Assessing Outcomes

Steven R. Cummings, MD

Director, San Francisco Coordinating Center
California Pacific Medical Center Research Institute
Professor of Medicine, Epidemiology, and Biostatistics, UCSF

Outline: Outcomes

Part 1 - Endpoints

- Primary and secondary aims
- Biomarkers and “surrogate markers”
- Co-primary endpoints
- Composite endpoints
- Phases of trials
- Summary

Morbone

A new (fictitious) treatment for osteoporosis

- Morbone improves bone mass and bone strength in animal models
- Early phase trials in postmenopausal women showed that it may improve bone density (BMD)
- The company is planning a trial to obtain FDA and EMA approval to market Morbone
- The company wants your advice

The most important decision

- What will be the primary endpoint?

The most important decision

- What will be the primary endpoint?

Possibilities

- Improved bone density (BMD)
- Vertebral fractures (by spine x-rays)
- Nonvertebral fractures
- Hip fractures

Why one primary outcome?

- To calculate sample size
- To test statistical significance of the result without a penalty
- The FDA requires that a primary outcome to approve a drug for that indication
  - Beneficial effects on secondary outcomes cannot be indications for prescribing the drug

Which primary endpoint for the Morbone trial?

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

Considerations in choosing the primary outcome

- **Clinical importance**
- Feasibility
  - Sample size and cost
- Regulatory requirements

Clinical importance

- Hip fracture: causes death, disability, and most of the costs of fractures
- Nonvertebral fractures: common, cause disability – less then hip fracture
- Vertebral fracture by x-ray: about 1/3 cause cause pain and disability
- Bone mineral density (BMD): no symptoms but low BMD causes an increased risk of fracture

Considerations in choosing the primary outcome

- Clinical importance
- **Feasibility**
  - **Sample size and cost**
- Regulatory requirements

Sample size depends on incidence of the endpoint and the efficacy of the treatment

Sample size for the trial

Alternatives	*Sample size*
• Improvement in BMD	• 200 / 2 yrs
• Vertebral fractures	• 2,500 / 3 yrs
• Nonvertebral fracture	• 5,000 / 3 yrs
• Hip fractures	• 8,000 / 4 yrs

Sample size and cost

Alternatives	*Sample size*	*Cost*
• Improvement in BMD	• 200 / 2 yrs	$5M
• Vertebral fractures	• 2,500 / 3 yrs	$50M
• Nonvertebral fracture	• 5,000 / 3 yrs	$100M
• Hip fractures	• 8,000 / 4 yrs	$200M

Why not make BMD the primary outcome?

- The best choice to minimize cost
- The issue: is it a valid “surrogate marker” for the clinical outcome, reduction in risk of fracture?

Definitions

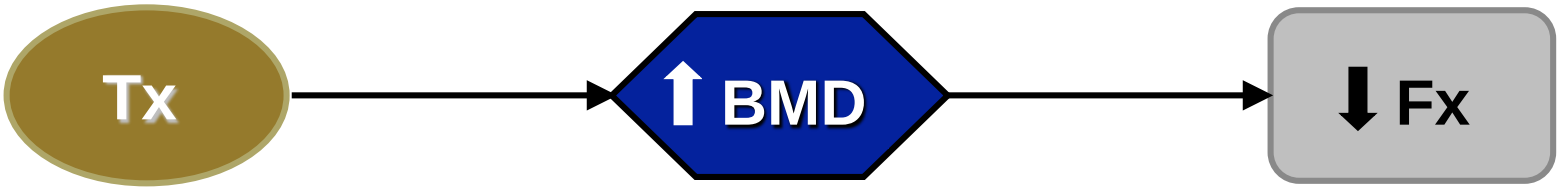
- Not all markers are ‘surrogates’
- Biomarkers
  - Measurement of a process or state.
- Surrogate marker
  - Often misused
  - It can substitute for clinical outcome
  - The of treatment on marker has been established to predict the clinical outcome.

Criteria for validating a surrogate marker for treatments

- Biologically plausible
  - It is the pathway by which treatment improves the clinical outcome
- The marker strongly predicts the clinical outcome
- Treatment changes the marker

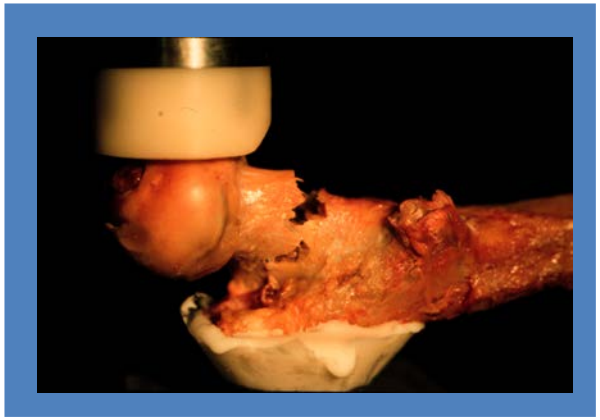
** Prentice, 1989*

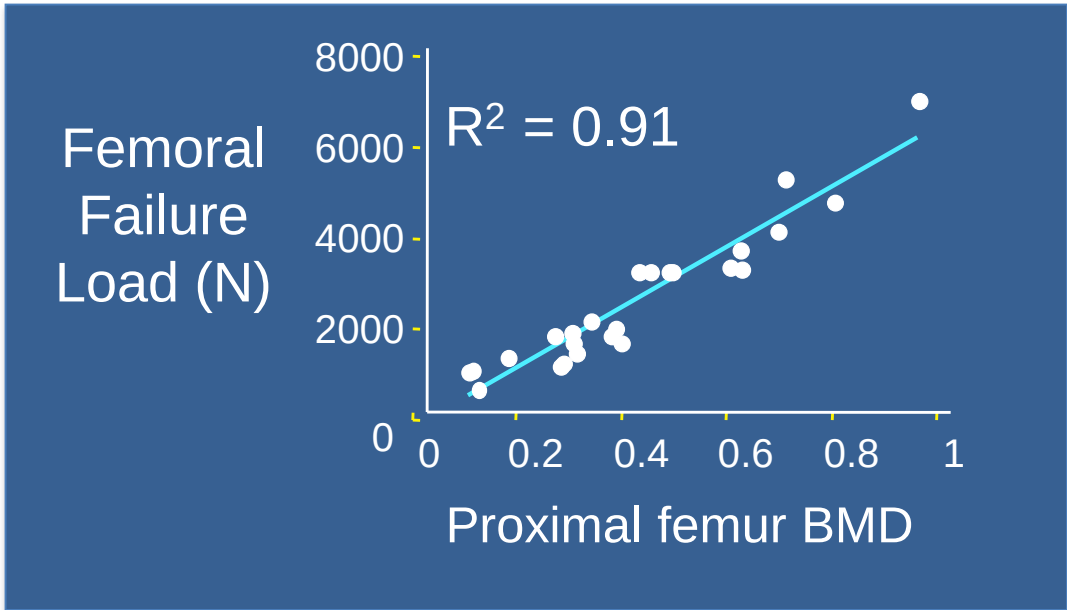
The perfect surrogate is *the* causal pathway by which treatment works



