

Biostat 208 Session 7 Outline

- Predictor Selection
 - three inferential goals
 - DAGs
 - model selection algorithms

Predictor selection

- Suppose 50 or 100 predictors have been measured
- Which ones should be included in a model?
- Depends on “inferential goal” :
 1. predict future outcomes
 2. evaluate causal effect of a primary predictor
 3. identify important causal risk factors for an outcome

Goal 1: Prediction

- Includes diagnosis and prognostic risk stratification
- Prediction problems often involve a practical decision
- Causal relationships useful, but not the focus
- Only strong predictors useful: OR for binary predictor with sensitivity and specificity of 90% is *81*
- Model validation in terms of prediction error (PE)

Prediction vs. resubstitution error

- Prediction error: how well does model predict outcome for new observations not used in fitting model?
- Cf. R^2 : same data used to estimate parameters, evaluate fit
- Model selection using such *resubstitution* measures leads to over-fitting, optimistic estimates of clinical utility
 - hence insistence on external validation
- See Altman & Royston, What do we mean by validating a prognostic model? *Stat Med*, 2000;**19**:453-73

Prediction error measures by outcome type

- Continuous: predicted residual sum of squares (PRESS)
- Binary: Brier score; area under ROC curve (C-statistic) for discrimination; Hosmer-Lemeshow statistic for calibration
- Survival: C-index (analog of C-statistic); extensions of Hosmer-Lemeshow statistic

Prediction error measures by outcome type

- Diagnostics: sensitivity/TPF and specificity/FPF
- Risk stratification: PPV of high risk classification, NPV of low risk classification
- New measures:
 - Net Reclassification Index (NRI) (Pencina, *Stat Med*, 2008)
 - cost-benefit curves (Vickers, *Med Dec Making*, 2006)

Estimating prediction error

- Use different data to estimate parameters, evaluate PE
- Generalized cross-validation (GCV): learning set/test set (LS/TS), leave-one-out (jack-knife), 5 or 10-fold cross-validation
- Bootstrap to estimate optimism
- Resubstitution measures penalized for number of predictors: adjusted R^2 , AIC, BIC
- Easiest in Stata: LS/TS, adjusted R^2 , AIC, BIC
(`estat ic postestimation` command)

Bias-variance tradeoff

- Parameter estimates, predicted values:
 - biased when important predictors omitted from model
 - noisier when unimportant predictors are included
- Too many predictors → *over-fitting*
- Smaller models often better for prediction
- Modern methods aggressively sift models using GCV

Prediction error example

- Walter prognostic index for 1-year mortality in older adults after hospitalization
- Point system based on sex, number of dependent ADLs, CHF, cancer, creatinine > 3 , albumin < 3 (all easy to evaluate)
- Included predictors with $p < .05$ after adjustment
- Validated using follow-up data from a different hospital

Walter et al, *JAMA*, 2001;**285**;2987-94.

PREDICTS analysis plan

- Prospective cohort of post-MI patients with ICDs
- Only 15-20% of patients with ICDs ever need a shock
- Multiple potential predictors from baseline tests
- Develop model using aggressive search algorithms, GCV within learning set ($N \approx 800$)
- Validate PPV, NPV in test set ($N \approx 400$)

Shrinkage estimators

- Newer method to improve calibration
 - smaller coefficients as an alternative to smaller models
 - examples: ridge regression, penalized likelihood, LASSO
- Useful tutorial on prognostic modeling, including simple methods for shrinkage estimation:
 - Harrell FE *et al.*, *Stat Med*, 1996;15:361-87

