Web-based enrollment and other types of selection: consequences for generalizability

Niels Keiding

Section of Biostatistics
University of Copenhagen

nike@sund.ku.dk

12th IBS-Italian Region conference

Napoli 10 July 2019

Main references

Keiding, N. & Louis, T.A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys (with discussion). *J.Roy.Statist.Soc. A* **179,** 319-376.

Keiding, N. & Louis, T. A. (2018). Web-based enrollment and other types of selection in surveys and studies: consequences for generalizability. *Annu. Rev. Stat. Appl.* **5**, 25-47**.**

Prologue:

Simpson (1951): Conditional or marginal effect measures

E.H. Simpson (1951). The interpretation of interaction in contingency tables.
J. Roy. Statist. Soc. B **13**, 238-241.

M.A. Hernán, D. Clayton, N. Keiding (2011). The Simpson's paradox unraveled.
Int. J. Epid. **40**, 780-785.

E.H. Simpson (1922-2019) is also famous for his work at Bletchley Park (Turing!) during World War 2 cracking the Germans' codes. See

E.H. Simpson (2010). Bayes at Bletchley Park. *Significance* **7,** 76-80.

Simpson (1951): Conditional or marginal effect measures

	B=1	B=0
A=1	20	20
A=0	6	6

OR = 1

C = 1

	B=1	B=0
A=1	5	8
A=0	3	4

OR = 5/6

C = 0

	B=1	B=0
A=1	15	12
A=0	3	2

OR = 5/6

Simpson: baby playing cards

	B=1	B=0
A=1	20	20
A=0	6	6

OR = 1

C = 1

	B=1	B=0
A=1	5	8
A=0	3	4

OR = 5/6

C = 0

	B=1	B=0
A=1	15	12
A=0	3	2

OR = 5/6

A = 0 court cards (B, D, K) B = 0 red (heart, diamond)
A = 1 not court cards (A, 2, 3,…, 10) B = 1 black (spade, club)

C = 1 card dirty because baby played with it
C = 0 card clean
Is colour independent of court status?

Yes, marginally OR = 1
No, conditionally on dirtiness OR = 5/6 for C = 1 and for C = 0.
Relevant effect measure: Marginal

Simpson: medical treatment

	B=1	B=0
A=1	20	20
A=0	6	6

OR = 1

C = 1

	B=1	B=0
A=1	5	8
A=0	3	4

OR = 5/6

C = 0

	B=1	B=0
A=1	15	12
A=0	3	2

OR = 5/6

A = 0 not treated
A = 1 treated

B = 0 not dead
B = 1 dead

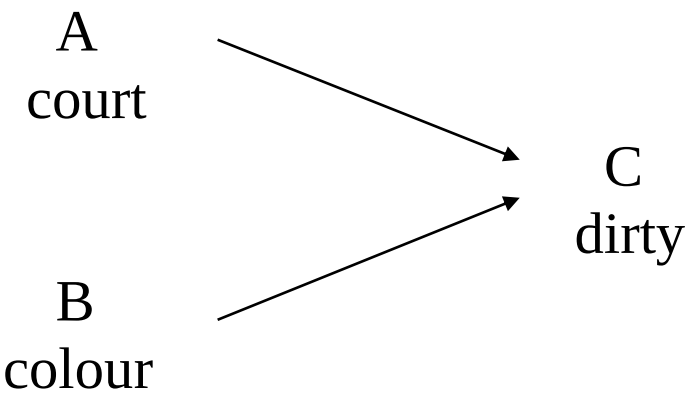
C = 1 male
C = 0 female

Does treatment affect death?

Marginally:No (OR = 1) Males:Yes (OR of dying = 5/6) Females: Yes (OR of dying = 5/6)

Relevant effect measure: Conditional on sex
(females die more, females are treated more).

