Probabilistic Pitfalls in Pursuit of Rigor and Reproducibility

Kevin W. Dodd

Biometry Research Group, Division of Cancer Prevention

NIH NATIONAL CANCER INSTITUTE

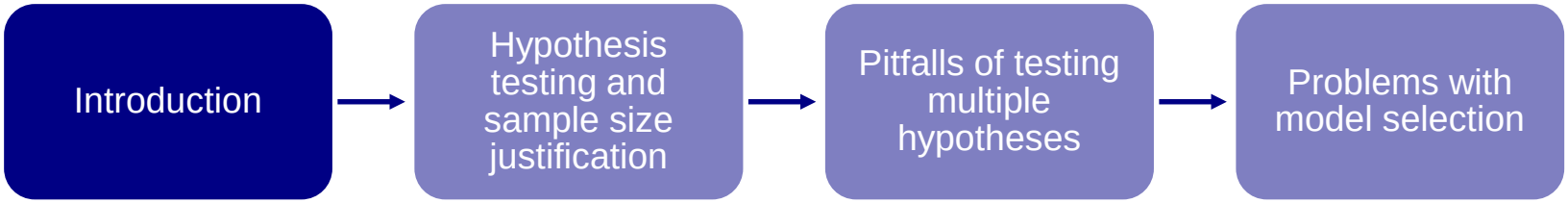
Collaborators

Victor Kipnis

Douglas Midthune

Grant Izmirlian

Lev Sirota



INTRODUCTION

Reproducibility crisis in biomedical research

"A growing chorus of concern, from scientists and laypeople, contends that the complex system for ensuring the reproducibility of biomedical research is failing and is in need of restructuring"

– Collins and Tabak, **Nature**, 2014

Guidance: Rigor and reproducibility in grant applications

https://grants.nih.gov/reproducibility/index.htm

Training

https://www.nih.gov/research-training/rigor-reproducibility

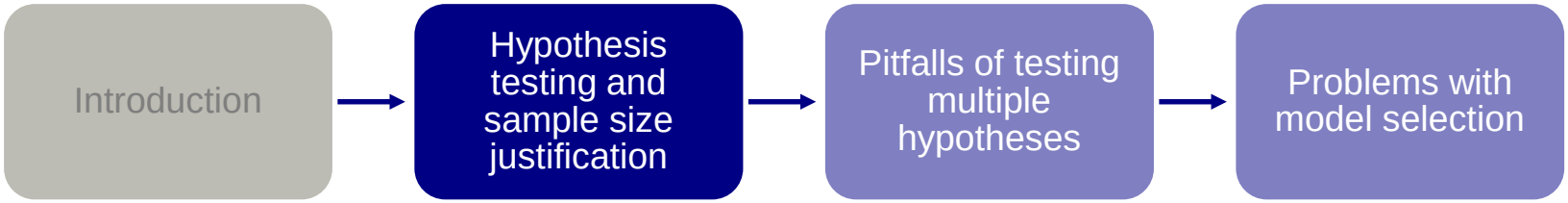
https://www.nih.gov/research-training/rigor-reproducibility

- Primary focus is on transparency in
  - Study design
  - Data handling
  - Proposed analysis
- Experimental designs and analysis methods should be chosen to avoid bias and reduce noise
  - Consideration of potential confounders (e.g., sex)
  - Authentication of biological/chemical resources

A statistician reads a grant proposal

Most relevant sections to a statistician

1. Study Design
   - How will my study be set up?
   - What data will I collect?
2. Analysis Plan
   - How will the collected data be analyzed?
3. Sample Size Justification
   - How much data should I collect?



