Indirect comparisons for single-arm trials or trials without common comparator arms

Elyse Swallow¹, James Signorovitch¹, Yong Yuan², Anupama Kalsekar²

1. Analysis Group, Inc.
2. Bristol-Myers Squibb

May 18, 2015
Workshop: ISPOR 20th Annual International Meeting, Philadelphia, PA, USA

Overview

- Single-arm trials and trials without common comparators
  - When and why do they arise?
  - Challenges for HTA evaluations
- Methods for indirect comparisons
  - How does each method work?
  - Example applications
- Comparison of methods and criteria for evaluation
- Concluding remarks
Randomized controlled trials (RCTs) are the gold standard – but are not always feasible or ethical

- Rare/orphan diseases
- Breakthrough therapies
- High unmet medical need

Single-arm trials can be used “when patient populations are extremely small, as in some orphan diseases, and the natural history of the disease is well-characterized and the drug’s beneficial effects are large” FDA

44% of EMA oncology approvals in the last decade were based on single-arm trials

> 50% of FDA accelerated approvals have been based on single-arm trials

Traditional pathway for health technology assessment using randomized trials

Novel treatment granted regulatory approval

Comparator(s) designated by reimbursement authority

Comparative effectiveness via indirect comparisons of randomized trials, e.g., network meta-analysis

Cost-effectiveness via economic modeling

Informed reimbursement decision or negotiation
Traditional indirect comparisons use RCTs

“Anchor-based” indirect comparisons rely on common comparators

Challenge: single-arm trials

Traditional anchor-based methods cannot be applied
Challenge: trials without common comparators

Traditional anchor-based methods cannot be applied for all comparisons

This often occurs due to rapidly evolving standards of care

Challenge: trials linked through > 1 comparator

Key comparisons may suffer from instability and limited precision

This often occurs due to rapidly evolving standards of care
Current HTA guidance on use of single-arm trials is limited – except to say that “naïve comparisons” are discouraged

IQWiG

“the Institute can also consider indirect comparisons to assess cost-benefit relations... [however, IQWiG] disapproves of the use of non-adjusted indirect comparisons (i.e. the naïve use of single study arms); it accepts solely adjusted indirect comparisons”

NICE

“inferences about relative treatment effects drawn from non-RCT evidence will necessarily be more circumspect than those from RCTs with properly controlled evidence”

Reimbursement submissions based on single-arm trials have been reviewed

- Access/reimbursement has been possible with only single-arm trials
- Perceived methodological strengths/weakness of indirect comparisons do not directly correlate with approval/rejection
- Other considerations are efficacy, unmet need, economic model and price

<table>
<thead>
<tr>
<th>NICE</th>
<th>pCODR</th>
<th>PBAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 submissions between 2009 and 2014</td>
<td>7 submissions between 2011 and 2014 for oncology therapies</td>
<td>5 submissions in 2007 and for oncology therapies</td>
</tr>
<tr>
<td>1 received a positive recommendation</td>
<td>4 received a positive recommendation</td>
<td>1 received full approval, 2 restricted approval</td>
</tr>
<tr>
<td>The accepted submission used clinical efficacy based on multiple single-arm trials and demonstrated a lack of alternative treatment regimens and significant potential benefits</td>
<td>Accepted submissions demonstrated limited treatment options and infeasibility of RCTs</td>
<td>Approved submissions were based on 'side by side' uncontrolled indirect comparisons to historical controls and/or other trial data</td>
</tr>
</tbody>
</table>
Uncertainty for manufacturers – what evidence should be generated for comparative effectiveness?

- Novel treatment granted regulatory approval
- Comparator(s) designated by reimbursement authority

Comparative effectiveness via indirect comparisons without common comparators

Cost-effectiveness via economic modeling

Informed reimbursement decision or negotiation

Methods for indirect comparison of single-arm trials

1. Naïve comparison
2. Benchmarking with historical controls
3. Simulated treatment comparison
4. Matching-adjusted indirect comparison

Appendix
- Comparison to self-controls
- Comparison to non-responders
Naïve indirect comparison

- **Methods**
  - Side-by-side comparison of treatment outcomes across separate trials
  - No adjustment for cross-trial differences in patient populations or control arm outcomes

- **Limitations**
  - Confounding bias due to cross-trial differences
  - Widely deprecated

