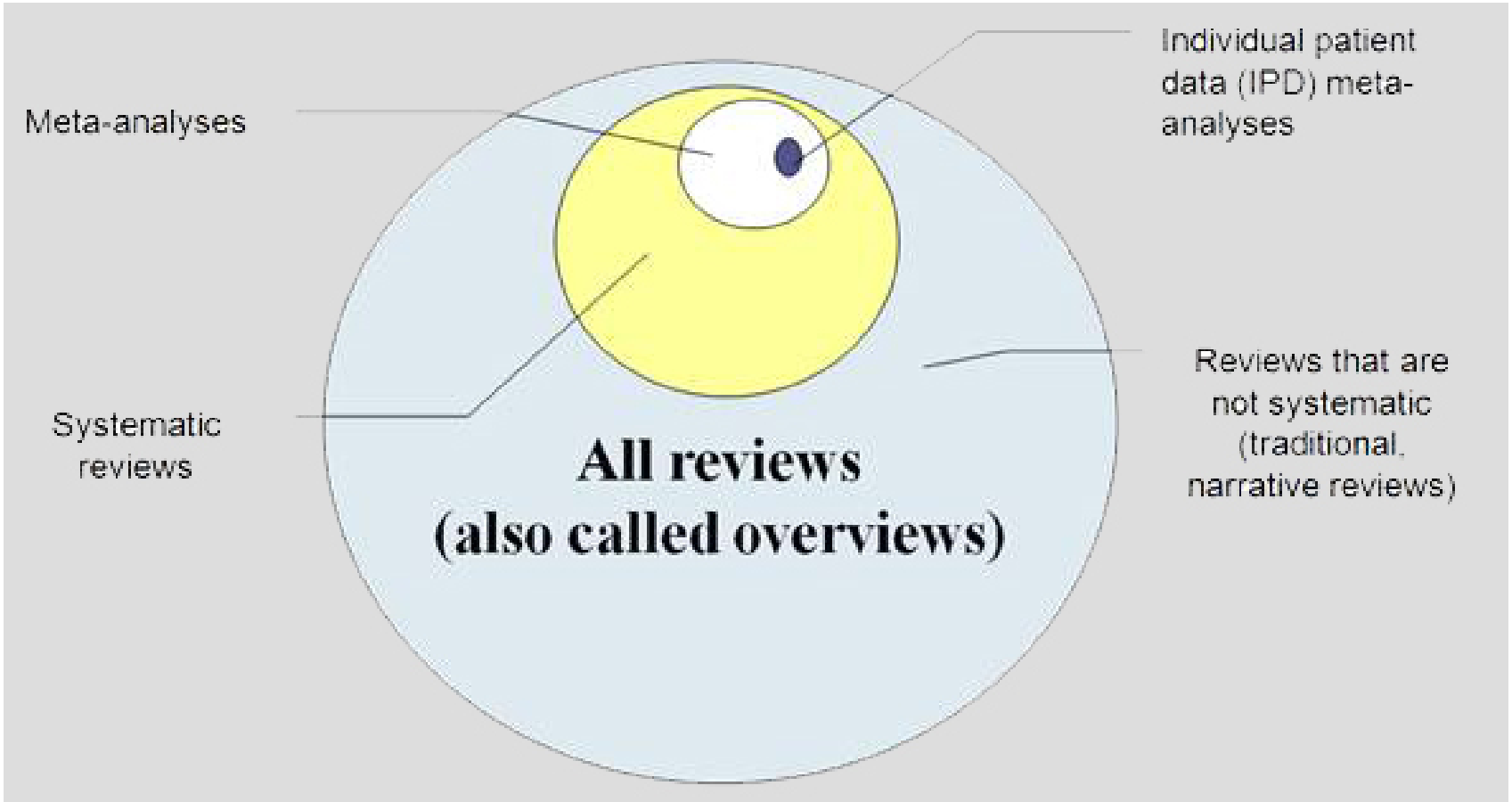
Applied Biostatistics

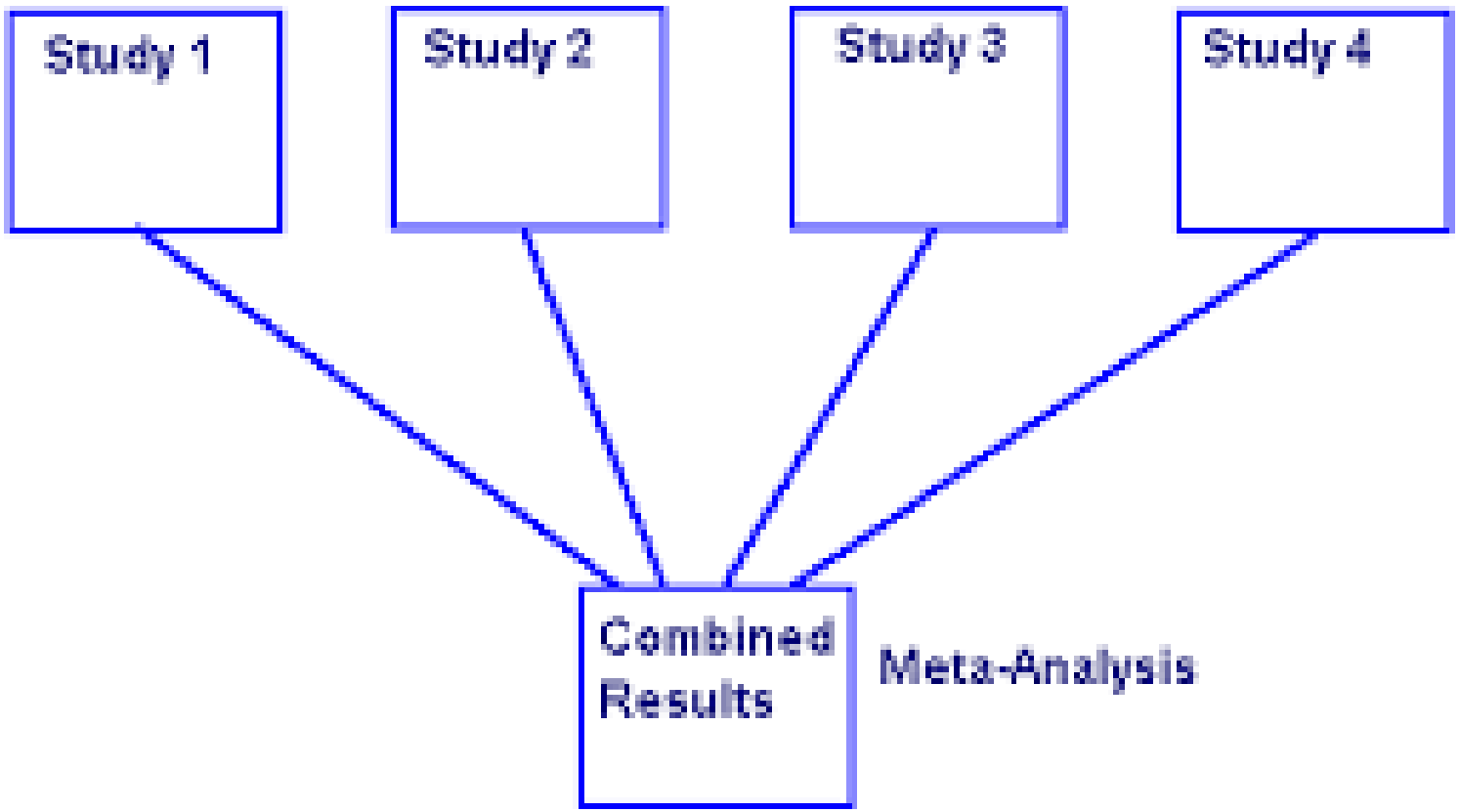
https://moodle.epfl.ch/course/view.php?id=15590

- Types of reviews
- Meta-analysis and combining information
- Bias/funnel plots
- Statistical analyses
- Power/sample size analysis
- Simulation studies

Types of reviews



Meta-analysis of independent studies



Meta-analysis : What is it ?

