

Assessing Outcomes and Safety

Steven R. Cummings, MD
Director, SF Coordinating Center

Outline: Outcomes

- Primary and secondary aims
- 'Surrogate markers'
- Composite outcomes
- Adverse experiences

Morbone

- A company wants help designing a trial of Morbone new treatment for osteoporosis
- Animal models: improves bone mass and bone strength
- Planning a clinical trial
- Would like FDA approval to market Morbone

Morbone Trial

Potential outcomes

- Bone density (BMD)
- Vertebral fractures (by spine x-ray)
- Nonvertebral fractures (by clinical dx)
- Hip fractures

How to start?

- Designate one “primary” and the others as “secondary” outcomes

Why one “primary” outcome?

- To calculate sample size
- For testing statistical significance without penalty
- Greater credibility
- The FDA requires that an outcome be “primary” in order to approve a drug for that indication
 - Beneficial effects on secondary outcomes can't be used for 'indication'

Which primary for Morbone?

- Bone density (BMD)
- Vertebral fractures (by x-ray)
- Nonvertebral fractures (by clinical dx)
- Hip fractures

Considerations in choosing the primary outcome

- Clinical importance
- Feasibility
 - Sample size and cost
- Scientific / biological interest
- (For new drugs: What does FDA need in order to approve an indication for prescribing the drug?)

Clinical importance

- Hip fracture: causes almost all of the deaths, and 75% of the costs of fractures
- 'Nonvertebral' fractures: the most common kind of fracture.
- Vertebral fracture by x-ray: about 1/3 cause recognized pain and disability. Mild changes might not be real fractures.
- BMD: loss leads to greater risk of fractures

Which Primary Outcome? Sample size

<i>Alternatives</i>	<i>Sample size/duration</i>
• Improvement in BMD	• 200 / 1 year
• Vertebral fractures	• 2,500 / 3 yrs
• Nonvertebral fracture	• 5,000 / 3 yrs
• Hip fractures	• 8,000 / 4 yrs

Why not make BMD the primary outcome?

- Best choice to minimize cost
- Issue: is it a valid surrogate marker of clinical outcomes?

Clean up your language!

- Not all markers are 'surrogates'
- Biomarkers
 - Measurement of a process or state.
- 'Surrogate' marker
 - Substitute for clinical outcome
- Validated surrogate
 - You can trust it. Effect of treatment on marker has been established to consistently represent the effect on clinical outcome.

Surrogate markers for trials

- 'A laboratory or physical sign that is used in trials as a substitute for a clinically meaningful endpoint.'

The perfect surrogate is the causal pathway by which tx affects the disease



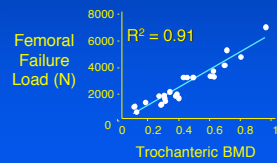
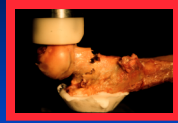
Criteria for validating a surrogate marker for treatments

- Biologically plausible
- Marker strongly predicts the clinical outcome
- Treatment changes the marker
 - Treatment changes the rate of disease in the predicted direction

* Prentice, 1989

Is BMD a valid surrogate?

- Biologically plausible
- Highly correlated with bone strength in destructive testing
- $R^2 = 0.7 - 0.9$

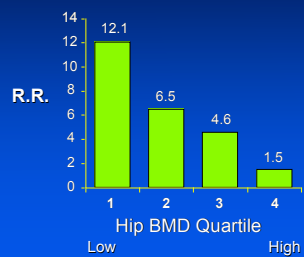


Criteria for validating a surrogate marker for treatments

- Biologically plausible
- Marker strongly predicts the clinical outcome

** Prentice, 1989*

Observational studies BMD predicts fracture



Bone density

- Biologically plausible: YES
- Marker strongly predicts the clinical outcome: YES
- Treatment changes the marker*
 - YES: treatment improves BMD~5%

** Prentice, 1989*

Treatment improves BMD



BMD predicts fracture



Are we there yet?

- BMD does all the right things a surrogate marker should do.
- Anything else?

