# Research with older people
# Problems and pitfalls

Leon Flicker.

Western Australian Centre for Health & Ageing,
Schools of Medicine & Pharmacology, University of Western Australia
Department of Geriatric Medicine Royal Perth Hospital

WESTERN AUSTRALIAN CENTRE FOR HEALTH AGEING

# Why Research?

- Curiosity, Investigator driven research
- We do research because we are driven to do it – we like finding answers to the questions that interest us.
- The wider community helps us by providing us with
  - 1) Interesting questions
  - 2) Resources to answer these questions
- We, the researchers are in a privileged position, to be able to follow our ideas rather than dwell on our careers.
- We should remember that ageing research is comparatively hopelessly under funded.
- We do need more ageing research in many disciplines, spanning all major interests of human endeavour.
- We need to translate research evidence into policy for Practical, Strategic and Moral reasons

# What distinguishes Standard Medical Practice from Alternative Medicine?

To quote Tim Minchin,

"By definition ... alternative medicine ... has either not been proved to work, or has been proved not to work. You know what they call alternative medicine that's been proved to work? Medicine."

# Which then leads to the question of "proof" and "science"

"The criterion of the scientific status of a theory is its falsifiability, or refutability, or testability." Karl Popper

We never prove anything, we fail to disprove a theory. LF

"No amount of experimentation can ever prove me right; a single experiment can prove me wrong."

Albert Einstein

"Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem which it was intended to solve."

Karl Popper

WESTERN AUSTRALIAN CENTRE FOR HEALTH AGEING

WHAT IS THE QUESTION?

# Why study humans?? Why not *hard sciences*?

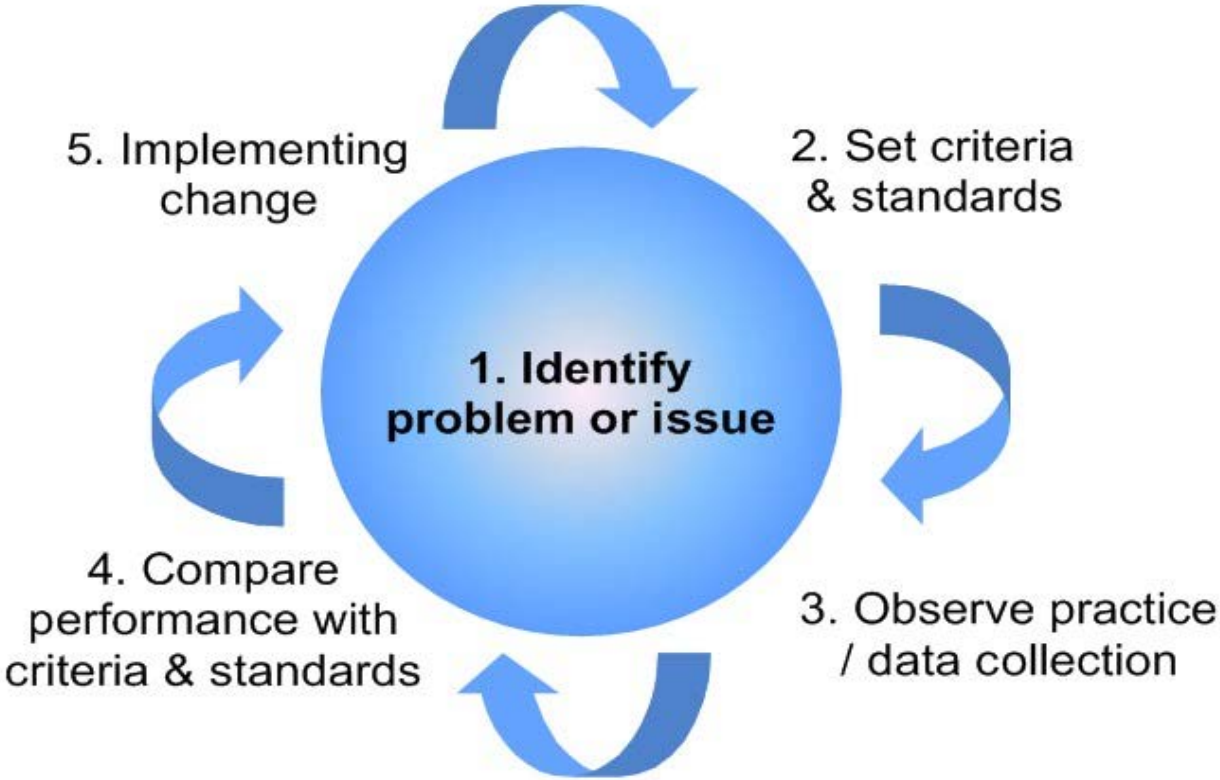
1) Benefits with genes don't necessarily translate to cells which don't necessarily translate into intact animals

2) Risks cannot be determined from studies on other than humans. In particular carcinogenesis is usually species specific.