HYPOTHESIS TESTING AND SAMPLE SIZE JUSTIFICATION

Example: Aspirin and cardiovascular mortality

Study design

- Randomize 1:1 to placebo/low-dose aspirin
- Collect data on cardiovascular mortality and potential confounders

Analysis plan

- Logistic regression

Sample size justification

- With *n* participants, there is 80% power to detect a 20% difference in CV mortality at $\alpha = 5\%$

Example: Aspirin and cardiovascular mortality

Study design

- Randomize 1:1 to placebo/low-dose aspirin
- Collect data on cardiovascular mortality and potential confounders

Analysis plan

- Logistic regression: outcome is CV death (yes/no)

Sample size justification

- With *n* participants, there is 80% power to detect a 20% difference in CV mortality at $\alpha = 5\%$

Foundation of hypothesis testing

		Your decision	
		Discovery	**No Discovery**
True state of nature	**Something to Discover**	√	False Non-discovery
	Nothing to Discover	False Discovery	√

- True state of nature vs. your **decision** based on data
- Your decision can be **discovery** or **no discovery**
- Wrong decisions may have different consequences

Similar to a trial by jury in US law

		Jury decision	
		Conviction	**Acquittal**
True state of nature	**Guilty**	√	Improper Acquittal
	Not Guilty	False Conviction	√

- True state of nature is **guilty** or **not guilty**
- Jury decision can be **conviction** or **acquittal**
- False conviction worse than improper acquittal

Rationale for hypothesis testing

		Your decision	
		Discovery	**No Discovery**
True state of nature	**Something to Discover**	√	False Non-discovery
	Nothing to Discover	False Discovery	√

- Null hypothesis set up according to state of nature (here, Nothing to Discover) where worse mistake can be made
- Power = chance of correctly identifying a Discovery
- α = chance of False Discovery = “Type I Error Rate”

Relationship between power and sample size

- For fixed *n* and *α*, power depends on the magnitude of the “distance” from the null hypothesis to the true state of nature
- For fixed *α* and distance, power increases with *n*
- For fixed *α* and power, smallest detectable distance decreases with *n*

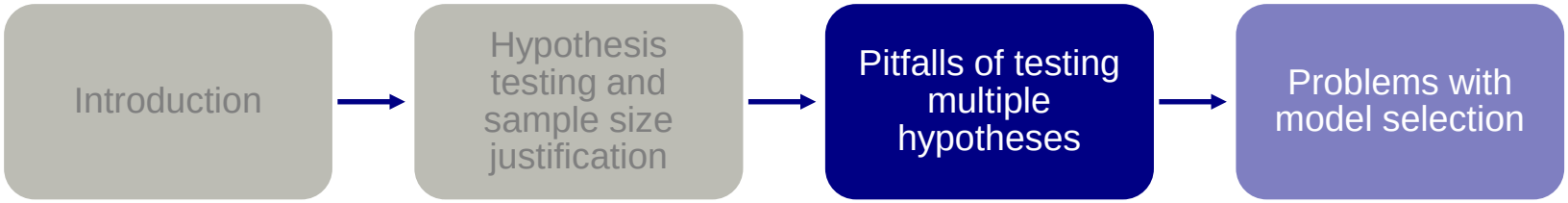
Physician's Health Study (1982-1990)

- 2×2 factorial study design:
  - Aspirin vs placebo for cardiovascular mortality
  - Beta-carotene vs placebo for cancer incidence
- Power not calculated for cancer endpoint
- 22,000 participants yields 80% power to detect a 20% difference in CV mortality at $\alpha = 5\%$

PHS: Final Report on Aspirin Component (1989)

- No reduction in CV mortality (RR = 0.96, p = 0.87)
- Reduced risk of myocardial infarction (RR = 0.56, p < 0.00001)
- Increased risk of stroke (RR = 1.22, p = 0.15)
- Increased risk of hemorrhagic stroke (RR = 2.14, p = 0.06)

"This trial of aspirin for the primary prevention of cardiovascular disease demonstrates a conclusive reduction in the risk of myocardial infarction."