- **However – naïve comparisons have been used in accepted HTA submissions**

### Cross-trial differences can bias indirect comparisons

<table>
<thead>
<tr>
<th>Category</th>
<th>Potential cross-trial differences (not exhaustive!)</th>
</tr>
</thead>
</table>
| **Trial design and conduct**    | • Blinding/concealment  
• Duration of follow-up  
• Visit frequency  
• Drop out procedure  
• Allowance for cross-overs | |
| **Trial setting**               | • Geography  
• Health system  
• Study centers  
• Year(s) of trial  
• Background standards of care  
• Disease definitions / diagnostic criteria | |
| **Patient characteristics and inclusion/exclusion criteria** | • Demographics  
• Genetic factors  
• Baseline biomarkers  
• Baseline disease severity  
• Treatment history  
• Concomitant therapies at baseline or allowed/disallowed  
• Concomitant therapies used post-baseline  
• Baseline comorbidities  
• Inclusion/exclusion criteria | |
| **Outcome measures**            | • Definitions  
• Assessment criteria & time points  
• Assessment frequency  
• Imputation of missing data | |

- Does a factor differ between trials?
- Could it be associated with outcomes?
- Could there be unobserved differences?
Methods for indirect comparison of single arm trials

1. Naïve comparison
2. Benchmarking with historical controls
3. Simulated treatment comparison
4. Matching-adjusted indirect comparison

Benchmarking with historical controls

- Historical context helpful when designing a clinical trial, but can also be incorporated into an analytical framework to estimate superiority vs. control (or non-inferiority)
- Pre-specified level of outcome for control arm and associated variability provide threshold of outcome for treatment arm to exceed
  - Can be estimated using appropriate comparator trial of interest or meta-analysis of comparator trials adjusted for differences in design
- Frequently used in proof of concept, Phase 4 and Device studies
- Intuitive to communicate and effective method to demonstrate superiority when effect sizes are greater than estimated variability
How to construct a benchmark

Evidence from a single study, with uncertainty due to *within-study variation*

Response rate

How to construct a benchmark

Combining evidence from multiple studies introduces *between-study* variation

Response rate
How to construct a benchmark

- Combining multiple historical controls, and considering within- and between-study variation, one can construct the distribution of expected outcomes for a new study of the same treatment.

How to construct a benchmark

- For a new study of a new treatment, how high does the response rate need to be to say that it is significantly different from the historical controls?
- Bayesian calculation
Case study in HCV: background and objectives

- Rapidly advancing therapeutic area with availability of all-oral regimens and cure rates approaching 100%
- Most ongoing clinical trials for newer all-oral regimens are single-arm studies
- Objectives:
  1. Establish historical control response rates for current standards of care (peginterferon alpha+ribavirin, telaprevir+ribavirin, boceprevir+ribavirin)
  2. Propose threshold for response superiority for new regimen based on estimated trial-to-trial variability
- Hypothesis: Given large improvements in efficacy/safety expected with first all-oral regimens vs. current standards of care, a head to head trial is not required to demonstrate superiority vs. the current standard of care

Case study in HCV: methods

- Systematic Literature review to identify trials of interest
- Bayesian random-effects model
  - Outcome: Sustained virological response (SVR)
  - Covariates: prior treatment history, presence of HIV coinfection, genotype, treatment group
- Assume that historical controls, sampled from multiple studies, represent the same distribution from which our treatment population was sampled.
- Characterize within- and between trial heterogeneity and set appropriate thresholds for comparison.
In treatment-naive genotype 1a, a new therapy would need to have a SVR24 rate of 84% to be deemed superior to pegIFN alfa plus ribavirin plus telaprevir.

Estimated SVR24 rates in treatment naïve subjects:

- Genotype 1a: 73%
- Genotype 1b: 70%
- Genotype 2: 75%
- Genotype 3: 80%
- Genotype 4: 48%

DCV/SOF: 98% (trial 040, treatment naïve 1a subgroup)

Methods used for indirect comparisons of single arm trials:

1. Naïve comparison
2. Historical controls
3. Simulated treatment comparison
4. Matching-adjusted indirect comparison
Simulated treatment comparison (STC)

**Treatment A Trial**
*Individual patient data*

**Treatment B Trial**
*Published summary data*
- Mean outcome for treatment B
- Mean baseline characteristics

---

Prediction model for outcome of treatment A
Simulated treatment comparison (STC)

