Generalization from self-selected epidemiological studies

Niels Keiding

Section of Biostatistics
University of Copenhagen

nike@sund.ku.dk

Developing a Comprehensive, Integrated Framework for
Advanced Statistical Analyses of Observational Studies

Banff International Research Station (BIRS),
Banff, Canada, July 6, 2016

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.* This is sometimes taken as a dogma, sometimes as a working hypothesis open to empirical critique.

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?

Outline

Historical example on Time To Pregnancy (TTP) (1985)

Snart Gravid (2007-)

Miettinen's declaration (1985)

Why representativeness should be avoided (2013)

Validation from population-level data bases (Nordic countries)

Methodology?

Pizzi & Richiardi: Concrete studies of realistic effects

Pearl and Bareinboim: Transportability

Chatterjee et al.: Model calibration

Conclusion

Main reference: 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.

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.

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 so far

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)

SnartGravid: basic analytical strategy

Delayed entry (left truncation): many couples were recruited some menstrual cycles after start of attempt, only risk sets after recruitment included in analysis

(care was taken to only include recently started couples to avoid hazard ratio attenuation)

Right censoring: when couples were

lost to follow-up,

initiated fertility treatment,

gave up trying

SnartGravid: 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.

"Internet-based recruitment of volunteers has raised concerns among critics because the demographics (e.g., age, socio-economic status) of those with ready internet access differ from those without it. Furthermore, among those with internet access, those who choose to volunteer for studies may differ considerably in lifestyle and health from those who decline,"

"Volunteering to be studied via the Internet does not, however, introduce concerns about validity beyond those already present in other studies using volunteers. Differences between study participants and non-participants do not affect the validity of internal comparisons within a cohort study of volunteers, which is the main concern. Given internal validity, the only problems with studying Internet users would occur if the biologic relations that we are studying differed between Internet users and non-users, a possibility that seems unlikely.

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."

(Remember the last sentence, we shall return to that later)

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. ***(to follow)***

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.

Why representativeness should be avoided

Discussion in *Int. J. Epid.* **42** (2013) by

Rothman, Gallacher & Hatch; Elwood; Nøhr & Olsen; Richiardi, Pizzi & Pearce;

Ebrahim & Davey Smith (editors of IJE); and rebuttal by Rothman et al.

Main attitude: Miettinen's 1985 declaration.

'Representativeness' interpreted as simple random sampling which discussants generally considered unnecessary or even counterproductive.

General attitude: perform careful confounder control (which it is hoped does not depend on representativeness of sample) to justify conditional associations which are hoped to be more generalizable than marginal associations in existing populations.

This reliance on 'careful confounder control' places much (too much?) responsibility on the statistical analysis and leaves unmeasured confounders unattended.

Richiardi, Pizzi & Pearce

Specify three Criticisms against using non-representative populations in internet-based birth cohorts:

Criticism 1: Non-representative cohorts lack heterogeneity

Criticism 2: If the exposure of interest is associated with the probability of selection, the exposure-outcome associations estimated in a non-representative cohort may be biased

Criticism 3: If an intermediate variable in the causal pathway from the exposure to the outcome is associated with the selection, exposure-outcome associations estimated in a non-representative cohort may be biased

Ebrahim & Davey Smith (editors)

Clearly felt that the discussion had run out of hand for them, tried to calm down the strong, almost unanimous opinion that representativeness is usually unnecessary and may be counterproductive.

They

- Contradicted the Miettinen dictum: not all epidemiology is 'abstract science', also noting that epidemiology contains rather more complex confounding patterns than researchers accustomed to randomized trials can imagine
- Claimed that non-representative study groups may produce biased associations
- Termed Rothman's call for skillful confounder control in non-representative studies over-optimistic
- Voiced concern about epidemiology in the big data world

but concluded very cautiously

We feel that representativeness should neither be avoided nor uncritically universally adopted, but its value evaluated in each particular setting

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.