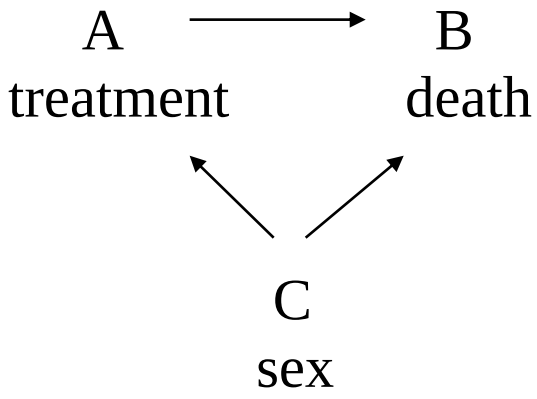
Baby playing cards



Even if A and B are independent, they become artificially associated by conditioning on the collider C

Collider bias

Medical treatment



Treatment as well as death depends on sex which is a possible confounder that should be controlled for.

Conditional and marginal effect measures

N. Keiding, D. Clayton (2014). Standardization and control for confounding in observational studies: a historical perspective. *Statist. Sci.* **29,** 529-558.

***Marginal**:* apply age specific rates to a *target* age structure and compare the predicted *marginal* summaries in this target population

- corresponds to handling confounders by making sure their distribution is the same in study and control population

Direct standardization, randomized trials

***Conditional**:* compare covariate-specific rates

- corresponds to handling confounders by stratification or restriction

Indirect standardization, regression analysis

Background

The internet is an attractive resource for enrolling and following *volunteer participants* in observational epidemiological studies. Should we be concerned about this deviation from classical ambitions of drawing *representative samples*?

Epidemiologists discuss this assuming that *representative sampling = simple random sampling* and generally downplay the role of sampling in favour of careful *confounder control*. However, they maintain a keen interest in the possibility of *selection bias* in the composition of the study group.

A central issue is whether *conditional effects in the study group may be transported to desired target populations.*

Prevalence studies vs. analytic epidemiology

Prevalence studies concern the distribution in a population of people with a particular disease (e.g. asthma) or health behaviour (e.g. smoking) and perhaps variation of the prevalence across subgroups (age, sex, occupation, calendar time).

Nobody questions the necessity of obtaining *representative* information here (often from surveys). Surveys may be based on *stratified random sampling*, and then *reweighting* may be used to estimate the marginal distribution in the population.

Analytic epidemiology is about relating the occurrence of an outcome (often: disease incidence) to an exposure. Such studies are done on study groups that are sometimes well-defined samples of specified populations. There is a lively debate on the role of representativity in analytic epidemiology. Important questions are:

- Does the study group have to be representative of some well-defined population?
- Do we need to worry about the composition of the *(target)* population for which we want to use the results?

Do such topics belong to basic epidemiological-biostatistical methodology?

Mortality for participants vs. non-participants: Direct validation from population-level databases

Andersen, L.B., Vestbo, J., Juel, K., Bjerg A.M., Keiding, N., Jensen, G., Hein, H.O. & Sørensen, T.I.A. (1998). A comparison of mortality rates in three prospective studies from Copenhagen with mortality rates in the central part of the city, and the entire country. *Eur.J.Epid.* **14**, 579-585.

Andersen et al. (1998) compared mortality of participants in 3 cohorts recruited in the Copenhagen area to the general mortality in that area since

> there is a risk of bias if other causes for the disease under study or confounders not taken into account in the analysis are differently distributed among the participating subjects and in the population that is target for generalization . Many factors associated with disease and death differ between participants and non-participants either because they are implicit in the selection criteria or because of the self-selection.

The analysis showed survivor selection in all cohorts (recruited participants being healthier at baseline than non-recruited individuals), which persisted beyond ten years of observation for most combinations of age and sex.

Observational studies:

Historical example

D.D. Baird, A.J. Wilcox (1985). Cigarette smoking associated with delayed conception. Preliminary report. JAMA **253,** 2979-2983.

*Pregnant women…were informed of the study in presentations at early pregnancy classes, through posters in the offices of obstetricians, or by obstetrics clinic nurses. Women were encouraged to volunteer for a 15-minute telephone interview if they had stopped birth control in order to get pregnant and had taken no more than two years to conceive. Of 762 volunteers****….****35 were not married throughout the noncontracepting time to pregnancy… leaving 678 women for analysis.*

After adjusting for potential confounding variables by Cox…, fertility of smokers was estimated to be 72% of the fertility of nonsmokers. Heavy smokers experienced lower fertility than did light smokers. Fertility was not affected by the husband's smoking.