Is BMD a biologically plausible "surrogate" marker?

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



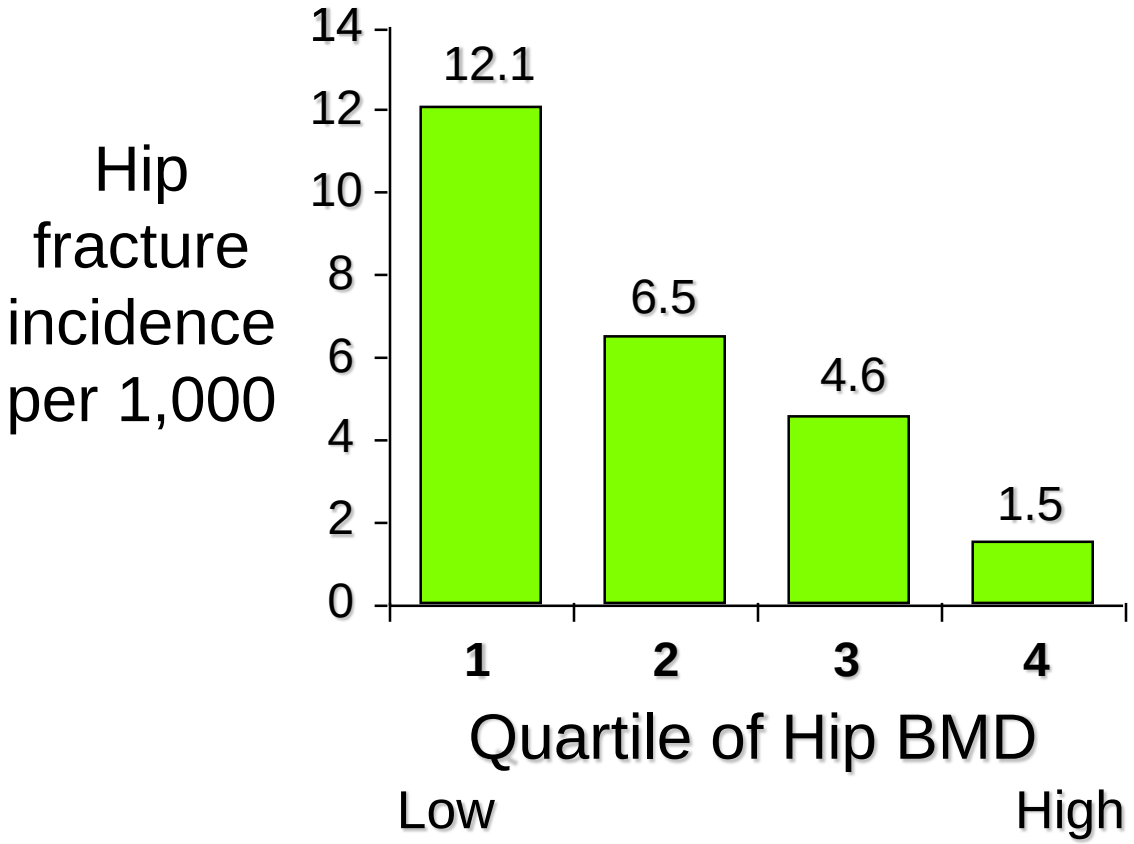


Criteria for validating a surrogate marker for treatments

- Biologically plausible
- The marker strongly predicts the clinical outcome

** Prentice, 1989*

Observational studies: BMD strongly predicts fracture

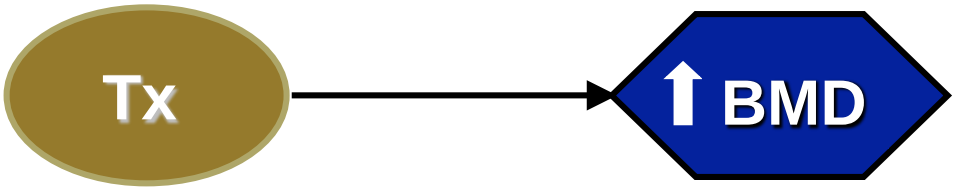


Bone density

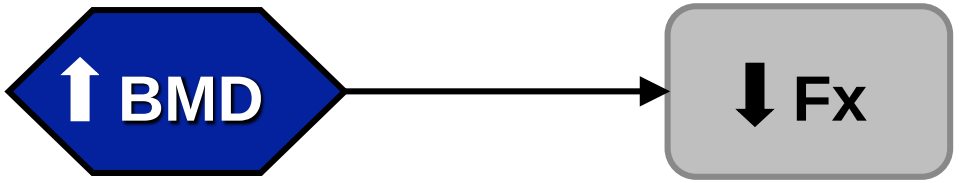
- Biologically plausible: YES
- Marker strongly predicts the clinical outcome: YES
- Treatment changes the marker: YES
  - Treatments improve hip BMD 3-5%

** Prentice, 1989*

Treatment improves BMD



BMD predicts fracture



Is BMD a good surrogate?

- BMD is biologically plausible, predicts fracture, and responds to treatment
- Anything else?

Is BMD a good surrogate?

