

Study Design, Data Analysis, Reporting, and Interpretation: Choices and Consequences

Amy H. Herring
Professor of Statistical Science & Global Health
Duke University

April 30, 2018

Most Important Consideration in Design, Analysis, Reporting, & Interpretation

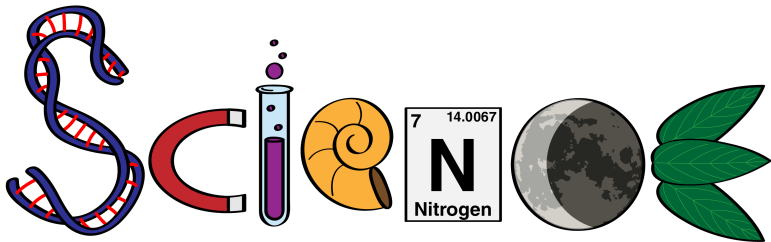


Figure courtesy of NOAA

Study Design Choices and Consequences

Design decisions affect the nature and interpretation of results. Design choices and consequences include the following.

Gold standard in many fields: randomized controlled trial

- ▶ Useful to examine effects of controlled exposures on biomarkers and subclinical outcomes
- ▶ Infeasible (\$) to recruit large number of participants
- ▶ Not ethical in every setting
- ▶ Data sharing often possible (controlled exposures+randomization reduce need for identifiers to estimate exposure or adjust for confounding)
- ▶ Example: MOSES (data available May 2018)

Study Design Choices and Consequences

Observational cohort studies

- ▶ Primary workhorse for studying associations with longer-term exposures
- ▶ Involve active (e.g., personal monitors, questionnaires) or passive (records-based) surveillance of individuals
- ▶ The better-characterized the cohort (e.g., personal information on potentially confounding contexts and behaviors, well-characterized exposures, clinically-confirmed outcomes), the more limitations on data sharing due to federally-mandated personal privacy protections (Common Rule, HIPAA Privacy Rule, informed consent restrictions)
- ▶ Example: CPS-II, Zigler HEI study (shareable data on Dataverse)

Challenges in Data Sharing: Locations and Dates

While data sharing limitations should not trump scientific principles, all other things being equal, we prefer to make data accessible as much as possible while safeguarding individual privacy.

Privacy considerations play a critical role in obtaining access to public data detailed enough to facilitate good science in the study of air pollution and health effects.

- ▶ Direct identifiers (e.g., lat/long, dates) typically not provided; access sometimes available in data enclaves (depends on resourcing; \$\$\$\$\$ to maintain); typically IRB approval and data use agreements required

Challenges in Data Sharing: Locations and Dates

While data sharing limitations should not trump scientific principles, all other things being equal, we prefer to make data accessible as much as possible while safeguarding individual privacy.

Privacy considerations play a critical role in obtaining access to public data detailed enough to facilitate good science in the study of air pollution and health effects.

- ▶ Direct identifiers (e.g., lat/long, dates) typically not provided; access sometimes available in data enclaves (depends on resourcing; \$\$\$\$\$ to maintain); typically IRB approval and data use agreements required
- ▶ Indirect identifiers (e.g., SES+race in limited geographic area) often must be modified before data sharing; data sometimes coarsened, or error added, before release

Challenges in Data Sharing: Locations and Dates

While data sharing limitations should not trump scientific principles, all other things being equal, we prefer to make data accessible as much as possible while safeguarding individual privacy.

Privacy considerations play a critical role in obtaining access to public data detailed enough to facilitate good science in the study of air pollution and health effects.

- ▶ Direct identifiers (e.g., lat/long, dates) typically not provided; access sometimes available in data enclaves (depends on resourcing; \$\$\$\$ to maintain); typically IRB approval and data use agreements required
- ▶ Indirect identifiers (e.g., SES+race in limited geographic area) often must be modified before data sharing; data sometimes coarsened, or error added, before release
- ▶ Try aboutmyinfo.org: Sweeney estimates 87% of US pop likely uniquely ID'ed by 5-digit ZIP, gender, MM/DD/YYYY of birth; 18% by county, gender, DOB. Using personal genome project data (voluntarily uploaded), they accurately matched $\geq 121/579$ files based on gender, 5-digit ZIP, & DOB.