- *Meta-analysis* consists of statistical methods for combining *results* of independent studies addressing related questions
- Several different methods, including
  - Comparative binary outcomes : combining *odds ratios*
  - Continuous outcomes : combining parameter estimates via *fixed effects* or *random effects models*
  - Any outcome type : combining (transformed) p-values from hypothesis tests about the data
- In some situations it makes sense to instead combine *data* for the analysis
- This is not always appropriate - *Simpson's paradox*

Simpson's paradox

Hospital	Mild	Severe	Total
A	60/100	1/10	61/110
B	9/10	30/100	39/110

- *Which hospital is better ? ?*
- Hospital *B* has a higher success rate *for each disease type*
- **But :** Hospital *A* has higher *overall* success ! !
- This type of story occurs quite frequently in medical
- Moral of the story (short version) : **Don't combine this type of data set across different studies**

Meta-analysis : Why do it ?

- To obtain *increased power*
- Studies with small sample sizes are less likely to find effects even when they exist
- 'Integration-driven discovery' (IDD ; Choi *et al.*)
- Given the small (but increasing) size of many microarray experiments, meta-analysis might be considered a 'natural' approach to the problem of integrating results

What/how to combine

- Avoid pooling data prior to analysis : make comparisons *within study*
  - Compare like with like
  - Avoid Simpson's paradox
- Consider analysis goals : *which deviations* from the null you want to detect
  - Genes doing the *same thing* across studies (*e.g.* genes associated with increased survival)
  - Genes doing *different things* across studies (*e.g.* platform comparison)
- Use available information *efficiently*
  - Increase power

Combining information

Can consider a 'spectrum' of possible analyses for combining information – can combine at the level of :

- (Raw or adjusted) data
- Parameter estimates
- (Transformed) p-values
- Ranks
- Decision (*e.g.* in gene list or not)

Loss of information as move from more 'raw' to more 'processed' quantities

Meta-analysis : finding studies

- Publication databases
- Congresses
- Internet searching

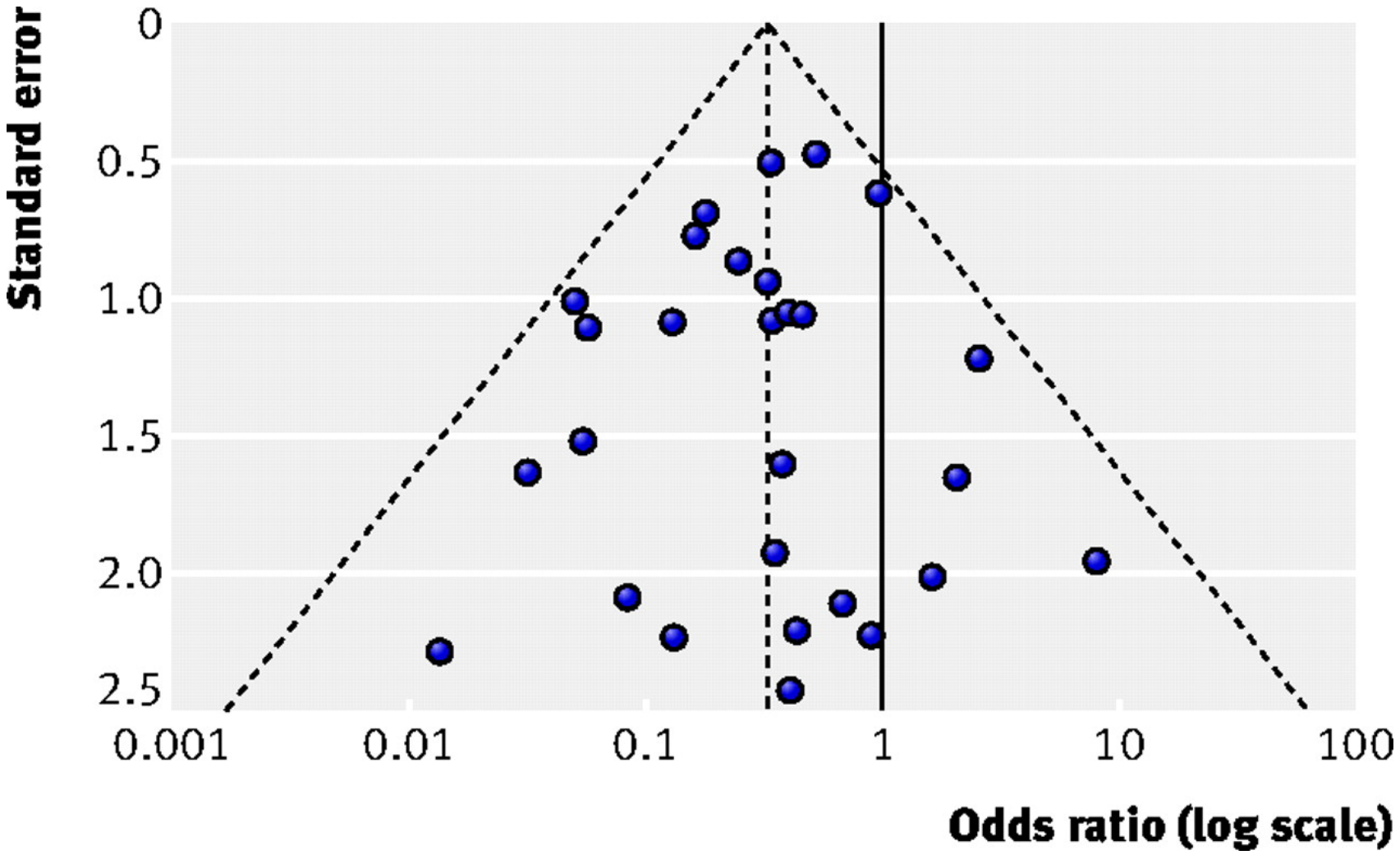
Meta-analysis : bias

- Bias is generally due to studies selected for inclusion being *insufficiently representative* of the totality of research being carried out
- Most commonly discussed is *publication bias* ('file drawer problem') : when the probability that a result is published depends on the the result
- Other information dissemination biases include :
  - language bias
  - availability bias
  - cost bias
  - familiarity bias
  - outcome bias