Recommendations for Goal 1

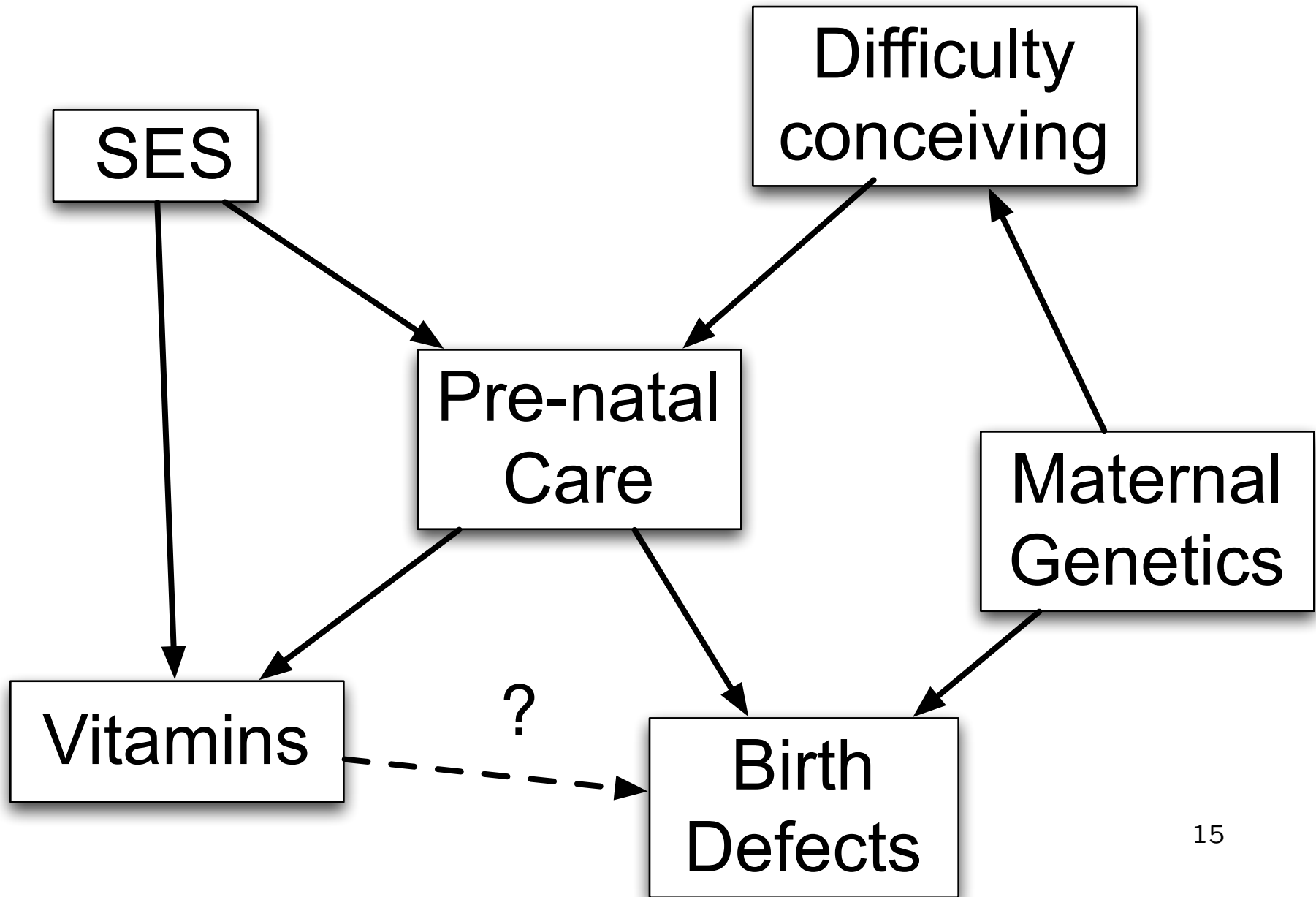
- Identify important potential predictors
- For clinical use, select easily available predictors
- Evaluate prediction error in a broad selection of models
- Select model minimizing GCV measure of PE, AIC, or BIC
- Validate model in external test set, or using bootstrap

Goal 2: assessing a predictor of primary interest

- Research question focuses on a single predictor
 - does HCV infection speed up progression of CKD?
- Ruling out confounding is key
- DAGs useful for determining what to adjust for, what to omit
- Example: maternal vitamin use and birth defects

Maternal vitamin use and birth defects: assumptions about causal pathways

- Prenatal care (PNC) increases vitamin use
- PNC prevents birth defects by other pathways
- Difficulty conceiving may lead mother to seek PNC
- Maternal genetics affect difficulty conceiving and birth defects
- SES affects both access to PNC and vitamin use