Example: CMS Data “Levels” of Access

Easiest level to access: public use files

- ▶ Non-identifiable data
- ▶ Aggregated summary data (e.g., state or county level)
- ▶ Free and completely public
- ▶ Can download in seconds at cms.gov
- ▶ No requirements for use
- ▶ Your kid could use these data in a science project!

Challenges with Completely Public Data

While one may wish for completely public data, such data have generally been de-identified to the point that they provide little useful information for research on air pollution.

- ▶ Coarse resolution introduces significant measurement error into exposure estimate \implies noise may dominate signal

Challenges with Completely Public Data

While one may wish for completely public data, such data have generally been de-identified to the point that they provide little useful information for research on air pollution.

- ▶ Coarse resolution introduces significant measurement error into exposure estimate \implies noise may dominate signal
- ▶ Lack of information on individual-level potential confounders and risk factors (e.g., smoking, SES) can lead to biased risk estimates

Challenges with Completely Public Data

While one may wish for completely public data, such data have generally been de-identified to the point that they provide little useful information for research on air pollution.

- ▶ Coarse resolution introduces significant measurement error into exposure estimate \implies noise may dominate signal
- ▶ Lack of information on individual-level potential confounders and risk factors (e.g., smoking, SES) can lead to biased risk estimates
- ▶ Averaging exposures and outcomes over heterogeneous areas and groups can obscure true effects, leading to ecologic fallacy and disallowing valid inference

Challenges with Completely Public Data

While one may wish for completely public data, such data have generally been de-identified to the point that they provide little useful information for research on air pollution.

- ▶ Coarse resolution introduces significant measurement error into exposure estimate \Rightarrow noise may dominate signal
- ▶ Lack of information on individual-level potential confounders and risk factors (e.g., smoking, SES) can lead to biased risk estimates
- ▶ Averaging exposures and outcomes over heterogeneous areas and groups can obscure true effects, leading to ecologic fallacy and disallowing valid inference



Example: CMS (+ other cohorts) Higher Levels of Access

	Limited	Research Identifiable
Individual-level claims	Yes	Yes
Exact dates	Limited	Yes
ZIP code	No	Yes
Coarsened/eliminated fields?	Yes	Mostly No
Data use agreement banning redistribution	Yes	Yes
Study protocol required	2-3 pgs	Extensive
Robust data mgmt/storage plan req'd	No	Yes
Proof of funding & IRB approval req'd	No	Yes
Payment	\$\$\$	\$\$\$\$
Processing Time	3-4 wks	3-4 mos

Multiple Testing: When to Worry

- ▶ In null hypothesis significance testing, p-values are uniformly distributed under the null hypothesis
- ▶ This means that *when the truth is that nothing is happening*, if we have a study with one test at significance level 0.05, we should expect to see $p < 0.05$ 5% of the time
- ▶ These “false” significant results are called type I errors
- ▶ The more tests we conduct *when the truth is nothing is happening*, the more type I errors we expect to see

Example: Null Association

Consider a simple simulation study to illustrate the problem with multiple comparisons under the null hypothesis.

Hypothetical Scenario: Rashid opens French bakery and shares treats

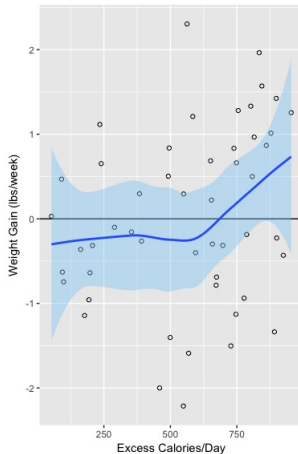
Exposure: Excess caloric intake/week of $n=50$ HEI staff members and friends: $U(0, 1000)$

Outcome: Weight gain per week of $n=50$ HEI partners and committee members not eating Rashid's treats, matched alphabetically: $N(0, 1)$

Exposure is independent of outcome so any $p < 0.05$ identified are type I errors.