3) The treatment of human beings usually require placing an individual in wider socio-cultural context and increasingly brings in questions of resource utilisation. For information on treating people in the real world we require studies with human beings.

# Quality Improvement

- Clinical audit is a process that has been defined as "a quality improvement process that seeks to improve patient care and outcomes through systematic review of care against explicit criteria and the implementation of change" What you are doing is asking the question "does my practice meet these criteria?




IAN CENTRE FOR AGEING

Broadly, research in humans has relied on the evidence from a hierarchy of study designs.

However, different methodologies have specific uses – e.g. when you are trying to work out barriers to uptake of interventions qualitative methods are usually optimal

## Types of study designs

(increasing validity or trust)

- Description – often using qualitative methods
- Case-control
- Cohort
- Randomized (control) (clinical) Trials

## Evaluation of Diagnostic Tests

# Case or problem description

- Has been the basis of hypothesis generation for over 2000 years.
- Relies on strong (gross) or obvious effects.
- Tends to be better in the sensational rather than the common.
- Often achieves more credibility as a case series than as an individual report but watch out for selection bias
- Inevitably retrospective in type.

# Case Control

- Allows a look at the effect of specific exposures on the incidence of disease or problem.
- Usually very efficient in the number of subjects needed (often only 1 or 2 controls per case).
- Allows us to look at the effects of exposures which may be unethical to explore in a more experimental type of design. (dementia and head injury)
- Major problem is CONFOUNDING. This is where you are misled because of the strong association between 2 exposures and the other exposure is actually involved in the causation of disease. Eg Lung cancer with coffee drinking could be due to smoking.

# Case Control

- Confounding can be dealt with by stratification or statistical modelling but with both methods you need more numbers
- Because you identify cases don't have to wait around for the disease to happen.
- Generally credibility is not high and if the effect is small (OR or RR <2) people will want more evidence.
- Recall bias etc
- Has been used to look at the effects of drugs eg benzos and hip fractures
- Usually selection of the controls is the major difficulty - must look like the cases but not have the disease or a disease associated with it.

# Cohort

- The concept is of a group of individuals marching through time.
- Several situations where it may be unethical to expose individuals to a particular exposure eg Workers in the petroleum industry, survivors of Hiroshima, children and lead.
- You are usually comparing rates of disease and so need to know the background rates i.e. in the control population.
- Because of the prospective nature of this type of study there is no recall bias and you have greater possibility of defining potential confounders with more precision (still the problem of the confounders you don't think about)

# Cohort

- Often there are big time periods – e.g. cognitive changes with ageing

- Has been used to observe the beneficial and deleterious effects of drugs. Eg bezos and hip fractures
- Always the possibility of selection bias (which really is a form of confounding) eg Breast cancer is associated with higher SES and HRT tends to be taken by women with higher SES and therefore may produce an association. Children of lower SES are exposed to more lead and may produce the decrease in cognitive functioning

# Randomised Clinical Trial

- This is the definitive study design for experimental proof in humans.
- Randomization produces a balance between groups for all the potential confounders - both the ones we know about and the ones we don't know about. No selection bias either intrinsic to the individual or extrinsic
- The problems with RCT are logistics and ethics.

# Randomised Clinical Trials

## ETHICS

- You can't expose humans to things you think may be bad for them
- Therefore you can only evaluate potentially beneficial agents eg new drugs
- However you can't deny humans any treatment which you think is beneficial for them
- This leaves only a small window of opportunity where you have access to an agent which you think may be beneficial but you are not sure - equipoise.
- People will alter ethics to suit their arguments

# Randomised Clinical Trials

## LOGISTICS

- Always difficult to recruit
- You have to level with the patient and tell them you don't know which treatment is the best (the usual answer is to ask you to send them to somebody who does know how to treat this condition).
- Adaptive design
- You try to get an homogenous group and end up excluding 80% of potential subjects

# Randomised Clinical Trials

## LOGISTICS

- The treatment effect you may be looking for may be small but important eg CVS disease after a stroke or heart attack need several thousand to observe a reduction in mortality of 25%

- You may do a lovely study and find out it only applies to a few people with this condition.