Bone density

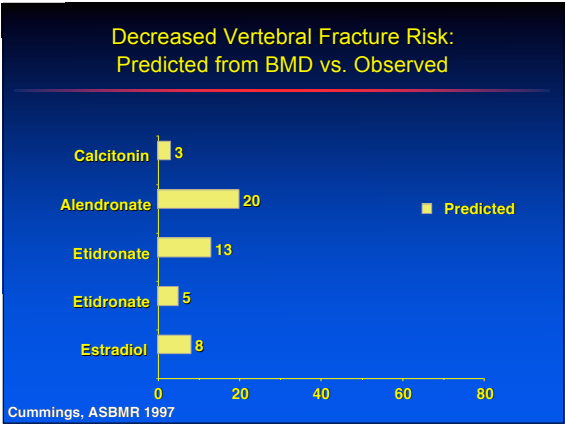
- Do changes in the surrogate (bone density) *account for* changes in reduction in the outcome (fractures)?

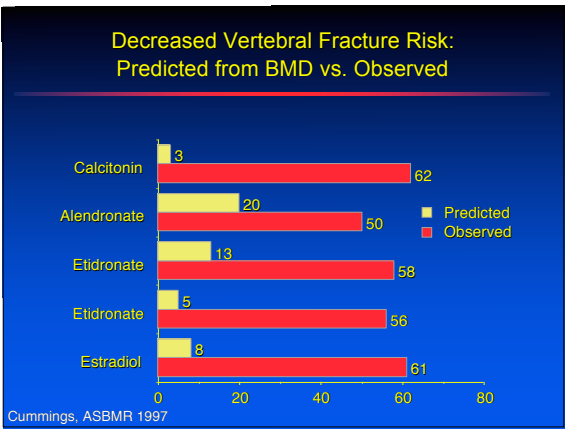
* Prentice, 1989

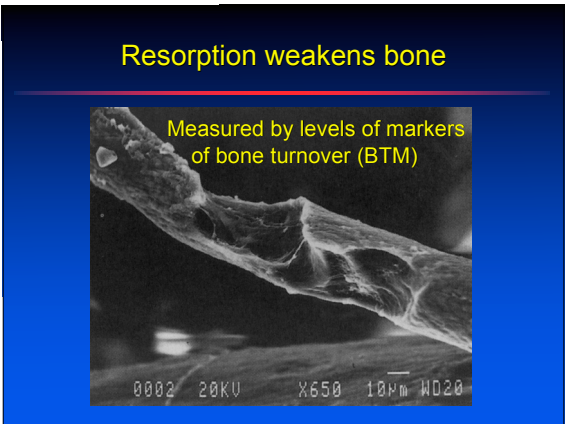
Flouride

- Increased BMD ~10%
- *Increased* the risk of fractures

* Prentice, 1989







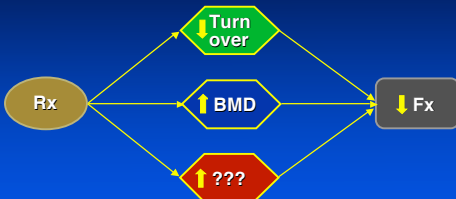
"Bone turnover"

- Bone resorption: osteoclasts dig 'pits' in bone
- Bone formation: osteoclasts make new bone in the pits
- More resorption -> faster loss
- More pits -> weaker bone
- Proteins produced by the process can be measured in blood: "Bone turnover markers"

The perfect surrogate is the causal pathway by which tx affects the disease



But treatments may have other effects that could also influence the disease



Change in one marker will account for only part of the effect.
Some changes might also be harmful.

Proving that a change in measurement predicts effect of treatment on fracture risk

Two approaches

- Individual level
 - How well does *change* in the marker account for the effect in people?
- Trial level
 - How well does the *change* in measurement predict the clinical results from trials?

Validating that a marker is a good surrogate

#1

- Individual level
 - How well does change in the measurement account for the decrease in fracture risk in people?
 - The test: What percent of effect of decrease in fracture risk 'explained' by change in the measurement

Does BMD "Explain" the Reduction in Fracture Risk?