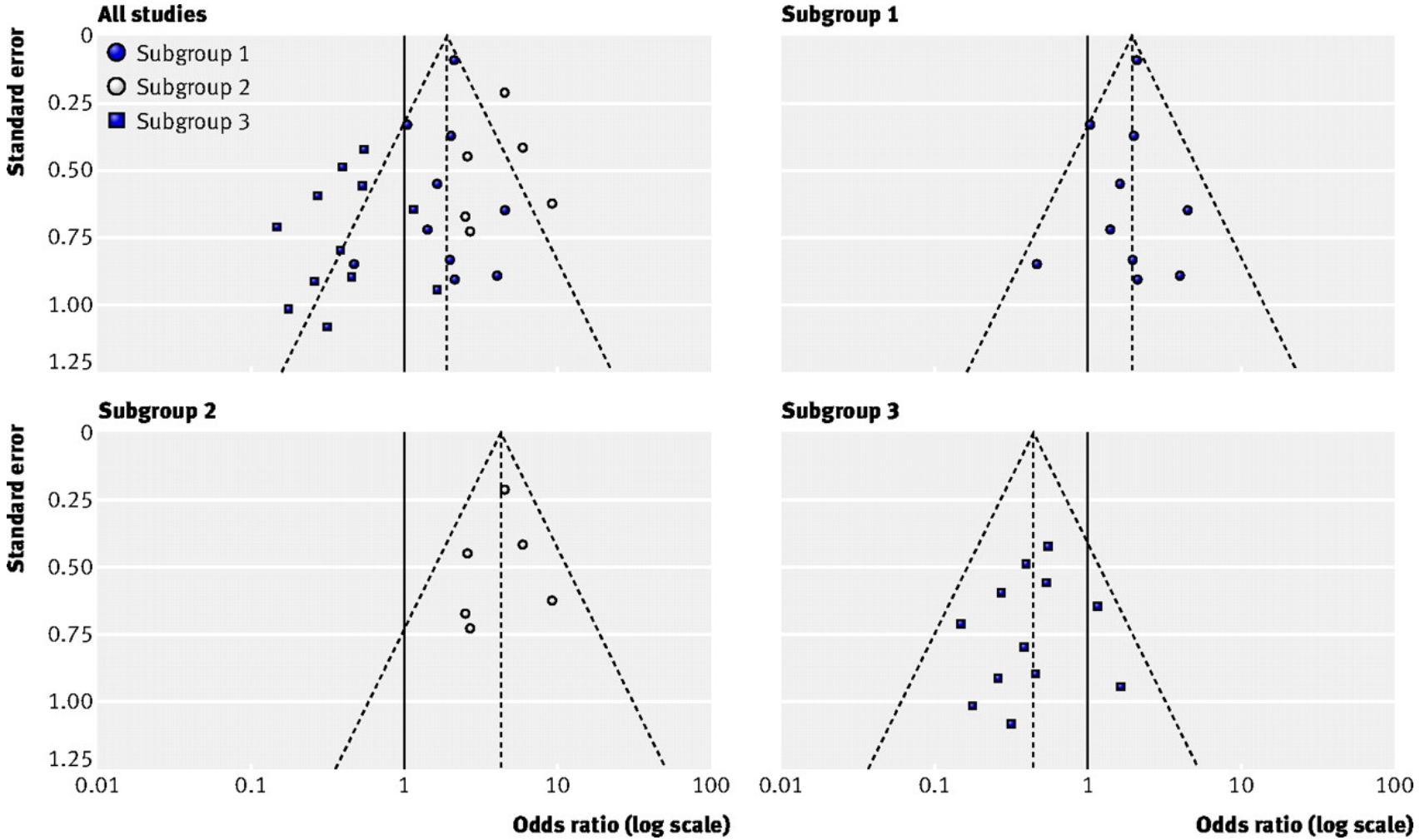
Graphical exploration of bias : funnel plot

- A *funnel plot* is a scatter plot of the effect estimates from individual studies compared to a measure of study size/precision (typically SE)
- Effect estimates from smaller studies should scatter more widely
- In the absence of bias and between study heterogeneity, the scatter will be due to sampling variation alone and the plot will resemble a symmetrical funnel
- A triangle centered on a fixed effect summary estimate and extending 1.96 standard errors either side will include about 95% of studies if no bias is present and the fixed effect assumption (that the true treatment effect is the same in each study) is valid