- The definition of a true surrogate
- Does the change in BMD accurately predict the change in risk of fracture by treatment

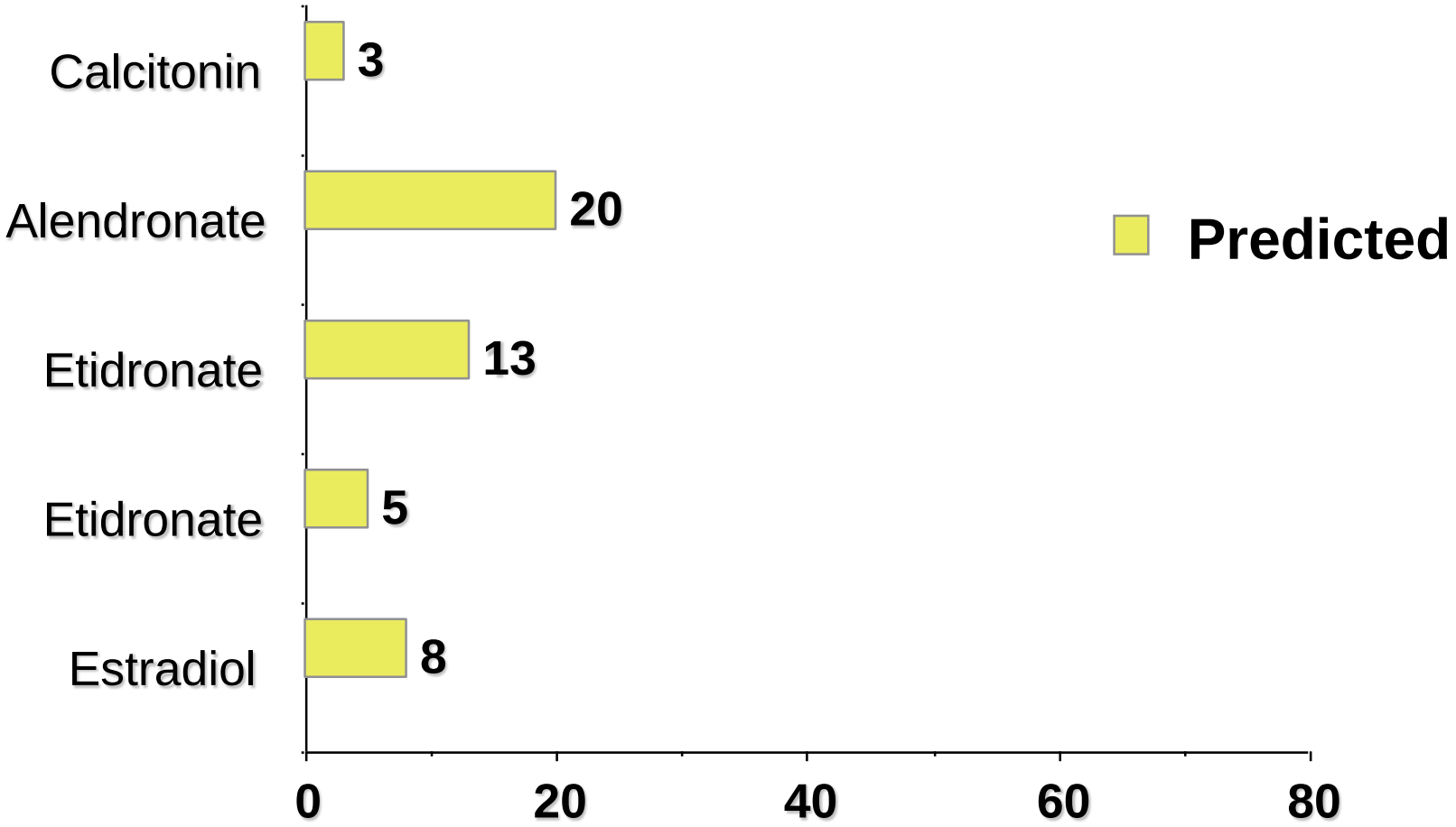
** Prentice, 1989*

Sodium Flouride

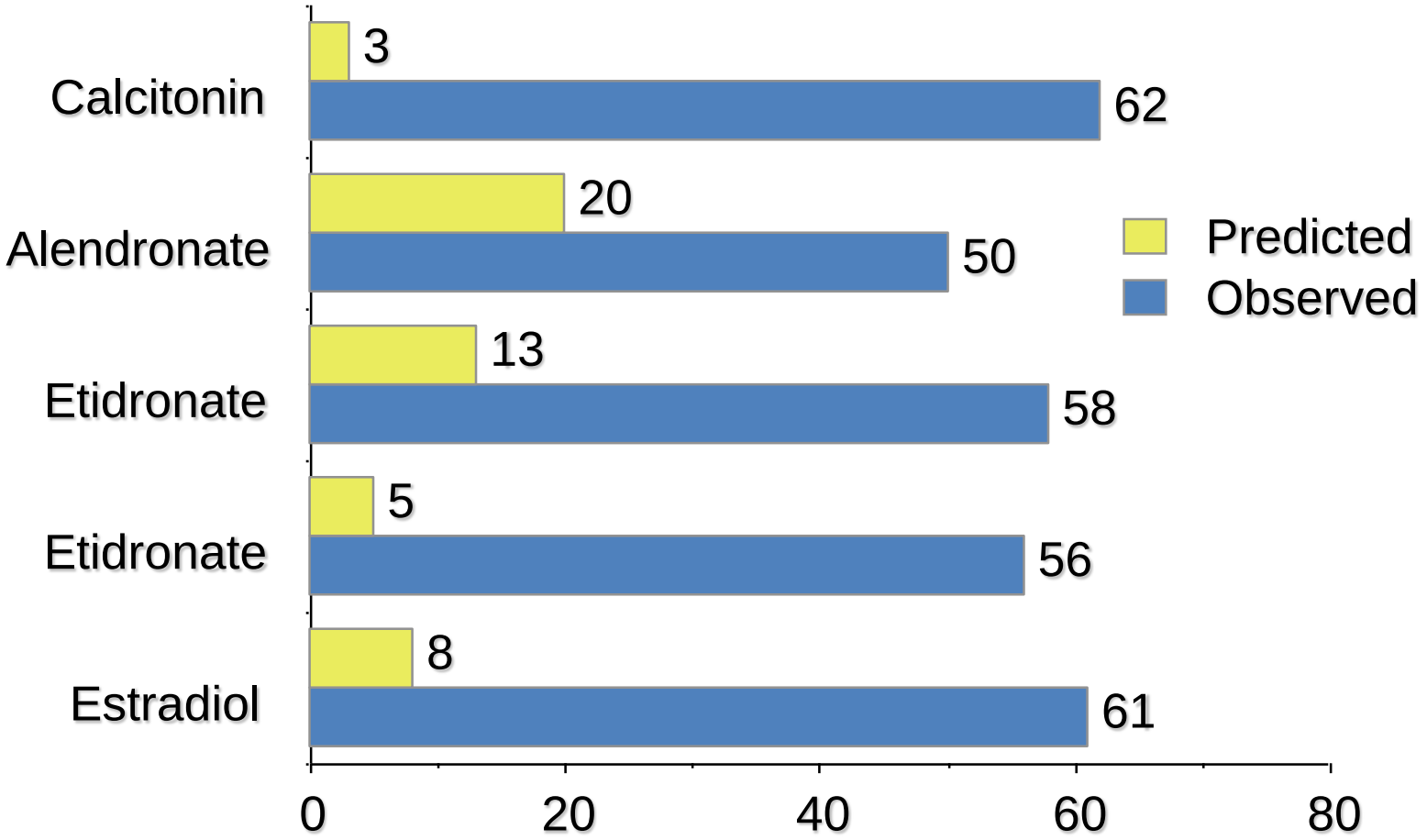
- A treatment used in the 1960s
- Increased BMD ~10%
- But in a randomized trial, it increased the risk of fractures!