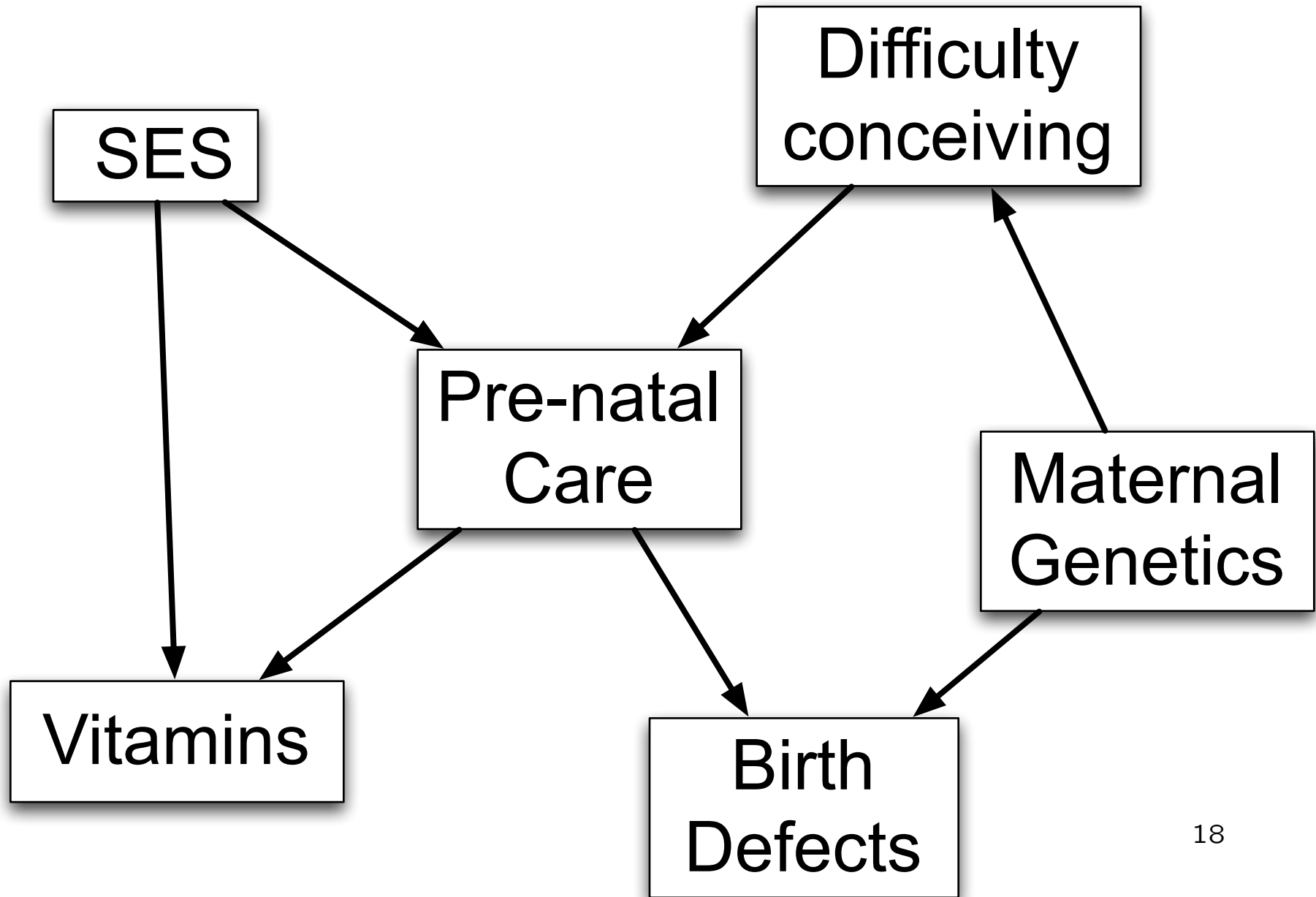


Additional assumptions in this DAG

- SES has no effect on difficulty conceiving
- SES affects birth defects only via PNC, vitamin use
- Difficulty conceiving affects vitamin use only through PNC
- No other common causes of vitamin use and birth defects
- No excluded causal connections

What do we need to adjust for?

- Step 1: check for mediators of vitamin use
- Step 2: delete all direct effects of vitamin use
- Step 3: check for common causes of vitamin use and birth defects