PITFALLS OF TESTING MULTIPLE HYPOTHESES

Testing multiple hypotheses

- Testing multiple hypotheses means there are more chances to make mistakes (of both types)

- Leads to
  - Increased chance to make a false discovery when all nulls are true – Familywise error rate (FWER)
  - Decreased chance to make all discoveries when all nulls are false – Global power (GP)

Familywise error rate and global power

# Tests	Per Test		All Tests	
	α	Power	FWER	GP
1	5%	90%	5%	90%
2	5%	90%	10%	81%
4	5%	90%	19%	66%
10	5%	90%	40%	35%
100	5%	90%	99.8%	0.003%

- A sample size sufficient for a study with a single endpoint is underpowered for multiple endpoints
- False discoveries and low power across multiple tests may be a significant obstacle to reproducibility of a study

Sample size justification for multiple hypotheses

To ensure a specified global power (GP) with FWER ≤ α

- For m hypotheses, test each at significance level α/m
- Calculate sample size for per-test power $(GP)^{1/m}$
- Required sample size balloons quickly as m gets large

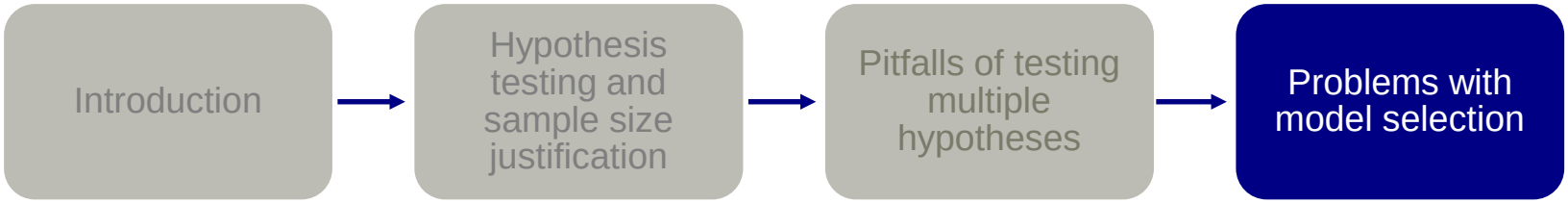
# Tests	α/m	$(80\%)^{1/m}$	n
1	0.05	80%	102
2	0.025	89.4%	168
4	0.0125	94.6%	240
10	0.005	97.8%	342
100	0.0005	99.8%	616

Alternative approach to controlling error rate

- Bonferroni methods can be used to control probability to reject at least one true null hypothesis (FWER)
- Benjamini and Hochberg (1995) proposed to control the expected proportion of errors among rejected hypotheses (FDR)
  - Weaker control of Type I errors, but more powerful
  - Sample size calculations more difficult

Takeaways

- Restrain the urge to test many hypotheses
- If you want to test multiple hypotheses, power the study for all of them
  - Justify the increased sample size up front as enhancing reproducibility
  - Classification of endpoints as “primary”, “secondary”, and “exploratory”, but powering the study for only primary endpoints is suspect
- Consult your statistician early and often!



PROBLEMS WITH MODEL SELECTION

Ongoing changes in landscape of cancer research

- Increased interest in biological targets
- Aided by advances in multiple molecular disciplines
- Assessments of comprehensive sets of biological molecules
  - DNA, RNA, proteins, metabolites, and more…
- “OMICS” refers to experimental analysis of these types of molecules
  - genomics, proteomics, metabolomics, etc.

OMICS studies

"Omics research generates complex high-dimensional data which are used to produce a model defined as series of steps in data processing, as well as the mathematical formula(s) to convert data into prediction of the phenotype of interest."

– Institute of Medicine, **Evolution of Translational Omics: Lessons Learned and the Path Forward**, 2012

Application: Panel of biomarkers for cancer screening

Idea: Look for differences in the biomarker profiles of cancerous vs. normal tissue/blood/serum

Goal is twofold:

1. Find the subset of markers that actually differ

2. Create a rule, based on the chosen subset, that best separates cancerous from non-cancerous profiles

   – "best" = optimizing a desirable quantity, e.g., AUC

Application: Panel of biomarkers for cancer screening

Potential problems:

- OMICS data profiles typically have many more variables per sample than the number of biological samples comprising the data set
- Looking for the subset of markers to use in panel analogous to making multiple hypothesis tests
- Final rule has additional estimation error