** Prentice, 1989*

Decreased vertebral fracture risk (%): Predicted from change in spine BMD

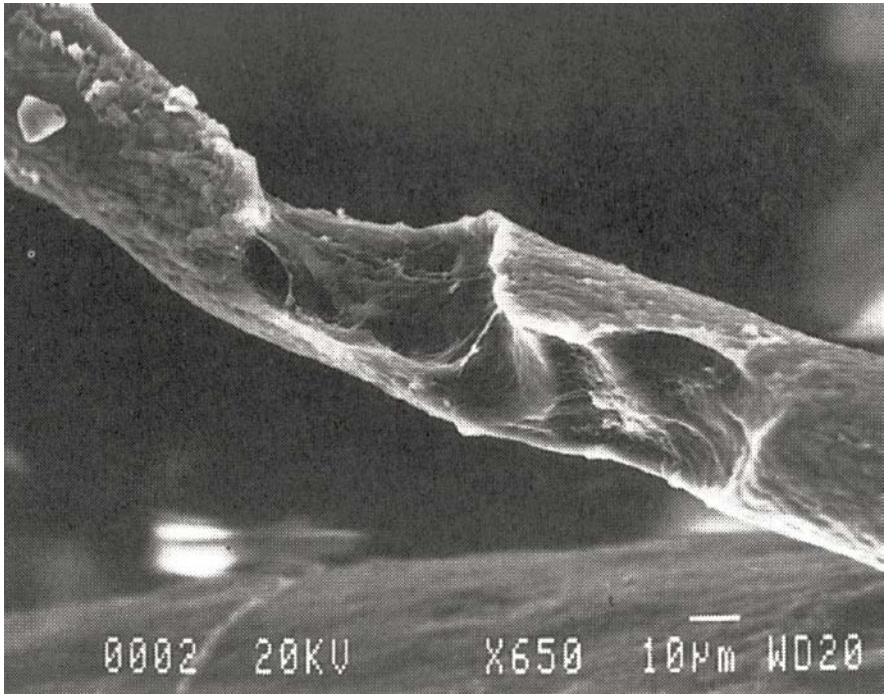


Decreased vertebral fracture risk (%): Predicted from BMD vs. observed



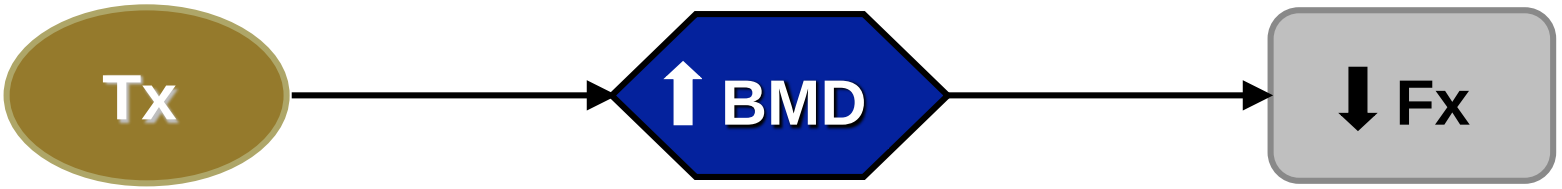
Why?