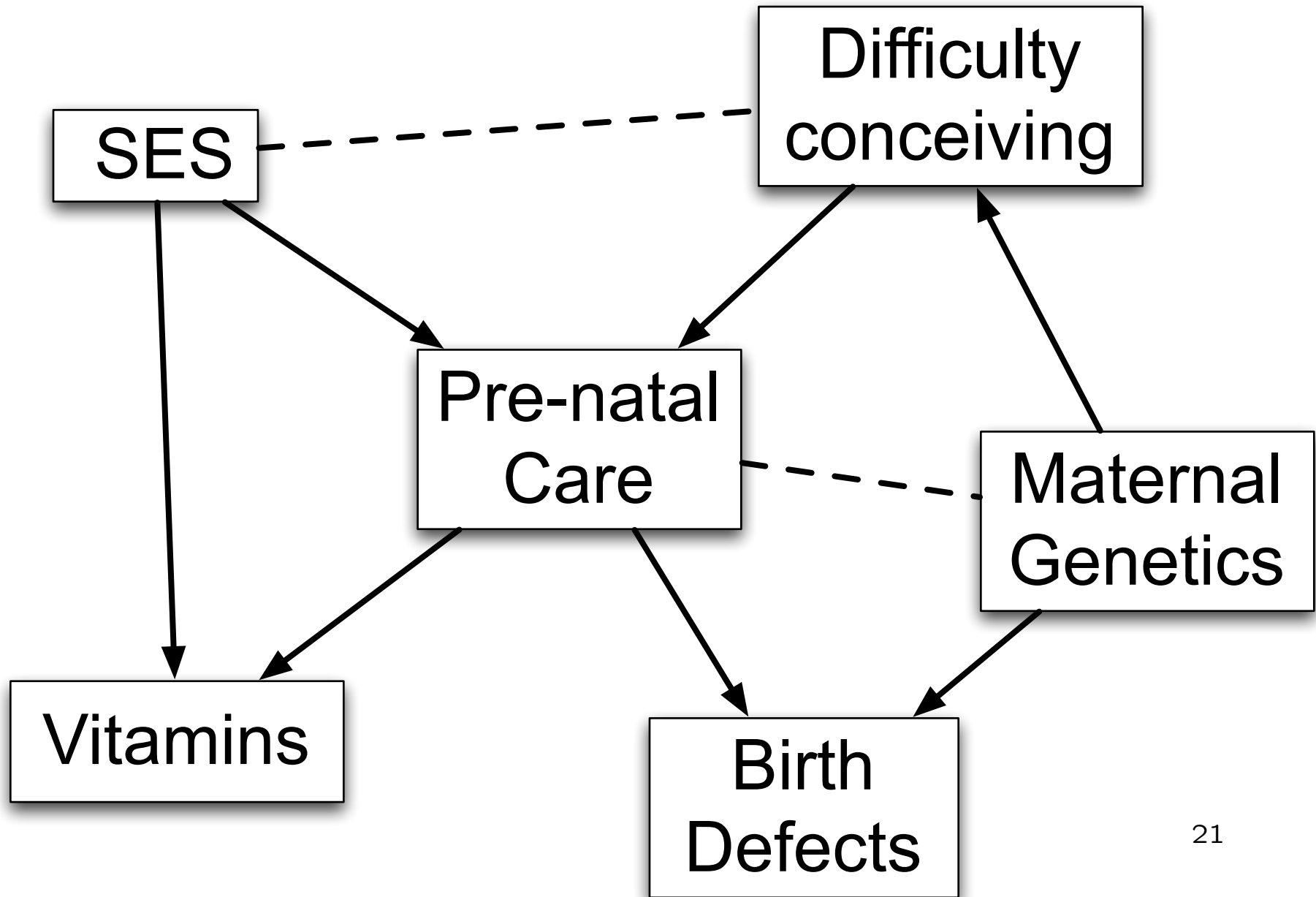


What do we need to adjust for?

- Four common causes of vitamin use and birth defects:
 - SES, PNC, difficulty conceiving, maternal genetics
- PNC blocks the pathways from SES, maternal genetics, and difficulty conceiving to birth defects
- Is it enough to adjust for PNC?

PNC is a confounder and a collider

- Step 4: connect any two causes sharing a common effect
 - i.e., direct ancestors of colliders
- Rationale: controlling for the collider induces an association

