Some Considerations for Medical Devices: Historical Controls and Performance Goals

Greg Campbell, Ph.D.
Director, Division of Biostatistics
Center for Devices and Radiological Health
Food and Drug Administration

greg.campbell@fda.hhs.gov
Outline

- Regulatory Considerations
- Non-Randomized Controls and Propensity Scores
  - Example: Berlin Heart Effectiveness
- Observational Uncontrolled One-Arm Studies
  - Objective Performance Criteria (OPC) and Performance Goals
  - Example: Berlin Heart Safety
- Closing Remarks
Valid Scientific Evidence for PMAs (PreMarket Approval Applications)

- Statutory directive for the FDA’s CDRH:
  - rely upon valid scientific evidence to determine whether there is reasonable assurance that the device is safe and effective.

- Valid scientific evidence is evidence from:
  - well controlled studies
  - partially controlled studies
  - objective trials without matched controls
  - well documented case histories
  - reports of significant human experience (21 CFR 860.7)
Valid Scientific Evidence for PMAs

The valid scientific evidence used to determine the effectiveness of a device shall consist principally of well-controlled investigations.

21 CFR 860.7(e)(2)
Humanitarian Device Exemption

- Different evidentiary standard than PMAs:
  Safety and probable benefit.
Room for More Innovative Approaches

- FDA Guidance on the Use of Bayesian Statistics in Medical Device Clinical Trials was finalized in 2010.
- Adaptive Designs is another area of innovation
Bayesian Guidance

Guidance for Industry and FDA Staff

Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials

Document issued on: February 5, 2010

The draft of this document was issued on 5/23/2006

For questions regarding this document, contact Dr. Greg Campbell (CDRH) at 301-796-5750 or greg.campbell2@fda.hhs.gov or the Office of Communication, Outreach and Development, (CDER) at 1-800-835-4709 or 301-827-1800.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Devices and Radiological Health

Division of Biostatistics
Office of Surveillance and Biometrics

Finalized February 5, 2010.

http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm
Pivotal Clinical Study
Design Draft Guidance

- Discusses several concepts that are fundamental to Good Device Development Practices with respect to clinical trials.
- Some of these concepts have always been true, but have not been promulgated widely by the Agency
Non-Randomized Controls

- Two types:
  - concurrent, non-randomized control;
  - historical control.

- Always a concern about how comparable the groups are without randomization. Even if comparable, is there the same expectation of benefit?

- Such studies are observational and comparative statistical inference is compromised.

- Example: some early hip studies
Studies with Non-randomized Controls

- From a scientific standpoint, the randomized controlled trial (RCT) offers the strongest form of evidence.

- It may be possible to match the non-concurrent control group to the investigational arm in all observed measures but there is no assurance for any unobserved ones. In contrast, randomization balances for observed as well as unobserved measures.

- Historical controls can be especially problematic due to temporal bias.
Historical Controls

- A one-arm study is not a clinical trial, not an experiment; it is just an observational study.

- Since there is no randomization, all comparative statistical inference is compromised.

- It may be possible to match the historical control group to the current study in all observed measures but there is no assurance for any unobserved ones.

- In contrast, randomization balances for observed as well as unobserved measures.
Statistical Methodology for Non-Randomized Controlled Studies

- Counterfactuals and Causal Inference
  - Propensity scores
  - Sensitivity analysis
- A propensity score approach can have some scientific validity (although not as scientifically valid as a well conducted RCT) but such validity is possible only with patient-level data.
- There is a risk for the company to do a historically controlled study or one with a concurrent non-randomized control.
Propensity Score Approach

- Basic idea: Use other measured variables besides outcome such as demographic and baseline variables to try to predict the probability of being in the treatment group compared to the historical control even though no randomization has been done.

- Then compare patients in the two groups that have the same or similar propensity scores.

- In an RCT with 1:1 randomization, each subject has a propensity of 0.5.
Prospective Planning for Propensity Scores Modeling

- It is extremely important that the propensity score modeling be done planned prospectively and that it be performed without access to any of the outcome data.

- An inherent danger is of non-randomized controls is that the propensity score modeling of the individual patient baseline characteristics may indicate that the two groups are not comparable, in which case the comparison of outcomes is completely invalid.
Example: EXCOR Pediatric from Berlin Heart

- Control group was from the ELSO registry of 771 ECMO patients. For the 24 patients in Cohort 1, challenge was to find 2 close matches for each patient based on propensity score (and the same for 24 Cohort 2 patients).

