Choosing Monitoring Boundaries: Balancing Risks and Benefits

Pamela Shaw

shawp@upenn.edu

Department of Biostatistics, Epidemiology and Informatics

University of Pennsylvania

April 19, 2017

Outline

- What do we monitor?
- How do we monitor it?
- A challenging example
- Some new proposals
- Discussion

What do we monitor?

- Clinical trial architecture is typically defined by a primary efficacy outcome
- A fundamental role of a DSMB is to assess the benefit/risk ratio
- Prior studies can yield a list of potential risks/benefits
  - May be symptoms (nausea, pain, etc.) or may be risks of severe outcomes (elevated stroke, cancer, death)
  - May also have important secondary efficacy endpoints (e.g. fractures in WHI Hormone Replacement Trial)
  - Trials structure also set up to capture unanticipated adverse events

Some of the challenges to assessing benefit/risk

1. Multivariate outcomes need to be considered
   - These outcomes may be of varying severity
2. Risks may change over time
3. Risks may be infrequent/rare
4. For novel therapies, risks may be largely unknown
5. Expect the unexpected…

1 & 2 imply that in order to evaluate risk/benefit one has to prioritize the outcomes and prioritize the importance of early/late events

(explicitly or implicitly, formally or informally)

Two approaches for monitoring risk/benefit

1. Multiple outcomes assessed separately
   - Primary endpoint may have a formal monitoring boundary
   - DSMB is presented with analyses of several separate endpoints: primary, key 2nd-ry, important safety outcomes
   - DSMB weighs totality of evidence, a subjective judgment is made for overall balance of risk/benefit

2. A statistic summarizing risk/benefit is assessed
   - Composite endpoint determined prior to start of trial
   - Risk/benefit p-value calculated/compared to a boundary
   - Subjective judgment still needed to weight totality of evidence

Issues that complicate evaluation of the benefit/risk trade off

- Severity of health outcomes affected by the treatment may be very different
  - Assessing overall benefit means giving relative weights to these risks/benefits
  - Patients/clinicians may have differing opinions on these weights
- Frequency of health outcomes affected by the treatment may be very different
  - When does the increased risk of a rare, but serious side effect offset the benefit of a treatment?
- Tolerance of a side effect depends on whether it is in a healthy population or sick population
- Timing of endpoints may differ: early harm, later benefit or vice versa

WHI Example

- Women's Health Initiative (WHI) conducted two hormone therapy (HT) trials
- Trials were unique in the amount of data collected on HT prior to the trial start
  - Expecting 40-50% decrease in heart attacks
  - Observational studies raised concern over increase in breast cancer
- A formal monitoring plan was put into place for both efficacy and harm for both HT trials
  - Considered 8 outcomes of roughly equal importance. Most thought to be related to efficacy
  - Had a global index of benefit/risk (Z=-1)

(Wittes et al. 2007; Freedman et al. 1996)

WHI HT monitoring plan

- Primary efficacy endpoint: Coronary heart disease (CHD)
- Primary safety endpoint: invasive breast cancer
- Formal analyses used weighted log-rank statistic to further down-weight early events
  - Motivated by expected early CHD benefit and late BCA harm. Also, drug needed time to have an effect
  - Unweighted useful in case there were early harms, don't want to down weight them

WHI Trial: Unexpected outcomes

- Discrepancy between expected and observed efficacy and safety endpoints
  - Early on, an increased risk of CHD/stroke/PE for active arm emerged in both trials
  - Later on, divergent effects appeared for breast cancer
- Debate ensued whether and how the safety endpoint be modified (Wittes et al 2007)
- Level of significance and direction of effect varied based on weighted vs unweighted log-rank statistics

WHI Example: Lessons learned/affirmed

- Monitoring multivariate outcomes is complex
- Reliably assessing risk and harms means knowing which endpoints are which
- Difficult to rely on a single p-value when considering a multivariate outcome
- Decision-making is ultimately a subjective activity

WHI example highlights monitoring duality

- Pre-specified boundaries protect against inflating p-values by defining risk categories after a difference is observed
- Formal boundaries however can ”lock thinking” and need to be flexible in the face of unexpected risks
  - A desire to stick to pre-specified boundaries
  - Ironically, statisticians can be quicker to ditch the boundaries than clinical colleagues
- Desire to have a clear, data-driven statistic do the work, but interpretation needs to bring in a global perspective
  - Data from other trials
  - Leanings of other trends in data
  - Uncertainty in assumptions behind monitoring boundary

Need for better statistical approaches to assess benefit/risk?