Treatment A Trial
*Individual patient data*

Prediction model for outcome of treatment A

Simulated outcome for treatment A in trial B

Treatment B Trial
*Published summary data*

- Mean outcome for treatment B
- Mean baseline characteristics

Compare A and B
Example: axitinib for renal cell carcinoma

- The pivotal trial for axitinib included a randomized comparison to sorafenib, both as second-line treatments
- A key comparator for market access was best supportive care among patients refractory to first-line sunitinib

Simulated treatment comparison (STC)

Example: axitinib for renal cell carcinoma

- Fit parametric survival models to axitinib data and BSC (Weibull and log-normal)
- Adjusted for a composite risk score (MSKCC) and age
- Compared fitted median and mean outcomes for BSC vs. predicted outcomes for axitinib

Proskorovsky et al. 2012, EU ISPOR poster
Conclusions

- Evidence of longer OS and PFS for axitinib vs. best supportive care in sunitinib refractory patients
- Main limitations are confounding due to
  - Systematically differing factors between trials (number of prior therapies, definition of risk scores, reason for discontinuation of first-line therapy)
  - Potential unobserved factors

Additional challenge for STC

- Method assumes that the predicted outcome for the average patient equals the average of predicted outcomes across patients
- While this is true for linear models, it is not true in general
- Many outcomes typically require non-linear models
  - Time-to-event (e.g., PFS, OS)
  - Binary outcomes (e.g., response rates)
  - Counts (e.g., seizure frequency, hospitalization rate)
  - STC may introduce bias for these types of outcomes
Significant bias can arise when STC is used for binary outcomes

Bias in estimated treatment effects for comparisons of single-arm trials

Methods used for indirect comparisons of single arm trials

1. Naïve comparison
2. Historical controls
3. Simulated treatment comparison
4. Matching-adjusted indirect comparison
Matching-adjusted indirect comparison

- Adjust the population receiving the new treatment to match average baseline characteristics with a reference population
- Compare outcomes across balanced populations

Matching-adjusted indirect comparison (MAIC)

Adjustment for baseline differences in MAIC

- Adjustment should aim to address all baseline characteristics available from both trials
  - Objective selection of factors to adjust for
  - Includes key prognostic factors used in trial reporting
- Matching is based on propensity score weighting

Need to increase the weight of males relative to females to match the sex distribution of the published trial
Case study in HCV genotype 3

- HCV genotype 3
  - Accelerated progression of liver damage; higher risk of liver cancer
  - Considered the most challenging HCV genotype to treat
- Sofosbuvir + ribavirin (SOF + R) for 24 weeks
  - Recently approved for HCV genotype 3
  - Recommended by guidelines
- Daclatasvir + sofosbuvir (DCV + SOF) for 12 weeks
  - Under investigation in HCV genotype 3
- No head-to-head randomized trial of DCV+SOF vs. SOF+R

Pivotal trials of DCV+SOF and SOF+R in HCV genotype 3

DCV+SOF
Individual patient data (ALLY-3 trial)

SOF+R
Published summary data (VALENCE trial)

No placebo arms or other common comparator arms
Matching-Adjusted Indirect Comparison (MAIC)

**1. Assess cross-trial similarities/differences** in inclusion criteria, baseline characteristics, clinical definitions, outcome definitions, and study procedures

**2. Match average baseline characteristics** between trials by applying propensity score-based weighting to patients in ALLY-3

**3. Compare multiple outcomes** across balanced trial populations: sustained virologic response (SVR) at week 12 and adverse event (AE) rates

Before matching, observed baseline differences could bias comparisons of outcomes