Main methods

- Freidman
 - For individual data from a trial
 - Estimates p = proportion of treatment effect "explained" by *change* in the marker
 - β = coefficient for treatment
 - β^* = "adjusted for change in the marker"
 - $(1 - \beta^* / \beta)$

Does BMD “Explain” the Reduction in Fracture Risk?

For example

- RRR for tx = 0.5
 - Adjusted for BMD, RRR = 1.0
 - Explains 100%
 - Adjusted for BMD, RRR = 0.5
 - Explains 0%
 - Adjusted for BMD, RRR = 0.6
 - Explains ~15%

The Math: Li

- The Li method
 - (Percent of treatment effect explained; PTE)

$$\text{PTE} = \frac{1 - e^{(\beta_{\text{marker}} * \Delta \text{marker})}}{1 - e^{(\beta_{\text{marker}} * \Delta \text{marker} + \beta_{\text{treatment}})}}$$

- Confidence intervals are usually very wide
- Requires many outcomes and large effects

Does BMD “Explain” the Reduction in Fracture Risk?

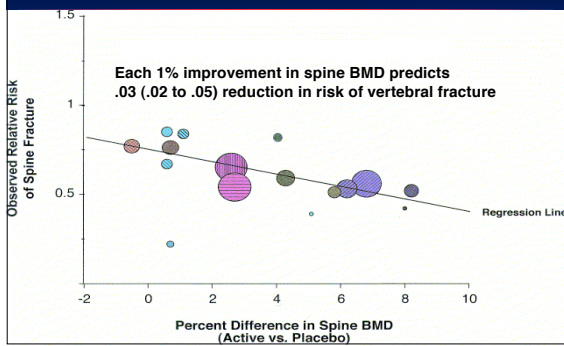
- This method applied to studies
 - FIT trial (alendronate): $p = 0.16$
 - MORE trial (raloxifene): $p = 0.05$
- Very little of the treatment effects are due to individual improvements in spine BMD, as measured by DXA.

Proving that a change in measurement predicts effect of treatment on fracture risk

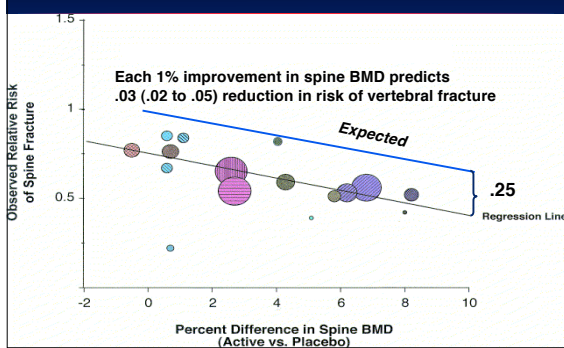
2nd approach

- Individual level
- Trial level
 - How well does the change in measurement predict the fracture results from trials?
 - ‘Meta-analysis’ of many trials

“Meta-analysis” of trials of antiresorptives



“Meta-analysis” of trials of antiresorptives



Implications

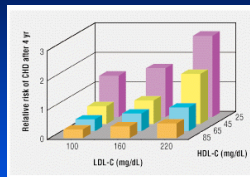
- You can't trust changes in BMD in your patients as an index of whether treatment is working.
- You can't trust that a drug that improves BMD will reduce the risk of fracture.
- FDA still requires fractures as the endpoint of trials for registering drugs.

Surrogate markers that failed

- Anti-arrhythmic drugs decreased the frequency of ventricular arrhythmias - and increased the risk of death (CAST Trial)

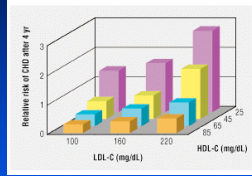
Torcetripib

- HDL-C predicts CHD
- Torcetripib increases HDL-C by 50-60%



Torcetripib

- HDL-C predicts CHD
- Torcetripib increases HDL-C 50-60%
- ILLUMINATE trial: Torcetripib + lipitor increased mortality and CVD events vs. lipitor alone



Surrogate markers that failed

- Anti-arrhythmic drugs decreased the frequency of ventricular arrhythmias - and increased the risk of death (CAST Trial)
- Torcetripib + lipitor (Pfizer) improved HDL cholesterol vs. lipitor alone but increased overall mortality
- Rosiglitazone improved fasting glucose and HgA1c levels but increased risk of CHD