- Expensive

- Nobody wants to know about negative results

# Types of Trials

## Primary Prevention

- vaccination
- low fat diet

## Early Detection

- hypothyroidism
- breast ca
- technology assessment

## Treatment

- drugs
- surgery
- psychotherapy

# Hypothesis

- Groups to be compared
- Outcome
- Time of Outcome
- Magnitude of predicted effect
- Variability of outcome
- Adverse Effects

# Miscellaneous Problems

- External validity
- Comparability of groups
  - Did randomization work?
- Blind assignment and blind assessment - best double blind
- Sample size!!!
  - In general normally distributed variable approx. 50
  - If dichotomous variables several hundreds
- Compliance
  - monitor

# Types of Error

## Type 1

Reject the null hypothesis falsely

Rests on alpha and no. of tests

## Type 2

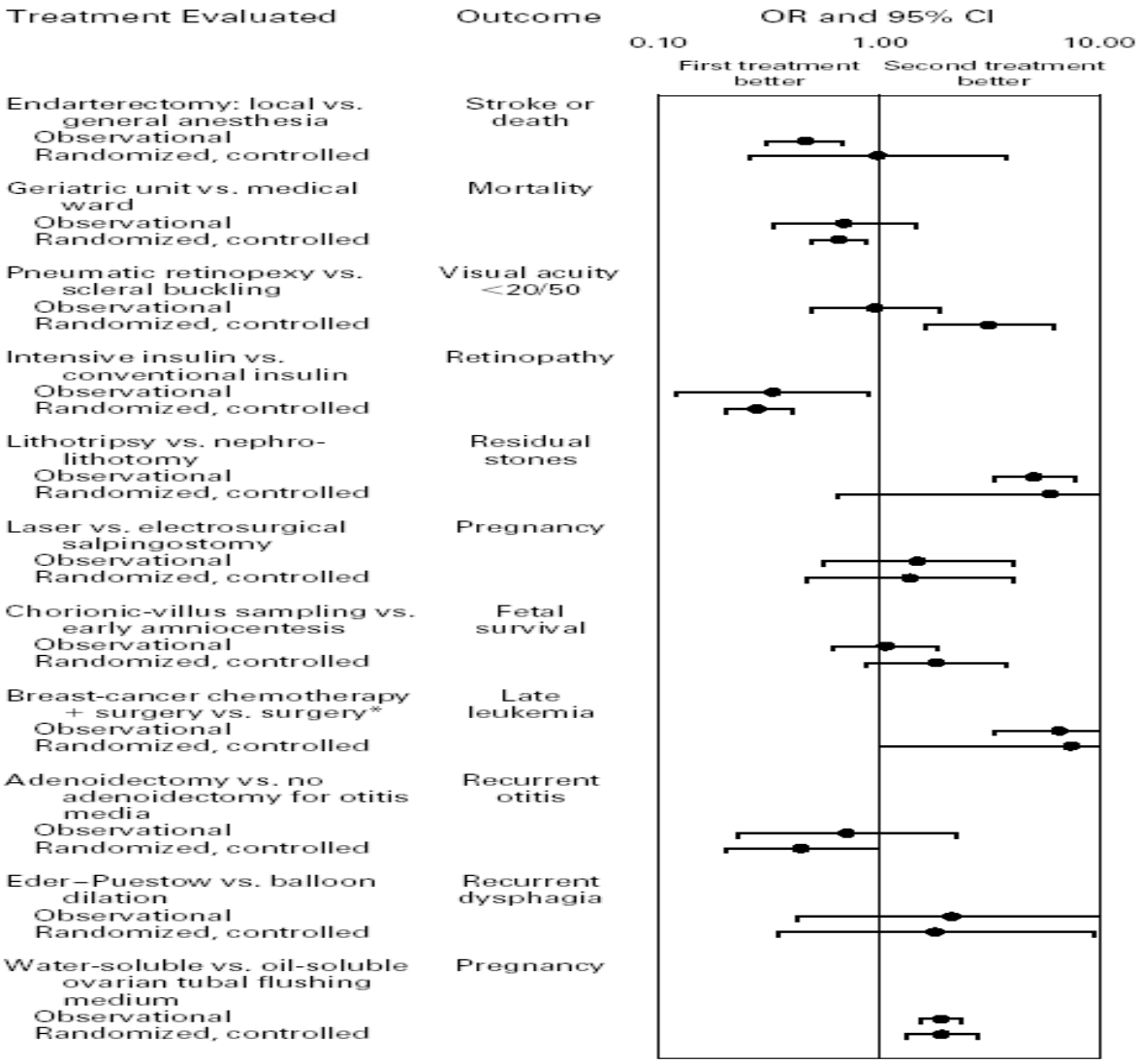
Probability called Beta

Accept the null hypothesis when it is false

usually inadequate sample size

# A COMPARISON OF OBSERVATIONAL STUDIES AND RANDOMIZED, CONTROLLED TRIALS

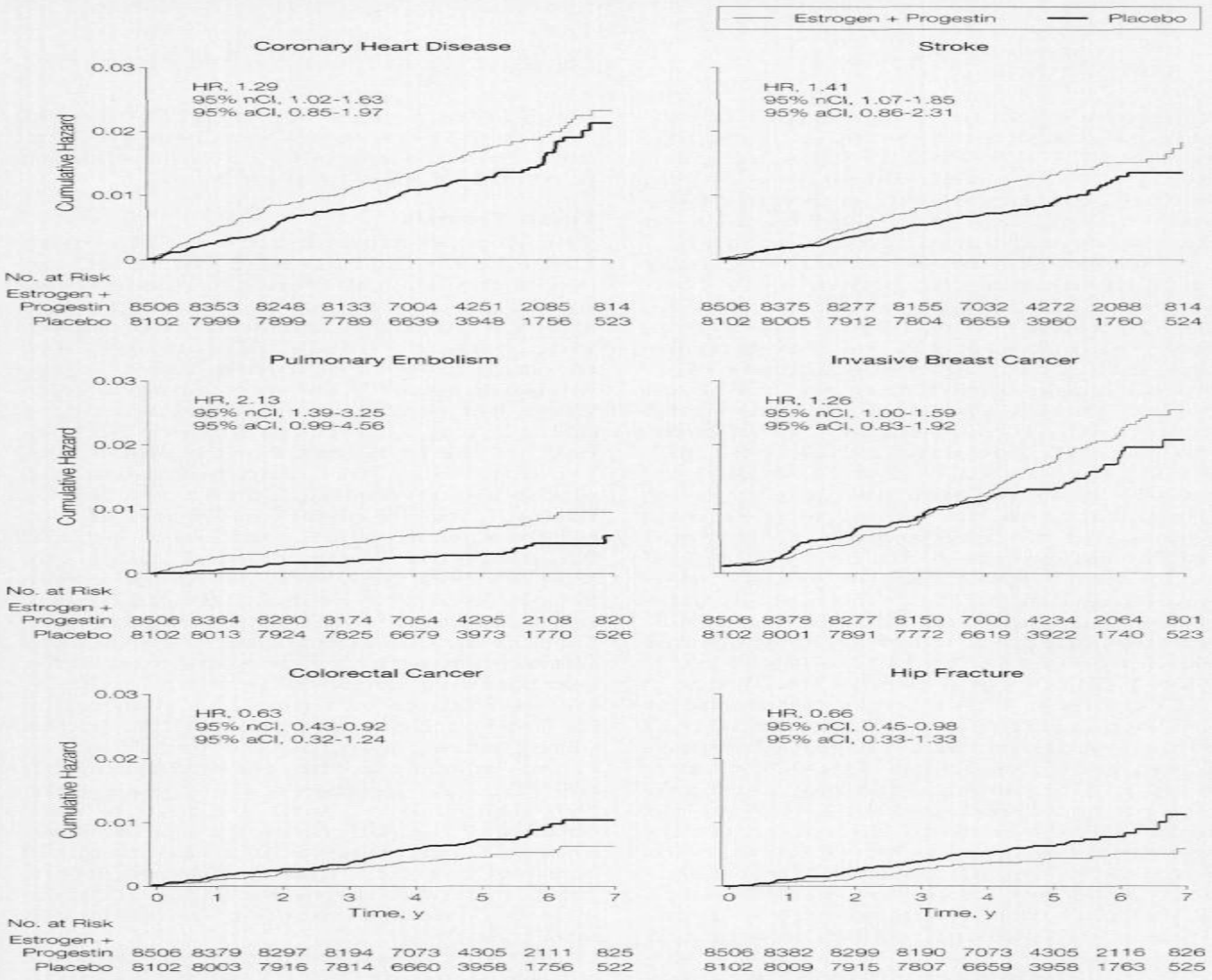
Benson, Hartz (N Engl J Med 2000; 342:1878)