<table>
<thead>
<tr>
<th>Baseline Characteristics</th>
<th>DCV+SOF Before Matching (n = 144)</th>
<th>SOF+R (n = 250)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, mean</td>
<td>52.02</td>
<td>48.00</td>
</tr>
<tr>
<td>Body mass index, mean</td>
<td>26.89</td>
<td>25.00</td>
</tr>
<tr>
<td>Male, %</td>
<td>59.7%</td>
<td>62.0%</td>
</tr>
<tr>
<td>White, %</td>
<td>90.3%</td>
<td>96.4%</td>
</tr>
<tr>
<td>HCV severity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasma HCV RNA log₁₀ IU/mL, mean (SD)</td>
<td>6.29 (0.84)</td>
<td>6.30 (0.70)</td>
</tr>
<tr>
<td>HCV RNA ≥ 6 log₁₀ IU/mL, %</td>
<td>69.4%</td>
<td>71.2%</td>
</tr>
<tr>
<td>Alanine aminotransferase &gt; 1.5 ULN, %</td>
<td>72.2%</td>
<td>74.4%</td>
</tr>
<tr>
<td>IL28B genotype CC, %</td>
<td>41.0%</td>
<td>34.4%</td>
</tr>
<tr>
<td>Cirrhosis, %</td>
<td>20.8%</td>
<td>24.0%</td>
</tr>
<tr>
<td>Prior interferon-based treatment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment-naive, %</td>
<td>70.1%</td>
<td>42.0%</td>
</tr>
<tr>
<td>Treatment-experienced, %</td>
<td>29.9%</td>
<td>58.0%</td>
</tr>
<tr>
<td>Relapse or breakthrough</td>
<td>17.4%</td>
<td>37.6%</td>
</tr>
<tr>
<td>No response</td>
<td>7.6%</td>
<td>16.4%</td>
</tr>
<tr>
<td>Intolerant</td>
<td>4.2%</td>
<td>4.0%</td>
</tr>
<tr>
<td>Indeterminate</td>
<td>0.7%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

~2x more with treatment experience in the SOF+R trial

*Highlighted values differed significantly before matching*
After matching there are no differences in baseline averages

<table>
<thead>
<tr>
<th>Baseline Characteristics</th>
<th>DCV+SOF Before Matching (n = 144)</th>
<th>DCV+SOF After Matching (n = 144)</th>
<th>SOF+R (n = 250)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, mean</td>
<td>52.02</td>
<td>48.00</td>
<td>48.00</td>
</tr>
<tr>
<td>Body mass index, mean</td>
<td>26.89</td>
<td>25.00</td>
<td>25.00</td>
</tr>
<tr>
<td>Male, %</td>
<td>59.7%</td>
<td>62.0%</td>
<td>62.0%</td>
</tr>
<tr>
<td>White, %</td>
<td>90.3%</td>
<td>96.4%</td>
<td>96.4%</td>
</tr>
<tr>
<td>HCV severity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasma HCV RNA log₁₀ IU/mL, mean (SD)</td>
<td>6.29 (0.84)</td>
<td>6.30 (0.70)</td>
<td>6.30 (0.70)</td>
</tr>
<tr>
<td>HCV RNA ≥ 6 log₁₀ IU/mL, %</td>
<td>69.4%</td>
<td>71.2%</td>
<td>71.2%</td>
</tr>
<tr>
<td>Alanine aminotransferase &gt; 1.5 ULN, %</td>
<td>72.2%</td>
<td>74.4%</td>
<td>74.4%</td>
</tr>
<tr>
<td>IL28B genotype CC, %</td>
<td>41.0%</td>
<td>34.4%</td>
<td>34.4%</td>
</tr>
<tr>
<td>Cirrhosis, %</td>
<td>20.8%</td>
<td>24.0%</td>
<td>24.0%</td>
</tr>
<tr>
<td>Prior interferon-based treatment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment-naïve, %</td>
<td>70.1%</td>
<td>42.0%</td>
<td>42.0%</td>
</tr>
<tr>
<td>Treatment-experienced, %</td>
<td>29.9%</td>
<td>58.0%</td>
<td>58.0%</td>
</tr>
<tr>
<td>Relapse or breakthrough</td>
<td>17.4%</td>
<td>37.6%</td>
<td>37.6%</td>
</tr>
<tr>
<td>No response</td>
<td>7.6%</td>
<td>16.4%</td>
<td>16.4%</td>
</tr>
<tr>
<td>Intolerant</td>
<td>4.2%</td>
<td>4.0%</td>
<td>4.0%</td>
</tr>
<tr>
<td>Indeterminate</td>
<td>0.7%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Highlighted values differed significantly before matching.
The DCV+SOF regimen was associated with significantly lower rates of AEs compared to SOF+R after matching.

