Statistical issues in HIV trial design

Andrew Hill
Senior Visiting Research Fellow
Liverpool University
Type 1 error

It is statistically significant!

(after 100 different statistical tests)
Type 1 error:  
‘significant’ $p$ value after many tests

When you throw two dice, the chance of getting a double six: $p=0.028$
Type 1 error: atenolol

Randomised trial was submitted to the Lancet

Reviewers asked for several subgroup analyses

Authors did not agree and so they tested for treatment effect by zodiac sign

Patients with zodiac sign of Sagittarius showed a strong benefit, while others did not have a significant benefit
Type 1 error in HIV trials?

50–100 statistical tests on safety endpoints
Enfuvirtide: bacterial pneumonia?

Interim and final analyses:

Raltegravir – significant benefit over
EFV in STARTMRK only seen after 4 analyses

Darunavir – significant benefit over LPV/r
In ARTEMIS only seen after 4 analyses

MONET trial – non-inferiority seen at Week 48, then not seen in later analyses

Are these real differences, or Type 1 errors?
Type 1 error: checks

Double six: p=0.028

A clinical trial only has the power to test the primary endpoint

Everything else is secondary and should be validated in another trial

Count how many tests and p values shown in the paper? Be careful in the safety sections

How significant was the result:

p<0.05 (2 sixes), p<0.01 (3 sixes), p<0.001 (4 sixes)?

You can correct for multiple tests, but p value drops: p<0.005 for 10 tests, p<0.001 for 50 tests
Clinical significance?

The difference is statistically significant, but is it clinically significant?
Change in LDL to Week 48
TAF versus TDF

Study 104 (n=867)
Study 111 (n=866)

Will there be more myocardial infarctions on TAF than TDF?
Data from meta-analysis needed

LDL cholesterol (mg/dL)

- TAF/FTC + ELV/c
- TDF/FTC + ELV/c
- TAF/FTC + ELV/c
- TAF/FTC + ELV/c

Study 104: TAF/FTC + ELV/c
Study 111: TAF/FTC + ELV/c
Change in eGFR to Week 48
TAF versus TDF

Study 104 (n=867)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>eGFR (mL/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAF/FTC + ELV/c</td>
<td>-10.4</td>
</tr>
<tr>
<td>TDF/FTC + ELV/c</td>
<td>-10.4</td>
</tr>
</tbody>
</table>

Study 111 (n=866)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>eGFR (mL/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAF/FTC + ELV/c</td>
<td>-11.9</td>
</tr>
<tr>
<td>TDF/FTC + ELV/c</td>
<td>-11.9</td>
</tr>
</tbody>
</table>

Will there be more renal failure on TAF than TDF?
Data from meta-analysis needed

Gilead press release, September 2014
Type 2 error

There is no difference between the arms

But the study was too small to show a difference
Sample size calculations

1. You can look at cohort studies to get estimates of endpoints, to use in clinical trial design.

2. For example, how many people would need to be enrolled in a clinical trial of an HIV vaccine to lower transmission rates by 70%?