AN CENTRE FOR AGEING

# Women's Health Initiative

**Figure 3.** Kaplan-Meier Estimates of Cumulative Hazards for Selected Clinical Outcomes



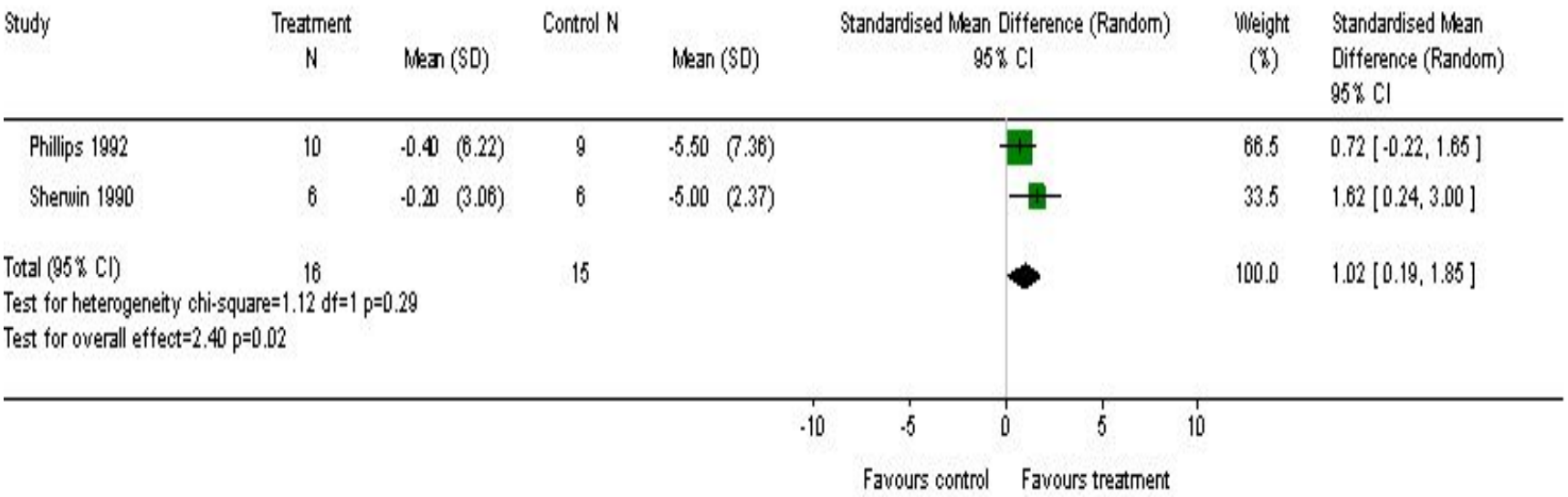

HR indicates hazard ratio; nCI, nominal confidence interval; and aCI, adjusted confidence interval.

# Results of Prospective studies of Estrogen Replacement on Outcome of AD

| Study | Outcome | RR or OR | 95%CI or *P*-value |
|---|---|---|---|
| Kawas *et al.* (1997) | AD | 0.46 | 0.21–1.00 |
| Tang *et al.* (1996) | AD | 0.40 | 0.22–0.85 |
| Waring *et al.* (1999) | AD | 0.42 | 0.18–0.96 |

# HRT for Cognitive Function in Women
# Effect on Verbal Memory

Review: Hormone replacement therapy for cognitive function in postmenopausal women
Comparison: 01 The effect of HRT vs placebo on verbal memory tests
Outcome: 01 Paired Associates (immediate recall) after 2 months of E2 i.m. or placebo