### Rates of AEs (after weighting)

<table>
<thead>
<tr>
<th>Symptom</th>
<th>DCV+SOF</th>
<th>SOF+R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discontinuation due to AE</td>
<td>0.0%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Any AE</td>
<td>75.0%</td>
<td>91.6% *</td>
</tr>
<tr>
<td>Any serious AE</td>
<td>0.0%</td>
<td>4.0% *</td>
</tr>
<tr>
<td>Arthralgia</td>
<td>0.0%</td>
<td>10.0%</td>
</tr>
<tr>
<td>Asthenia</td>
<td>0.0%</td>
<td>21.2% *</td>
</tr>
<tr>
<td>Cough</td>
<td>0.8%</td>
<td>10.8%</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>11.1%</td>
<td>12.0%</td>
</tr>
<tr>
<td>Dry skin</td>
<td>0.0%</td>
<td>12.4% *</td>
</tr>
<tr>
<td>Dyspnea</td>
<td>0.0%</td>
<td>10.8% *</td>
</tr>
<tr>
<td>Fatigue</td>
<td>18.6%</td>
<td>30.0%</td>
</tr>
<tr>
<td>Headache</td>
<td>23.3%</td>
<td>29.6%</td>
</tr>
<tr>
<td>Insomnia</td>
<td>11.0%</td>
<td>16.4%</td>
</tr>
<tr>
<td>Irritability</td>
<td>1.3%</td>
<td>10.4%</td>
</tr>
<tr>
<td>Nasopharyngitis</td>
<td>5.5%</td>
<td>14.4% *</td>
</tr>
<tr>
<td>Nausea</td>
<td>19.3%</td>
<td>13.2%</td>
</tr>
<tr>
<td>Pruritus</td>
<td>0.2%</td>
<td>26.8% *</td>
</tr>
</tbody>
</table>

* P-value < 0.05.

### Conclusions

- After adjusting for cross-trial differences in baseline characteristics:
  - Rates of SVR12 were similar between DCV+SOF (12 week) and SOF+RBV (24 week) treated patients
  - DCV+SOF (12 week) was associated with lower rates of certain AEs compared with SOF+RBV (24 week)
  - Some of these AEs may be associated with RBV
- The main limitation is potential for unobserved confounding
- Matching adjustment ensured that comparative outcomes were not biased by observed baseline differences, particularly the % treatment naïve
Practical considerations for selecting an approach

<table>
<thead>
<tr>
<th>Requires individual patient data for one treatment</th>
<th>Naïve</th>
<th>Bench-mark</th>
<th>STC</th>
<th>MAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allows adjustment for cross-trial differences</td>
<td>no</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Valid for continuous outcomes (linear models)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Valid for binary, count &amp; time-to-event outcomes (non-linear models)</td>
<td>✓</td>
<td>✓</td>
<td>?**</td>
<td>✓</td>
</tr>
<tr>
<td>Requires one adjustment model per outcome (which can be applied to multiple comparators)</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Requires one adjustment model per comparator (which can be applied to multiple outcomes)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Allows extrapolation of outcomes beyond the observed study period</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓***</td>
</tr>
</tbody>
</table>

*Can adjust for a limited number of characteristics at the trial-level (meta-regression)
**Can be biased with non-linear outcome models; approximation of the bias and additional simulations and assumptions to reduce the bias are proposed by Ishak et al. 2015, Value in Health
***May require additional parametric modeling

Criteria for evaluation

Criteria for evaluation of indirect comparisons with single-arm trials

1. **Includes all relevant data**
2. Populations and key assessments are similar
3. Adjustment procedure is transparent and objective
4. Adjustment is comprehensive
5. Method has demonstrated statistical validity

**Best case scenario**
- Study populations identified in a systematic review
- Relevant outcomes for decision making are included (e.g., efficacy and safety)
- Key baseline characteristics and study design features are available
Criteria for evaluation of indirect comparisons with single-arm trials

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Best case scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Includes all relevant data</td>
<td>• Similar or equivalent study designs and definitions, e.g.</td>
</tr>
<tr>
<td>2. Populations and key assessments are similar</td>
<td>• Outcome definitions</td>
</tr>
<tr>
<td>3. Adjustment procedure is transparent and objective</td>
<td>• Diagnostic criteria</td>
</tr>
<tr>
<td>4. Adjustment is comprehensive</td>
<td>• Background therapies</td>
</tr>
<tr>
<td>5. Method has demonstrated statistical validity</td>
<td>• Study locations</td>
</tr>
<tr>
<td></td>
<td>• Time periods</td>
</tr>
<tr>
<td></td>
<td>• Blinding</td>
</tr>
</tbody>
</table>