Historical example, cont.

Careful Comment (nowadays called Discussion):

…asked to volunteer only if they had planned pregnancies, and volunteers were generally affluent and educated. These characteristics of the study design and study population raise questions about the generalizability of the findings.

Of primary concern is any source of bias that might result in finding an association in our study population even if no true association exists in the general population. ……. .....the exclusion of unplanned pregnancies. If smokers use less effective birth control or use birth control less carefully than nonsmokers, they would have more accidental pregnancies…. (which) naturally tend to occur among the most fertile women, which selectively removes them from the pool of women who go on to have planned pregnancies. Thus, by selecting only those who planned their pregnancies, we would have selected the less fertile women. If this occurred more often with smokers than with nonsmokers, we would overestimate the smoking-associated reduction in fertility.

This issue was handled through what we now call a sensitivity analysis.

Historical example, sensitivity analysis

…by developing a hypothetical population in which smokers and non-smokers had similar fertility but differed in their use of birth control. We assumed that 30% of pregnancies were accidental (a recent study found 27% of pregnancies attributed to careless use of birth control or birth control failure) and that smokers were 1.5 times more likely to have accidental pregnancies than non-smokers (smokers in our study were about 1.5 times more likely than non-smokers to use birth control sporadically in the initial months of their times to pregnancy). With these assumptions, smokers with planned pregnancies in the hypothetical study population showed a conception rate of 0.91 relative to non-smokers. This is a much smaller effect than we observed in our data, suggesting that the association between smoking and fertility is not attributable to this bias.

Lessons learned from the historical example

Two possibilities for selection bias were mentioned:

1. *volunteers were generally affluent and educated.*

Standard problem with surveys. Sometimes reweighting is possible, cf. later.

2. *by selecting only those who planned their pregnancies, we would have selected the less fertile women*.

A specific problem in this context, requires subject matter insight for detection as well as for handling (statistical brilliance not enough – but as we shall see, sometimes quite important)

Selection bias

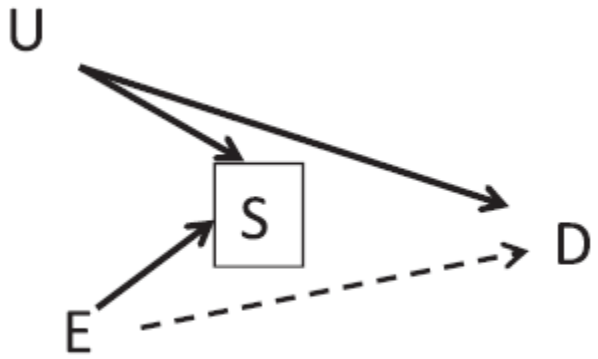
'Selection biases are distortions that result from procedures used to select subjects and from factors that influence study participation. The common element of such biases is that the relation between exposure and disease is different for those who participate and for all those who should have been theoretically eligible for study, including those who do not participate. Because estimates of effect are conditioned on participation, the associations observed in a study represent a mix of forces that determine participation and forces that determine disease occurrence.'

Rothman, Greenland & Lash (2008). Modern Epidemiology, 3rd Edition, p. 134.

The current development in causal inference is picking up these issues with the basic reference on selection bias being

M.A. Hernán, S. Hernández-Diaz, J.M. Robins (2004). A structural approach to selection bias. *Epidemiology* **15,** 615-625.

Collider bias



If selection S depends on an unobserved confounder U as well as exposure E, S will be a *collider* which generates *an artificial connection* between exposure E and outcome D in the conditional analysis given S.

Time to pregnancy (TTP)

The time from a couple decides they want to become pregnant (“initiation”) until they succeed. This is regarded as one of the most precise indicators of biological fecundity.