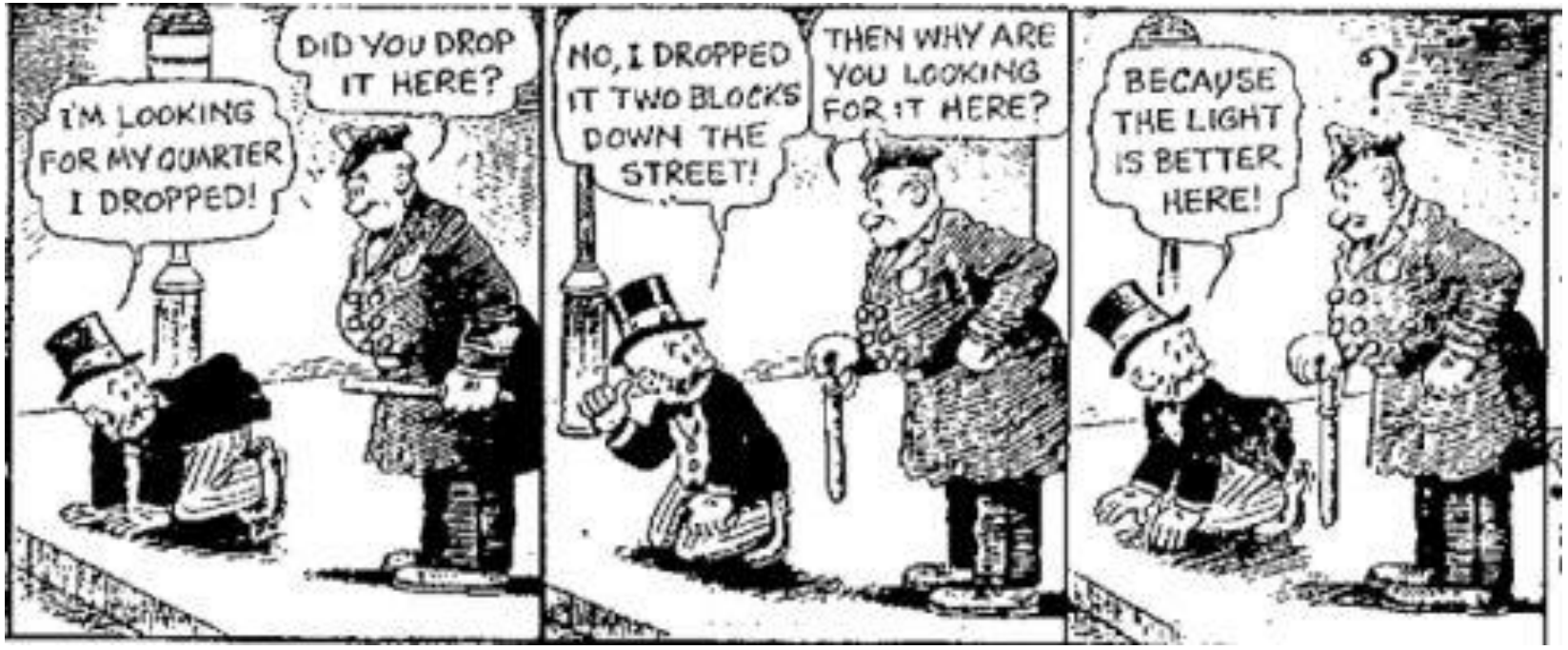
Snart-Gravid and representativity

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.

Studies 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). Compares these relations between the SnartGravid participants and the full Registry for the relevant years and finds 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.

Basic limitation in using population-level databases for validation: what if our target of interest (or essential exposure variable) is not registered there?



Validation from population-level databases: mortality in cohort

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.

Validation from population-level databases: are results from clinical trial on breast-conserving operations of breast cancer applicable to all Danish women?

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.

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.

Methodology??

For many years influential epidemiologists seem to have been discouraged by the Miettinen declaration and its fall-out; this attitude still very much alive.

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.

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*

Reweighting results from clinical trial to fit target population

An influential series of papers on generalization to target population started with

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).

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.

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

The NINFEA study

Pizzi, De Stavola, Pearce, Lazzarato, Ghiotti, Merletti, Richiardi (2012). Selection bias and patterns of confounding in cohort studies: the case of the NINFEA web-based birth cohort. *J. Epid. Comm. Health* 66, 976-981.

Of 36,092 pregnancies of Italian mothers in Torino in 2005-08, 1105 self-selected to participate in the birth cohort NINFEA. Can one generalize the information from this sample to the population?

There is computerized information available for all births in Torino in the Piedmont Birth Registry (PBR).