3. Assume 80% power, 5% significance level, one primary analysis
### Cohort studies and HIV vaccine trials

<table>
<thead>
<tr>
<th>Country</th>
<th>Population</th>
<th>HIV transmission (% per year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>STD clinic / Heterosexuals</td>
<td>0.5%</td>
</tr>
<tr>
<td>Canada</td>
<td>Men having Sex with Men</td>
<td>0.6%</td>
</tr>
<tr>
<td>Brazil</td>
<td>IV drug users</td>
<td>0.8%</td>
</tr>
<tr>
<td>Tanzania</td>
<td>Factory workers</td>
<td>1.5%</td>
</tr>
<tr>
<td>UK</td>
<td>Men having Sex with Men</td>
<td>1.5%</td>
</tr>
<tr>
<td>USA</td>
<td>IV drug users</td>
<td>2.0%</td>
</tr>
<tr>
<td>Australia</td>
<td>Men having Sex with Men</td>
<td>2.1%</td>
</tr>
<tr>
<td>Kenya</td>
<td>Commercial Sex Workers</td>
<td>3.5%</td>
</tr>
<tr>
<td>Congo</td>
<td>Discordant couples</td>
<td>4.0%</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>New mothers</td>
<td>4.8%</td>
</tr>
<tr>
<td>China</td>
<td>IV Drug Users</td>
<td>5.0%</td>
</tr>
<tr>
<td>Thailand</td>
<td>IV Drug Users</td>
<td>5.8%</td>
</tr>
<tr>
<td>Cote D'Ivoire</td>
<td>Commercial Sex Workers</td>
<td>6.0%</td>
</tr>
<tr>
<td>Uganda</td>
<td>Rural adults</td>
<td>6.5%</td>
</tr>
<tr>
<td>USA</td>
<td>STD clinics / MSM</td>
<td>7.1%</td>
</tr>
<tr>
<td>Thailand</td>
<td>IV Drug Users</td>
<td>17%</td>
</tr>
<tr>
<td>South Africa</td>
<td>Women / Antenatal clinics</td>
<td>17%</td>
</tr>
<tr>
<td>Tanzania</td>
<td>Police Officers</td>
<td>20%</td>
</tr>
</tbody>
</table>
Sample sizes for HIV vaccine trials

Sample size for an HIV vaccine trial (2 arms, 70% reduction in transmission)

The sample size for this trial could range between 200-8200 patients, depending on the cohort study used to estimate HIV transmission rates.

Make sure to assess as many cohorts as possible.
Non-inferiority trials: key points

• Non-inferiority trials are powered to show that the new treatment is NO WORSE than the control

• The trial is powered with a ‘delta’ estimate, so we can conclude that the new treatment is no worse than control by this delta (10–12% for HIV trials)

• The control treatment needs to be the recognised standard of care, and show a typical level of efficacy versus reference trials
Non-inferiority trials: key terms

- Difference between treatment arms
- Point estimate
- Delta, or non-inferiority margin
- Point of no difference between treatments
- 95% confidence intervals

Non-inferiority trials: outcomes

-10% or -12%  0

Control arm better  New drug better

- Inferior
- Significantly worse, but non-inferior
- Non-inferiority not demonstrated
- Non-inferior
- Superior

HIV trials where non-inferiority was not shown

HIV RNA <50 copies/mL at Week 48
difference between treatment arms

Control arm better | New drug better

Trial: new drug and control

BICOMBO: ABC/3TC vs TDF/FTC

2NN: NVP vs EFV

MERIT: MVC vs EFV

Gilead 903: TDF/3TC vs d4T/3TC

* Assumes symmetrical confidence intervals
Underpowered studies – DTG and TB

Study is too small to establish non-inferiority for DTG vs EFV in TB co-infection
Strong interaction between DTG and rifampicin – does double dosing cover this?
How do we use DTG in Africa without clinical validation of efficacy?

www.clinicaltrials.gov
Type 2 error: checks

• ‘Not significantly different’ does not mean the same

• A new treatment needs to be NON-INFRINGEMENT to the standard of care to be acceptable: look at the 95% confidence intervals of the difference between the arms

• Nevirapine has never shown true non-inferiority versus efavirenz in naïve patients

• Abacavir has never shown true non-inferiority versus tenofovir in treatment naïve or experienced patients
Intent to Treat analyses: a reality check

• Intent to Treat analysis (FDA TLOVR): includes all patients randomised, even if not compliant with the protocol

• Per Protocol analysis: patients with major protocol violations are excluded (not well standardised)

• As Treated analysis: only virological endpoints are included

• Results need to be consistent across these analyses
Per Protocol analysis – why do we do it?

• Example 1 – a patient was randomised to Arm A, but wanted to take the drugs in Arm B. He took the Arm B drugs for 3 years and had HIV RNA <50 copies/mL at the end of the trial.

• Example 2 – a drug addict entered a trial and took cocaine for 2 weeks, then HIV drugs for 2 weeks, for three months. The patient showed virological failure.

• Example 3 – a patient was randomised but then died in a car accident before taking the first dose of medication
2NN trial: EFV is significantly better than NVP when only the treated patients are included.