Resorption weakens bone more than measured by BMD

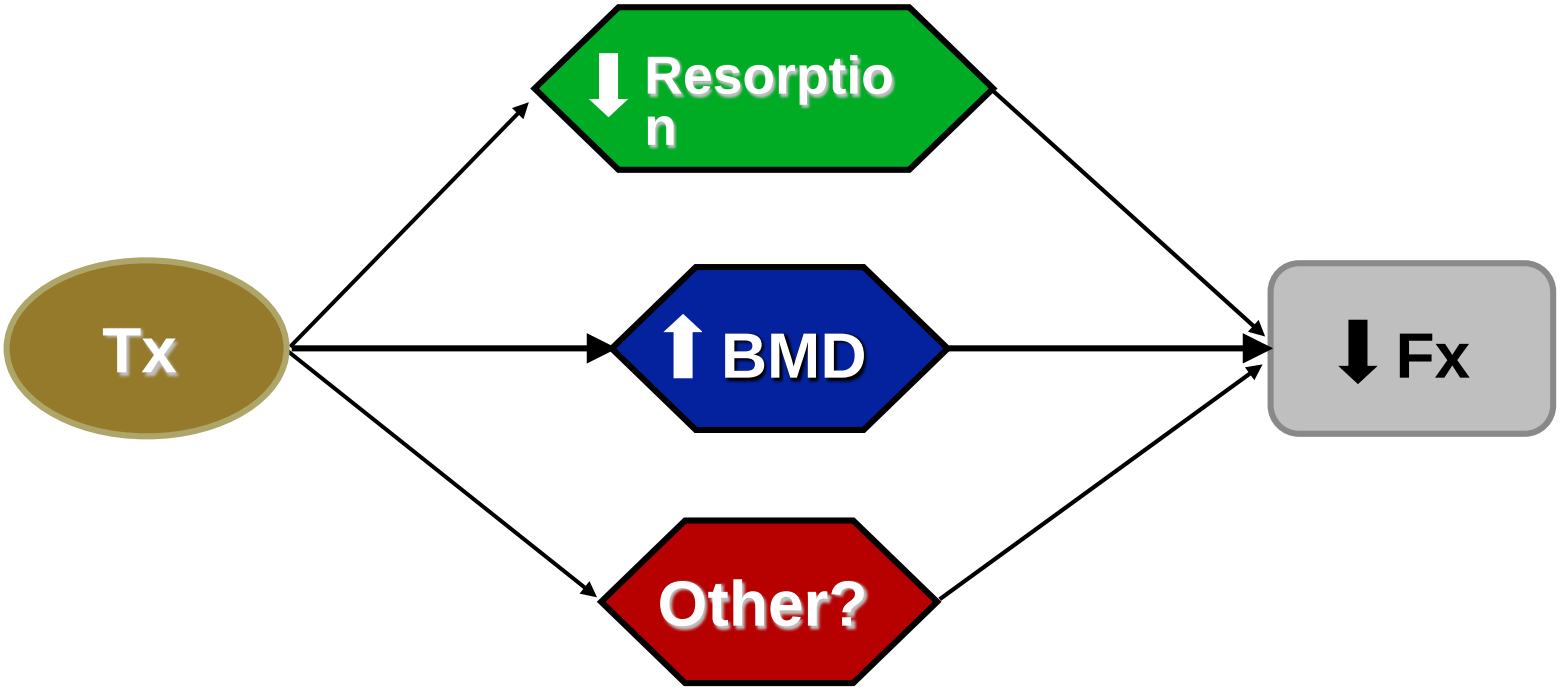


Treatments for osteoporosis decrease resorption
This improves strength more than BMD

The perfect surrogate is *the* causal pathway by which treatment works



The treatments works by other pathways



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

Showing that a change a marker predicts effect of treatment on fracture risk

Two approaches

- Individual level
  - How well does *change* in the marker account for the effect in individual people?

- Trial level
  - How well does the *change* in measurement predict the clinical results from trials?

Validation

- 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
  - Percent of Treatment Explained: PTE

How to estimate PTE

Simplistically:

How much does statistical adjustment for change in the marker (BMD) ‘explain’ the decreased risk of the outcome (fracture)?

See: Li Z, et al. Stat Med. 2001;20:3175-3188.

A simple example

- Treatment reduces risk by 50%

Multivariate model

	Risk reduction
Treatment	0.5

A simple example

- Treatment reduces risk by 50%

Multivariate model

	Risk reduction
Treatment	0.5
Adjusted for BMD	

- Adding BMD does not change the association between treatment and decreased fracture
- Change in BMD explains 0% of the effect
- PTE = 0%

A simple example

- Treatment reduces risk by 50%

Multivariate model

	Risk reduction
Treatment	1.0
Adjusted for BMD	

- Adjusting for BMD -> no association between treatment and decreased risk of fracture
- Change in BMD explains 100% of the effect
- PTE = 100%

State of the Art

- More sophisticated approaches
- Modeling and graphic representation
- Biostatistical expertise

Does BMD explain the effect of treatments on fracture risk?

- Alendronate PTE = 16%
- Raloxifene PTE = 5%
- Regulatory bodies, EMA and FDA require reduction in fracture risk, not just change in BMD, to approve a treatment for osteoporosis