Funnel plot : symmetry



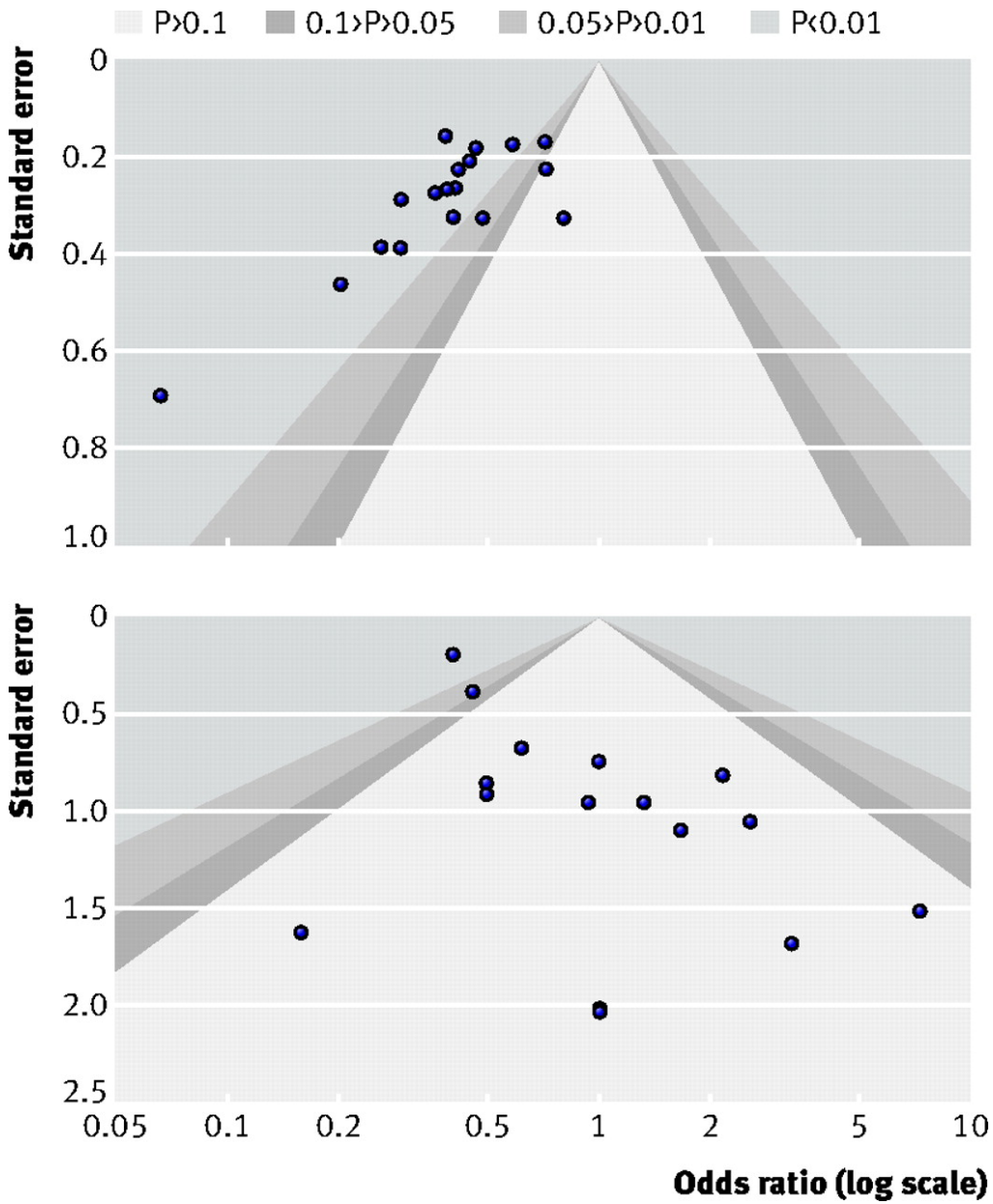
Funnel plot : subgroup problem



Possible sources of asymmetry in funnel plots I

- Reporting biases
  - Publication bias/file drawer problem
  - Delayed publication (time lag or pipeline) bias
  - Location biases (eg, language bias, citation bias, multiple publication bias)
  - Selective outcome reporting
  - Selective analysis reporting
- Poor methodological quality → spuriously inflated effects in smaller studies
- Poor methodological design
- Inadequate analysis
- Fraud
- Heterogeneity between studies of differing size
- Artifacts/batch effects : association between effect and its SE
- Chance error → motivates assessing plot for symmetry

Funnel plot : examination for publication bias



Steps in Meta-Analysis

1. Define the research question and specific hypotheses
2. Define the criteria for including and excluding studies
3. Locate research studies
4. Determine which studies are eligible for inclusion
5. Classify and code important study characteristics (e.g., sample size; length of follow-up; definition of outcome; drug brand and dose)
6. Select or translate results from each study using a common metric
7. Aggregate findings across studies, generating weighted pooled estimates of effect size.
8. Evaluate the statistical homogeneity of pooled studies
9. Perform sensitivity analyses to assess the impact of excluding or down-weighting unpublished studies, studies of lower quality, out-of-date studies, etc.

Problem : study heterogeneity

In general, studies may vary in

- scientific research goals
- population of interest
- design
- quality of implementation
- subject inclusion and exclusion criteria
- baseline status of subjects (even with the same selection criteria)
- treatment dosage and timing
- management of study subjects
- outcome definition or measures
- statistical methods of analysis

Test of homogeneity

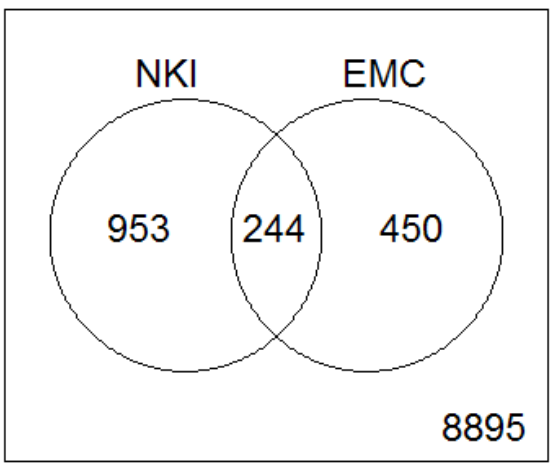
- Cochran test for homogeneity tests for equality of estimates against the alternative that at least one is different
- Test statistic $Q = \sum_{i=1}^{k} w_i(\hat{\beta}_i - \bar{\beta}_.)^2$
- $\hat{\beta}_i$ estimates the treatment effect (the HD coefficient in the linear model for a given gene) in study i
- w_i is the weight for study i (most commonly taken as the reciprocal of the variance of the outcome estimate)
- $\bar{\beta}_. = \sum_i w_i \hat{\beta}_i / \sum_i w_i$ is the weighted average treatment effect
- Under the null, $Q \sim \chi^2_{k-1}$

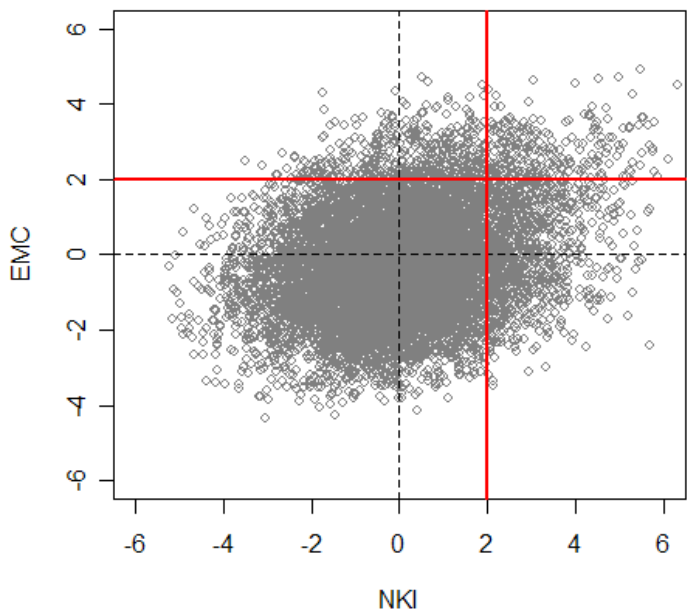
Popular methods of combination

- Combine *decisions* : 'Venn diagram'
- Combine *parameter estimates* :
  - Fixed effects meta-analysis (FEMA)
  - Random effects meta-analysis (REMA)
- Combine *p-values* : Fisher p-value combination
- Combine *test statistics* (or p-values) : Combining z-scores

Venn diagram

- Selects genes significant *in both (all) studies*
- This rule seems intuitive for biologists
- **Problem :** *what does 'reproducible' mean ?*
- At the top are signal (true +) *and* noise (false +)
- **This method has very low power, and is NOT recommended**





Combining estimates : heterogeneity analysis

- Before combining estimates from different studies, verify that they are *homogeneous*, *i.e.* do they all seem to be estimating the same underlying population parameter
- Graphical methods (*e.g.* forest plots) are useful when there are several *single outcome studies* to be combined
- For a *microarray study*, need one plot for each gene
- => Use numerical assessment

Fixed effects model

- Each individual study estimate $\hat{\beta}_i$ receives weight w_i *inversely proportional to its variance*
- The weighted estimates are combined to yield an overall effect estimate $\bar{\beta}_\cdot = \frac{\sum_i w_i \hat{\beta}_i}{\sum_i w_i}$
- The variance of the weighted estimator is $1/\sum_{i=1}^{k} w_i$

Random effects model

- If there is heterogeneity between studies, then assume *no single underlying value of the effect*
- Instead, there is *distribution of values*
- Differences among study results are considered to arise from both *between-study variation* of true effect size and *chance variation*

FE vs. RE meta-analysis

- FE and RE are both ways to obtain a *single, combined par. est.* from a *set* of estimates obtained from different studies
- The combined estimates are *weighted averages*
- FE assumes there is *no heterogeneity between results* of the different studies
- In FE meta-analysis, each individual study estimate receives weight inversely proportional to its variance
- RE meta-analysis assumes that individual studies may be estimating *different* treatment effects
- Study weights adjusted to take into account additional variability τ^2 between studies : $w_i^* = \frac{1}{(1/w_i)+\hat{\tau}^2}$ (DerSimonian-Laird)
- When the additional variability between studies is 0, then the RE model reduces to the FE model
- If we assume *normality* of the estimates, we can get p-values