ITT exposed = Per Protocol analysis: only includes patients who actually took at least one dose of study medication.
Endpoints in HIV clinical trials – a mixture

Virological failure

versus

Discontinuation for adverse events or other reasons
The FDA Snapshot endpoint

- Success or failure depends only on the HIV RNA level at the time of the endpoint (e.g. Week 48)

- If HIV RNA is below 50 copies/mL at this time, the patient is a success

- If HIV RNA is at or above 50 copies/mL FOR ANY REASON, the patient is a failure

- **Advantages**
  - Simple endpoint to work with
  - Can be easily analysed across trials

- **Disadvantages**
  - HIV RNA assays are variable, HIV RNA can be 50-200 copies/mL by chance
  - Patient could have HIV RNA <50 before and after Week 48, but could still be classified as failure

- It is more accurate to use two or more HIV RNA measures
What is virological failure?

• In regulatory trials, only 25–30% of endpoints are virological

• Discontinuation for adverse events or other reasons can also be classified as ‘virological failure’

• True virological failure and discontinuations may not be balanced across the arms
## Summary efficacy data from registrational trials (data source: FDA drug labels)

Percentage of patients with ‘virological failure’ by FDA TLOVR endpoint, in three categories. HIV RNA endpoint of 400 copies/mL at Week 48.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Arm</th>
<th>N</th>
<th>% of patients with VF</th>
<th>d/c AE</th>
<th>d/c other</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMP-006</td>
<td>ZDV/3TC/EFV</td>
<td>422</td>
<td>6</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>FTC301A</td>
<td>FTC/ddI/EFV</td>
<td>286</td>
<td>3</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>FTC301A</td>
<td>d4T/ddI/EFV</td>
<td>285</td>
<td>11</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>EPV2001</td>
<td>ZDV/3TC/EFV</td>
<td>278</td>
<td>8</td>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td>EPV2001</td>
<td>ZDV/3TC/EFV</td>
<td>276</td>
<td>8</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>Gilead 903</td>
<td>TDF/3TC/EFV</td>
<td>299</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Gilead 903</td>
<td>d4T/3TC/EFV</td>
<td>301</td>
<td>4</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>CNA30021</td>
<td>ABC/3TC/EFV</td>
<td>384</td>
<td>5</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>CNA30021</td>
<td>ABC/3TC/EFV</td>
<td>386</td>
<td>5</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>APV 3002</td>
<td>ABC/3TC/fAPV</td>
<td>322</td>
<td>6</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Abott 863</td>
<td>d4T/3TC/LPV/r</td>
<td>326</td>
<td>9</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Gilead 934</td>
<td>TDF/FTC/EFV</td>
<td>244</td>
<td>2</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Gilead 934</td>
<td>ZDV/3TC/EFV</td>
<td>243</td>
<td>4</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>KLEAN</td>
<td>ABC/3TC/fAPV/r</td>
<td>434</td>
<td>6</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>KLEAN</td>
<td>ABC/3TC/LPV/r</td>
<td>444</td>
<td>7</td>
<td>5</td>
<td>17</td>
</tr>
</tbody>
</table>

VF = true virological failure; d/c AE = discontinuation for adverse events; d/c other = discontinuation for other reasons
CASTLE trial: 96-week efficacy analysis
ATV/r is superior by ITT, not by On Treatment analysis

**ITT analysis**

- ATV/r: 74% (N=440)
- LPV/r: 68% (N=443)

+6.1%, 95% CI: +0.3, +12.0%

**On Treatment analysis**

- ATV/r: 89% (N=440)
- LPV/r: 88% (N=443)

+1.6%, 95% CI: -3.1, +6.2%

Molina et al. ICAAC 2008; Abstract H1250-d.
SINGLE: FDA Snapshot, Week 144

<table>
<thead>
<tr>
<th>Categories of response and failure</th>
<th>ABC/3TC/DTG N=414</th>
<th>TDF/FTC/EFV N=419</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV RNA &lt;50 copies/mL</td>
<td>72%</td>
<td>63%</td>
</tr>
<tr>
<td>Virological failures</td>
<td>10%</td>
<td>7%</td>
</tr>
<tr>
<td>Discontinuation of treatment</td>
<td>18%</td>
<td>30%</td>
</tr>
<tr>
<td>Drug resistance</td>
<td>0%</td>
<td>1.4%</td>
</tr>
</tbody>
</table>