Usual statistical approaches have some limitations:

- Time to first ignores subsequent and potentially more severe occurring endpoints
- HR can over emphasize increases in small absolute risk
- HR –precision limited by number of events
- Uno et al. 2015 discuss advantages of risk difference, percentile difference, restricted mean survival difference in non-inferiority trials
- Multiplicity

Many recent proposals for assessing benefit/risk*

- **Win-ratio: Pocock et al. 2012,** Finkelstein and Schoenfeld 1999; Bebu and Lachin 2016, Oakes 2016
- **Severity ranking: Shaw and Fay 2016**
- **Total Assessment: Evans et al 2015 (DOOR),** Berry et al. 2013
- Outcome Weighting: Bakal et al. 2013
- Proportion favoring treatment: Buyse 2010
- Joint test: Finkelstein and Schoenfeld 2014

Approaches to assessing benefit/risk

1. Create an aggregate score from a weighted sum of outcomes
   - Interpreted as a global assessment of patient outcome
   - Naturally incorporates multiple events
2. Order outcomes in terms of a preferred importance and rank/classify patients using the highest ordered outcome possible
   - For censored event times often means ranking patients over a common follow-up time
   - Essentially creates a weighted combination of score statistics, where the rates relate to the probability of the events of higher order being observed

Differing opinions on whether to create separate safety and efficacy composites

- Evans and Follmann 2016 advocate a unified composite of benefit and risk as a pragmatic endpoint of effectiveness
- Kip et al 2008 recommend against lumping safety and efficacy limits interpretability in setting of cardiovascular disease
  - Frequently dominated by a subclass of endpoints
  - Too susceptible to providing misleading evidence
- “Although numerous approaches and frameworks have been proposed in recent years, there is no single approach or framework that can be applied and utilized in every setting.” (Ch 8, Jiang, He 2016)

Win ratio[1]

- Patients in treatment and control groups are placed into matched pairs according to their risk profiles
- Determine prioritization of outcomes
  - Example: two endpoints: death or MI Hospitalization, consider time to death first then time to hospitalization
- Within each pair, a tx subject is labeled a winner or loser using the outcome of highest priority possible
  - Compare time to death if possible; otherwise compare time to hospitalization; otherwise tied
- The win ratio is the ratio of wins/loss for treatment arm
  - P-value and CI are readily obtainable

1. Pocock 2012

Useful features of win ratio

- Can consider all observed events on a patient
  - Allows more severe events to have higher priority
  - Particularly useful in cases where first event is expected to be the less severe event
- Potentially higher power than any single endpoint
  - Particularly if treatment effect similar across endpoints
- Easy to calculate and make inference[1,2,3]
  - Unpaired version is available using a U-statistic derived from all possible tx-control pair

1. Pocock 2012 2. Bebu and Lachin, 2016; 3. Oakes 2016

Win ratio example: The SOLVD trial (NEJM 1991)

Background

- SOLVD included a RCT of a novel treatment for prevention of mortality/ hospitalization in patients with congestive heart failure (CHF) and weak left ventricle ejection fraction (EF)
- In 1986-89, 2569 patients randomized to enalapril or placebo
- Enalapril found beneficial for mortality ($p = 0.0036$) and time to first hospitalization/death ($p < 0.0001$)

Analysis

- Considered a subset of 662 diabetic subjects
- Compute usual time-to-first (TTF) endpoint
- Compute win ratio for control-treatment patients pairs formed using a baseline Cox model risk score for death