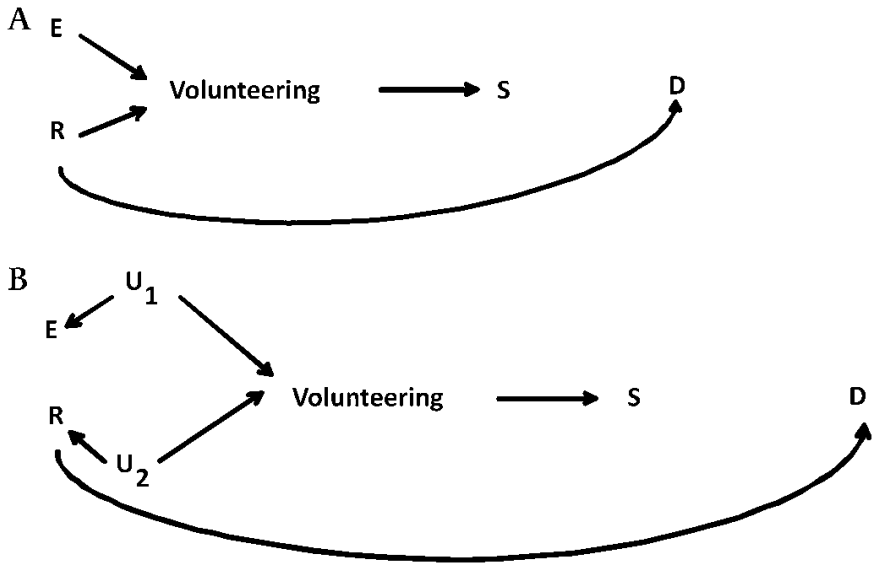
Two birth events (caesarean section; LBW: low birth weight for gestational age) were studied as end-points, with standard determinants (parity, maternal education, smoking during pregnancy, alcohol during pregnancy, infertility treatment, folic acid intake, maternal age, previous miscarriages, pregnancy weight gain) .

The web-based sample was not representative (more nullipari (79% vs. 56%), better education (34% vs. 18%), more folic acid intake (86% vs. 80%)).
Did that matter?

The NINFEA study:

example of stronger confounding in sample

Maternal education and folic acid intake are independent (OR=1.01) in the background population PBR, but associated (OR=1.44) in the sample NINFEA. Considering the effect of maternal education on outcomes this means that folic acid becomes a confounder in the sample, which it is not in the full population.



The NINFEA study:

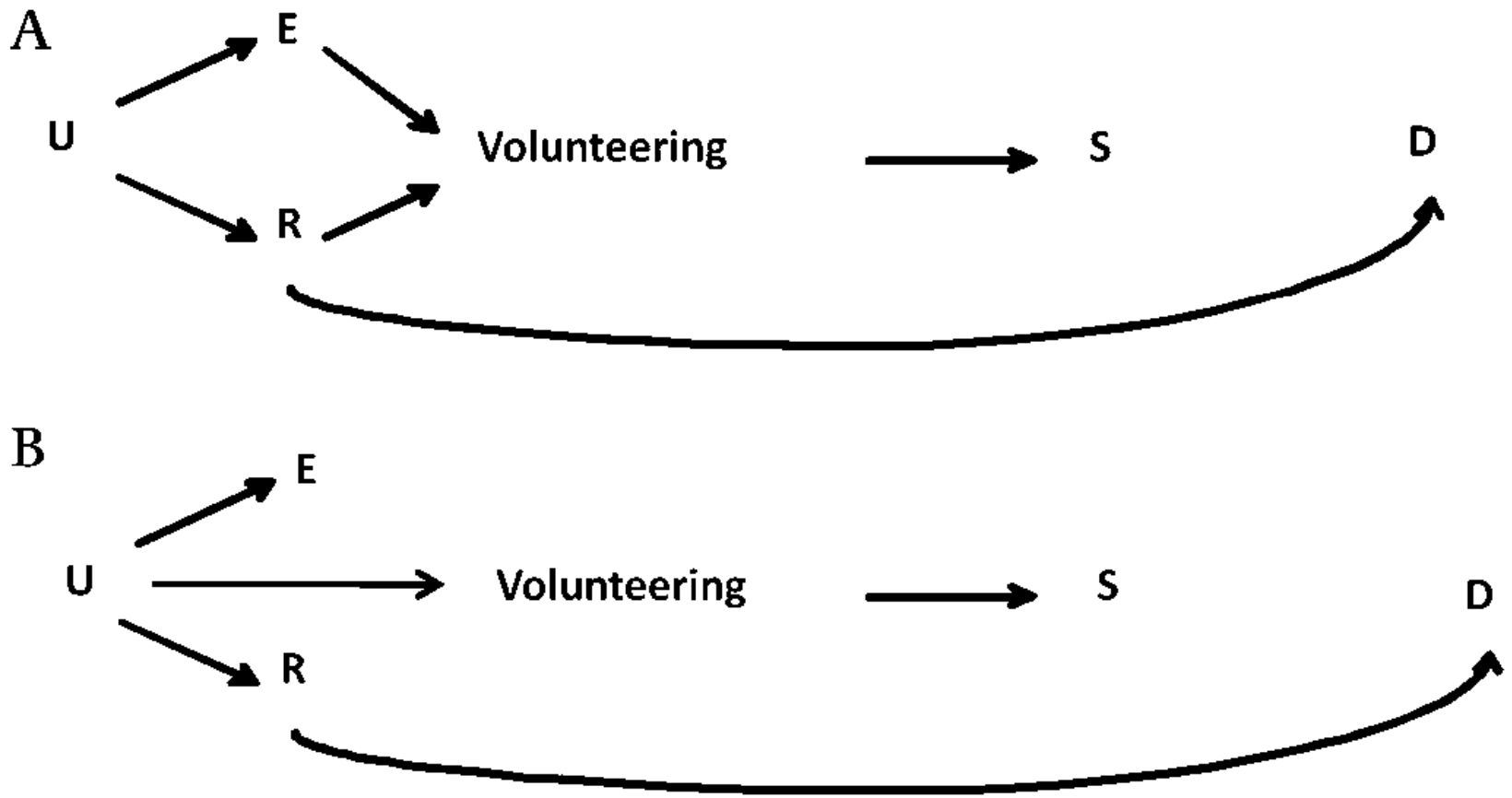
example of weaker confounding in sample

Parity is positively and maternal age is negatively associated with participation in NINFEA. They are positively associated in the population PBR (OR=2.45). In the sample NINFEA they are even more strongly associated (OR=3.17).

Maternal education and maternal age are both positively associated with being selected for NINFEA. They are highly associated in the population PBR (OR=2.09), but not much in NINFEA (OR=1.22).

Maternal age is a residual confounder *in the population PBR* for the association between maternal education and caesarean section (OR=1.14) when adjusted for all except maternal age, OR=1.07 when adjusted for all).

But *in the sample NINFEA* maternal age is no longer much of a (residual) confounder for the association between maternal education and caesarean section (OR=0.98 when adjusted for all except maternal age, OR=0.97 when adjusted for all).


A
E
U
Volunteering
S
D
R
B
E
U
Volunteering
S
D
R

Likely ranges of bias due to non-representativity of study group (simulation study)

Pizzi, De Stavola, Merletti , Belocco, dos Santos Silva, Pearce, Richiardi (2011). Sample selection and validity of exposure-disease association estimates in cohort studies. *J. Epid. Comm. Health* 66, 976-981.

Result: even with major effects of the selection probabilities on exposure and outcome the bias in the effect estimates is not large.

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.