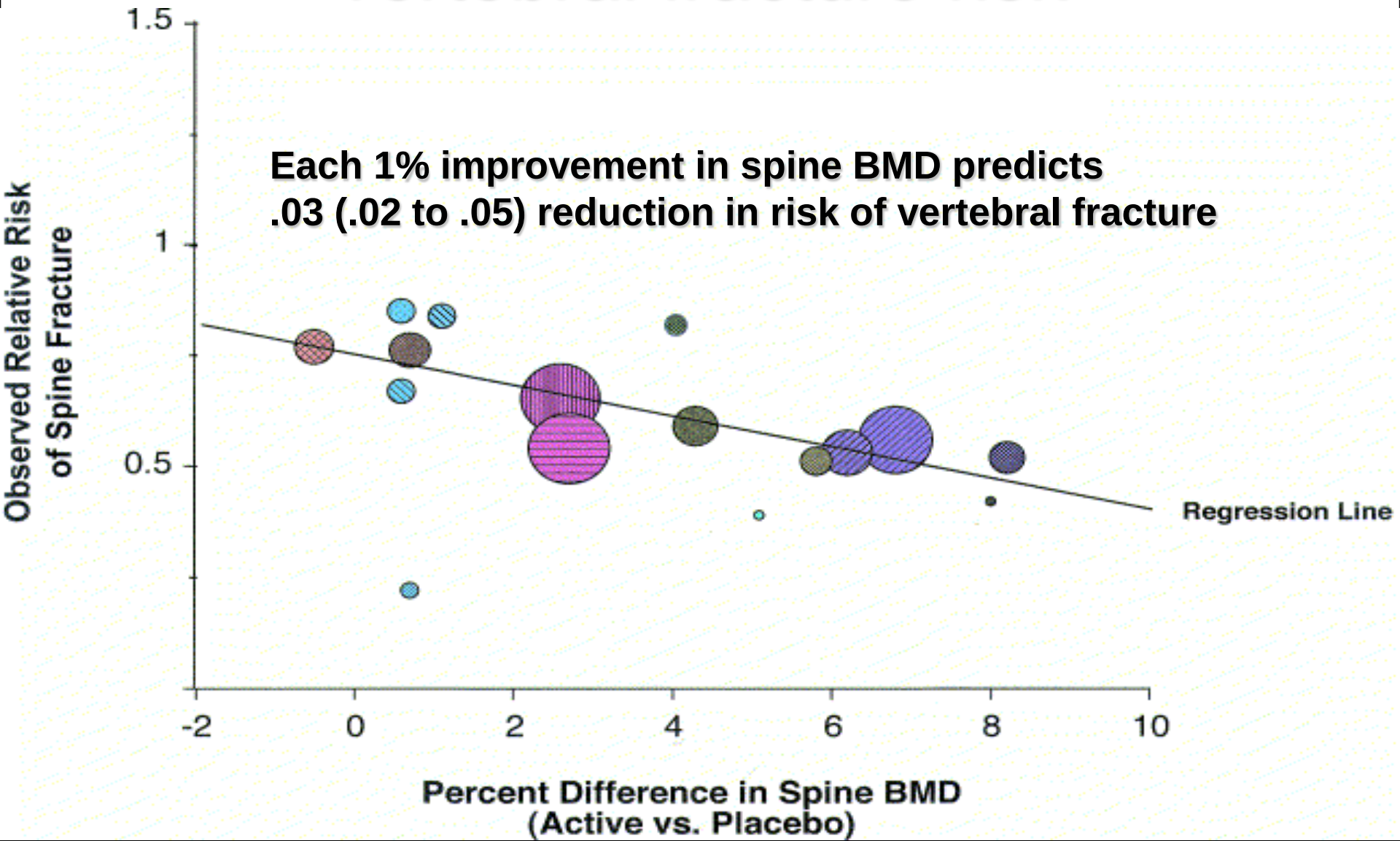
Implications

- The effect of treatment on BMD in patients is a poor predictor of the effect on risk of fractures

A surrogate marker for the effects in trials?

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

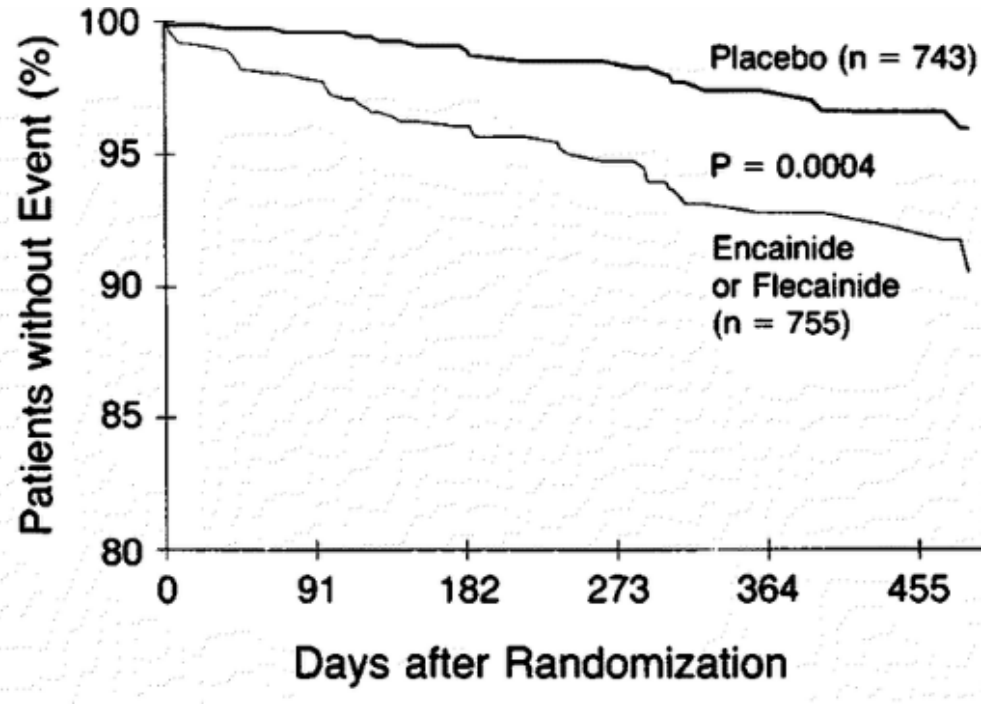
Meta-regression of trials for vertebral fracture risk



Cardiac Arrhythmia Suppression Trial (CAST)