More virological failures in the DTG arm
No significant difference in drug resistance between the arms
MERIT: percentage of patients with undetectable HIV-1 RNA at Week 48 (primary endpoint)

-400 copies/mL

- EFV + CBV
- MVC + CBV

-50 copies/mL

- EFV + CBV
- MVC + CBV

*Difference (adjusted for randomisation strata)
†Lower bound of 1-sided 97.5% confidence interval; non-inferiority margin = -10%
Per-protocol analysis: <400 copies/mL difference = -4.1 (-10.5†), <50 copies/mL difference = -4.4 (-11.2†)
MERIT Study 48 weeks

Intent-to-treat (ITT) analysis
# MERIT trial: summary of discontinuations through 48 weeks

Includes all patients who received at least one dose of study medication

<table>
<thead>
<tr>
<th>Reason for discontinuation</th>
<th>EFV + CBV N=361</th>
<th>MVC + CBV N=360</th>
</tr>
</thead>
<tbody>
<tr>
<td>All, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adverse event, n (%)</td>
<td>49 (13.6)</td>
<td>15 (4.2)</td>
</tr>
<tr>
<td>Lack of efficacy, n (%)</td>
<td>15 (4.2)</td>
<td>43 (11.9)</td>
</tr>
<tr>
<td>Other reason, n (%)</td>
<td>9 (2.5)</td>
<td>14 (3.9)</td>
</tr>
<tr>
<td>Withdrew consent or lost to follow-up, n (%)</td>
<td>18 (5.0)</td>
<td>25 (6.9)</td>
</tr>
</tbody>
</table>
What are the consequences of virological failure?

- Patients could ‘fail’ with high level drug resistance, or no resistance

- Patients could discontinue the randomised treatment with HIV RNA <50, and switch straight onto an alternative: not a failure in ACTG trials

- A patient could decide to go on holiday, stop taking ARV’s, then come back and be re-suppressed on original Rx

- Patients could have a serious drug toxicity (e.g. Stevens-Johnsons syndrome)

- Why do we count all these endpoints equally???
Percentage of genotyped first-line HAART failures with NNRTI, M184V, TAMS and PI resistance

Gupta et al. BHIVA 2008.
# TAF Phase 3 trials: FDA Snapshot, Wk 96

<table>
<thead>
<tr>
<th>Categories of response and failure</th>
<th>TAF/FTC/ELV</th>
<th>TDF/FTC/ELV</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV RNA &lt;50 copies/mL at Week 96, n (%)</td>
<td>N=866</td>
<td>N=867</td>
</tr>
<tr>
<td>HIV RNA &lt;50 copies/mL</td>
<td>87%</td>
<td>85%</td>
</tr>
<tr>
<td>Virological failures</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>Grade 3 or 4 clinical adverse events</td>
<td>12%</td>
<td>12%</td>
</tr>
<tr>
<td>Grade 3 or 4 lab abnormalities</td>
<td>28%</td>
<td>25%</td>
</tr>
</tbody>
</table>

No difference in overall efficacy or safety between TAF and TDF

German IQWIG rejects application for TAF

Value versus generic TDF?
Are these patients virological failures?

- FDA Snapshot algorithm: failure is simply having HIV RNA >50 copies/mL at Week 48, or being off study drug

- TLOVR algorithm: failure is stopping randomised treatment, rebound of HIV RNA >50 copies/mL at any time, or never suppressed <50 copies/mL

- ACTG trials: failure is rebound of HIV RNA >200 copies/mL at any time, or never suppressed. Switches in treatment allowed
SENSE: HIV RNA versus time, patient 3
Treatment arm: efavirenz

Patient did not take randomised treatment daily, between Day 1 and Week 6

Phenotype:
- 3TC resistant
- EFV resistant
- NVP resistant
- ETR sensitive

Patient switched to TDF/ZDV/ATV/r

Listing EFF: 1  25/02/11
Patient in MONET trial
Treatment arm: DRV/r

ARV treatment started: 1999
Prior antiretrovirals taken: ABC, ddi, 3TC, ZDV, EFV, NFV, LPV/r
Nadir CD4 count: 191
Hepatitis C: Antibody Positive
HCV RNA PCR: <50
NRTI intensification: No
Status: Completed trial

Adverse events / Investigator comments:
*Dr Pulido, Spain*
Poor adherence: Patient interrupted treatment by her own decision, for a month.