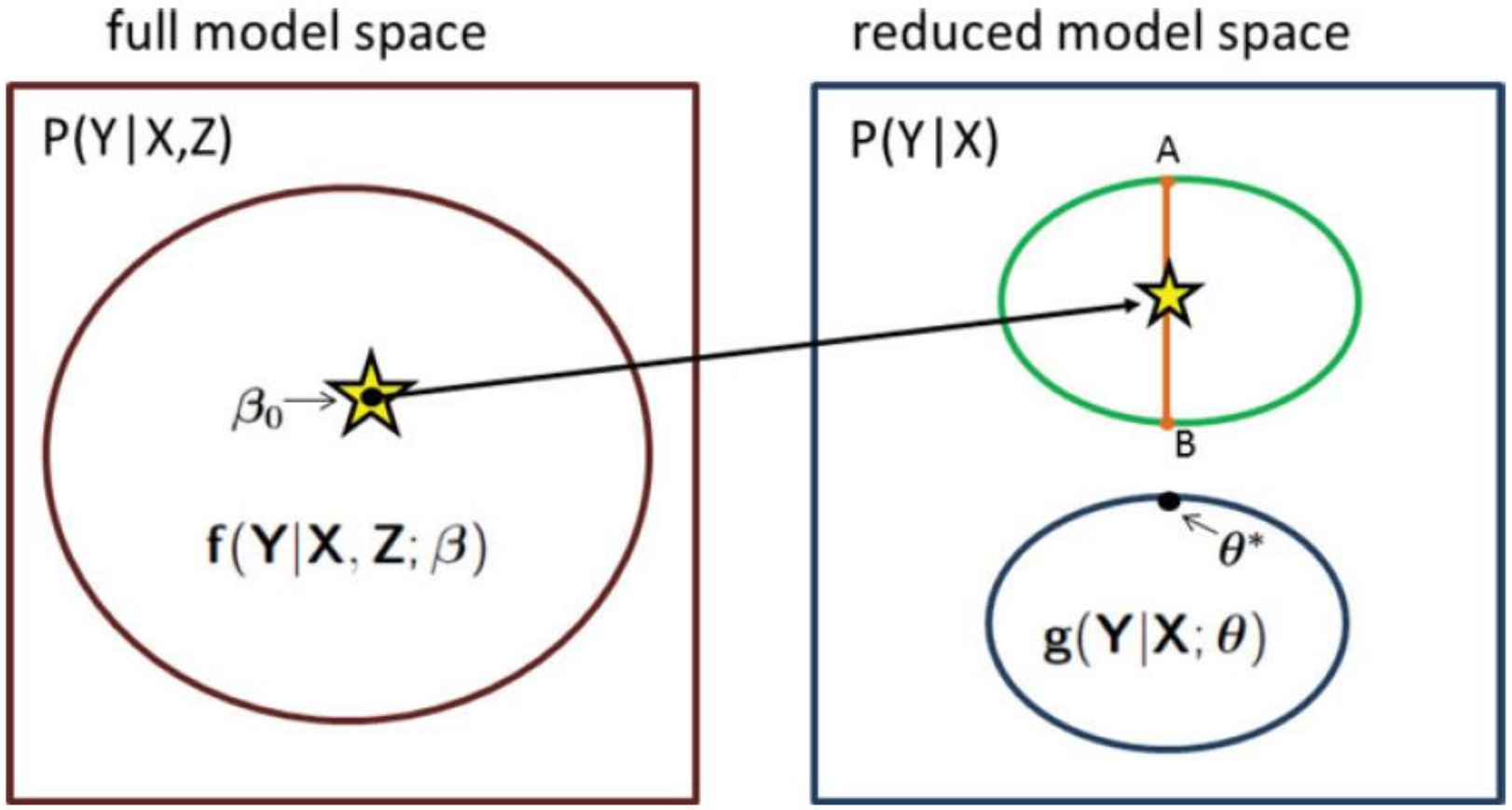
Model calibration using summary-level information from external big-data sources with discussion

Chatterjee, Chen, Maas, Carroll (2016) *JASA* **111**, 107-131.

- A model $g_\theta\left(y|x\right)$ has been derived from 'external' Big Data linking outcome Y to covariates X.
- A more detailed 'full' model $f_\beta\left(y|x,z\right)$ based on 'internal' data is built on standard data on Y and X containing also an additional covariate Z.

It is desired to *calibrate* the detailed regression model f_β using the Big Data model g_θ. (The possibility of misspecified g_θ is allowed along the lines of White (1982)).

Note: Here the framework is minimum mean-squared-error, where *accuracy* may be improved despite some moderate bias if *precision* is greatly improved.



To the surprise of several discussants the authors seem to claim that these tools allow generalization of the validity of the full model beyond the population from which it was derived.

The feminist complaint

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

I asked the authors whether their formulae solved this issue, but they did not respond, beyond a general statement in a different context:

If the external study represents a broader population of interest than the internal study, a reasonable goal for building predictive models could be to use the internal study to learn about parameters associated with the new ‘features’ and use the external parameters to improve generalizability of the models to the broader population for which prediction is desired.

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."

Conclusion

Epidemiological generalization is not an abstract issue, but a very concrete one requiring attention to the world around us.

Methodological developments have been delayed and work is needed at several levels.

To quote Huybrechts et al. (2010): The relevant issue is whether the factors that distinguish studied groups from other groups somehow modify the effect in question.

Age-specific frequencies of nulliparous women in Denmark 1980-2010

(Unpublished data acquired from the Fertility Database of Statistics Denmark, October 2014)

Age	1980	1985	1990	1995	2000	2005	2010
20	.88	.93	.94	.946	.952	.963	.964
25	.45	.56	.64	.68	.74	.77	.79
30	.18	.25	.29	.32	.35	.40	.42
35	.104	.129	.17	.18	.18	.19	.20
40	.092	.094	.113	.139	.144	.138	.138

The table shows a dramatic increase (particularly for ages 25 and 30) over both period and cohort of the frequencies of nulliparous women. The much more dominant presence of definitely low-fecund women among the 30-year old nullipari in the beginning than the end of the period would very likely induce an apparent increase in fecundity among nullipari in this age group over the years studied.

The data indicate that there are *strong secular trends in age at initiation of the pregnancy attempts* which are indeed very likely to generate bias from the resulting survivor selection.

Keiding, N. & Scheike, T.H. (2016). Fertility behavior and studies of fecundity trends. *Epidemiology* **27,** 459-461.