Criteria for evaluation of indirect comparisons with single-arm trials

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Best case scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Includes all relevant data</td>
<td>• Similar patients, e.g.</td>
</tr>
<tr>
<td>2. Populations and key assessments are similar</td>
<td>• Demographics</td>
</tr>
<tr>
<td>3. Adjustment procedure is transparent and objective</td>
<td>• Comorbidities</td>
</tr>
<tr>
<td>4. Adjustment is comprehensive</td>
<td>• Disease severity</td>
</tr>
<tr>
<td>5. Method has demonstrated statistical validity</td>
<td>• Biomarkers</td>
</tr>
<tr>
<td></td>
<td>• Treatment history</td>
</tr>
<tr>
<td></td>
<td>• Ideally the population with individual patient data is broader than the population without</td>
</tr>
</tbody>
</table>
## Criteria for evaluation of indirect comparisons with single-arm trials

### Criteria

1. Includes all relevant data  
2. Populations and key assessments are similar  
3. **Adjustment procedure is transparent and objective**  
4. Adjustment is comprehensive  
5. Method has demonstrated statistical validity

### Best case scenario

- Accounting for any patient exclusions  
- Adjustment for all available baseline characteristics  
  - Or justifications and sensitivity analyses for limited adjustments  
- Justification of parametric outcomes models (STC)

---

### Criteria

1. Includes all relevant data  
2. Populations and key assessments are similar  
3. Adjustment procedure is transparent and objective  
4. **Adjustment is comprehensive**  
5. Method has demonstrated statistical validity

### Best case scenario

- Key prognostic factors have been adjusted for  
- Limited potential for unmeasured confounding based on expert clinical input
Criteria for evaluation of indirect comparisons with single-arm trials

Criteria

1. Includes all relevant data
2. Populations and key assessments are similar
3. Adjustment procedure is transparent and objective
4. Adjustment is comprehensive
5. Method has demonstrated statistical validity

Best case scenario

- Statistical justification within framework for causal inference
- Statistical performance validated in simulations with realistic scenarios
  - Bias reduction
  - Valid confidence intervals
- Empirical validation with real data and RCTs

Simulation studies have validated MAIC performance: Reduction of bias – even for small samples

MAIC

<table>
<thead>
<tr>
<th>n</th>
<th>Severe Confounding</th>
<th>Mild Confounding</th>
<th>No Confounding</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=50</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Naïve

<table>
<thead>
<tr>
<th>n</th>
<th>Severe Confounding</th>
<th>Mild Confounding</th>
<th>No Confounding</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=50</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MAIC included adjusted for 10 baseline characteristics
Simulation studies have validated MAIC performance:
Correct coverage for 95% CIs – even for small samples

Criteria for evaluation

STCs and MAICs have been used in multiple disease areas

<table>
<thead>
<tr>
<th>Disease</th>
<th>Treatment evaluated</th>
<th>Single-arm</th>
<th>Method(s)</th>
<th>Publically-available HTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renal cell carcinoma</td>
<td>Axitinib</td>
<td>Y</td>
<td>STC, MAIC</td>
<td>Y</td>
</tr>
<tr>
<td>Multiple myeloma</td>
<td>Bortezimib</td>
<td>Y</td>
<td>MAIC</td>
<td>Y</td>
</tr>
<tr>
<td>Pancreatic neuroendocrine tumors</td>
<td>Everolimus</td>
<td>Y</td>
<td>MAIC</td>
<td>Y</td>
</tr>
<tr>
<td>Chronic myelogenous leukemia</td>
<td>Imatinib</td>
<td>N</td>
<td>MAIC</td>
<td>Y</td>
</tr>
<tr>
<td>Non-small cell lung cancer</td>
<td>Ceritinib</td>
<td>Y</td>
<td>MAIC</td>
<td></td>
</tr>
<tr>
<td>Mantle cell lymphoma</td>
<td>Ibrutinib</td>
<td>Y</td>
<td>MAIC</td>
<td></td>
</tr>
<tr>
<td>Hepatitis C</td>
<td>Daclatasvir</td>
<td>Y</td>
<td>MAIC</td>
<td>Y</td>
</tr>
<tr>
<td>Hepatitis C</td>
<td>Simeprevir</td>
<td>Y</td>
<td>MAIC</td>
<td>Y</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>Vildagliptin</td>
<td>N</td>
<td>MAIC</td>
<td>Y</td>
</tr>
<tr>
<td>ADHD</td>
<td>Guanfacine</td>
<td>N</td>
<td>MAIC</td>
<td>Y</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>Adalimumab</td>
<td>N</td>
<td>MAIC</td>
<td></td>
</tr>
<tr>
<td>Psoriatic arthritis</td>
<td>Adalimumab</td>
<td>N</td>
<td>MAIC</td>
<td></td>
</tr>
<tr>
<td>HIV</td>
<td>Atazanavir</td>
<td>N</td>
<td>MAIC</td>
<td></td>
</tr>
<tr>
<td>Cystic fibrosis</td>
<td>Tobramycin</td>
<td>Y</td>
<td>MAIC</td>
<td></td>
</tr>
</tbody>
</table>