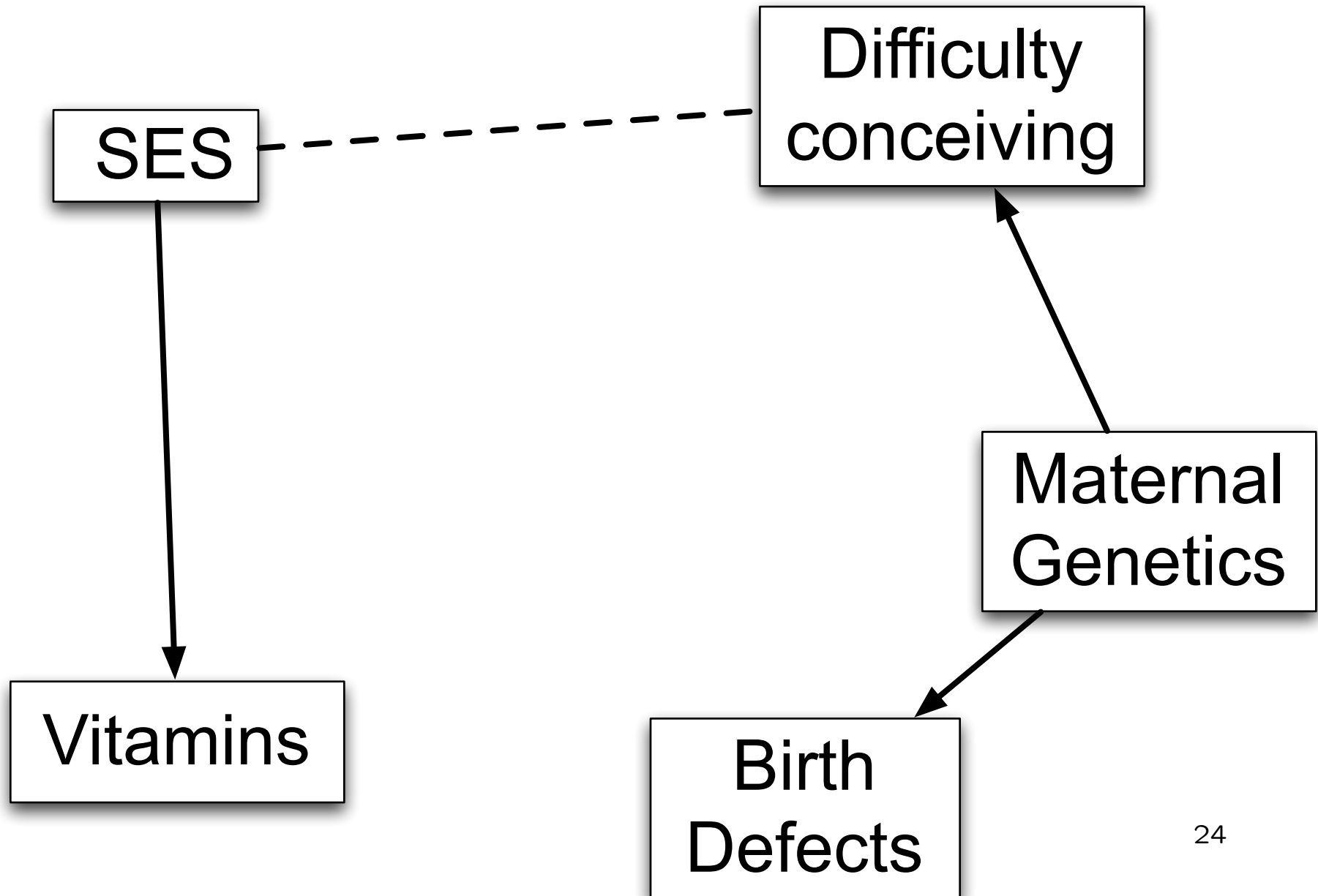


Controlling for a collider induces an association

- Controlling for PNC is tantamount to stratifying on it.
- *Within PNC user stratum:*
 - SES, difficulty conceiving competing explanations for PNC
 - higher SES women less likely to have had difficulty conceiving, and vice versa
- Controlling for PNC opens a backdoor path via induced association of SES and difficulty conceiving