Difficult to design:

Prospective: hard to recruit couples at initiation, hard to identify study base, analysis standard

Retrospective (e.g. at maternity clinic): easier to recruit, result conditional on success, harder to interpret

Current duration: recruit couples currently trying, analyse backward recurrence times, not yet widely used

So why not try recruiting via the Web?

SnartGravid - SnartForældre

Initiated in Denmark in 2007 by researchers from Boston University (K. Rothman, L. Wise et al.) and Aarhus University (H.T. Sørensen, E. M. Mikkelsen et al.). From 2011 both parents included ('SnartForældre').

Volunteer couples recruited via on-line advertisements (non-commercial health sites, social networks), press releases, blogs, posters, word-of mouth. Recruitment shortly after initiation, followed until pregnancy *or* giving up trying *or* 12 menstrual cycles after initiation. No attempt at representativity of the volunteers. Follow-up via web.

By June 1, 2014, more than 8,500 couples recruited. Fine follow-up (more than 80% of the cohort still included after 1 year).

American companion study: Boston University Pregnancy Study Online (PRESTO)
cf. Wise et al. (2015), *Paed.Perinat.Epid.* **29,** 360-371.

SnartGravid: selected results

Two intro-papers (Mikkelsen et al. IJE 2009; Huybrechts et al., Eur J Epid 2010)

Results on the association of *Exposure -> TTP*, with *Exposure*:

Body size (Wise et al., Hum.Repr. 2010)

Menstrual Characteristics (Wise et al., AJE 2011)

Caffeinated drinks, soda (Hatch et al. Epidemiology 2012)

Physical activity (Wise et al., Fertil.Steril. 2012)

Volitional factors and age (Rothman et al., Fertil.Steril. 2013)

Oral contraceptives (Mikkelsen et al., Hum.Repr. 2013)

Weight at birth (Wildenschild et al., PLOS ONE 2014)

Active and passive smoking (Radin et al., Fertil.Steril. 2014)

Woman's own gestational age (Wildenschild et al., Hum.Repr. 2015)

Folic acid supplementation (Cueto et al., Eur.J.Clin.Nutr. 2016)

as well as other outcomes (spontaneous abortion, adverse birth outcomes, birth weight)

Reflections on representativity (Hatch et al. Epidemiology 2016)

SnartGravid: (initial) attitude to self-selection via the internet

Huybrechts KF, Mikkelsen EM, Christensen T, Riis AH, Hatch EE, Wise LA, Sørensen HT, Rothman KJ (2010). A successful implementation of e-epidemiology: the Danish pregnancy planning study 'Snart-gravid'. *Eur J Epidemiol* **25**, 297–304.

"The primary concern should therefore be to select study groups for homogeneity with respect to important confounders, for highly cooperative behavior, and for availability of accurate information, rather than attempt to be representative of a natural population.

Scientific generalization of valid estimates of effect (i.e., external validity) does not require representativeness of the study population in a survey-sampling sense either. Despite differences between volunteers and non-participants, volunteer cohorts are often as satisfactory for scientific generalization as demographically representative cohorts, because of the nature of the questions that epidemiologists study. The relevant issue is whether the factors that distinguish studied groups from other groups somehow modify the effect in question."

The nature of the questions that epidemiologists study

In science the generalization from the actual study experience is not made to a population of which the study experience is a sample in a technical sense of probability sampling…In science the generalization is from the actual study experience to the abstract, with no referent in place or time

O. S. Miettinen (1985). *Theoretical Epidemiology.* Wiley.

paraphrased by

K.J. Rothman (1986). *Modern Epidemiology.* Little, Brown

K.J. Rothman & S. Greenland (1998). *Modern Epidemiology,* Second Edition. Lippincott Williams and Wilkins

K.J. Rothman, S. Greenland & T.L. Lash (2008). *Modern Epidemiology,* Third Edition. Wolters Kluwer.

K.J. Rothman et al. (2013). Why representativeness should be avoided. *Int. J. Epid.* **42,** 1012-1014.