- Using Body Surface Area (BSA),
  - Cohort 1 had BSA < 0.7 m² and
  - Cohort 2 had BSA at least 0.7 m² but < 1.5 m²

- The following EXCOR slides are courtesy of Terri Johnson, FDA statistical reviewer who presented them at the Advisory Committee panel in July, 2011.
Propensity Score Analysis for Primary Effectiveness Endpoint

- Primary effectiveness endpoint is time to cardiac transplantation, death or recovery (survival at 30 days post-explant or discharge with acceptable neurologic outcome, whichever is longer).

- Propensity score: probability of receiving EXCOR

- Covariates built in the propensity score model: age, weight, diagnosis, ventilator status, inotrope use, and prior cardiac arrest

- BSA as an important covariate not included in the propensity score model (since not measured in historical control)
Propensity Score Analysis for Primary Effectiveness Endpoint

**Historical control: ELSO registry (N=771)**

- Eligible Controls for Cohort 1 (n=640):
  - Excluded 10-16 years old and > 40 kg
  - 48 controls were selected (2 ECMO patients for each EXCOR patient)
Propensity Score Analysis: Cohort 1

[Box plot showing distribution of propensity scores for ELSO and EXCOR]
## Propensity Score Analysis: Cohort 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Category</th>
<th>Cohort 1 n=24</th>
<th>ELSO matches N=48</th>
<th>P-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age Group</strong></td>
<td>0-30 days</td>
<td>0 (0.0%)</td>
<td>4 (8.3%)</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>30 days – 2 years</td>
<td>20 (83.0%)</td>
<td>32 (66.7%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 to 10 years</td>
<td>4 (17.0%)</td>
<td>12 (25.0%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10 to 16 years</td>
<td>0 (0.0%)</td>
<td>0 (0.0%)</td>
<td></td>
</tr>
<tr>
<td><strong>Weight Group</strong></td>
<td>3-10 kg</td>
<td>16 (67.0%)</td>
<td>34 (70.8%)</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>10-30 kg</td>
<td>8 (33.0%)</td>
<td>14 (29.2%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30-60 kg</td>
<td>0 (0.0%)</td>
<td>0 (0.0%)</td>
<td></td>
</tr>
<tr>
<td><strong>Primary Diagnosis</strong></td>
<td>Cancer</td>
<td>0 (0.0%)</td>
<td>0 (0.0%)</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Congenital Heart Disease</td>
<td>3 (12.5%)</td>
<td>8 (16.7%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coronary Artery Disease</td>
<td>0 (0.0%)</td>
<td>0 (0.0%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dilated Myopathy</td>
<td>19 (79.2%)</td>
<td>39 (81.3%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hypertrophic Cardiomyopathy</td>
<td>1 (4.2%)</td>
<td>0 (0.0%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Restrictive Myopathy</td>
<td>1 (4.2%)</td>
<td>0 (0.0%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Valvular Heart Disease</td>
<td>0 (0.0%)</td>
<td>1 (2.1%)</td>
<td></td>
</tr>
<tr>
<td><strong>Ventilator Use</strong></td>
<td>Yes (pre-implant)</td>
<td>20 (83.3%)</td>
<td>36 (75.0%)</td>
<td>0.55</td>
</tr>
<tr>
<td><strong>Inotrope Use</strong></td>
<td>Yes (pre-implant)</td>
<td>22 (91.7%)</td>
<td>43 (89.6%)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td><strong>Cardiac Arrest</strong></td>
<td>Yes (pre-implant)</td>
<td>7 (29.2%)</td>
<td>14 (29.2%)</td>
<td>&gt;0.99</td>
</tr>
</tbody>
</table>

*Fisher’s exact test*
Statistical Results for Primary Effectiveness Endpoint: Cohort 1

- Unadjusted hazard ratio = 0.043
  \( (P-value = 0.004) \)

- Adjusted hazard ratio = 0.099
  \( (P-value = 0.03) \)

- The objective for Cohort 1 seems to be met.
Propensity Score Analysis for Cohort 2

- Historical control: ELSO registry (N=771)

  - Eligible controls for Cohort 2 (n=682)

  - Appropriate selection of the 48 controls cannot be verified
Propensity Score Analysis: Cohort 2
Statistical Results (Cohort 2): Primary Effectiveness Endpoint

Primary effectiveness endpoint for Cohort 2 cannot be evaluated since matched controls cannot be found.