Nasal culture for staphylococcal aureus

- Marker: + culture for staph in the nose
- Staph bacteria in the nose:
 - A source of bacteria for staph infections.
- Excellent predictor of postoperative wound infection with staph aureus
- Large trial of nasal mupirocin:

	Placebo	Mupirocin
+ staph culture	21%	5%
Staph infections	2.4%	2.3%

Biomarkers and safety

- Biomarkers markers are useful indices of safety when abnormal. Normal values provide limited confidence in the safety of a drug.
- Problems with reliance on biomarkers of 'safety' in trials of drugs:
 - Markers cover only a few systems.
 - Trials are often too small and too short to detect important adverse effects

Psaty, JAMA 2008;299:1474

Believing in biomarkers assumes that we understand pathophysiology

Limitations of biomarkers as indicators of effectiveness and safety are the main reason for relying on trials with clinical outcomes

Which Primary Outcome? Sample size

<i>Alternatives</i>	<i>Sample size/duration</i>
• Improvement in BMD	• 200 / 1 year
• Vertebral fractures	• 2,500 / 3 yrs
• Nonvertebral fracture	• 5,000 / 3 yrs
• Hip fractures	• 8,000 / 4 yrs

Wait a minute...

a bright idea...

Lets combine the endpoints!

	Rate per 3 yrs
• Nonvertebral fractures	12%
• Vertebral 'fractures'	4%
• Combination	16%

Likely effects of treatment

	% reduction
• Nonvertebral fractures	20%
• Vertebral 'fractures'	50%
• Combination	~25%

Pros and Cons of Composite Endpoints

Pro

- More events (smaller sample size?)
- Data about several outcomes

Con

- Heterogeneous biology
- Treatment may have different effects on each of the outcomes

Which Primary Outcome? Sample size

<i>Alternatives</i>	<i>Sample size/duration</i>
• Improvement in BMD	• 200 / 1 year
• Vertebral fractures	• 2,500 / 3 yrs
• Nonvertebral fracture	• 5,000 / 3 yrs
• Combined	• 2,000 / 3 yrs

Multiple Outcomes

FAT

Adding a second “co-primary” aim?

- For examples
 - Both vertebral fracture and hip fracture
 - Add a breast cancer endpoint
- Why “co-primary aims” of a study?
 - Increases the credibility of the 2nd result
 - FDA may require that an endpoint be a ‘primary aim’ to receive an indication for that outcome?
- Having 2 primary endpoints has a cost

Sharing Alpha

- Setting $p=.05$ for all of the primary outcomes of trial.
 - Result will be called “significant” (drug approved) if any result is “statistically significant” at the specified level

Adding breast cancer

- FDA requires “sharing” a total $\alpha = 0.05$
- Co-primary outcomes; power = 0.9 for both

	<u>alpha</u>
– Breast cancer	.048
– Vertebral deformities	.003
- Keeps the overall chance of a falsely ‘significant result’ at $P < 0.05$

Sharing Alpha

- Setting $p=.05$ for all of the primary outcomes of trial.
 - Result will be called “significant” (drug approved) if any result is “statistically significant” at the specified level
- Other approaches
 - Stepwise testing of outcomes: If vertebral fractures are reduced, then you can test for hip fracture...

Assessing Safety

Adverse Experiences: AEs

- Hugely expensive & time-consuming
 - May account for 1/4 of the expense of drug trials
- Poorly done
- Poorly studied

Issues re: Adverse Events

- Elicited vs. volunteered
- Nuisance AEs
- Attribution of cause
- MedRA system
- Validation of events

Open-ended interview

- Record any symptoms or conditions the subject has experienced:

- *What's wrong with this approach?*

Check list approach

“Since your last visit, has a doctor told you you had (check all that apply)”

A blood clot in the leg or lung (venous thrombosis)

An ulcer

...for all possible diseases

What's wrong with this approach?