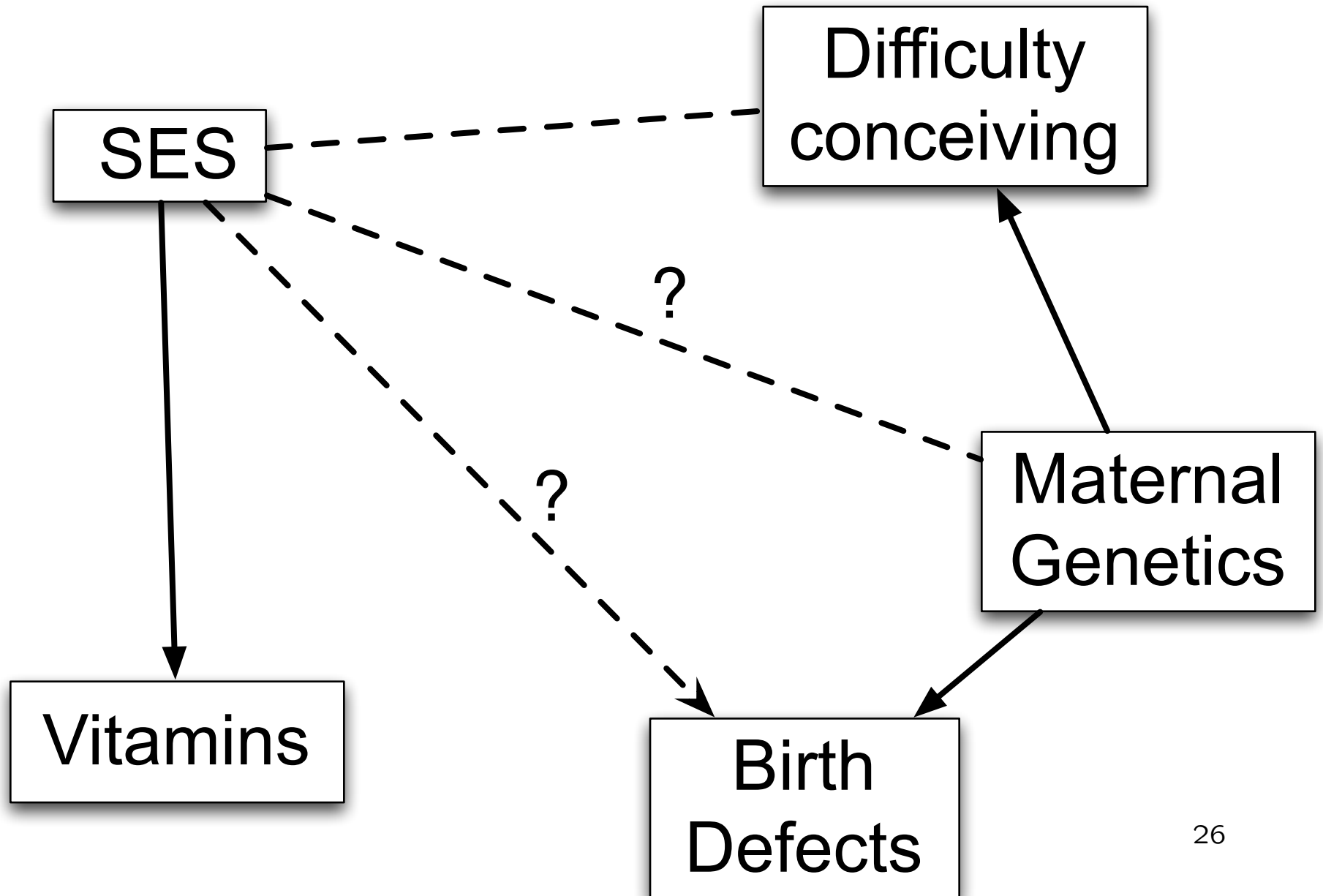
Controlling for PNC

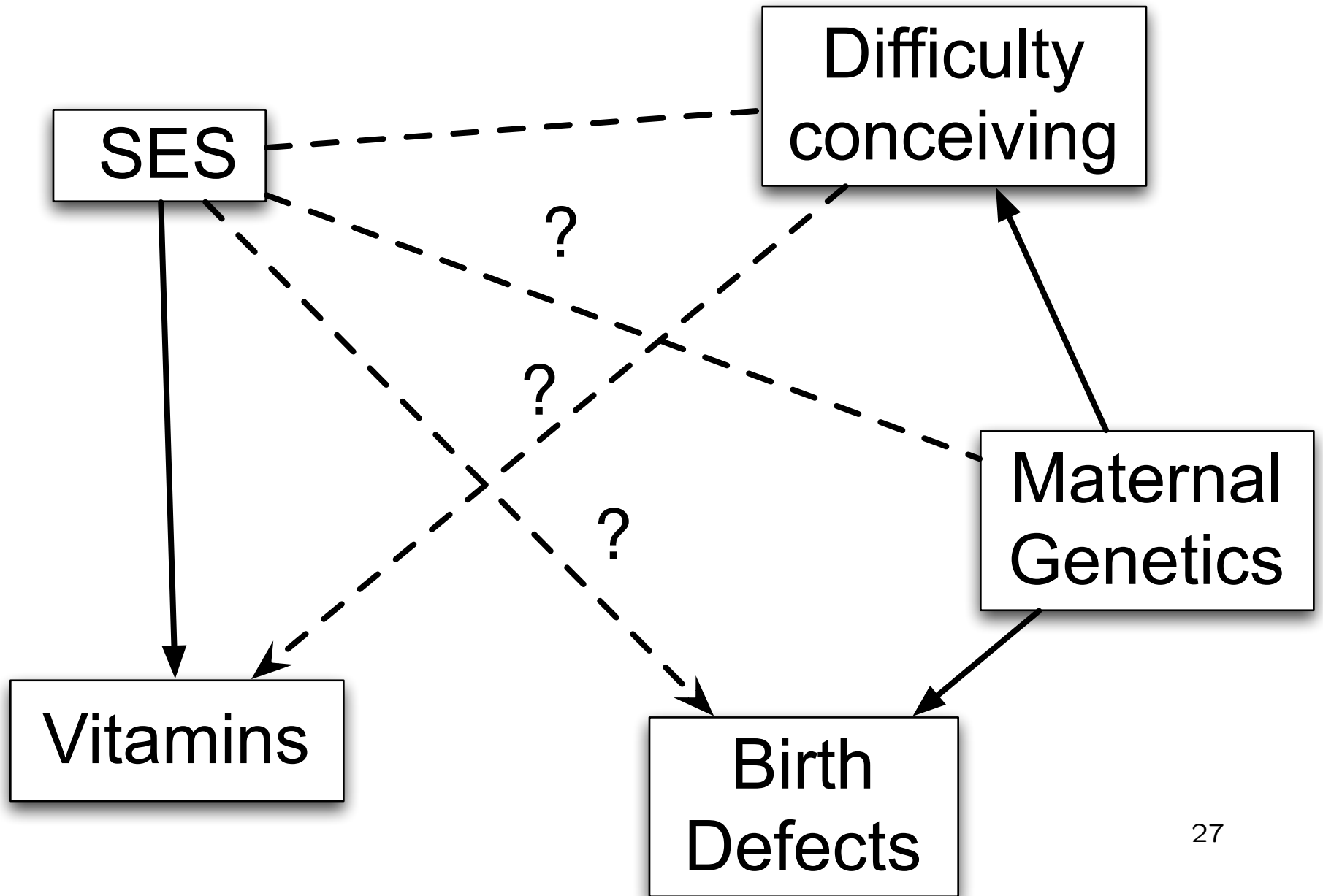
- Need to control for PNC because it is a common cause of vitamin use and birth defects
- Controlling for PNC effectively removes it from DAG
- Step 5 delete PNC from DAG, blocking that path
- Step 6: re-check for backdoor paths



