PH 458/231
Evidence & Policy
Lecture 6

John Worrall
LAK 3.02
Office Hours:
Monday 13.30-14.30
Thursday 12.30 – 13.30

Ethics and Stopping Rules

- “The investigators were sensitive to the individual ethics of seeking parental consent and randomization for the next newborn infant .. However, with only 19 patients this does not represent strong evidence of the superiority of ECMO and provides little scope for making reliable judgments on the benefits of this treatment for universal use in such newborn infants in the future.”

Ethics and Stopping Rules

- “Thus collective ethics may have been compromised by such early stopping…. [I]f ECMO really is effective the prolonged uncertainties maintained by lack of really substantial evidence may well have led to fewer newborn infants worldwide receiving it than would have been the case had the trial continued longer.”

Ethics and Stopping Rules

- Two points:
- A. WHY were “the investigators sensitive to the individual ethics of seeking parental consent and randomization for the next newborn infant”?
- B. An implicit further argument here of a ‘social epistemological’ kind, that is interesting and well worth teasing out.

Ethics and Stopping Rules

- Re A: Surely the investigators had this ‘sensitivity’ because they believed that
- the historical evidence *plus*
- the earlier trial result *plus*
- the 0 deaths out of 9 on ECMO compared to 4 out of 10 on CT in this 2nd trial
- Provided reasonably compelling OBJECTIVE evidence of the superiority of ECMO.

Ethics and Stopping Rules

- In other words they implicitly rejected the orthdox, frequentist statisticians' view that the only objective, scientific evidence is that garnered from officially sanctioned RCTs.
- Were they right?

Ethics and Stopping Rules

- Re B: Implicit in Pocock's treatment (and made explicit elsewhere) is the claim that:
- Whatever the objective rights and wrongs of the epistemic situation, it is just a matter of fact about the medical community that it will only accept that a new treatment is really superior (and therefore agree to its general introduction and use) if that superiority has been 'established' by a ' proper' RCT.

Ethics and Stopping Rules

- Hence the argument that application of the stopping rule “may well have led to fewer newborn infants worldwide receiving it than would have been the case had the trial continued longer.”
- BUT this is extraordinary – it seems to concede that there may at least sometimes be some good evidence that a treatment is effective that is not produced by a fully accredited RCT;

Ethics and Stopping Rules

- Yet in order to *convince* the medical community, a ‘proper’ RCT should still be performed
- even though there is already - objectively good but intersubjectively unappreciated - evidence that the new treatment is superior; and therefore
- even though there is good evidence ahead of the trial that those in the control arm are being given a decidedly inferior treatment

Ethics and Stopping Rules

- Surely an alternative that is both ethically and epistemologically more attractive would be
- To stop teaching medics that RCTs are the ‘gold standard’ and stop encouraging the view that only they provide real scientific evidence; and
- Encourage a more sensible view of objective evidence.

Ethics and Stopping Rules

- In any event clear that there's lots of work to be done here; but
- Surely clear, as I claimed, that you can't develop a serious ethical view about the matter of stopping rules without getting seriously involved in epistemic, evidential issues

Ethics and Stopping Rules

- IF you hold that you really don't know anything objectively about a new treatment unless it has been validated in a ' proper' RCT – one with a pre-stated significance level and acceptable power
- THEN the stopping rule used in the 2nd ECMO trial was certainly obfuscatory and arguably itself unethical

Ethics and Stopping Rules

- WHEREAS
- IF you hold that there can be real evidence short of a ‘proper’ RCT and that in the ECMO case the interim results of the 2nd trial (at the latest) provided it, then you would regard it as unethical to have continued the trial

The ethics of ‘informed consent’

- In the 2nd & 3rd trials, the ‘Zellen method’ of obtaining “informed” consent was used.
- This involves seeking informed consent only after randomization and only if a patient is randomized to the experimental arm.
- The justification being that the normal treatment that a patient could expect was CT – so no need to ask for informed consent if actually assigned to CT

The ethics of ‘informed consent’

- Was this ethical?
- Again much will depend on your attitude to an epistemological issue

The ethics of ‘informed consent’

- IF statistical orthodoxy is correct that the results of the initial ‘historically controlled’ trial and the first ‘play the winner’ trial represent no sort of genuine information,
- THEN you may not find any ethical problems in not telling the ‘control’ parents about the possibility that their baby could have been assigned to the ECMO arm
- (Some issues about paternalism here.)