Red graph numbers = Optical density readings for HIV RNA samples <50 copies/mL

Week 4    Week 48    Week 96    Week 144
DrV PK    RTV PK    Adherence
9110      65        >95%
257       19        ≤95%
3310      120       >95%
2380      392       >95%

Listing EFF 1, PK 1  08/03/11
SENSE: HIV RNA versus time, patient 3 (etravirine arm)

- **Genotype:** No mutations
- **Phenotype:** ETR sensitive

**HIV RNA copies/mL**
- 10,000,000
- 1,000,000
- 100,000
- 10,000
- 1,000
- 100
- 10

**Time - Weeks**

- FDA Snapshot = success
- TLOVR = failure
- ACTG = success

Listing EFF: 1 25/02/11
Endpoints: checks

• If the treatment arms had equivalent efficacy, was this also true looking only at the virological endpoints?

• If there was a difference between the arms, what was the reason for this?
  – One treatment could be more effective (HIV RNA suppression)
  – One treatment could be better tolerated
  – People might stop treatment in one arm for other reasons

• ACTG trials analysis: uses only virological endpoints: discontinuation for adverse events or other reasons are not endpoints. This raises efficacy rates for ACTG trials versus TLOVR algorithm

• MRC PIVOT Trial: should resistance at failure be the new endpoint?
Cohort studies

Why do we believe the results, when they are not randomised?
Real-life evaluation of treatments

There could be patients who are included in the final approved drug label, but who were excluded from the Phase 3 trials (e.g. CNS abnormalities on dolutegravir).

There may be unresolved safety issues which can only be answered by long-term follow up. For example, does treating HIV+ pregnant women with antiretrovirals lead to toxicity in their children?

Sometimes you cannot randomise
Percentage of Estimated Total PLHIV on ART VS Global Peace Index, Weighted by Epidemic Size

Bubble colour:
- Red = GPI = 2.88 - 4.00
- Orange = GPI 2.83 - 2.38
- Yellow = GPI = 2.37 - 1.92
- Green = 1.91 - 1.43
- Blue = 1.43 - 1.19
Real-life evaluation of treatments

Clinical trials often have strict inclusion criteria – so will not represent all patients who would be treated with a drug.

Screen failure rates in Janssen HIV clinical trials (naïve)

<table>
<thead>
<tr>
<th>Trial</th>
<th>Screened</th>
<th>Screen Failures (n, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>THRIVE (RPV)</td>
<td>947</td>
<td>267 (26%)</td>
</tr>
<tr>
<td>ECHO (RPV)</td>
<td>948</td>
<td>254 (27%)</td>
</tr>
<tr>
<td>ARTEMIS (DRV/r)</td>
<td>843</td>
<td>154 (18%)</td>
</tr>
<tr>
<td>SENSE (ETR)</td>
<td>193</td>
<td>36 (19%)</td>
</tr>
</tbody>
</table>
Detecting rare adverse events in cohorts

Rare toxicities are often only seen after a drug has been approved – large numbers needed to detect rare but serious events

**Number of patients required to detect a 100% rise in risk**

- Attempted suicide, efavirenz (0.3%)
- Drug-related hepatitis, darunavir (0.5%)
- Severe GI bleeding, aspirin (1.8%)
Sources of bias in cohort studies

- Correlation to causation – false associations (no control group)

- Channelling bias – patients with certain risk factors are put onto particular treatments, creating a false association between the risk factor and the treatment

- Confounding factors – factors are not corrected for in multivariate analyses

- Lead-time bias – an intervention changes the time when a disease is first detected

- Behavioural disinhibition – the behaviour of people can counteract the potential benefits of a treatment

- Regression to the mean – people who are ill at baseline can get better with no intervention at all
Vitamin A – antioxidant to prevent cancer?
Vitamin A: is it good for you?

In cohort studies, Vitamin A supplements were associated with improved survival.