SOLVD Trial: Time-to-first analysis

	Enalapril (N=319)		Placebo (N=343)		Cox PH	Score Test
Endpoint	Yes	No	Yes	No	HR	(P-value)
Death	137	182	145	198	0.99	(0.91)
Hospitalization	94	225	148	195	0.60	(< 0.0001)
TTF	174	145	229	114	0.71	(0.0007)

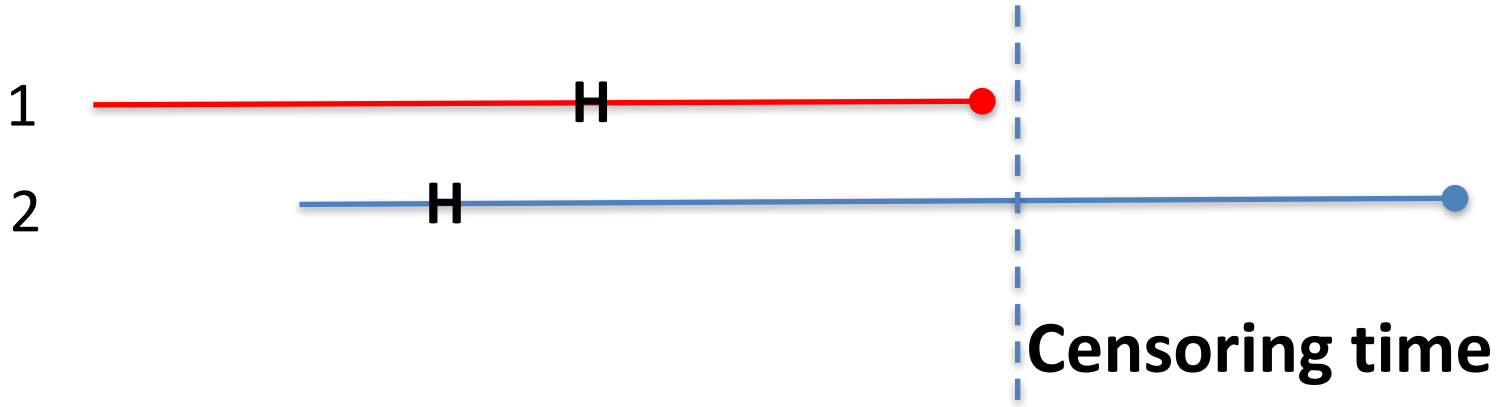
SOLVD: Win ratio

- 343 on Placebo arm, 319 on active arm
  - 24 patients go unused in the paired analysis
- 145 wins on active; 112 wins on placebo
- WR=145/112 = 1.29 (p=0.038)
  - 189 ranked on death: 98 wins for active, 91 wins for placebo
  - 68 ranked on hospitalization: 47 wins active; 21 wins placebo

A few key points about win ratio

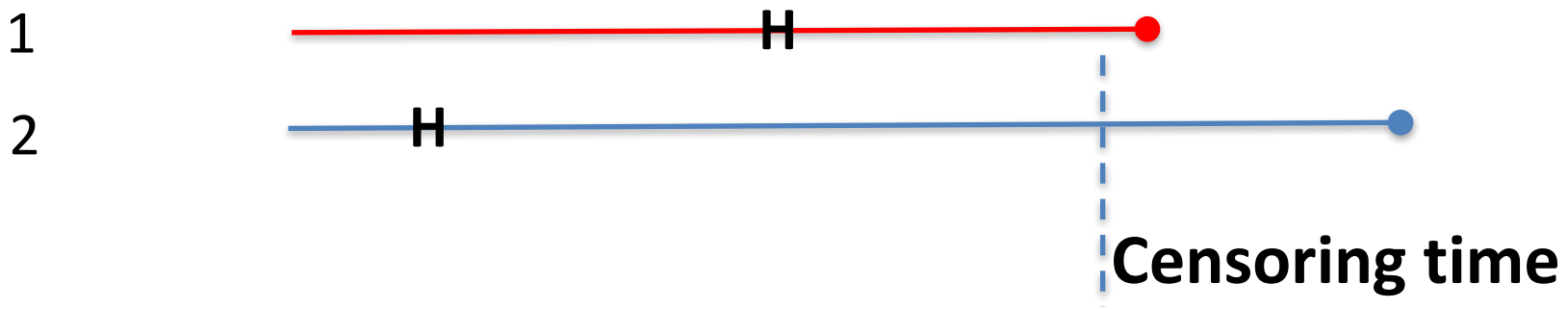
- The parameter the WR estimates depends on the censoring distributions of the endpoint
  - Important consideration if early and late risks
- Trials of different lengths will generally be estimating a different effect estimate
- When patients have varying follow-up lengths the WR becomes more difficult to interpret
  - SOLVD follow-up: 1 day to 4.6 years in example
- If death determines severity, then is ranking by other less severe endpoints gaining information or a means of potential misclassification?