The ethics of ‘informed consent’

- BUT IF you think that these earlier results did supply at least some sort of objective evidence, THEN
- In so far as you think that it is an ethical requirement that patients give genuinely informed consent
- You will definitely regard these trials as ethically questionable.

Summary

- You won't do biomedical ethics/health policy properly unless you dig into evidential issues.
- I argued this by considering 3 questions arising from the ECMO case:
- 1. When is it ethical in general to perform a trial?
- 2. Are 'stopping rules' ethically mandated?
- 3. When should patients be regarded as having given their *informed* consent?

So, what's so telling about an RCT?

- So some people in clinical trials are willing to effectively identify 'evidence' with 'evidence from an RCT'
- And willing to make testing ethical decisions on that basis
- Moreover not just in medicine that RCTs are nowadays regarded as setting the 'gold standard' of evidence
- Also across a whole range of social sciences
- The Campbell Collaboration http://www.campbellcollaboration.org/
- mimicking the Cochrane Collaboration http://www.cochrane.org/

So, what's so telling about an RCT?

- What are *randomized* controls?
- What other forms of controlled studies are there?
- What is randomizing supposed to achieve in epistemological terms? – why should randomization produce stronger evidence of effectiveness?

Why Control?

- Suppose we want to know if taking Vitamin C cures colds
- Do a trial
- Give a bunch of people with colds vitamin C
- Say they all get better within 7 days
- So what?
- Might have got better anyway
- What we would really like
- What we settle for: comparison group = *control* group

Why Control?

- Suppose we do nothing to those in control
- = ‘natural history’ group
- Suppose none in the control group get better within a week
- Is this good evidence that taking vitamin C is good for (“cures”) colds?
- What if those in the ‘experimental group’ were young, fit, otherwise heathy, with relatively mild colds
- While those in ‘control’ were older, overweight, had concomitant pathology and, on average, heavier colds??

Why Control?

- Would ideally like the two groups to be alike in all respects except for the treatment
- Can at least move towards the ideal by *matching* for features – like age, general fitness level, severity of cold, etc
- (notice how well this ties in with the general 'intuitive' principle that I emphasised at the beginning)
- Suppose we have matched and more of the vitamin C group get better within a week
- Do we now at last have good evidence for the efficacy of vitamin C?

Why Control?

- Well, certainly stronger evidence , BUT
- The two groups may be matched, but there is certainly at least one remaining difference – namely one group gets treatment, while the other does not
- What if getting treatment (*any* treatment) has an effect?
- (perhaps through raising expectations which in turn have some biochemical/physiological effect)
- Then there wouldn't yet be evidence that it was the *specific part* of the treatment – i.e. the pharmacological effects of the vitamin C – that was effective
- So we give the control group a 'placebo'

Why Control?

- But if those in the control group know they are getting a placebo, then maybe that will make it ineffective
- Hence ‘single blind’
- Finally: perhaps if the clinicians knew which group a patient was in, they could communicate their expectations that the experimental patients will do better, etc
- Hence ‘double blind’

Why Control?

- Notice how it all ties in
- If vitamin C is effective then we would expect the experimental arm to do better in all these trials
- But the successive controls make it less likely that we would be seeing such a positive result IF vitamin C were in fact INeffective
- And they do so by eliminating (or providing evidence against) alternative explanations of the same ‘positive’ outcome.

Why Control?

- “’Known’ confounders’
- Remaining spectre: “‘unknown’ confounders’
- ‘Historically controlled trials’
- RCTs
- (remember ECMO case)

Why does the (DB) RCT set the 'gold standard'?