PAUSE

Fisher combined p-values

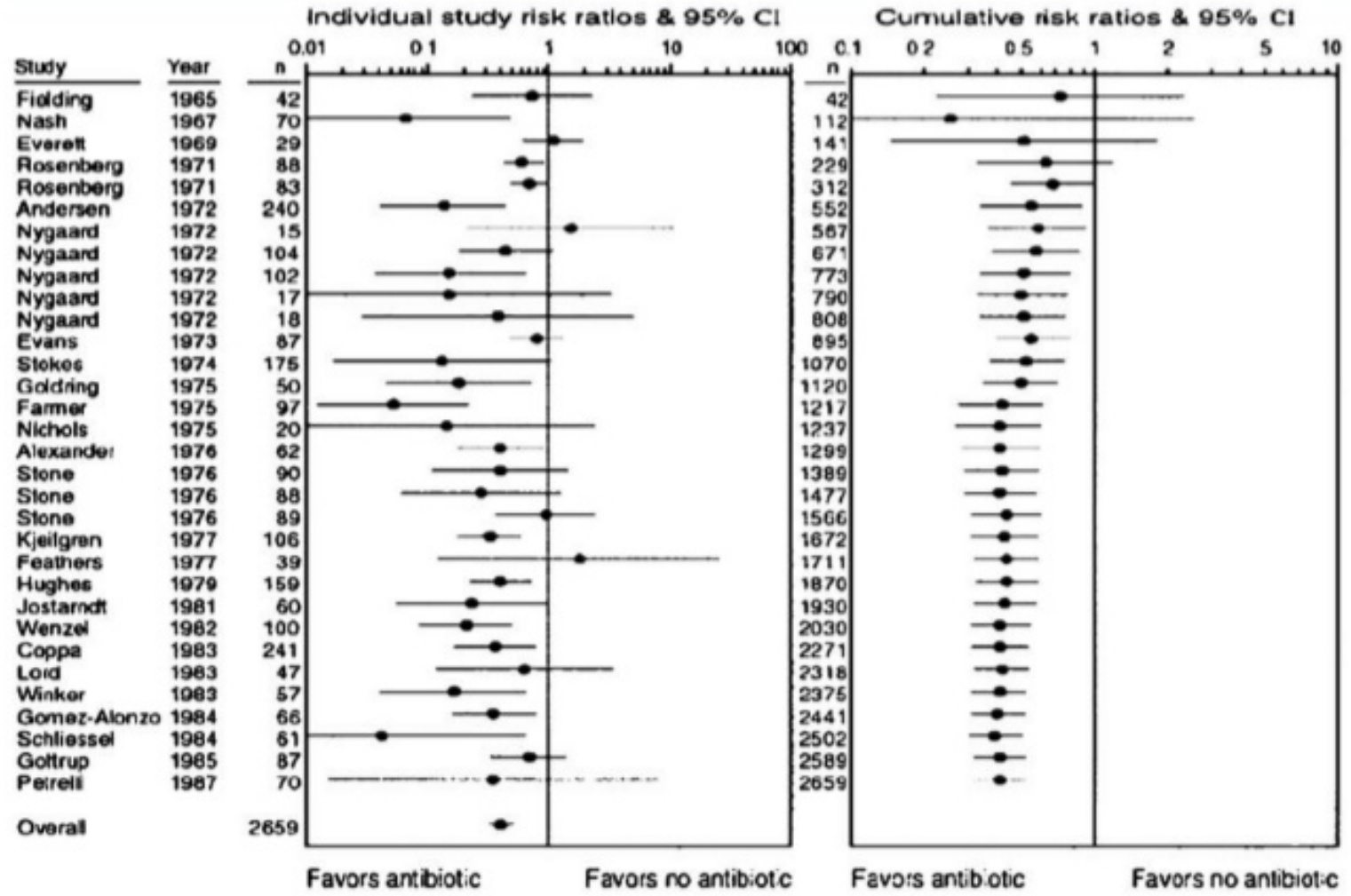
- Other methods for combining results focus on *p-values*
- Usually preferable to combine parameter estimates, but sometimes this is impossible – for example, if only p-values and no parameter estimates are given
- There are several possibilities for combining p-values, an old (1930s) and commonly used method is due to Fisher
- The Fisher summary test statistic $S = -2 \sum_{i=1}^{k} \log(p_i)$
- The theoretical null distribution of S should be χ^2_{2k}
- Can also obtain a p-value for S by *resampling*

Method of combining z-scores

- Can use when all test statistics have a *normal distribution*
- Can also be considered as part of class of methods based on *p*-value transformation (Stouffer's method)
  - *BUT :* not generally efficient if have original test statistics and these are not normal
  - In particular, *should not use* to combine χ^2 statistics
- Weighted or unweighted (*i.e.* equal weights) versions
- Simplest (unweighted) case : Combined $Z = \sum Z_i/\sqrt{k}$ has a standard normal distribution under the null

Forest plot

Forest plots of the meta-analysis addressing the use of antibiotic prophylaxis compared with no treatment in colon surgery



Example : Identifying genes associated with breast cancer survival

- Many gene expression (microarray) studies have been carried out in breast cancer patients
- Typically, these studies are looking for genes whose expression is associated with some outcome of interest :
  - stage/grade of tumor
  - response to treatment
  - time to relapse/metastasis
  - survival outcome
- Different studies find different genes
- *How to make sense of the results ?*

Methodology for genome-scale survival data

- Need *raw (or suitably processed) data*, not just p-value from previous study
- Response variable : metastasis-free survival, no covariates
- Multiple probes of the same genes *made unique* by choosing the most variable
- Do *NOT* need to consider only the common probes : *missing data readily accommodated* in this framework
- For each gene fit a separate Cox model :

$$h(t) = h_0(t)exp\{\beta_0 + \beta_j x_{ij}\}$$

(i = sample, j = gene)

- Can do p-value adjustment for multiple testing (*e.g.* FDR)

Difficulties with public data sources

- Lack of *independent* patient cohorts
- No standard variable names or representation of values
  - same name, different things
  - different name, same thing
  - need to document measurement technology (*e.g.* ER receptor status : immunohistochemistry, ligand binding assay, RT-PCR, microarray)
- Difficulty maintaining *consistent mapping* of probes to genes
- *Selective inclusion* of information
  - *e.g.* only data from a specific type of microarray
- Unclear or differing study design and patient selection criteria
  - tumor bank samples (population sampling)
  - patients selected for clinical trials
  - longitudinal data