Win ratio: Gaining information from hospitalization or misclassifying?



Patient 1 died at 3 years;
Patient 2 censored at 2.5 years ; died at 4 years

Win Ratio: Gaining information from hospitalization or misclassifying?



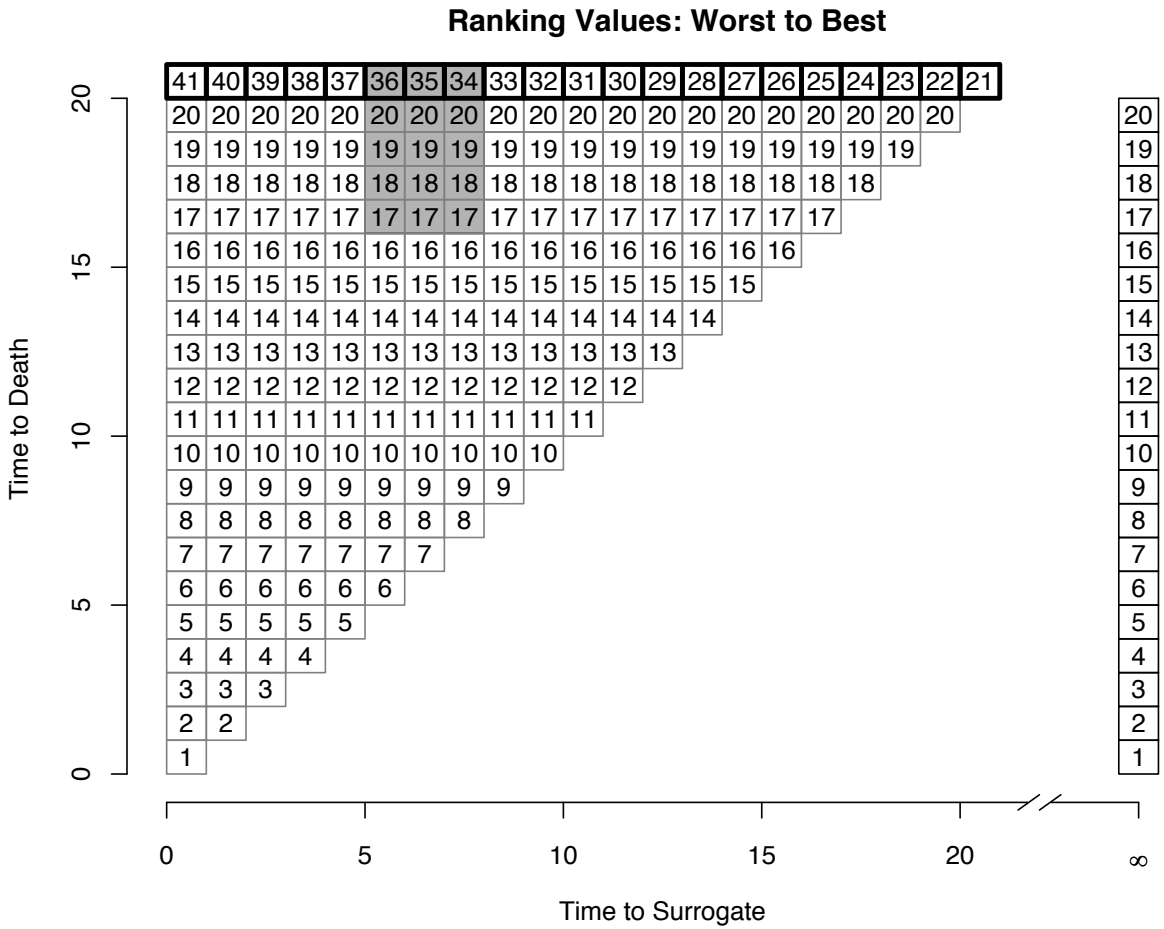
Patient 1 died at 3 years;
Patient 2 censored at 2.5 years ; died at 4 years