WESTERN AUSTRALIAN CENTRE FOR HEALTH AGEING

# Science Changes a Good Story

## Effect of Estrogen Plus Progestin on Global Cognitive Function in Postmenopausal Women

The Women's Health Initiative Memory Study:
A Randomized Controlled Trial

WESTERN AUSTRALIAN CENTRE FOR HEALTH AGEING

# Results for WHIMS – HRT is bad for your brain

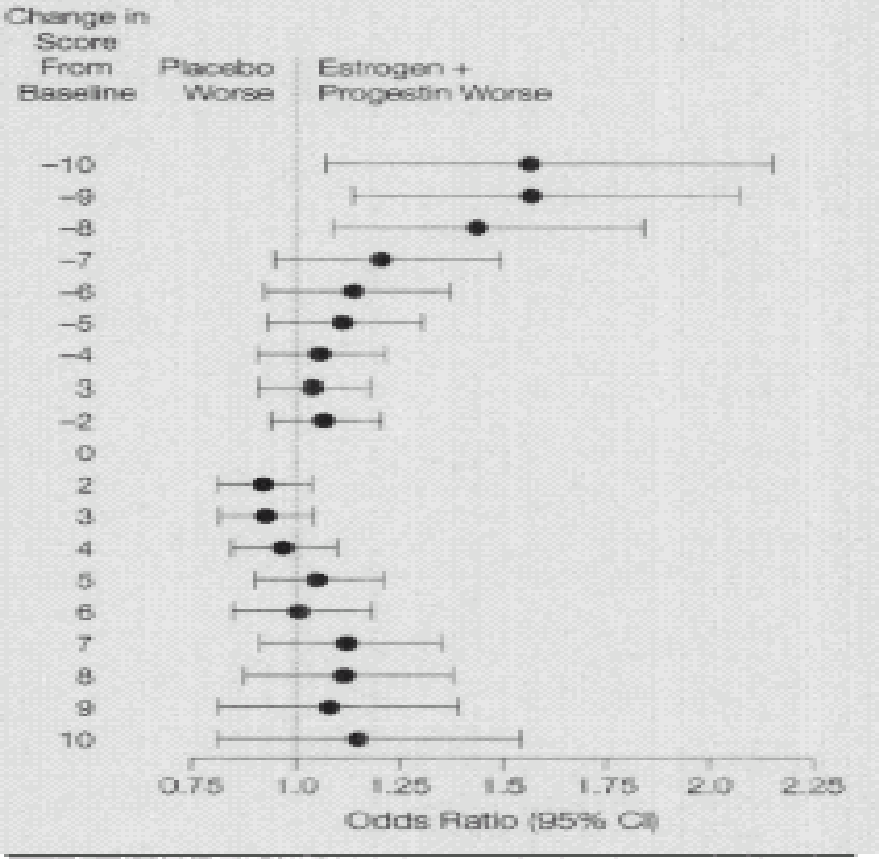
**Table 2.** Mean Modified Mini-Mental State Examination Scores by Time From WHI Randomization and Treatment Assignment

| No. of Years Since Randomization | Treatment Assignment: Estrogen + Progestin, No. of Patients | Estrogen + Progestin, Mean (SD) | Placebo, No. of Patients | Placebo, Mean (SD) | Difference Between Treatments, Mean (95% CI) |
|---|---|---|---|---|---|
| 0 | 2132 | 95.50 (4.21) | 2215 | 95.63 (3.87) | −0.13 (−0.37 to 0.11) |
| 1 | 2102 | 95.98 (4.10) | 2188 | 96.20 (3.70) | −0.22 (−0.46 to 0.01) |
| 2 | 2022 | 96.37 (3.92) | 2095 | 96.50 (3.91) | −0.13 (−0.37 to 0.11) |
| 3 | 2005 | 96.38 (4.36) | 2083 | 96.74 (3.72) | −0.36 (−0.61 to −0.11) |
| 4 | 1814 | 96.38 (4.96) | 1875 | 96.90 (3.92) | −0.50 (−0.79 to −0.21) |
| 5 | 833 | 96.71 (4.66) | 901 | 96.87 (4.26) | −0.16 (−0.58 to 0.26) |
| 6* | 31 | 96.81 (2.39) | 44 | 96.16 (8.63) | 0.65 (−2.53 to 3.82) |

Abbreviations: CI, confidence interval; WHI, Women's Health Initiative.
*Includes 9 women who were tested at 7 years after randomization.

# Results for WHIMS – HRT is bad for your brain

**Figure 4.** Odds Ratio (95% Confidence Intervals) for Various Magnitudes of Modified Mini-Mental State Examination Score Changes From Baseline (Across All Follow-up Visits): Estrogen Plus Progestin vs Placebo