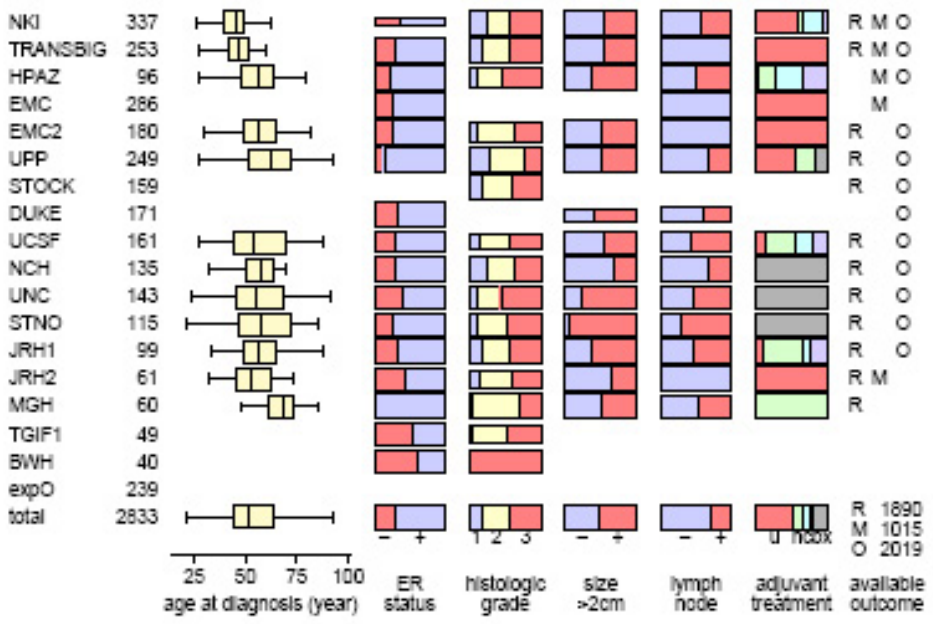
SwissBrod : Swiss Breast Oncology Database

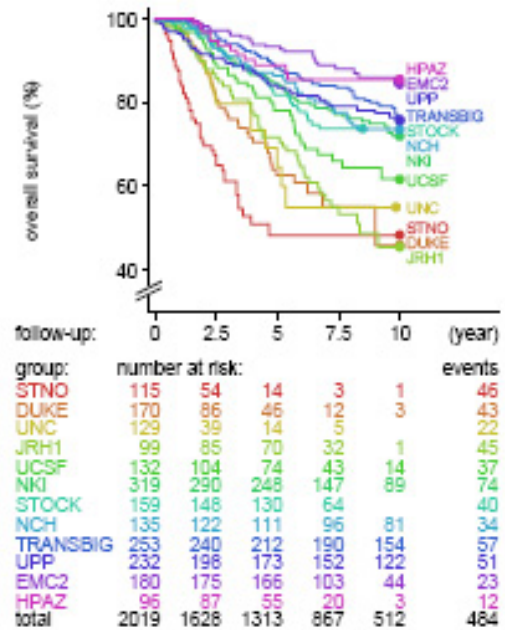
- *SwissBrod* provides curated clinical and expression data
- Aim to avoid these problems, facilitate data mining and integration, ensure high data quality
- Need to identify *actual sampling units* (patients, tissues, *etc.*) and *design* (patient selection criteria)
- Contains primary data on breast cancer (raw or normalized matrix of expression values)
- Data curation
  - primary dataset acquisition : public repositories, supplementary materials, author websites, *etc.*
  - quality control
  - reconfiguration to independent patients
  - annotate study design, selection criteria
  - stable probe identifiers

Publicly available breast cancer survival datasets

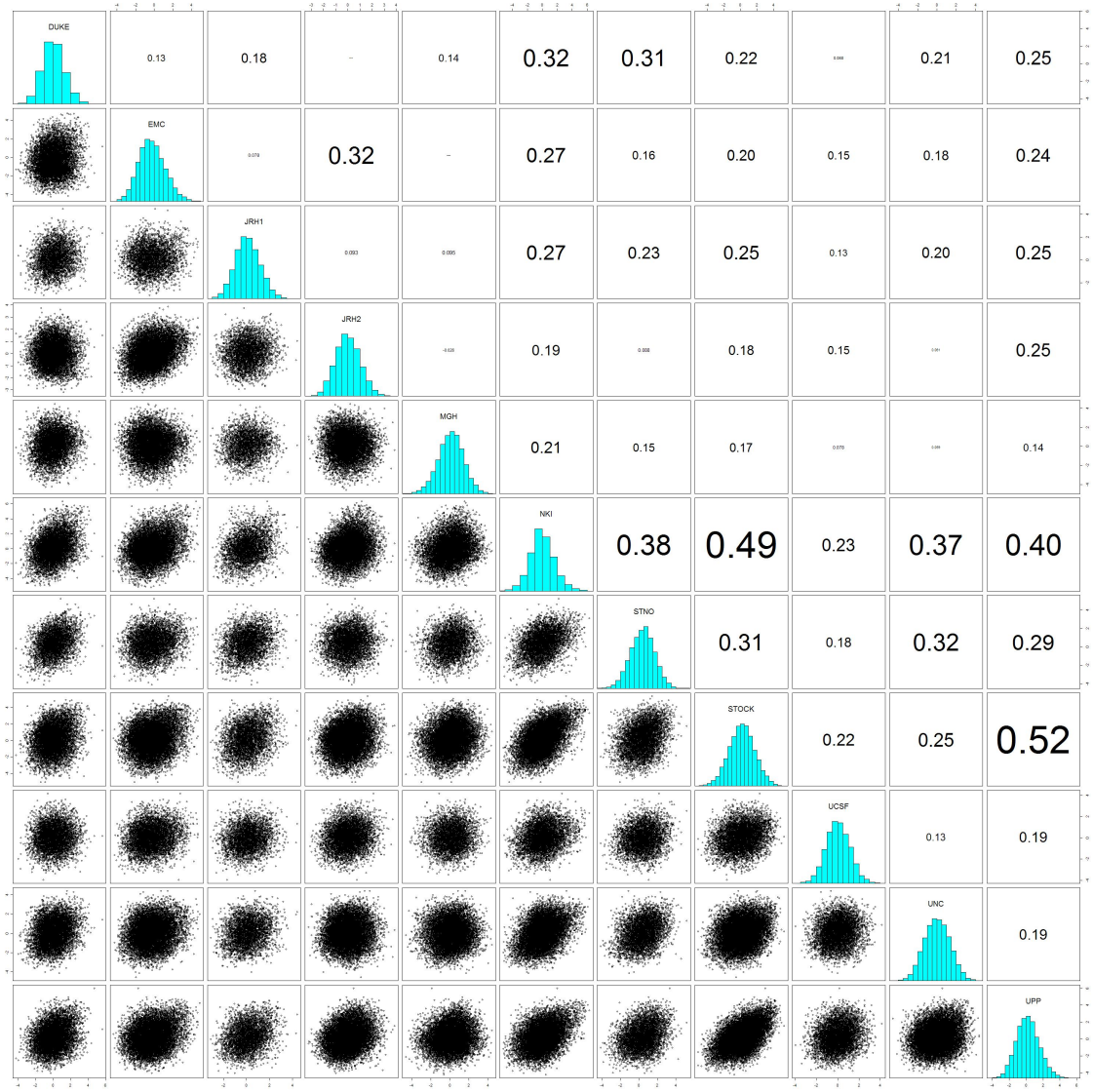
Dataset symbol	No. of arrays	Institution	Platform	Data source	No. of GeneIDs
NKI	337	Nederlands Kanker Instituut	Agilent	author website	13120
EMC	286	Erasmus Medical Center	Affy U133A	GEO :GSE2034	11837
UPP	249	Karolinksa Institute (Uppsala)	Affy U133A,B	GEO :GSE4922	15684
STOCK	159	Karolinska Institute (Stockholm)	Affy U133A,B	GEO :GSE1456	15684
DUKE	171	Duke University	Affy U95Av2	author website	8149
UCSF	161+8	UC San Francisco	cDNA	author website	6178
UNC	143+10	University of Carolina	Agilent HuA1	author website	13784
NCH	135	Nottingham City Hospital	Agilent HuA1	AE :E-UCON-1	13784
STNO	115+7	Stanford + Norwegian Radium Hosp.	cDNA	author website	5614
JRH1	99	John Radcliffe Hospital	cDNA	journal website	4112
JRH2	61	John Radcliffe Hospital	Affy U133A	GEO :GSE2990	11837
MGH	60	Massachusetts General Hospital	Agilent	GEO :GSE1379	11421
Total	2530	= 2505 carcinomas		Total # GeneIDs :	17198
		+ 25 non-malignant breast tissues		# common GeneIDs :	1963

