Role of Integrated Analysis in Regulatory Applications*

H.M. James Hung
Division of Biometrics I, OB/OTS/CDER, FDA

Presented in New Jersey Chapter of ASA Symposium, Union, NJ, October 10, 2008

*The views presented in this presentation are not necessarily of the U.S. Food and Drug Administration

Integrated analysis is a broad term. In regulatory applications, integrated analysis contains multiple dimensions and components

Dimensions	Components
- efficacy (e.g., ISE)	- summary
- safety (e.g., ISS)	- inference
- benefit/risk	- exploration

Integrated analysis in its current form needs to be expanded
- non-inferiority margin relies heavily on guidance from extensive integrated analyses
- bridge information from one geographical region to another
- provides potential formal statistical inference for secondary endpoints
- provide better dose-response information
- understand dropouts' impact on interpretation of trial results
- understand duration of treatment effect

Regulatory Evidence of Efficacy

Efficacy assessments mostly rely on <u>at least two</u> positive adequate and well controlled clinical trials (AWCT)
- derive positive findings from individual trials
- positive findings are replicated
- negative trials are also considered but ***not*** in a formal integrated analysis of totality of data base

Evidence of efficacy is based on ≥ 2 positive AWCT plus "collective evidence" analysis

As a general principle, an integrated analysis without positive individual studies is unlikely to be accepted as a support for a claim unless a pooled analysis is specified as the primary endpoint.

Even if a pooled analysis is a pre-specified primary endpoint, the question of whether the pooled study is acceptable as one adequate and well controlled study equivalent remains. Reproducibility is still an issue.

Failed studies will discount strength of positive studies in "collective evidence" assessment

Difficulty of "collective evidence" assessment
- require selection of trials for consideration
- require identification of 'similar' trials for assessing homogeneity of treatment effects
- pre-specification of statistical decision rule for assessing collective evidence is difficult
- post hoc evaluation of chance of false positive is hard to defend

Ex. In an application, only two of 12 very similar trials showed a statistically significant drug effect
- Raise serious concern of whether the positive findings are real
- Identifying reasons for such exclusive heterogeneity in trial results becomes crucial
- How to interpret the positive results in the sea of so many negative trials
- Collective evidence: worth of 1.5?, < 1?, ≈ 0? positive trial
- No post hoc formal integrated analyses can help

If all the studies share the same effect size, under H_0: no drug effect, the probability of falsely rejecting H_0 in exactly two out of M studies is

$$P = \binom{M}{2}(.025)^2(.975)^{M-2}$$

M	2	4	6	8	12
P	.000625	.0036	.0084	.015	.032

What is the minimum number of positive studies (1p < 0.025) in order to have the probability of false positive error close to .000625?

$$P(X \geq 3 \mid M = 12) = 0.0029$$
$$P(X \geq 4 \mid M = 12) = 0.00016$$

The answer seems to be "somewhere between 3 and 4 studies"

If all the studies are planned to detect the same effect size with the same power, 95% say, then the chance of winning 2 or fewer studies is

$$P(X \leq 2 \mid M = 12) = \sum_{x=0}^{2} \binom{12}{x} (.95)^{x} (.05)^{12-x}$$
$$<< 10^{-8}$$

The chance of winning exactly 2 studies in

$$P(X = 2 \mid M = 12) << 10^{-8}$$

Utility of ISE

Assess collective evidence
- Strengthen/weaken actual claim
- Re-evaluate whether positive results are truly replicated

Improve estimation of overall treatment effect
- Identify sources of bias
- Enhance precision

Utility of ISE

Help assess treatment effects on secondary endpoints or on components of composite endpoints (e.g., mortality)

Help identify informative bridging studies
- regional differences
- subgroup differences

Help investigate treatment effects by subgroups, regions, intrinsic/extrinsic factors

Statistical Considerations in ISE

Existence of at least two positive studies?
- treatment effect in overall population conclusive?