- 5 arguments can be found in the literature purporting to show that RCTs provide more powerful evidence than other studies of clinical effectiveness.
- 1. Randomization is necessary to underpin the validity of the significance test.
- 2. By randomizing you control for all possible confounders, known *and unknown*
- 3. It is just an empirical fact that non-randomized studies have tended systematically to exaggerate positive effects
- 4. Randomization controls for the specific possible confounder: 'selection bias'.
- 5. By randomizing, and only by randomizing, you get evidence that any observed positive effect in the trial is *caused* by the treatment rather than merely being associated with it.

Fisher's argument

- Essentially that randomization is necessary to underwrite the logic of the standard Fisherian significance test
- In particular, in order to be able legitimately to identify the (indeterminate) assumption that the hypothesis H at issue is false with the (determinate) Fisherian 'null hypothesis'
- The trial needs to have been randomized.

Fisher’s argument

- However
- 1. Bayesians have convincingly argued that Fisher’s argument fails even on its own terms (see Howson)
- 2. Even if it did succeed, it would only be a convincing reason to randomize for someone who accepted Fisherian statistical methods; and
- As we already saw, there are reasons to question this.

Argument 3: HCTs (aka “Observational studies”) exaggerate positive effects

- This an *empirical* argument
- Alleged to be just a matter of fact that “observational studies” routinely exaggerate the effects of treatment
- Response A: how would you know what the ‘true effect’ of treatment is?
- Only by identifying RCTs as definitive
- Hence argument circular
- (Data just as compatible with inference that RCTs routinely *under*estimate the ‘true effect’.)

Argument 3: HCTs (aka “Observational studies”) exaggerate positive effects

- Response B: Based on observational studies done in the 1980s that were methodologically deficient in anyone’s book
- Chalmers et al point out themselves that there were reasons to think that important prognostic factors were maldistributed across the two groups.

Argument 3: HCTs (aka “Observational studies”) exaggerate positive effects

- More recent studies (2000) seem to show that
- (a) differences between average effects based on a number of RCTs and based on a number of – *well performed* - historically controlled trials of the same treatment tend to be small
- (b) the variation in outcomes between individual trials on the same treatment is greater for RCTs than for historically controlled trials
- (c) sets of well-controlled HCTs seem to produce fewer ‘paradoxical’ results than do RCTs

Argument 3: HCTs (aka “Observational results”) exaggerate positive effects

- Certainly need to be careful about historical controls (Maintenance and management – nutrition, fluid balance, *etc*)
- On the other hand you get the full placebo effect with historical controls
- (reason why RCTs may systematically underestimate the effects of treatments ‘in the wild’ – at any rate for certain conditions)

Argument 3: HCTs (aka “Observational results”) exaggerate positive effects

- 1. Size of effect is clearly important
- 2. As is: how historical are the controls?
- 3. In general the question is always the Mill’s methods/ Popper one:
- Is there a plausible alternative explanation?
- (place to say a little about meta-analyses and systematic reviews)

Argument 2: “controlling for all factors”

- Thinking back to our general consideration of the value of ‘controlling’
- We see the big attraction of randomizing – the idea that it might control for all possible ‘confounders’: ‘known’ and ‘unknown’
- Mike Clarke: “In a randomised trial, the *only* difference between the two groups being compared is that of most interest: the intervention under investigation.”
- Rawlins: “The greatest strength of an RCT is that the allocation of the treatments is random so that the groups being compared are similar for baseline factors.”
- “Other study designs … can detect associations between an intervention and an outcome. But they cannot rule out the possibility that the association was caused by a third factor linked to both intervention and outcome. Random allocation *ensures* no systematic [?] differences between intervention groups in factors, known and unknown, that may affect outcome.” (Sibbald and Roland, *BMJ,*1998,201)

The RCT as gold standard?

- Wikipedia:
- As their name suggests, RCTs involve the random allocation of different interventions (treatments or conditions) to subjects. As long as numbers of subjects are sufficient, this ensures that both known and unknown confounding factors are evenly distributed between treatment groups.