In randomised trials, Vitamin A supplements were associated with worse survival.

Taking Vitamin A could be a marker of health awareness.

It is very hard to eliminate bias from analyses of cohort studies.
Tenofovir and survival in HIV: Africa

Randomised studies have shown improved tolerability for tenofovir versus zidovudine. Two studies have evaluated the effects of tenofovir versus ZDV or d4T on survival in African patients in large access programmes:

Study 1: 18,866 patients in PEPFAR, Zambia.
Tenofovir associated with a 43% increase in mortality, compared with zidovudine.
Chi et al 2011 (JAIDS, 58: 475-480)

Study 2: 6,196 patients in Johannesbourg, South Africa
Tenofovir associated with a 29% decrease in mortality, compared with zidovudine.
Velen et al, PLoS One 2013, 8(5)
Cohort studies versus randomised trials

• We can never eliminate the chance of bias from cohort studies

• Randomised trials should remain the gold standard for making decisions

• When looking at results from a cohort study:
  – Have the same results been seen in other cohorts?
  – Is there another explanation for the results (e.g. channelling bias)?
  – Remember the experience from cohorts in other disease areas
Absolute risk versus relative risk

• “Absolute risk is for making decisions. Relative risk should be strictly for research”

• A rise in risk from 1/1000 to 2/1000 is a doubling in relative risk

• Relative risk can be alarming for patients! e.g. abacavir is associated with a 50% rise in the risk of heart attacks

• The number one cause of death for people with HIV in the UK is late diagnosis

• Relative risk is often misused and misinterpreted

• Much better to express results for a population of 100 patients: what percentage would develop safety issues or resistance on a new treatment?
<table>
<thead>
<tr>
<th>Trial (n)</th>
<th>Treatment</th>
<th>M184 I/V mutations</th>
<th>Major NNRTI mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gilead 903</td>
<td>TDF/3TC/EFV (n=299)</td>
<td>12 (4.0%)</td>
<td>16 (5.4%)</td>
</tr>
<tr>
<td>Gilead 934</td>
<td>TDF/FTC/EFV (n=255)</td>
<td>2 (0.8%)</td>
<td>9 (3.5%)</td>
</tr>
<tr>
<td>CNA3024</td>
<td>ABC/3TC/EFV (n=324)</td>
<td>2 (0.6%)</td>
<td>4 (1.2%)</td>
</tr>
<tr>
<td>CNA3021</td>
<td>ABC/3TC/EFV (n=770)</td>
<td>15 (1.9%)</td>
<td>14 (1.8%)</td>
</tr>
<tr>
<td>Gilead 934</td>
<td>ZDV/3TC/EFV (n=254)</td>
<td>7 (2.8%)</td>
<td>16 (6.3%)</td>
</tr>
<tr>
<td>CNA3024</td>
<td>ZDV/3TC/EFV (n=325)</td>
<td>4 (1.2%)</td>
<td>4 (1.2%)</td>
</tr>
<tr>
<td>EPV20001</td>
<td>ZDV/3TC/EFV (n=554)</td>
<td>14 (2.5%)</td>
<td>19 (3.4%)</td>
</tr>
<tr>
<td>Abbott M03-613</td>
<td>ZDV/3TC/EFV (n=51)</td>
<td>1 (2.0%)</td>
<td>1 (2.0%)</td>
</tr>
<tr>
<td>Gilead 903</td>
<td>d4T/3TC/EFV (n=301)</td>
<td>8 (2.7%)</td>
<td>12 (4.0%)</td>
</tr>
<tr>
<td>2NN</td>
<td>ZDV/3TC/EFV (n=400)</td>
<td>10 (2.5%)</td>
<td>n.d.</td>
</tr>
<tr>
<td>ECHO/THRIVE</td>
<td>TDF/FTC/EFV (n=682)</td>
<td>9 (1.3%)</td>
<td>15 (2.2%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>1.6%</strong></td>
<td><strong>2.4%</strong></td>
</tr>
</tbody>
</table>

There are many examples of HIV clinical trials which were never published; often these had unfavourable outcomes.

- Some of these were presented at conferences but never written up.
- They can be very hard to find by MEDLINE searches.