What do we need to control for?

- Controlling for PNC is not enough
- Controlling for SES, difficulty conceiving, or maternal genetics required to block backdoor path
- Is just SES enough? why or why not?
- Any missing connections?





Goal 2: insights from DAGs:

- Exclude from model
 - mediators (unless estimating direct effect)
 - common effects (colliders)
 - redundant confounders (all on the same pathway)
- Adjusting for a confounder/collider (e.g., PNC) may require adjusting for additional factors
- More than one subset of confounders may suffice

Goal 2: DAGs

- Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*, 1999;10(1):37–48.
- Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*, 2004;15:615–625.

Goal 2: uncertainty in DAGs

- If you can't confidently exclude an arrow, include it
- To deal with uncertainty, also adjust for potential confounders
 - necessary for face validity
 - with $p < 0.2$ after adjustment and/or if inclusion changes coefficient for primary predictor by $>10\%$

Recommendations for Goal 2

- Use DAG to determine minimum set of confounders, exclude mediators, common effects
- Use liberal inclusion rules for potential confounders to minimize residual confounding
 - more later on number of predictors
- Check for interactions with main predictor
- Mediation, high correlation among control variables OK

Recommendations for Goal 2

- Hypothesis testing only of interest for primary predictor
 - so somewhat robust against inflation of type-I error
 - type-I errors for covariates are acceptable
 - results for covariates can be left in the background

RCTs – a special case of Goal 2

- Treatment the predictor of primary interest
- Confounding not a problem if randomization worked
- Include covariates to
 - account for clustering by clinical center
 - reduce variance using pre-specified stratification variables

RCTs – adjusting for baseline imbalances

- Not possible to specify what will be imbalanced *a priori*
- Problems: model error, shopping for significance
- One solution: do as a sensitivity analysis

Goal 3: multiple causal factors

- What are the risk factors for a condition?
- Most difficult of three inferential goals
- Instead of one predictor of primary interest, several

Goal 3: multiple causal factors

- Potential problems
 - many possible mediating, interaction relationships
 - causal pathways may be unclear
 - false positive findings, particularly for interactions
 - no single model will summarize causal relationships
 - * mediation is usually an issue

Recommendations for Goal 3

- Ruling out confounding is still central
- Select potential predictors
 - as identified by DAG
 - for face validity
 - if they meet liberal inclusion criteria

Recommendations for Goal 3

- Cautious interpretation of weaker results is key
- Multiple models may be necessary to address mediation
- Not recommended: treating each predictor as primary in turn
 - may not give a consistent picture

Parsimonious models

- Retain only variables with $p < 0.05$ for goals 2 and 3?
- Justified by
 - Occam's razor: simpler explanations likelier
 - less to explain in the discussion
 - limit type-I errors
- Drawback: residual confounding, especially in small samples

Parsimonious models

- Over-fitting mainly an issue for prediction
 - Goal 2: minor problem for primary predictor
 - Goal 3: interpret weak findings cautiously

How many predictors can safely be included?