Consistency in cross-study comparisons
- poolability issue

Multiplicity in statistical inference

Strength of statistical evidence

Statistical Considerations in ISE

Weighting each study in ISE
- pitfalls of D-L random-effects method

Controlling for design differences
- Stratification by studies might not be sufficient (need to understand real study-to-study differences)

Reduce bias by pre-planning how data will be integrated and analyzed prospectively

Additional Inconsistency Analysis

- Percentage of total variation that is due to heterogeneity rather than chance (Higgins et al, 2003, BMJ)

Q: Cochran's Q test statistic with *df* degrees of freedom

$$I^2 = \max(100\% \times \frac{Q - df}{Q} , 0)$$

This was proposed to handle meta-analysis

Fererora-Gozalez et al (2007, BMJ) examined 114 identified CV RC trials from literature

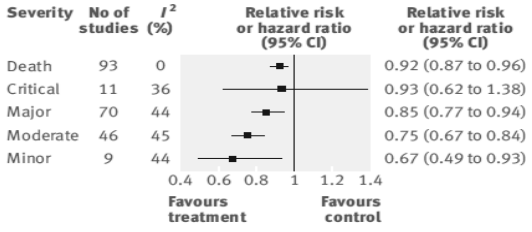
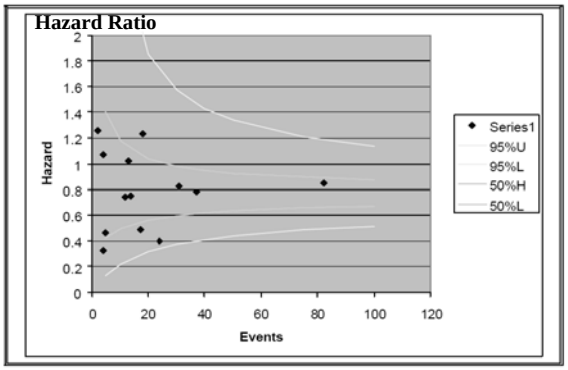


Fig 2 | Variability in magnitude of the effect of intervention across categories of importance to patients

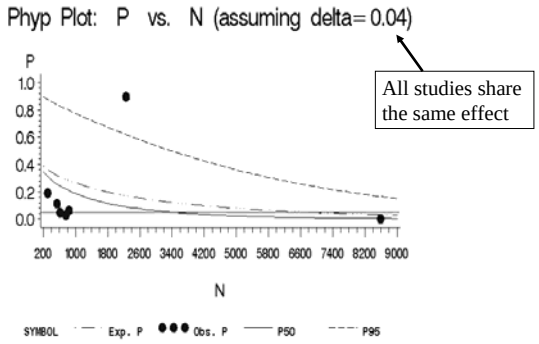
- Funnel plot
  - effect size estimate vs. N (or # of events) need CI curve?
- P*hyp* plot*
  - p-value & p-value quantile curve vs. N (or # of events), given delta
- Galbraith plot#: ln(OR)/se vs 1/se?
- Forest plot

*Hung, O'Neill, Bauer, Köhne (1997, Biometrics)
#Galbraith et al (1988, Stat. in Med.)

Funnel Plot of hazard ratio



Aspirin mortality effect in post-MI patients (data from Fleiss 1993)



- Analytic approach
  - testing treatment by study interaction?
  - testing qualitative interaction*
  - inconsistency assessment

***Gail & Simon, 1985, Biometrics**

Multiplicity Issue

Ex. In an application, three trials are run in parallel to study a drug effect as compared to the same control on the same clinical outcome composite 1° endpoint (including mortality) in three complementary patient populations.

It was proposed that if all three trials are positive, test all-cause mortality by pooling the three studies, if the mortality results are 'consistent'.

Type I error is program-wise, but *conditional* on 'all trials are positive' (an important condition)

Mortality evidence: worth 1?, < 1?, > 1? trial (this seems to be a *pseudo* trial)
- Mortality analysis correlated with clinical outcome endpoints in 3 separate trials
- Studywise type I error not part of consideration for mortality (***can be problematic***)

Same issues apply to integrated analysis for secondary endpoints, subgroup analyses, …

Concept of using an extreme p-value from one trial to lessen the evidence requirement of p-value for a positive replicate study is controversial.

Ex. Suppose that Trial 1 and Trial 2 are almost identical. Trial 1 gives p = 0.005. Trial 2 gives p = 0.07.
Q: Do we have one positive study or 2 positive studies?
It is difficult to argue that the evidence from Trial 1 is so strong that the second study with p = 0.07 will suffice to be said "positive" or "positively supportive".

Post-trial statistical power evaluated at observed effect $\Delta = d$ cannot help to determine whether nonsignificant result is truly negative or inconclusive due to insufficient sample size

$$\text{pr}\{ P < 0.025 \mid \Delta = d \} = \Phi(z_p - z_{0.025})$$

$p > 0.025 \Rightarrow$ Statistical power at $d < 50\%$

Hung and O'Neill (2003)

Therefore, it does not seem sensible to calculate the level of joint statistical significance by multiplying the two p-values (i.e., 0.005×0.07) or using some kind of multiplicity adjustment to make $p = 0.07$ statistically significant.