Smoking and Health

Miettinen’s standard example in the happy days at Harvard in the 1970s was the pathbreaking study by Doll and Hill of male British doctors showing that smoking is associated with lung cancer incidence. This study group was not representative.

The example is still often quoted by Miettinen’s former students.

SnartGravid and generalization, concretely:

Wise et al., Hum.Repr. 2010 (Body fat)

The proportion of couples in the Snart-Gravid study that conceived after 1 year was somewhat lower than that found in other prospective studies (…), and those interested in our study may have had lower fertility on average than the general population. (…)

Careful and credible comment on possible motivation for participation, generating participation bias.

… a non-negligible proportion of pregnancies may have been unplanned. If pregnancy intention was related both to the exposures studied here and to fertility, our results may not be generalizable to women with unplanned pregnancies.

Well-known problem that prospective TTP studies cannot catch accidental pregnancies – and remember Baird & Wilcox (1985)

Wise et al., AJE 2011 (Menstrual characteristics)

Finally, although this study enrolled a self-selected sample of pregnancy planners recruited via the Internet, there is little reason to believe that such women would differ from the general population of women at risk of pregnancy in ways that would lead to biased effect estimates.

Why not? We just heard an example of the contrary.

Snart-Gravid and representativity (after reflection)

Hatch, E.E., Hahn, K.A., Wise, L.A., Mikkelsen, E.M., Kumar, R., Fox, M.P., Brooks, D.R., Riis, A.H., Sorensen, H.T. & Rothman, K.R. (2016). Evaluation of selection bias in an internet-based study of pregnancy planners. *Epidemiology* **27**, 98-104.

Studied relations between routinely recorded variables in the Danish Medical Birth Registry (*exposures* such as age at delivery, smoking during pregnancy, parity at entry, maternal BMI, *outcomes* such as birth weight, pre-eclampsia, method of delivery). Compared these relations between the SnartGravid participants and the full Registry for the relevant years and found good agreement.

Problem: The main outcome in SnartGravid is TTP which is not registered in the Birth Registry. So Hatch et al. do not address the possible self-selection bias issue regarding TTP directly, but rather study the *representativity* of the SnartGravid sample for some other relations, hoping that this *by analogy* will cover the self-selection issue for TTP.

Hatch et al. concluded:

Without effect-measure modification by a factor, weighted selection by that factor will not influence the effect estimate in a study. In the presence of effect-measure modification, overall results depend on the distribution of the effect modifier in the study population.

The study result may differ from the corresponding value in the source population if sampling in the study is not proportional across subgroups of the modifying variable, and summary results are not standardized to the overall population.

If information is available on the effect modifier, the study should report important effect-measure modification, rather than present summary estimates.

Interlude on methodology

Basic rationale for randomization and representative sampling

M. Elliott (2016). Discussion of Keiding and Louis, *J.Roy.Statist.Soc.A,* **179,** 357.

- Randomization negates the influence of *unobserved confounders*
- Representative sampling negates the influence of unobserved *effect modifiers*

Transportability

J. Pearl and E. Barenboim. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science* **29,** 579-595.

Science is about generalization, and generalization requires that conclusions from the laboratory be transported and applied elsewhere, in an environment that differs in many aspects from that of the laboratory.

…the fact that most studies are conducted with the intention of applying the results elsewhere means that we usually deem the target environment sufficiently similar to the study environment to justify the transport of experimental results or their ramifications.

Remarkably, the conditions that permit such transport have not received systematic formal treatment.

(Note the difference from Miettinen's *In science the generalization is from the actual study experience to the abstract, with no referent in place or time*)

Transportability, cont.

Given judgments of how target populations may differ from those under study, the paper offers a formal representational language for making these assessments precise and for deciding whether causal relations in the target population can be inferred from those obtained in an experimental study. When such inference is possible, the criteria provided by Theorems 2 and 3 yield transport formulae, namely, principled ways of calibrating the transported relations so as to properly account for differences in the populations.

Pearl and Bareinboim's development was formulated in terms of Pearl's graph-based approach to causal analysis and yielded graphical criteria for deciding transportability and estimating transported causal effects.