The true state of information here is that the patient 1 severity relative to patient 2 is interval censored.

Clinical severity ranking
Shaw and Fay SIM 2016

- Rank individuals according to clinical severity, using information on both the surrogate and true endpoint
  - Ranking function of the two event times can vary by setting
- Setting of interest: XDR-TB: sputum conversion/death
  - Rank patients by time of death if observed
    - Earlier is worse
  - Rank time to sputum conversion for the survivors
    - Earlier is better
  - Conversion time irrelevant if patient later dies
- Perform two sample test on an interval-censored clinical severity which incorporates bivariate survival information

Shaw and Fay SIM 2016



Grey box: Severity score for patient who converted in weeks 6-8, inclusive, but dropped out after week 16. Interval censoring in disjoint intervals

Further musings on tests of severity using joint survival distribution

- Take advantage of all the information regarding the survival time (not limited to common follow-up times for pairs)
- Including the true uncertainty about the severity of a patient
- Test statistic for composite still has the problem that that parameter estimated depends on length of trial
  - Fay and Shaw showed that the resulting test statistic is a weighted sum of a test statistic on death and on the surrogate

DOOR Ranking Evans et al. 2015

- Collection of possibilities of clinical outcomes of a patients are ranked according to preferred to least preferred outcome
  - Rate all possible clinical paths on an ordinal scale
  - Rating/ranking can be done by expert clinical panel, potentially also including patients
  - Then a U-type statistic could be used to examine if the outcome on tx better than that for a patient
- Proposed a blinded adjudicated committee could evaluate clinical severity based on patient chart
  - Not practical for larger trials or very reproducible
- Similar ideas discussed by a number of authors, including Chuang-Stein et al. 1991

DOOR hypothetical example (Evans et al. 2015)

1. Clinical benefit without AE
2. Clinical benefit with AE
3. Survival w/o clinical benefit or AE
4. Survival w/o clinical benefit + AE
5. Death

- In setting of anti-infective, break ties using length of antibiotic regimen (DOOR/RADAR)

DOOR: Advantages and limitations

Advantages

- Simple and intuitive measure
- Ranking cognitively easier than weighting
- Can incorporate different ranking systems

Limitations

- Varying length of follow-up can be a challenge
- Loss of information through ties can be a problem for ordinal
- Will be difficult to adapt to unexpected benefits or risks. Would need to reconvene outcome ranking panel

- Perhaps best used along-side Individual components for interpretation

Some pragmatic considerations

- For composite endpoints: create group(s) based on similar severity
  - Some settings may want to pool safety and risk for net clinical benefit
  - Added interpretability if individuals outcomes occur with similar frequency
- Sensitivity analysis to see impact of value systems
  - If using outcome weighting, can be used identify the “value breaking point”
- Practice Run decision scenarios: Valuable exercise to hone the needed value judgements (some can be pre-specified) and statistical decision boundaries
- Clear presentation and visualization of data (estimates) for DSMB report will aid in assessment of totality of evidence

Conclusions

- No one approach will work for every setting
- Good to remember all approaches involve subjectivity
- Specific endpoint + composites that summarize effect on multiple endpoints seems like a flexible and powerful combination
- Statistical properties of composites need rigorous examination and thorough numerical investigation before start of trial for expected scenarios
- Practice run decision scenarios: Valuable exercise to to hone the needed value judgements and statistical decision boundaries
- A prior development of risk-benefit statistic and boundary is a useful decision tool but cannot be prescriptive

Thank you!