Based on searching websites for NICE, SMC, PBAC, CADTH/pCODR, IQWiG, as well as PubMed
Final Remarks

- Limited practical guidance currently exists from HTAs for indirect comparisons of single-arm trials
- However this setting is common for high-priority treatments
- Benchmarking, STC and MAIC:
  - Have been used successfully in single-arm settings
  - Have foundations in traditional statistical methods
  - Can be evaluated similarly to traditional observational studies
  - Provide improvements over naïve/unadjusted comparisons to a single historical control
- Multiple methods can be complimentary when used in combination

Thank You
Appendix
Comparison to self-controls
Comparison to non-responders

Self-controlled study design

- Typically used to evaluate the relationship between a transient exposure (e.g., drug treatment) and an acute event (e.g., hospitalization rate, progression)
- Best suited for independent recurrent events
- A single treatment is evaluated by comparison of a patient’s status in two distinct periods (e.g., before and after treatment)

<table>
<thead>
<tr>
<th>Untreated</th>
<th>Treated</th>
</tr>
</thead>
</table>

Event rate without treatment vs. event rate on treatment

- Time-invariant confounders, such as sex and genetics are implicitly controlled
- Key assumption: expected outcomes without the new treatment are equal to (or worse than, to be conservative) observed outcomes during the prior treatment
Comparison to non-responders

Case example: idealisib for refractory FL

- Refractory follicular lymphoma (FL) patients
  - Idealisib 150mg bid (monotherapy)
  - Primary endpoint: Overall response rate (ORR)
  - Secondary endpoints: Progression-free survival (PFS) and overall survival (OS)

- The manufacturer and TLV believed a randomized controlled trial was unethical for this population

- Best supportive care (BSC) was an important comparator

- Idealisib non-responders were used as a proxy for BSC

Comparison of the full treatment cohort to non-responders

<table>
<thead>
<tr>
<th>Full 101-109 study sample</th>
<th>N=125</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idealisib responders</td>
<td>N=58</td>
</tr>
<tr>
<td>Idealisib non-responders</td>
<td>N=67</td>
</tr>
</tbody>
</table>
Comparison of the full treatment cohort to non-responders

Claimed to be conservative

- **Progression-free survival**
- **Overall survival**

- Idealisib non-responders
- All study participants

This methodology is not necessarily conservative

- **Key assumption:** non-responders receiving treatment have equal or better outcomes vs. all patients without treatment
- **Thought experiment** with two plausible assumptions
  - Suppose treatment has no effect, yet by chance some tumors satisfy responder criteria and others do not
  - In addition, suppose that response is associated with longer OS/PFS
  - The non-responder group would be expected to have worse PFS/OS compared to the full cohort, even though treatment is not effective – this contradicts the key assumption for comparison to non-responders
- **Other potential concerns**
  - Impact of timing of response assessment
  - Differences in baseline characteristics for non-responders vs. full cohort
Additional critiques

From TLV

“The clinical documentation has major shortcomings … [but] it may be prudent to use the data for those who did not respond … as a proxy for best supportive care”

From NICE ERG (different submission)

“The effect on ORR, PFS, and OS is uncertain. It is not self-evident that the design employed will necessarily lead to an overestimate of the treatment effect [for best supportive care], although superficially tempting to conclude that this is the most likely impact of bias”