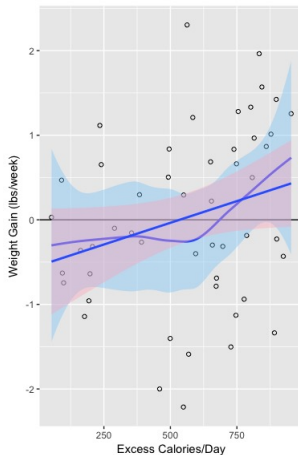


Truth=No Association Spline



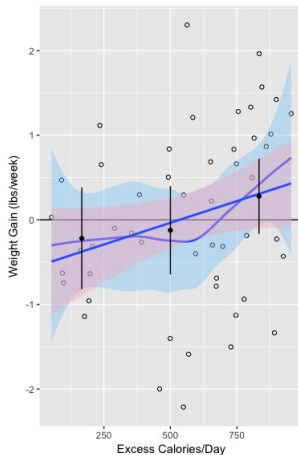
Interpretation: no association

Truth=No Association Linear Model



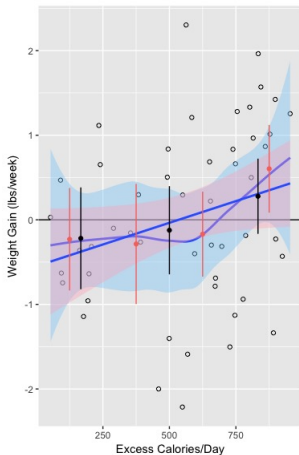
Interpretation: no association

Truth=No Association Tertiles



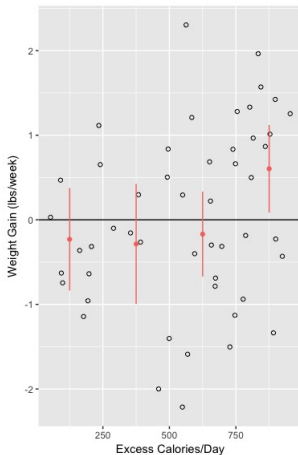
Interpretation: no association

Truth=No Association Quartiles



- ▶ $p = 0.08$ from overall F-test on quartiles
- ▶ q4 vs. q1 $p = 0.04$
- ▶ What about model uncertainty?
- ▶ Interpretation: no association!

What's the Truth?



Generally do not learn “truth” from a single study, but showing this figure and not acknowledging the number of (null) exposure parameterizations that preceded this one is like...

Perfect Match?

© 2013 Ted Goff



**“You can’t keep adjusting the data
to prove that you would be the best
Valentine’s date for Scarlett Johansson.”**

Limiting Risk of Being Misled by Type I Errors

Many strategies can be employed to limit risk, including

- ▶ Acknowledgement of all analyses conducted (would prevent mistake in this case)
- ▶ Reliance on overall tests rather than individual contrasts alone (attempting to limit multiple comparisons); also works here
- ▶ Comparison with results from similar studies; also works here

Disclosure: it took me a while (simulation hacking!) to generate this example; most datasets gave $p > 0.05$ in all models; most type I errors with $p < 0.05$ less believable (e.g., only 2nd quartile “significant”); “lucky data” + multiple testing here

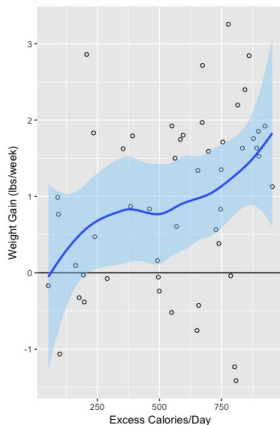


Example: True Association

Now we simulate data with an association between caloric intake and weight gain (no type I errors).

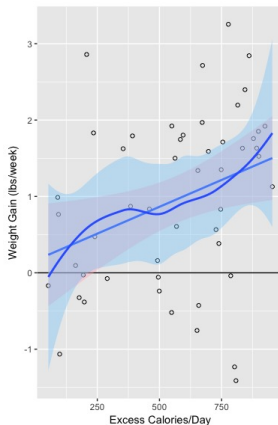
- ▶ Caloric intake: $n=50$ from $U(0, 1000)$
- ▶ Weight gain: generated from normal distribution with $SD=1$ and mean equal to the excess caloric intake divided by 500
- ▶ Truth is linear regression model with $\sigma^2 = 1$, $\beta_0 = 0$, and $\beta_1 = 0.002$: 500 extra calories/week translates into an extra pound on average

True Association: Spline



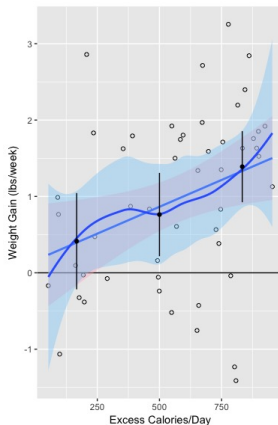
Interpretation: excess calories are associated with weight gain