Patient characteristics in breast cancer studies

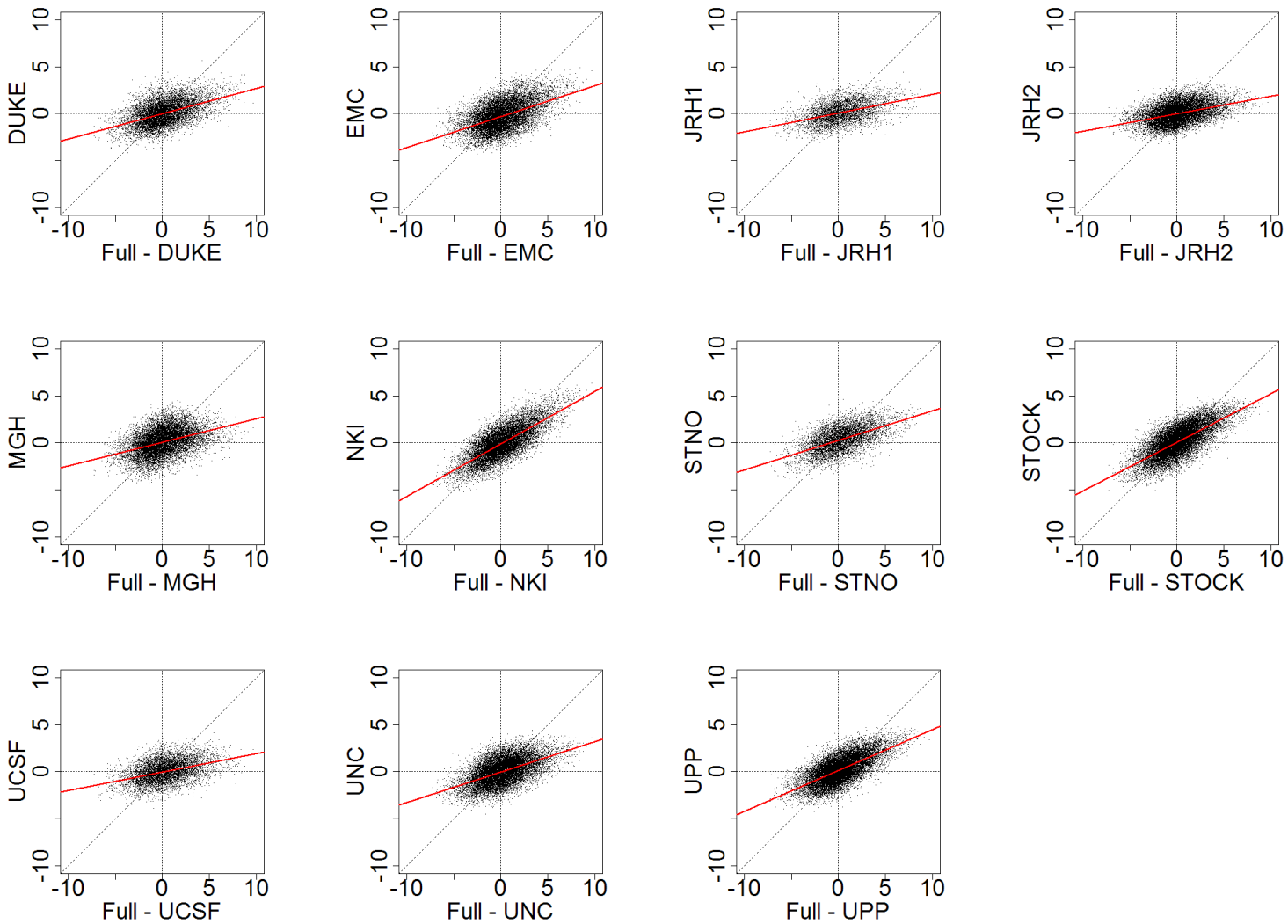




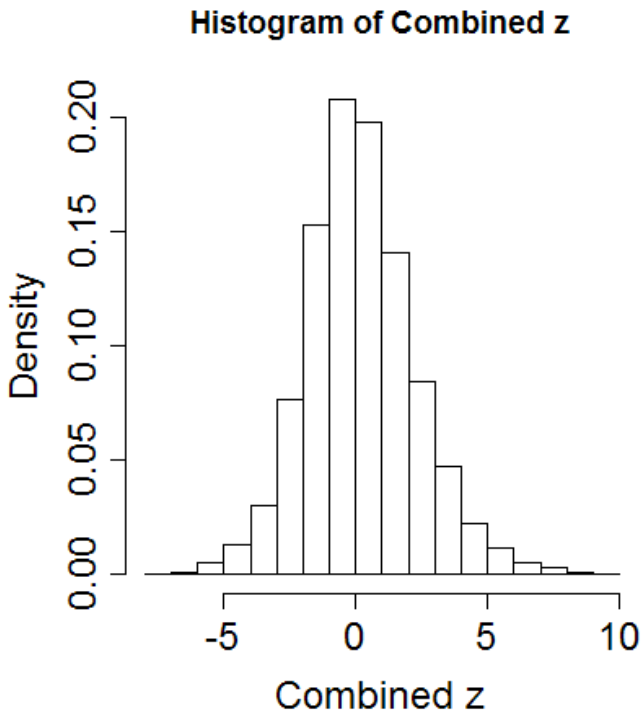
Pairwise scatter plots

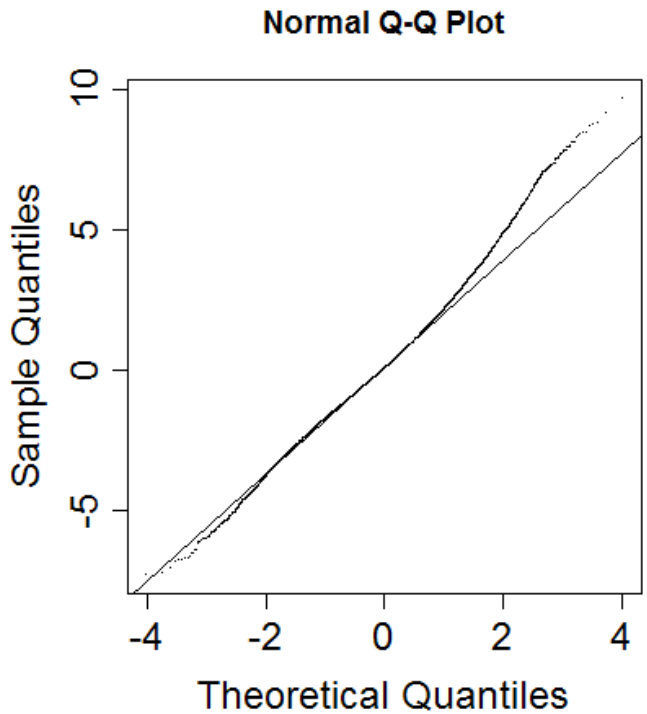


One set vs. z-score combination of the rest



Distribution of combined z

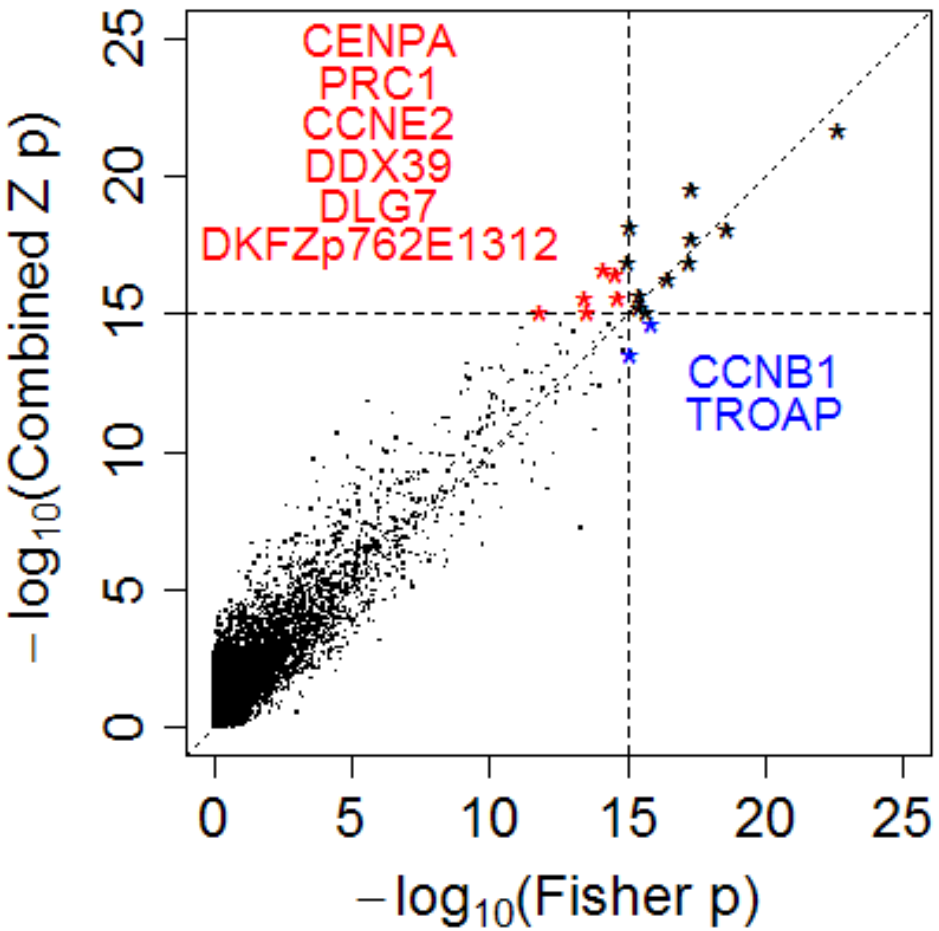




Preliminary results – Top 25 genes

symbol	Z	NKI	DUKE	UCSF	STNO	JRH1	MGH	UPP	STOCK	EMC	UNC	JRH2
*AURKA	9.67	6.33	1.09	2.33	3.05	1.83	1.56	3.38	3.28	4.52	3.55	1.16
*CCNB2	9.17	5.56	3.95				1.17	3.67	4.18	3.64	2.70	1.05
*MELK	8.82	4.51	4.10			2.77		3.64	3.84	3.31	2.11	0.66
*MYBL2	8.79	4.94	3.20	0.56	3.38	2.73	1.23	4.37	3.02	2.61	3.01	0.11
*BUB1	8.70	4.43	1.15	1.24	3.65	2.63	0.79	2.88	4.24	3.37	2.78	1.69
*AURKB	8.47	5.01	4.12	−0.12	3.56	2.09		3.44	3.71	1.15	3.00	0.84
*RACGAP1	8.47	5.48					0.48	4.24	3.76	4.91	1.99	1.56
CENPA	8.40	5.75	2.43	2.35				3.41	3.70	2.84	2.19	1.09
DDX39	8.35	5.49	3.29				1.09	3.53	4.49	2.71	1.15	1.89
*UBE2C	8.32	5.63	3.56	1.15	2.07	0.66		3.68	3.48	3.43	1.70	0.94
*FEN1	8.15	5.31	1.43	0.81	1.92	1.99		4.49	3.28	2.47	3.05	1.00
DLG7	8.13	4.31	2.64	0.88	3.14	1.27		3.18	3.96	3.75	1.81	0.77
p762E1312	8.12	6.10					1.68	4.00	3.72	2.52	2.73	0.74
*TRIP13	8.02	4.97	3.11	0.53	2.90	0.71		4.33	3.79	1.34	2.68	1.01
*GPI	7.97	4.12	3.16	0.75	3.77	1.76	1.75	3.61	3.34	0.16	3.58	0.45