Approaches to AEs Volunteered vs. elicited

- | | |
|--|---|
| Pro elicited/check list | Pro volunteered |
| <ul style="list-style-type: none">• More sensitive?• Easier to code | <ul style="list-style-type: none">• Catch unexpected AEs• Fewer AEs |
| Con: | Con: |
| <ul style="list-style-type: none">• More AEs | <ul style="list-style-type: none">• Hard to code; costly• Less sensitive for real adverse effects? |

STEP Trial

- Randomized comparison of open-ended vs. open-ended (at least 1 day limited activity) vs. check list for adverse events
- 70 men in each group
- Treatment had no effect on AEs

STEP Trial

	# of AE reports
• Open-ended	11
• Open-ended (limited activity)	14
• Check list	214

An approach?

- Standardized questions to elicit AEs suspected to be related to drug.
- Open ended questions to capture other AEs.

FDA AE classifications

- Serious AEs
 - Deaths
 - Hospitalized (or prolonged stay)
 - Cancer (except skin cancer)
 - Birth defects
- Reported within 24 hours
- Collect documentation

FDA AE classifications

- Non-serious AEs
 - Anything symptom or clinical event
 - Regardless of importance or potential relevance

The Bunion Problem in Large Trials

- FIT Trial of alendronate in 6,400 women for 4 years
- Recorded over 20,000 episodes of URIs (and thousands of reports of bunions!)
- Enormous data management effort and cost
- How could this be avoided?

An approach to 'nuisance' AEs

- Limit collection of non-serious AEs to samples of subjects
 - 1st 300, then stop collecting
- Collect all other AEs (and SAEs)
- New FDA policy
 - After initial small trials, no need to collect AEs

Attribution

- SAEs must be classified as
 - Definitely
 - Probably
 - Possibly, or
 - Not...
- ...related to the study drug

Did the treatment cause the AE?

- 67 year old Morbone volunteer starts taking the study drug.
- She reports a pruritic blistering rash on her arms that started 7 days after starting and disappeared 2 days after stopping the drug.
- Your attribution?

Attribution

- Attributions to the drug are usually wrong
 - (Unless the treatment is known to cause the effect)

The MedRA system

- Commercial system used by all trials for FDA approval
- Several types of terms
 - Preferred terms (pneumococcal pneumonia, cellulitis): what the MD writes
 - Higher order terms (skin infections)
 - Group terms (pulmonary disease)

The MedRA *morass*

- Commercial system used by all trials for FDA approval
- Several types of terms
 - Preferred terms (pneumococcal pneumonia, cellulitis): what the MD writes
 - Higher order terms (skin infections)
 - Group terms (pulmonary disease)

The MedRA *morass*

- May miss real adverse effects because of how they are coded by sites
- Event: acute pneumococcal lobar pneumonia can be classified as
 - Pneumonia
 - Pneumococcal pneumonia, or
 - Acute pneumonia, or
 - Lobar pneumonia, or
- Analyzed separately: no chance of finding that pneumonia is a significant effect

The MedRA *morass* A real example

- Drug D SAEs

	Placebo	Active D	P-value
Cellulitis	1	7	0.07
Erysipelas	0	6	0.06
Combined	1	13	<0.01

Conclusion: "D *increases* the risk of serious skin infections"

Adjudication

- Self report is sometimes inaccurate
- Adjudication
 - Prospective definition of a case
 - Systematically collect essential medical information
 - Experts classify each case
- Expensive process
 - Collection of records
 - Central adjudication by experts
- Use for important conditions and plausibly related to the treatment

Summary Assessing AEs in the Morbidity Trial

- Elicit (check list)
 - Hospitalizations, birth defects..
 - Disability, new prescriptions and procedures
- Open-ended collection of minor AEs
 - (May not be necessary)
- No attribution
- The MedRA system
 - Review MedRA terms in advance to create medically and biologically meaningful groupings

Summary

- Choosing a primary outcome is the most important decision
- Ideally, choose a clinical outcome or 'validated' surrogates. Usually not feasible.
- Composite outcomes sometimes improve statistical power but can blur heterogeneity
- Focus your effort on careful assessment of serious or important adverse events
 - Minimize effort spent on minor AEs