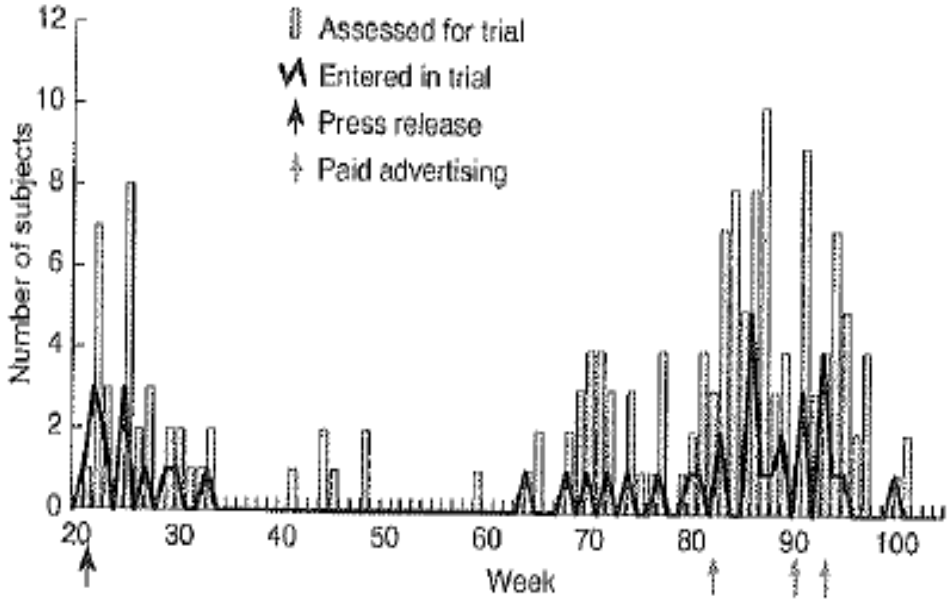

STERN AUSTRALIAN CENTRE FOR
EALTH AGEING

# Recruitment strategies for randomised clinical trials in elderly Australians

Leon Flicker and John D Wark

*Paid newspaper advertisements produced the best response*

established osteoporosis in women over the age of 60. There were many patient exclusion criteria: other mobility problems, calcium and bone disorders, thyroid disease, liver disease, active malignancy, treatment within six months for osteoporosis, corticosteroid therapy, renal dysfunction, diabetes mellitus requiring insulin, anticonvulsant treatment, unstable hypertension, congestive cardiac failure and recurrent renal calculi. Additionally, women unlikely to survive two years were excluded. The requirement for informed consent precluded participation of women with significant cognitive dysfunction. The recruitment stategies used are



# Practical Statistics??

- Definitely an oxymoron
- The single most important thing to learn is that the study design drives the statistical analyses.
- Most analyses if you really wanted to do them could be done with a hand calculator, certainly for the raw figures.

# Impractical Statistics

- The most important decision in designing the study is what is the outcome?
- If it is an all or nothing event (which usually is uncommon) you need large studies, often forcing you outside the country.
- As a compromise you choose a score on something, a surrogate and measure the difference between and intervention and control group.

# Evaluation of Diagnostic Tests

- Inevitably the major problem is the selection of the gold standard.
- For some conditions there is no completely acceptable gold standard eg Alzheimer's Disease
- For other conditions you wait along time for diagnoses to be made eg breast cancer

# Diagnosis

- The steps for evaluation of reports of the validity of diagnostic tests
  - 1) Was there an independent blind comparison with a gold standard?
  - 2) Was the test applied in an appropriate spectrum of patients?
  - 3) Was the reference standard applied independent of the results?

# Plasma and cerebrospinal fluid amyloid beta for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI) (Review)

Ritchie C, Smailagic N, Noel-Storr AH, Takwoingi Y, Flicker L, Mason SE, McShane R

**Main results**

Alzheimer's disease dementia was evaluated in 14 studies. Of the 1349 participants included in the meta-analysis, 436 developed Alzheimer's dementia. Individual study estimates of sensitivity were between 36% and 100% while the specificities were between 29% and 91%..... At the median specificity of 64%, the sensitivity was 81% (95% CI 72 to 87).

**Authors' conclusions**

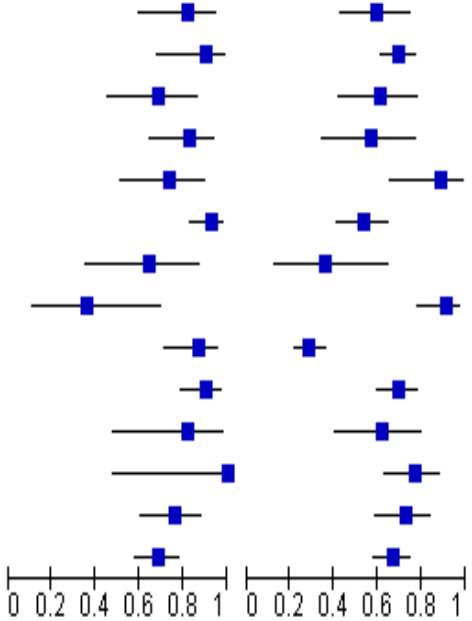
.. From our review, the measure of abnormally lowCSF Aß levels has very little diagnostic benefit …..We conclude that when applied to a population of patients with MCI, CSF Aß levels cannot be recommended as an accurate test for Alzheimer's disease.