Transportability:
generalizing evidence from clinical trials

J. Pearl (2015). Generalizing experimental findings. *J. Causal Infer*. **3,** 259-266.

Pearl studied conditions for generalization of result (average causal effect) of a clinical trial from the population P where it was conducted to a different population P*.
Compared this to the essential self-selection problem: generalize average causal effect from self-selected (possibly biased) sample S to full population P.

Formalized the classical confounder control approach (standardization-stratification) in the *post-stratification formula* which requires *S-ignorability* (there is a stratification variable Z for which the potential outcome Y_x of X is conditionally independent of the variable S that defines the difference between P and P* (resp. the sampling of S within P). This formula is essentially inverse probability weighting.
Pointed out that this will not suffice in certain situations (in connection with conditioning on post-treatment variables), where another condition called *S-admissibility* might help.

Wirth & Tchetgen Tchetgen on external validity

Wirth KE, Tchetgen Tchetgen EJ (2014) Accounting for selection bias in association studies with complex survey data. *Epidemiology* **25**, 444–453.

"It has been argued that, despite the unequal selection induced by the design of complex surveys, analyses that treat the sampled data as the population of interest remain valid. Using a DAG framework, we show that this will depend on knowledge about the relationships among determinants of selection, exposure, and outcome. If the determinants of selection are associated with exposure and outcome, failure to account for the sampling design may result in biased effect estimates. This includes settings where determinants of selection are the exposure or outcome under study."

A clear statement on the importance of collider bias

Generalization from clinical trials:

Introductory example
with direct validation from population-level databases

Ewertz et al. (2008) Breast conserving treatment in Denmark, 1989–1998. A nationwide population-based study of the Danish Breast Cancer Co-operative Group.

Acta Oncologica, **47**, 682–690.

Are results from clinical trials on breast-conserving operations of breast cancer applicable to all Danish women?

The Danish Breast Cancer Cooperative Group (DBCG) coordinates since 1978 breast cancer therapy in Denmark, where almost all women are treated for free at the public hospitals. Many randomized clinical trials on adjuvant therapy have been conducted with sampling frame: in principle all Danish women, suitably stratified e.g. by age and/or menopausal status. From 1982 to 1989 a randomized trial regarded breast conserving surgery against total mastectomy. Conclusion: breast conserving therapy offered as option to suited patients across Denmark.

The population-based registry of DBCG allowed population-based follow-up 1989-98: women younger than 75 years, and operated on according to the recommendations, had survival, loco-regional recurrences, distant metastases and benefit from adjuvant radiotherapy closely matching the results from the clinical trial.

Generalization of results from randomized trials

Imai et al. (2008), *JRSS* A **171,** 481-502, Cole & Stuart (2010), *Am.J.Epid.* **172,** 107-115,

Hartman et al. (2015), *JRSS* A **178,** 757-778, Lesko et al. (2017) *Epidemiology* **28,**553–61.

Study Sample S within population P. An outcome Y depends on an intervention ('treatment') which can take one of two values: a or a^l .

Population Averaged Treatment Effect:

$$PATE = E[Y(a)] - E[Y(a^l)].$$

In the study treatment is randomized within S and response recorded.

The *Sample (or Study) Averaged Treatment Effect SATE* is the mean difference in potential outcome between treatment and control. Randomization makes *SATE* directly estimable from the study.

The generalizability issue:

is *SATE* representative of anything beyond S?

1. Does *SATE* estimate *PATE* in the population P within which S was defined?
2. Can *SATE* be transported to a different population P' *(transportability)*?

Basic difficulty: Often S is not representative of P. But sometimes we may assume that effects are constant within strata, or that S is derived from a stratified random sample, and then for covariates *W* we may write

$$PATE = E\{E[Y(a)|W]\} - E\{E[Y(a^l)|W]\}$$

$$= E\{E[Y(a) - Y(a^l)|W]\},$$

recovering *PATE* as a weighted average of strata-specific effects.

Hernan & Robins (2006), *J.Epid.Comm.Health* **60,** 578-586 formalized the requirements for using standardization in terms of *exchangeability*.