Model selection methods

- Traditional methods for finding the optimal rule were developed (and well-studied) for the case when the set of covariates (e.g., markers) is prespecified

- Techniques have been developed to produce a rule when the set of covariates is unknown *a priori*
  - Statistical properties not well understood
  - Can be assessed via simulation studies

Simulation study

- Case/control study design:
  - # subjects (n) = 60 (30 cases, 30 controls)
  - # potential markers (m) = 30, 60, 120, 240
  - 1000 simulated data sets
- Markers X:
  - $X_j \sim \text{Normal}(0,1), \quad j = 1, \ldots, m$
  - Independent
- Binary response Y:
  - Simulate Y by logistic regression on first 10 markers
  - $\log\{p / (1 - p)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{10} X_{10}$

True regression coefficients

Marker	Scenario 1		Scenario 2	
	Beta	AUC	Beta	AUC
X_1	1.5	0.81	1.5	0.81
X_2	1.0	0.85	1.5	0.87
X_3	1.0	0.87	1.0	0.89
X_4	0.5		1.0	
X_5	0.5		1.0	
X_6	0.5		1.0	
X_7	0.5		1.0	
X_8	0.5		1.0	
X_9	0.5		1.0	
X_{10}	0.5	0.89	1.0	0.94

Simulation study

- Model selection techniques:
  - Forward selection (classic)
  - LASSO (machine learning)
- Estimated logistic regression parameters characterize the full ROC curve, and thus can be used to calculate AUC
- Model performance (AUC) computed on an independent validation data set

Simulation results

Mean AUC for selected model

	Scenario 1 Theoretical AUC = 0.89		Scenario 2 Theoretical AUC = 0.94	
# potential markers	Forward Selection	LASSO	Forward Selection	LASSO
30	0.74*	0.78	0.74	0.82
60	0.71	0.74	0.69	0.76
120	0.67	0.71	0.65	0.72
240	0.66	0.70	0.62	0.66

* Standard errors for all estimated means < 0.005

Simulation results

Median number of markers selected (Total/True)

	Scenario 1 Theoretical AUC = 0.89		Scenario 2 Theoretical AUC = 0.94	
# potential markers	Forward Selection	LASSO	Forward Selection	LASSO
30	3/3	8/6	5/4	13/9
60	4/3	8/5	6/4	13/7
120	4/2	6/4	5/3	10/5
240	3/2	5/3	5/2	10/4

Simulation results

Percentage of time most important* marker(s) selected

		Scenario 1 Theoretical AUC = 0.89		Scenario 2 Theoretical AUC = 0.94	
# potential markers	Important markers selected	Forward Selection	LASSO	Forward Selection	LASSO
30	At least 1	86	96	85	98
	Both	-	-	46	87
60	At least 1	84	93	84	96
	Both	-	-	44	74
120	At least 1	76	90	73	88
	Both	-	-	26	55
240	At least 1	71	83	61	82
	Both	-	-	15	41

* Marker(s) with regression coefficient $\beta = 1.5$

Conclusions from simulation study

- When selecting from a large number of potential markers, model selection techniques often
  - fail to select important markers
  - select unimportant markers
- This leads to poor performance/lack of reproducibility
- LASSO performs better than forward selection, but is still unsatisfactory with many potential markers

A note of caution

"In contrast, the biological rationale for the set of biomarkers in an omics-based test frequently is not well-defined scientifically. This puts an additional burden on the statisticians and bioinformatics experts involved in test validation to ensure that the biological data and computational model are scientifically sound."

– Institute of Medicine, **Evolution of Translational Omics: Lessons Learned and the Path Forward**, 2012

Takeaways

- Potentially hundreds of thousands of covariates
  - Only a few needles in very big haystacks
  - Historically, studies may have been grossly underpowered

- Let scientific rationale narrow the number of potential covariates and/or increase sample size, thus reducing likelihood of excessive overfitting

NATIONAL CANCER INSTITUTE

www.cancer.gov

www.cancer.gov/espanol