Are you only seeing part of the story about a treatment?

Look for trials in [www.clinicaltrials.gov](http://www.clinicaltrials.gov), or the European Summary of Product Characteristics (SMPC).

Systematic reviews often include all available results, including conference presentations.
DMP-005 trial: ZDV/3TC + EFV 200, 400, 600 mg o.d.
Presented, not published

Haas et al. 5th CROI 1998; Abstract 698.
Levels of evidence to support treatment guidelines

- System of grading evidence to support types of treatment and care

- Used by WHO, IAS and other groups in their guidelines documents

- Example
  - A1: supported by large prospective randomised trials
  - B and C: cohort studies down to expert opinion
Virological suppression comparing 3TC and FTC-including regimens

<table>
<thead>
<tr>
<th>Study</th>
<th>Date</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backbone regimen identical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanne</td>
<td>2002</td>
<td>[22]</td>
</tr>
<tr>
<td>Benson</td>
<td>2004</td>
<td>[23]</td>
</tr>
<tr>
<td>Mulenga</td>
<td>2013</td>
<td>[32]</td>
</tr>
<tr>
<td>Subtotal (I-squared = 55.8%, p = 0.104)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Backbone regimen comparable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Martinez</td>
<td>2009</td>
<td>[26]</td>
</tr>
<tr>
<td>Martin</td>
<td>2009</td>
<td>[25]</td>
</tr>
<tr>
<td>Smith</td>
<td>2009</td>
<td>[27]</td>
</tr>
<tr>
<td>Calza</td>
<td>2009</td>
<td>[24]</td>
</tr>
<tr>
<td>Sax</td>
<td>2011</td>
<td>[9]</td>
</tr>
<tr>
<td>Sax</td>
<td>2011</td>
<td>[9]</td>
</tr>
<tr>
<td>Nishima</td>
<td>2013</td>
<td>[30]</td>
</tr>
<tr>
<td>Raffi</td>
<td>2013</td>
<td>[31]</td>
</tr>
<tr>
<td>Martinez</td>
<td>2013</td>
<td>[29]</td>
</tr>
<tr>
<td>Martinez</td>
<td>2013</td>
<td>[29]</td>
</tr>
<tr>
<td>Campo</td>
<td>2013</td>
<td>[28]</td>
</tr>
<tr>
<td>Raffi</td>
<td>2013</td>
<td>[31]</td>
</tr>
<tr>
<td>Subtotal (I-squared = 0.0%, p = 0.865)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall (I-squared = 0.0%, p = 0.696)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


http://www.plosone.org/article/info:doi/10.1371/journal.pone.0079981
Overall, in 4 randomised trials of 1090 patients, PI/r + 3TC showed HIV RNA suppression rates 4% higher than PI/r + 2NRTIs. This difference was within the limits for non-inferiority (lower 95% confidence interval -1%). There was no evidence for heterogeneity between the trials (p=0.10).

### Table: Study or Subgroup Data

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>2-drug arm</th>
<th>3-drug arm</th>
<th>Risk Difference IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Events</td>
<td>Total</td>
<td>Events</td>
<td>Total</td>
</tr>
<tr>
<td>GARDEL</td>
<td>189</td>
<td>217</td>
<td>169</td>
</tr>
<tr>
<td>Ole</td>
<td>108</td>
<td>118</td>
<td>110</td>
</tr>
<tr>
<td>SALT</td>
<td>112</td>
<td>134</td>
<td>109</td>
</tr>
<tr>
<td>KALEAD</td>
<td>37</td>
<td>72</td>
<td>42</td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>541</td>
<td>549</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Total events 446 430

Heterogeneity: Tau² = 0.00; Chi² = 1.66, df = 3 (P = 0.65); I² = 0%
Test for overall effect: Z = 1.63 (P = 0.10)
Conclusions – 7 key questions to ask:

1. What was the primary endpoint of the study? What was this result?

2. Was the study randomized?

3. Was the study properly powered to show a difference?

4. What was the difference in true virological endpoints between the arms?

5. Is the difference between the treatments clinically meaningful?

6. Are you seeing all of the evidence on a treatment?

7. Do you believe results from a cohort study, if randomized trials say something else?