The results in the HDE is NOT interpretable.
Some Historical Control Dangers

- The two groups may not be comparable. With propensity score machinery, you may not know this until after you have identified the historical control enrolled the investigational patients.

- All the other considerations also apply; there could be a change in patient care, etc.

- The proper covariates were not measured in both groups and their omission may have made a difference.

- The subsequent analysis needs to address the propensity score usage.
Observational Uncontrolled One-arm Studies

- **Objective Performance Criteria (OPC)**
  - Usually a very well-described and publicly available control used to set the criterion.
  - **Examples in CDRH:**
    - IOLs, heart valves

- **Performance Goals**

In some instances, the OPC or PG may be based on the upper (or lower) bound of the confidence interval of an effectiveness (or safety) endpoint.
Objective Performance Criteria

- An OPC is a numerical target value derived from historical data from clinical studies and/or registries which may be used for comparison with safety or effectiveness endpoints.

- Currently there are very few OPCs since device technology needs to be fairly mature.

- An OPC can be carefully constructed from all available patient-level data on a particular type of device. This is usually not done by FDA nor a particular company.
Performance Goals

- A Performance Goal (PG) is a numerical value (point estimate) that is considered sufficient by FDA for use as a comparison for a safety or effectiveness endpoint.
- Generally PGs are inferior level of evidence to OPCs.
- It is recommended that (like OPCs) that PGs not originate from a particular sponsor nor from the FDA. It is often helpful if it is recommended by a scientific or medical society.
- A fundamental regulatory question is: When a device passes an effectiveness (safety) PG (or an OPC), does that provide evidence that the device is effective (safe)?
Often the Performance Goal is constructed so that it is not simply a matter of comparing point estimates but of assuring that the confidence interval lies completely in the acceptable Performance Goal range.
Performance Goal Example:
EXCOR Pediatric
Primary Safety Endpoint

To show that the SAE calculated as the number events per patient-day is less than 0.25 tested at one-sided significant level of 0.025 (using Poisson exact method)

\[ H_0: \lambda \geq 0.25 \text{ vs. } H_1: \lambda < 0.25 \]
Primary safety endpoint:

– The total time on device of the Cohort 1 subjects was 1,411 days.
– There were 96 SAEs for Cohort 1 yielding a rate of 0.068 events per patient-day.
– The two-sided upper 95% Poisson confidence interval was 0.083 (< PG of 0.25)
– *The safety objective seems to be met.*
Statistical Results: Primary Safety Endpoint: Cohort 2

Primary safety endpoint:

- The total time on support of the Cohort 2 subjects was 1,376 days.
- There were 107 SAEs for this cohort yielding a rate of 0.078 events per patient-day.
- The two-sided upper 95% Poisson confidence interval was 0.094 (< PG of 0.25).
- *The safety objective seems to be met.*
Closing Remarks

- The law that regulates devices is substantially different from pharmaceutical drugs.
- While not ideal, propensity score analysis can be helpful in assessing group comparability and in analyzing results.
- OPCs and Performance Goals can be helpful in uncontrolled investigations but should be used with great care.
Multiple Choice Q #1

- Propensity score analysis
  - A) is always a valid way of analyzing non-randomized studies.
  - B) can be improved based on the use of the outcome data in fitting the model.
  - C) not only answers how comparable the two groups are but allows for any subsequent analysis of the groups.
  - D) can in some cases allow for the comparison of non-randomized groups but is less preferred than the comparison of randomized groups.
A) Both OPCs and Performance Goals can help to provide a control group comparison.

B) If a study relies on an OPC or Performance Goal that is the satisfied, that provides the same level of evidence as a successful controlled investigation.

C) OPCs and Performance Goals can be helpful in cases where no control is possible but the ultimate question is whether satisfaction of such numerical targets provides evidence of safety or effectiveness.

D) OPCs and Performance Goals work best when FDA (and not the sponsor) proposes the numerical target.