CCNE2	7.97	5.31	2.90					2.46	3.01	4.27	1.55	1.58
PRC1	7.96	5.80			−0.01			4.35	3.72	3.50	2.16	1.54
CCNB1	7.84	4.76	3.23	−1.33	2.41	0.51		4.30	3.71	3.12	1.81	2.28
SEC61G	7.83	4.61	1.47	1.37	3.74	2.13	2.72	3.48	2.84	2.17	0.57	0.87
CENPF	7.83	3.44	1.53	1.41	2.93	1.93		2.90	4.37	2.65	2.13	1.46
GINS2	7.79	5.21						4.16	4.00	3.36	0.64	1.70
ZWINT	7.75	4.59	1.80	0.52			1.32	4.63	3.28	2.95	2.50	1.65
SPAG5	7.74	5.02	2.48	0.71			0.91	4.20	3.73	2.78	3.24	0.15
KIF23	7.69	3.53	2.02	−0.26	4.06	2.49	0.04	3.32	4.02	2.27	2.85	1.17
UBE2S	7.64	4.45	2.62	1.06	1.66	0.59		4.42	4.22	2.36	0.99	1.77

Combined Z compared to Fisher p



Concluding remarks

- Pooling raw data not always possible or desirable
- Integrating information across studies might not be straightforward even in the 'simplest' cases – several decisions required before data analysis can proceed
- Data adjustment does not necessarily remove artifacts/batch effects
- Between *and within* lab variability should be examined where possible
- These results have substantial implications for large studies, where patients are recruited over time, arrays not hybridized at the same time, ...
- Can compare *results* from different methods of analysis, but textitcan't assess method performance or robustness – 'known truth' not available (but can get an idea of this using simulation studies)