Whether a study can be counted as a "supportive" study or not is a *multi-dimensional* issue
- p-value
- estimate of treatment
- comparability of design characteristics, e.g. population
- consistency ……

Weighting Studies

- When there are substantial unexplainable study-to-study differences in the results, some type of random-effects analyses are necessary
- Weighting in D-L random-effects method may be hard to justify when sample sizes of study are quite various
  - shrink weight of large study and bolster weight of small study

Ex. Mortality effect of Aspirin in Post-MI patients

Study	Aspirin		Placebo		Relative Risk (95% CI)
	N	Death rate	N	Death rate	
MRC-1	615	8.0%	624	10.7%	0.74 (0.52, 1.05)
CDP	758	5.8%	771	8.3%	0.70 (0.48, 1.01)
MRC-2	832	12.2%	850	14.8%	0.83 (0.65, 1.05)
GASP	317	10.1%	309	12.3%	0.82 (0.53, 1.28)
PARIS	810	10.5%	406	12.8%	0.82 (0.59, 1.13)
AMIS	2267	10.9%	2257	9.7%	1.12 (0.94, 1.33)
ISIS-2	8587	18.3%	8600	20.0%	0.91 (0.86, 0.97)

Fleiss (1993)

The results indicate:
- the effect of aspirin, if any, attenuates over time
- in the two largest trials, AMIS shows a numerically adverse effect with aspirin, whereas ISIS-2 shows that the effect of aspirin is at best marginal though it appears to be statistically significant
- the relative risk in AMIS appears to be significantly different from the mean relative risk of the remaining studies ($p < 0.005$)

There is some question concerning the validity of pooling the results of AMIS with those of the remaining studies.

However, it would certainly be invalid to drop AMIS from the meta-analyses for the only reason that its measure of effect differed significantly from the measures in the remaining studies unless there are clinical or scientific reasons for why dropping is adequate.

If such heterogeneity were due to chance and one had assumed *a priori* that the studies' relative risks varied randomly one from another, then the meta-analysis of the seven studies may be based on some kind of random effects approaches.

Study	Weight by # of events	D-L weight
MRC-1	0.024	0.079
CDP	0.021	0.072
MRC-2	0.049	0.136
GASP	0.014	0.053
PARIS	0.028	0.089
AMIS	0.097	0.204
ISIS-2	0.766	0.367

Analyses using the two weighting schemes are needed to examine whether the results of integrated analyses are not quite different

As for CI generated from random-effects analyses, the coverage probability can be a problem [Follmann & Proschan, 1999]
e.g., for normal approximation is problematic until # of studies to integrate gets to at least 16 studies.

Coverage Probability Issue

- CI of treatment effect derived from D-L random-effect method needs conservative adjustment (Follmann & Proschan, 1999),
  e.g., use *t*-distribution instead of normal approximation
  - asymptotic normality requires a large number of studies, not sample sizes of studies

D-L random effects approach would yield that the point estimate of relative risk is 0.89 with 95% confidence interval (0.78, 1.02), suggesting that the effect of aspirin on prevention of death after myocardial infarction is inconclusive.
Weighting by (1/within-study variance) in random-effects approach gives a wider CI, so it is a moot point.

*** DerSimonian & Laird (1986)**
Follmann & Proschan (1999)

Prospective Planning for ISE

- To avoid potentially biased results
  - make clear the rationale for methodology
  - group similar studies, subgroups, etc.
  - control error rate
  - well thought out to reduce number of uninterpretable studies because of biases due to trial misconduct and incorrect inference

Additional Remarks

ISE should provide a comprehensive, detailed, in-depth analysis of the efficacy results in aggregate of all studies in submission

Focus on investigating study-to-study differences in the results

Need research on methodology for integrated dose-response analysis

Selected References

FDA Guidance for Industry Integrated Summary of Effectiveness (ISE), 2008

Gail & Simon (1985, Biometrics)

DerSimonian & Laird (1986, CCT)
Galbraith et al (1988, Stat. in Med.)
Hung, O'Neill, Bauer, Köhne (1997, Biometrics)
Fleiss (1993, SMMR)
Follmann & Proschan (1999, Biometrics)
Higgins et al (2003, BMJ)
Hung, O'Neill (2003, Biometrical Journal)