References

1. Bakal JA, Westerhout CM, Cantor WJ, Fernández-Avilés F, Welsh RC, Fitchett D, Goodman SG, Armstrong PW. Evaluation of early percutaneous coronary intervention vs. standard therapy after fibrinolysis for ST-segment elevation myocardial infarction: contribution of weighting the composite endpoint. European Heart Journal. 2013 Mar 21;34(12):903-8.Bebu

2. Bebu I, Lachin JM. Large sample inference for a win ratio analysis of a composite outcome based on prioritized components. Biostatistics. 2016 Jan 1;17(1):178-87.

3. Berry JD, Miller R, Moore DH, Cudkowicz ME, Van Den Berg LH, Kerr DA, Dong Y, Ingersoll EW, Archibald D. The Combined Assessment of Function and Survival (CAFS): a new endpoint for ALS clinical trials. Amyotrophic lateral sclerosis and frontotemporal degeneration. 2013 Apr 1;14(3):162-8.Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Statistics in Medicine. 2010 Dec 30;29(30):3245-57.

4. Chuang-Stein C, Mohberg NR, Sinkula MS. Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. Statistics in Medicine. 1991 Sep 1;10(9):1349-59.

5. Evans SR, Follmann D. Using Outcomes to Analyze Patients Rather than Patients to Analyze Outcomes: A Step Toward Pragmatism in Benefit: Risk Evaluation. Statistics in Biopharmaceutical Research. 2016 Oct 1;8(4):386-93.

6. Evans SR, Rubin D, Follmann D, Pennello G, Huskins WC, Powers JH, Schoenfeld D, Chuang-Stein C, Cosgrove SE, Fowler VG, Lautenbach E. Desirability of outcome ranking (DOOR) and response adjusted for duration of antibiotic risk (RADAR). Clinical Infectious Diseases. 2015 Jun 25:civ495.

7. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. Statistics in Medicine. 1999 Jun 15;18(11):1341-54.

References (2)

8. Finkelstein DM, Schoenfeld DA. A joint test for progression and survival with interval-censored data from a cancer clinical trial. Statistics in Medicine. 2014 May 30;33(12):1981-9.

9. Freedman L, Anderson G, Kipnis V, Prentice R, Wang CY, Rossouw J, Wittes J, DeMets D. Approaches to monitoring the results of long-term disease prevention trials: examples from the Women's Health Initiative. Controlled Clinical Trials. 1996 Dec 1;17(6):509-25.

10. Jiang Q, He W, editors. Benefit-Risk Assessment Methods in Medical Product Development: Bridging Qualitative and Quantitative Assessments. Chapman and Hall/CRC; 2016 Jul 7.

11. Kip KE, Hollabaugh K, Marroquin OC, Williams DO. The problem with composite end points in cardiovascular studies: the story of major adverse cardiac events and percutaneous coronary intervention. Journal of the American College of Cardiology. 2008 Feb 19;51(7):701-7.

12. Oakes D. On the win-ratio statistic in clinical trials with multiple types of event. Biometrika. 2016 Jul 25:asw026.

13. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. European Heart Journal. 2012 Jan 1;33(2):176-82.

14. Shaw PA, Fay MP. A rank test for bivariate time-to-event outcomes when one event is a surrogate. Statistics in medicine. 2016 Aug 30;35(19):3413-23.

15. Uno H, Wittes J, Fu H, Solomon SD, Claggett B, Tian L, Cai T, Pfeffer MA, Evans SR, Wei LJ. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. Annals of Internal Medicine. 2015 Jul 21;163(2):127-34.

16. Wittes J, Barrett-Connor E, Braunwald E, Chesney M, Cohen HJ, DeMets D, Dunn L, Dwyer J, Heaney RP, Vogel V, Walters L. Monitoring the randomized trials of the Women's Health Initiative: the experience of the Data and Safety Monitoring Board. Clinical Trials. 2007 Jun 1;4(3):218-34.