- Anti-arrhythmic drugs decreased the frequency of ventricular arrythmias

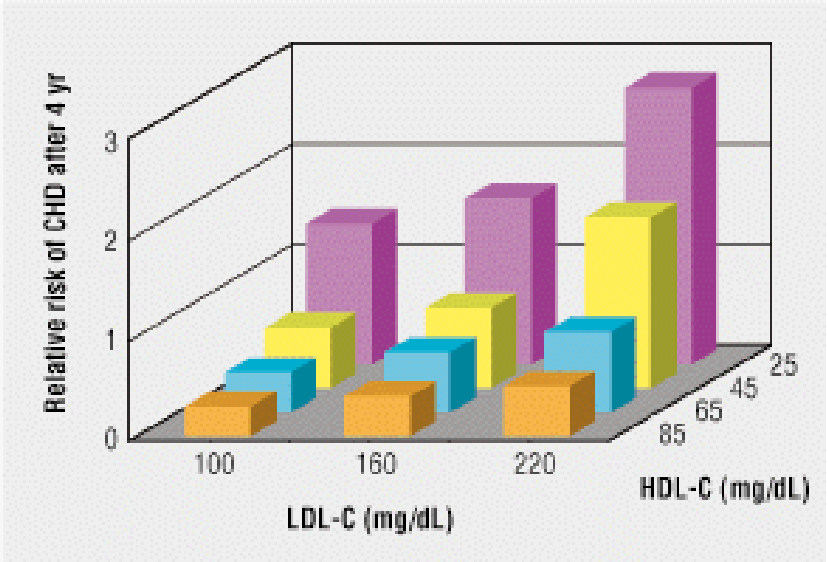
- Increased the risk of death



Echt, et al. N Engl J Med 1991; 324:781-788

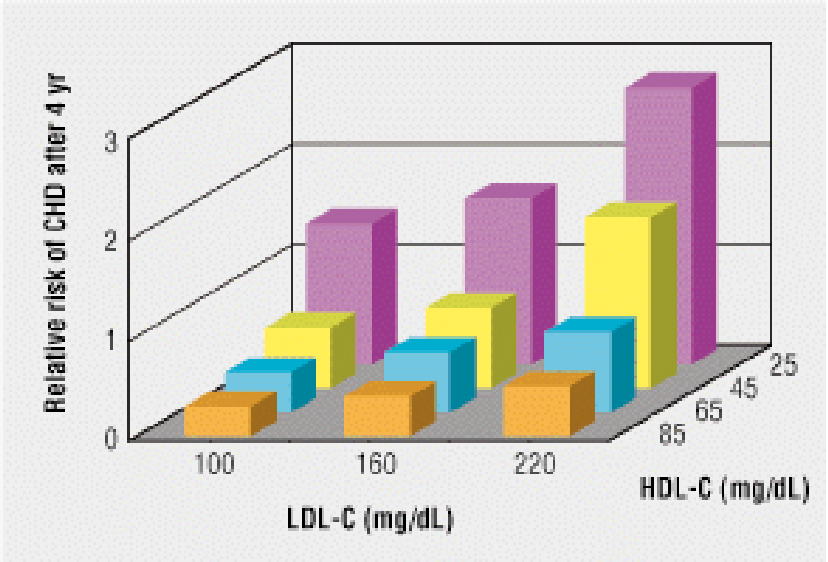
Torcetripib

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



Torcetripib

- HDL-C predicts CHD
- Torcetrapib increases HDL-C 50-60%
- ILLUMINATE trial: Torcetrapib + lipitor vs. lipitor alone *increased* mortality



Surrogate markers and humility

- Believing in a ‘surrogate marker’ 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

But trials with biomarker endpoints are useful and important

- Some biomarkers, such as blood pressure, are good surrogate markers
- Marker endpoints are an essential, affordable and good first step before undertaking more expensive trials with clinical endpoints.
- For some conditions, biomarkers are the only feasible endpoint for testing treatments

Morbone

A new (fictitious) treatment for osteoporosis

- Early phase trials in postmenopausal women showed that it may improve bone density (BMD)
- The company is planning a trial to obtain EMA and FDA approval to market Morbone
- The company wants your advice

Which primary outcome for the Morbone Trial?

Alternatives

- ~~Improvement in BMD~~
- Vertebral fractures
- Nonvertebral fracture
- Hip fractures

Sample size/duration

- 200 / 2 years
- 2,500 / 3 yrs
- 5,000 / 3 yrs
- 8,000 / 4 yrs

Change in BMD is not accepted as a valid surrogate marker

Composite endpoints

Combine the endpoints?

- Decrease the size and cost by increasing the number of endpoints

	Rate per 3 yrs
• Nonvertebral fractures	12%
• Vertebral ‘fractures’	4%
• Composite	16%

Sample size also depends on the effect size

	Expected % reduction
• Nonvertebral fractures	20%
• Vertebral ‘fractures’	50%
• Composite endpoint	~25%

Pros and Cons of Composite Endpoints

Pro

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

Con

- May have heterogeneous biology
- Treatment may have different effects on each of the outcomes
  - Might lose statistical power

Which Primary Outcome? Sample size

Alternatives	*Sample size/duration*
~~• Improvement in BMD~~	• 200 / 1 year
• Vertebral fractures	• 2,500 / 3 yrs
• Nonvertebral fracture	• 5,000 / 3 yrs
• *Composite*	• *2,000 / 3 yrs*

Little gain in power or reduction in cost

The trump card

EMA (and FDA)

Which Primary Outcome? What EMA requires

Alternatives

- ~~Improvement in BMD~~
- **Vertebral ‘fractures’**
- **Nonvertebral fracture**
- Hip fractures

Sample size/duration

- 200 / 1 year
- **2,500 / 3 yrs**
- **5,000 / 3 yrs**
- 8,000 / 4 yrs

Morbone may have other benefits

- Raloxifene for osteoporosis partially blocked the estrogen receptor
- A trial for vertebral fracture reduction found a 70% reduction in the secondary endpoint of risk of ER+ breast cancer
- FDA did not allow breast cancer as an indication for treatment
- A second trial was needed

Co-primary endpoints

- Allows testing more than one potential benefit of a treatment
- What is “statistically significant?”
- Two main approaches
  - Splitting alpha
  - Gated, or stepwise, testing of statistical significance (not covered)

“Sharing alpha”

- The overall level of statistical significance for “benefit” is maintained at $P<0.05$
- The 0.05 can be divided, in an additive way, between endpoints
- Vertebral fracture: significant if $P<0.04$
  - More lenient it requires less power to achieve that level
- Breast cancer: significant if $P<0.01$
  - More stringent because it requires less power