The RCT as gold standard?

- Similarity/identity?
- But in either case

The RCT as gold standard?

- ‘Otherwise equivalent’ claim obviously false
- Clearly not ‘ensured’ in a *single* randomized division
- Quietly conceded by the injunction to look for ‘baseline imbalances’ (in ‘known’ confounders)
- But if possible in known then clearly also in ‘unknown’

The RCT as gold standard?

- Amusing example
- Leibovici et al “Effects of remote, retroactive, intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial” *BMJ* 2001

The RCT as gold standard?

- 3393 patients having a bloodstream infection while being an inpatient at the Rabin Medical Centre during 1990-6 were identified.
- In July 2000 a random number generator was used to divide these patients into two groups and which of these two became the treatment group was decided by a coin toss.
- 1691 were randomized to the intervention group and 1702 to the control.

The RCT as gold standard?

- Checked for baseline imbalance with regard to main risk factors for death and severity of illness.
- The names of those in the intervention group were given to a person ‘who said a short prayer for the well being and full recovery of the group as a whole.’
- Results: both length of stay in hospital and duration of fever were significantly shorter in the intervention group ($p = 0.01$ and $p = 0.04$)!

The RCT as gold standard?

- The names of those in the intervention group were given to a person ‘who said a short prayer for the well being and full recovery of the group as a whole.’
- Results: both length of stay in hospital and duration of fever were significantly shorter in the intervention group ($p = 0.01$ and $p = 0.04$)!

The RCT as gold standard?

- Conclusion: ‘Remote, retroactive intercessory prayer said for a group is associated with a shorter stay in hospital and shorter duration of fever in patients with bloodstream infection and should be considered for use in clinical practice.’
- Somewhat tongue-in-cheek of course (‘No patients were lost to follow up’!!)

The RCT as gold standard?

- Natural reaction shows we are all Bayesian:
- “If the pre-trial probability is infinitesimally low, the results of the trial will not really change it, and the trial should not be performed. This, to my mind, turns the article into a non-study, though the details provided (randomization done only once, statement of a prayer, analysis, etc) are correct.”
- Fisher a natural but egregious mistake

The RCT as gold standard?

- Kosher trial:
- Proper randomization
- No data mining
- Baseline imbalances
- But …

The RCT as gold standard?

The RCT as gold standard?

- Nobody believes surefire guarantee?
- Instead ‘probabilistic guarantee’
- But what does this amount to?
- ‘Either groups are equivalent or a chance event has occurred’?

The RCT as gold standard?

- Again philosophy of science clarifies:
- 1.Slip from what is true in the indefinite long run of many repeated random divisions, to what is (allegedly) true in the single random division
- (Single case probabilities?)

The RCT as gold standard?

- RCTs cannot control for all possible confounders – nothing can
- The question to ask if whether or not there is some plausible alternative explanation of an apparently positive trial result
- The answer may be no even without randomization

“last men standing”: (a) the argument from selection bias

- So, remember, there were 5 arguments for the claim that RCTs carry special evidential weight
- 1.Fisher: randomizing essential to underpin the logic of the significance test
- 2. Randomizing controls for all possible confounders known *and unknown*
- 3. Historical fact that non-randomized studies exaggerate positive effects
- 4. Randomizing controls for the specific possible confounder – ‘selection bias’
- 5. By randomizing, and only by randomizing, you get evidence that any observed positive effect in the trial is *caused* by the treatment rather than merely being associated with it.
- 1- 3 arguably don’t stand up to critical scrutiny
- We are left with 4 and 5

“last men standing”: (a) the argument from selection bias

- Undoubtedly a solid argument
- Though it is not the randomization as such that is doing the work
- But rather the fact that the division into ‘experimental’ and ‘control’ *is taken out of the hands of the experimenters*
- Can’t selection bias, or any rate any great amount of it, be effectively ruled out by other means?
- Cp ECMO again

“last men standing”: (b) evidence from RCTs underwrites *causes*

- This takes us to our next topic