True Association: Linear Model



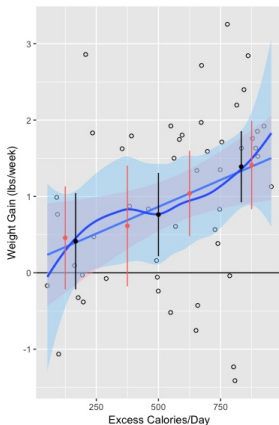
Interpretation: excess calories are associated with weight gain

True Association: Tertiles



Interpretation: excess calories are associated with weight gain

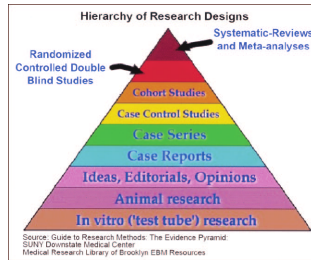
True Association: Quartiles



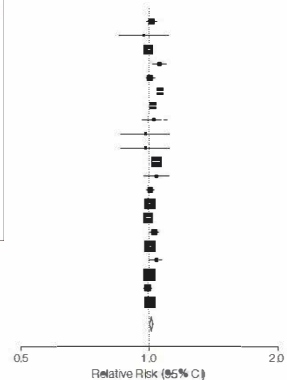
- ▶ Quartiles $p = 0.14$ in overall F test
- ▶ 4q vs. 1q $p = 0.04$
- ▶ Overall analyses are largely consistent, with significance of overall test in 3/4 models
- ▶ The quartile model should not reassure me about eating half of Rashid's croissant!

What is the Truth: “True Null” or “True Association”?

- ▶ Time well tell!
- ▶ Meta-analysis allows comparison of results over time, across populations, in different models ... all to inform the likelihood an estimated association reflects the truth. In this sense our lack of “true” replicates is a strength.



NO₂ and MI, Mustafic et al., 2012 *JAMA*



Sensitivity Analysis to Evaluate Robustness

Sensitivity analyses may include

- ▶ Use of different exposure models (e.g., EPA downscaler versus Brauer et al. data fusion approach)
- ▶ Multi- (or single-) pollutant analysis
- ▶ Stratification/restriction of cohort to focus on susceptible groups
- ▶ Application of causal inference methods (but note assumptions involved here as well so important to vary them across reasonable scenarios)
- ▶ **Caution:** BMA/ensemble methods great for prediction ignoring causal attribution but problematic for inference on correlated variables (Ghosh & Ghattas, 2015); consider probability of including groups of correlated variables

Other Means of Reassurance about Multiple Testing

Study registries publicly declare analysis intentions in advance; can reassure those with concerns regarding multiple testing.

ClinicalTrials.gov allows registration of observational studies (15-20% of total, including Zhang et al. HEI study on diesel exhaust and asthma), with declaration of primary outcomes and exposures, though details of analysis plans are often scant.

Generally quite detailed analysis plans, and planned sensitivity analyses, are presented at the time of application for funding (certainly for HEI funding) and could be published more broadly to enhance transparency (some obtainable via FOIA).

Easy to report planned analyses as well as unplanned sensitivity analyses.

What to Archive for Methods Reproducibility

Important to archive data and all code (not just from final model on processed data) so entire analysis process can be reproduced (budget for this in advance!). Some information published in journal; some on website or repository like GitHub.

- ▶ Data analysis plan as crafted in advance
- ▶ Rationale for modeling and other decisions
- ▶ Data cleaning and any exclusions
- ▶ How exposures are estimated
 - ▶ Pollution data sources used
 - ▶ Exposure models used
 - ▶ Methods used to handle spatial misalignment/linkage of exposure data to participant over space and time
 - ▶ Handling of missing data
 - ▶ All modeling steps with clearly documented (clean!) code
 - ▶ Final data used for analysis (+ raw data/processing steps)

Interpretation

- ▶ Be mindful of robustness to model assumptions evaluated through sensitivity analysis
- ▶ Honestly acknowledge challenges to validity and potential sources of uncertainty that may not be addressed in modeling
- ▶ Interpret results in the context of the broader literature subject to rigorous peer review, with thoughtful discussion of how current work adds to our knowledge, raises new questions, and provides directions for future work