Phases

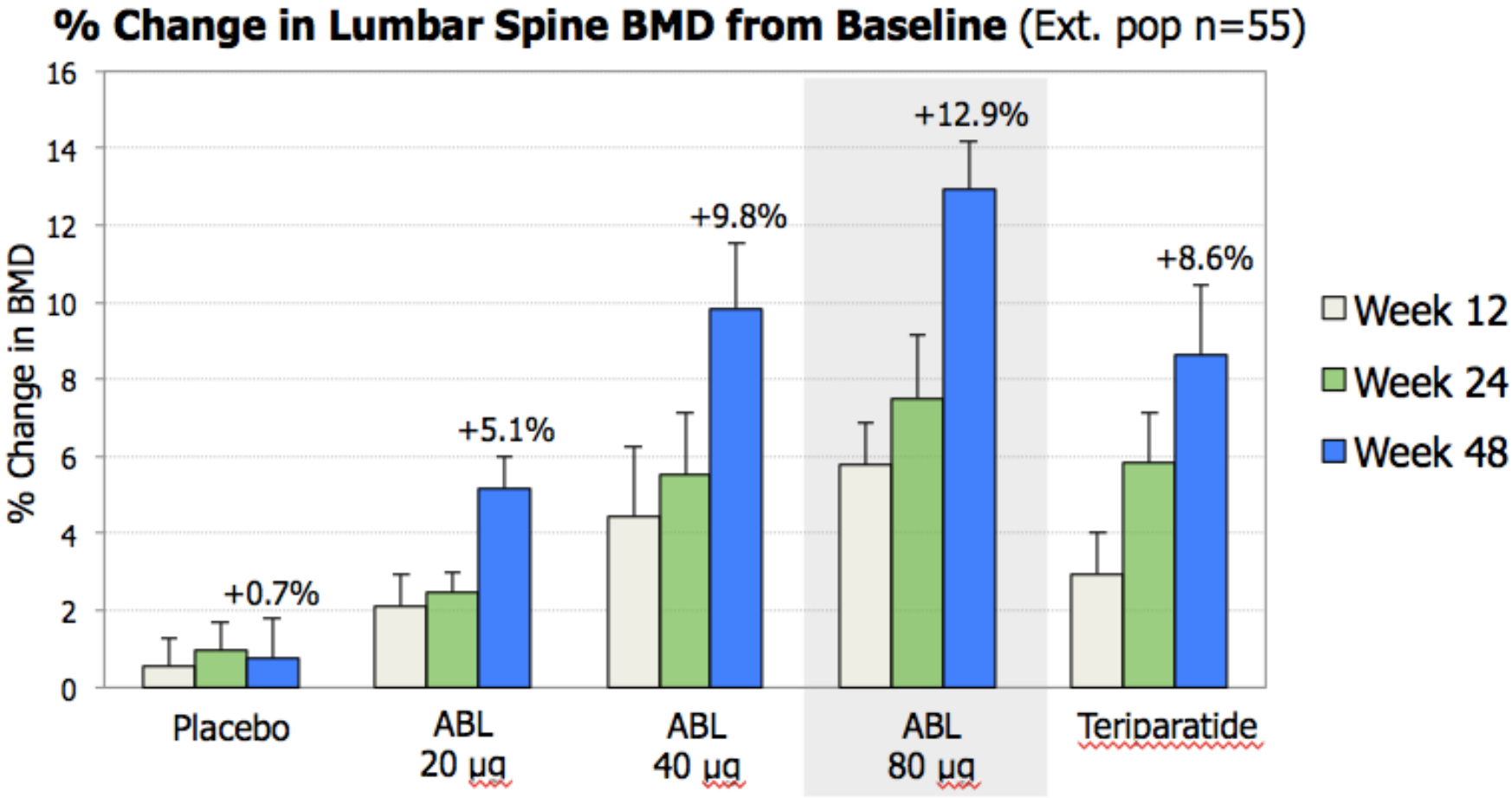
Phases

- Phase 1: “First in humans”
- First administration of the drug to human volunteers. Generally in special units.
- Healthy people
- Dose may be escalated to the highest potential level
- Endpoints:
  - Symptoms (tolerability)
  - Many markers of potential efficacy and adverse effects

Phases

- Phase 2: (usually) randomized trials of effect of the drug on biomarkers of efficacy and safety.
- Usually tests several doses and includes PK/PD studies to identify the best dose
- For example, placebo vs. 3 potential doses of Morbone

Effect of doses of Morbone compared with placebo and Teriparatide



Phases

- Phase 3: randomized trials powered to detect the effect on a clinical endpoint required by FDA for approval to prescribe
- Morbone trial with vertebral fracture endpoint
- Adverse event data collected to determine the ‘safety’ of treatment
- Results reviewed in detail by FDA and an advisory panel to decide on approval

Phases

- Phase 4: studies with the FDA-approved drug for new indications or safety
- May be ‘open label’ or without controls
- For example, Morbone tested in patients taking corticosteroids or in men
- Or an extension, without placebo, for 5 years after the main placebo–controlled trial ends
- For example, an obesity drug tested in a very large trial for incidence of CVD events

Outcomes: Summary

- Selecting a primary outcome is the most important decision
- Trials with biomarker endpoints may be valuable building blocks, essential for early phase trials. They sometimes mislead about the effect on clinical outcomes.
- Composite endpoints might reduce the size and cost of trials but risk heterogeneous results

Summary

- Choosing a primary outcome is the most important decision in designing a trial
- Ideally, choose a clinical outcome
  - Trials with ‘marker’ endpoints are useful building blocks, for example, Phase 1 and 2 trials
- Composite outcomes may improve power if the treatment has similar effects on the components
- An approach to AEs:
  - Query for important and likely AEs
  - Solicit others by open ended questions

Assessing Safety

AE classifications

- Serious Adverse Events (SAEs)
  - Deaths
  - Hospitalized (or prolonged stay)
  - Cancer (except skin cancer)
  - Birth defects
- Others: Adverse events (AE)

Outline

- Classifying Adverse Events (AEs)
- How to elicit AEs
- Coding AEs
- Adjudicating AEs

Adverse Experiences: AEs

- Very expensive & time-consuming
  - Collecting and monitoring AEs can account for 20-25% of the cost of trials
- Little research about the best approaches

Morbone Trial

- Standard approach:

  ‘Record any symptoms or conditions the subject has experienced:

  _____

  _____

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**

- More sensitive?
- Easier to code

Con:

- More AEs

Pro **volunteered**

- Catch unexpected AEs
- Fewer AEs

Con:

- 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
• Check list	214

Example: MeDRA System

- A standard dictionary for coding terms
- Categorized at 4 levels, from system (Pulmonary) to lower level (pneumococcal pneumonia).
- AE of “pneumococcal pneumonia” will also be categorized as pneumonia and pulmonary AE

Example: MeDRA System

- Usually accurate, but can occasionally be misleading
  - Cellulitis and impetigo: same condition, separate codes
- Needs medical judgment to group terms
  - Review the AEs

Expert Adjudication

- Expert or panel of experts
  - Neurologists for stroke
- Collect medical records for the event
  - Neurologic exam?
  - Brain MRI or CT
- One or more expert reviews the records and makes a diagnosis

Expert Adjudication

- Expensive and time consuming
- Expensive process
  - Consider for important AEs
- Does it improve the accuracy?
  - Depends on the event
    - Hip fracture: accurate, no need for adjudication?
    - Rib fracture: often wrong – needs adjudication
    - Heart attack / MI: may need expert adjudication

Summary

- Open ended questions to capture AEs
- Consider a check-list for AEs you don’t want to miss and likely to be related to the drug
- If the potential AE is very important and sometimes misclassified, then adjudicate by experts