>HR</th>
<th>TFR</th>
<th>All</th>
<th>Y</th>
<th>O</th>
<th>U</th>
<th>N</th>
<th>YU</th>
<th>YN</th>
<th>OU</th>
<th>ON</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0%</td>
<td>1000</td>
<td>120</td>
<td>880</td>
<td>400</td>
<td>600</td>
<td>48</td>
<td>72</td>
<td>352</td>
<td>528</td>
</tr>
<tr>
<td>100%</td>
<td>100%</td>
<td>93.1%</td>
<td>26%</td>
<td>91%</td>
<td>78%</td>
<td>65%</td>
<td>19%</td>
<td>13%</td>
<td>71%</td>
<td>56%</td>
</tr>
<tr>
<td>75%</td>
<td>75%</td>
<td>71.7%</td>
<td>17%</td>
<td>68%</td>
<td>51%</td>
<td>38%</td>
<td>12%</td>
<td>9%</td>
<td>43%</td>
<td>30%</td>
</tr>
<tr>
<td>50%</td>
<td>50%</td>
<td>37.0%</td>
<td>9%</td>
<td>36%</td>
<td>24%</td>
<td>19%</td>
<td>8%</td>
<td>6%</td>
<td>23%</td>
<td>15%</td>
</tr>
</tbody>
</table>

As we look at subgroups – the smaller groups clearly have very little power, and at less than 100% of effect, very few of our subgroups do. But perhaps the fear of less than 100% is because of dilution, perhaps the real expectation is one group will do well and the others won’t.
• Can adaptation help us make the sub-population analysis more powerful?
• Plan:
  – analyse the groups separately
  – And use a limited adaptation - monitor each subgroup and drop it if ineffectual, allowing more to recruit in the remaining arms.
  – Allow recruitment to last at most 1.5 years? But keep overall max study size of 1,000?
• After a critical amount of information has been obtained we monitor each group
• If the result is beyond doubt we stop recruitment and declare futility or success
• Otherwise we continue until 1,000 subject has been recruited or we have been recruiting 1.5 yrs – which ever happens first
• Remaining groups are followed up until subjects complete
• We then analyse each group that didn’t stop to see if its successful
• Stopping groups early for futility or success
  – Saves time
  – Saves budget
  – Reduces subjects exposed to ineffective therapies
  – Speeds availability of effective treatment to patients / shortens time to market

• It allows us to re-assign trial budget to the remaining groups which may improve the power of the trial in those groups
• How we set the stopping criteria:
  – Simulate the “null” scenario and set final success criteria as low as possible for each group whilst keeping the number of successes < 0.025 (assumed required level of type-1 error control)
  – Then set early success criteria as low as possible whilst keeping the overall number of successes < 0.025 (i.e. introduce almost no new errors)
  – Simulate the “effective” scenario and similarly set the early futility criteria as high as possible while reducing the success rate < 0.01
  – [could recheck success criteria and possibly lower them further – but not if FDA have to approve the type-1 error control]
  – Note here: final futility is defined as not achieving final success (no inconclusive)
• Simulating each group independently
• With its sample size, recruitment rate, and expected response rates
• First interim at 36 weeks (approx 50% recruited) and then every 8 weeks thereafter

<table>
<thead>
<tr>
<th>Group</th>
<th>Max</th>
<th>Rec rate per week</th>
<th>HR – AC</th>
<th>HR – Trt</th>
<th>TFR – AC</th>
<th>TFR – Trt</th>
</tr>
</thead>
<tbody>
<tr>
<td>YU</td>
<td>72</td>
<td>1</td>
<td>0.4</td>
<td>0.25</td>
<td>0.3</td>
<td>0.25</td>
</tr>
<tr>
<td>YN</td>
<td>108</td>
<td>1.5</td>
<td>0.25</td>
<td>0.18</td>
<td>0.1</td>
<td>0.06</td>
</tr>
<tr>
<td>OU</td>
<td>528</td>
<td>7</td>
<td>0.25</td>
<td>0.15</td>
<td>0.15</td>
<td>0.1</td>
</tr>
<tr>
<td>ON</td>
<td>792</td>
<td>10</td>
<td>0.1</td>
<td>0.07</td>
<td>0.15</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Treatment failure rate here is **excluding** hospitalisation
• Stopping & success rules to prevent > 0.025 false positive rate in any group
• In terms of estimated probability that Utility is > 1

<table>
<thead>
<tr>
<th></th>
<th>Stop for futility if &lt;</th>
<th>Stop for success if &gt;</th>
<th>Declare success at the end if &gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>YU</td>
<td>0.2</td>
<td>0.99975</td>
<td>0.9875</td>
</tr>
<tr>
<td>YN</td>
<td>0.2</td>
<td>0.9995</td>
<td>0.9775</td>
</tr>
<tr>
<td>OU</td>
<td>0.2</td>
<td>0.995</td>
<td>0.89</td>
</tr>
<tr>
<td>ON</td>
<td>0.15</td>
<td>0.95</td>
<td>0.675</td>
</tr>
<tr>
<td></td>
<td>Ppn early success</td>
<td>Ppn late success</td>
<td>Ppn late futility</td>
</tr>
<tr>
<td>----------</td>
<td>------------------</td>
<td>------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>YU – null</td>
<td>0.015</td>
<td>0.010</td>
<td>0.645</td>
</tr>
<tr>
<td>YN – null</td>
<td>0.013</td>
<td>0.010</td>
<td>0.418</td>
</tr>
<tr>
<td>OU – null</td>
<td>0.014</td>
<td>0.011</td>
<td>0.219</td>
</tr>
<tr>
<td>ON – null</td>
<td>0.014</td>
<td>0.009</td>
<td>0.186</td>
</tr>
<tr>
<td>YU – 100%</td>
<td>0.079</td>
<td>0.133</td>
<td>0.703</td>
</tr>
<tr>
<td>YN – 100%</td>
<td>0.043</td>
<td>0.068</td>
<td>0.572</td>
</tr>
<tr>
<td>OU – 100%</td>
<td>0.422</td>
<td>0.332</td>
<td>0.151</td>
</tr>
<tr>
<td>ON – 100%</td>
<td>0.288</td>
<td>0.318</td>
<td>0.299</td>
</tr>
</tbody>
</table>
• There no improvements in power in the successful group if we simulated 3 groups with null response and one with 100% response.

• The adaptive design has not increased the power of the trial in the YU and YN groups, the power of the trial test in the OU and ON groups have increased a few percentage points 56% to 61%, 71% to 75%.

• This is likely to be due to a follow-up time the same length as the recruitment time.

• However the adaptive design does offer earlier decisions in null or 100% cases – with average decision times in the range 70 to 90 weeks instead of 104.
What might be next

• Add analysis of the U/N and or the Y/O groups
  – Do both?
  – Do these as well as the individual sub-groups?

• Suggest to the team that the Y group sample size could be even bigger (take longer) if the O group is dropped

• Can we get better power from the two endpoints?
  – Need more discussion on what is expected / plausible and what payers require

• Can we get power from modelling? (E.g. Bayesian hierarchical model of treatment difference, or Bayesian Augmented Control)
  – Depends if assumptions are felt justified
• Technical problem – very slow simulations
• Technical problem – no 1 tool that does both multiple endpoint and sub-population analysis
• Utility function
• Early results appear to be less useful than expected

• Use of R to post-process simulation results – e.g. to determine early stopping rules and success threshold very powerful (but in this case based on only 1,000 sims – should be 10,000 ... or more) and then needs validating