THE COCHRANE COLLABORATION®

WESTERN AUSTRALIAN CENTRE FOR HEALTH AGEING

# Study results of cerebrospinal amyloid beta 42 for detection of Alzheimer's disease dementia.

| Study | TP | FP | FN | TN | Abeta 42 Threshold | Threshold Unit of measurement | Sensitivity (95% CI) | Specificity (95% CI) |
|---|---|---|---|---|---|---|---|---|
| Galluzzi 2010 | 18 | 17 | 4 | 25 | 500.0 | ng/l | 0.82 [0.60, 0.95] | 0.60 [0.43, 0.74] |
| Bjerke 2009 | 18 | 43 | 2 | 99 | 512.0 | ng/l | 0.90 [0.68, 0.99] | 0.70 [0.61, 0.77] |
| Zetterberg 2003 | 15 | 12 | 7 | 19 | 532.0 | ng/l | 0.68 [0.45, 0.86] | 0.61 [0.42, 0.78] |
| Hampel 2004 | 24 | 10 | 5 | 13 | 679.0 | ng/l | 0.83 [0.64, 0.94] | 0.57 [0.34, 0.77] |
| Chiasserini 2010 | 17 | 2 | 6 | 16 | 741.0 | ng/l | 0.74 [0.52, 0.90] | 0.89 [0.65, 0.99] |
| Hansson 2007 | 53 | 36 | 4 | 41 | 0.64 | ng/ml | 0.93 [0.83, 0.98] | 0.53 [0.42, 0.65] |
| Blom 2009 | 9 | 9 | 5 | 5 | 82.0 | pM | 0.64 [0.35, 0.87] | 0.36 [0.13, 0.65] |
| Parnetti 2006 | 4 | 4 | 7 | 40 | 500.0 | pg/l | 0.36 [0.11, 0.69] | 0.91 [0.78, 0.97] |
| Shaw 2009 | 32 | 113 | 5 | 46 | 192.0 | pg/ml | 0.86 [0.71, 0.95] | 0.29 [0.22, 0.37] |
| Hertze 2010 | 47 | 33 | 5 | 75 | 209.0 | pg/ml | 0.90 [0.79, 0.97] | 0.69 [0.60, 0.78] |
| Monge-Argiles 2011 | 9 | 10 | 2 | 16 | 320.0 | pg/ml | 0.82 [0.48, 0.98] | 0.62 [0.41, 0.80] |
| Papaliagkas 2009 | 5 | 11 | 0 | 37 | 472.0 | pg/ml | 1.00 [0.48, 1.00] | 0.77 [0.63, 0.88] |
| Kester 2011 | 32 | 16 | 10 | 42 | 495.0 | pg/ml | 0.76 [0.61, 0.88] | 0.72 [0.59, 0.83] |
| Vos 2013 | 62 | 41 | 29 | 82 | 500.0 | pg/ml | 0.68 [0.58, 0.78] | 0.67 [0.58, 0.75] |




What additional information does this provide and what is its value?

THE COCHRANE COLLABORATION®

WESTERN AUSTRALIAN CENTRE FOR HEALTH AGEING

# Diagnosis

- How do you define what is “abnormal” on a test? In virtually all tests there is an arbitrary division or “cut-point” dividing between the two. This is based
  - 1) Gaussian  mean± 2 SD
  - 2) Percentile
  - 3) Culturally  appropriate eg alcohol intake
  - 4)Risk factor
  - 5)Diagnostic - when the target disorder become highly probable
  - 6) Therapeutic - When treatment does more good compared to harm
- This obviously varies eg the evolution of treatment for hypertension (now 30% of the population) and cholesterol.

# Example: The Importance of Sampling to Determine Reference Ranges - BMD

- BMD is essentially a normally distributed variable with abnormality based on an arbitrary amount of increased fracture risk.
- To establish local reference data, 493 Australian female volunteers had bone mineral densitometry performed at a single medical centre at the proximal femur and lumbar spine.
- These data were compared with reference material from North American women compiled by Hologic, Inc.
- The Australian volunteers had, on average, 7% greater bone mineral density at the lumbar spine for the age range 25 to 55 years.
- Possible explanations for this include an actual population difference or the presence of a differential selection bias between the two samples.
- The weights of the two North American samples were significantly different for all decades.