PATT: Population averaged treated effect on the treated

A different conclusion from the non-representativity of study samples is to focus attention on the typical population of patients who are in practice being treated (Imai et al., 2008, Hartman et al. 2015). The latter authors also work with *PATC:* Population averaged treated effect on the controls.

Reweighting results from clinical trial to fit target population: influential example

S.A. Cole & E.A. Stuart (2010). Generalizing evidence from randomized clinical trials to target populations. *Amer.J.Epid.* **172,** 107-115.

In 1996-97 the ACTG 320 study was performed in USA testing a new highly active antiviral therapy against AIDS with conventional therapy as control. Patients were recruited from 40 clinical trial units in USA and Puerto Rico (577 in treatment group, 579 in control group). Intention to treat (ITT) estimate of treatment effect was 0.51 with 95% conf.iv. (0.33, 0.77).

BUT there was clear *treatment effect heterogeneity* (stronger effect for older, for males, for blacks).

It is desired to generalize the result from the trial to what it would mean for the estimated 54,220 HIV-infected people in USA in 2006. Treatment effect after reweighting

by *age*: 0.68 (0.39,1.17), by *race:* 0.46 (0.29,0.72) by *age, sex, race:* 0.57 (0.33, 1.00).

Cole & Stuart assumed
that conditional effects are directly valid in the target population and then the task reduces to versions of direct standardization.

Who wants such an average across a heterogeneous population?

In cases of clear treatment effect heterogeneity it is usually better to report the stratified results.

Star example: different effects for men and women cf. the 'feminist complaint':

'Clinical trials are often conducted only on men and the results generalized to women without direct evidence that this is justified'.

Example: Suicidality and RCTs of antidepressant drugs

Hammad et al. (2006). *Arch.Gen.Psychiatry.* **63,** 332-339; Greenhouse et al. (2008). *Statist. Med.* **27,** 1801-1813
Weisberg et al. (2009;2010). *Clin.Trials* **6,** 109-118; **7,** 118-119; Weisberg (2010). *Bias and Causation.* Wiley

Meta-analysis of 24 randomized placebo-controlled trials of antidepressant drugs among adolescents led FDA to issue a 'black box' warning that there might be an increased risk of suicides and suicidal behavior among pediatric patients taking these drugs.

Greenhouse et al. wrote in the abstract

'For the results of randomized controlled clinical trials (RCTs) and related meta-analyses to be useful in practice, they must be relevant to a definable group of patients in a particular clinical setting. To the extent this is so, we say that the trial is generalizable or externally valid.'

They compared results from trials to Youth Risk Behavior Survey: more whites in trials and suicidality rate in trials more than twice that in YRBS.

Weisberg et al. presented a model for potential outcomes (in the spirit of *principal stratification*) explaining the apparent excess suicidal rate among treated as a *selection effect* generated by excluding from the trials individuals with high suspected suicide risk under no treatment.

Generalization from clinical trials: Constructive approach

Ackerman, B., Schmid, I. Rudolph, K.E., Seamans, M.J., Susukida, R., Mojtabai, R., Stuart, E. A. (2019). Implementing statistical methods for generalizing randomized trial findings to a target population. *Addictive Behaviors* **94,** 124–132.

In trials related to substance use disorders (SUDs), especially, strict exclusion criteria make it challenging to obtain study samples that are fully "representative" of the populations that policymakers may wish to generalize their results to. In this paper, we provide an overview of post-trial statistical methods for assessing and improving upon the generalizability of a randomized trial to a well-defined target population. We then illustrate the different methods using a randomized trial related to methamphetamine dependence and a target population of substance abuse treatment seekers, and provide software to implement the methods in R using the "generalize" package. We discuss several practical considerations for researchers who wish to utilize these tools, such as the importance of acquiring population-level data to represent the target population of interest, and the challenges of data harmonization.

Conclusion

Generalization from observational as well as randomized studies may well benefit from careful statistical analysis coupled with detailed insight in the representativity of the study group. Perhaps the methodology for observational studies could take more inspiration from the developments for randomized studies.