- Too many predictors can
 - degrade precision
 - in smaller datasets, swamp a real association
 - induce bias in logistic and Cox models, other GLMs

A too-simple rule of thumb

- 10 observations or events per variable (Peduzzi et al, *J Clin Epidemiol*, 1996;**49**:1373-9)
- Relaxing the rule of 10 EPV:
 - may need to violate rule to rule out confounding
 - positive findings robust down to 5 EPV
 - power the big problem; interpret negative findings using CIs
- Vittinghoff & McCulloch, *Am J Epidemiol*, 2007;165:710-8

10 EPV and sample size calculations

- Point of 10 EPV is model reliability, not power
- 10 EPV *not* a method for sample size calculation
- Sample size depends mainly on effect size
- 10 EPV can be way too few if effect size is small

Covariate adjustment and sample size

- Estimate sample size ignoring adjustment
- Inflate estimated sample size N by $VIF = 1/(1 - \rho_j^2)$
- ρ_j is multiple correlation of x_j with covariates
- Hsieh *et al.* (*Stat Med*, 1998;**17**:1623-34) recommend assuming $\rho_j = 0.3$ if prior data unavailable
- Implemented in `stpower` but not `sampsiz`

Recommendations: number of predictors

- Use 10 EPV as a cautionary flag
- If you are close to 10 EPV, check for
 - high correlations between predictors
 - inflated SEs when a new covariate is added
 - inconsistency between Wald and likelihood ratio tests (logistic and Cox models)
 - gross inconsistency with smaller models

Recommendations: number of predictors

- If trouble is apparent:
 - tighten inclusion criterion to $p < 0.15$ or $p < 0.1$
 - omit variables included only for face validity
- With a binary primary predictor, many potential confounders, and a rare outcome, consider propensity scores
 - *Note:* propensity scores don't solve the problem with a small dataset, or control for unmeasured confounders

Table 5.1: Cox Models for DVT-PE

Predictor variable	Hazard Ratio (95% Confidence Interval)				<i>P</i> -values	
	11 Predictors		5 Predictors		Wald	LR
HT vs. placebo	2.7	(1.4–5.2)	2.7	(1.4–5.1)	0.002	0.001
≥ 53 at LMP	3.6	(2.0–6.4)	3.3	(1.8–5.8)	< 0.001	< 0.001
Inpatient surgery	4.3	(2.1–8.7)	4.7	(2.3–9.5)	< 0.001	< 0.001
Hospitalization	5.6	(2.9–11)	6.7	(3.6–13)	< 0.001	< 0.001
Hip fracture	5.9	(0.8–46)	6.6	(0.9–51)	0.09	0.18
Leg fracture	17.3	(5.1–58)	14.1	(4.2–47)	< 0.001	< 0.001
Cancer	4.1	(1.7–9.7)	3.5	(1.5–8.4)	0.002	0.006
Nonfatal MI	6.0	(2.3–16)	4.4	(1.7–11)	< 0.001	0.002
Stroke/TIA	0.9	(0.1–6.5)	0.9	(0.1–6.4)	0.88	0.88
Aspirin use	0.4	(0.2–0.7)	0.4	(0.2–0.6)	0.003	0.004
Statin use	0.4	(0.2–0.9)	0.4	(0.2–0.7)	0.02	0.02

Co-linearity

- Variance inflation factor: SE of $\hat{\beta}_j$ increases with ρ_j
- Can make coefficient estimates very imprecise, even when F or χ^2 test for combined effects is statistically significant
- Co-linear predictors give similar information about outcome
 - p -value machinery can't figure out where it's coming from

Diagnosing co-linearity

- High pairwise correlations between predictors: caution if $r > 0.8$, trouble if $r > 0.9$
- Check variance inflation factors (`estat vif postestimation` command); watch out for $VIF > 5$
- If either of two co-linear predictors is included one at a time, t -test for predictor is significant
- In model with both predictors, F -test for joint effect significant, t -tests are not

Recommendations: dealing with co-linearity

- See if the model solves the problem: e.g., one clearly dominates, or both are independently predictive. Otherwise:
- *Co-linearity between the predictor of interest and a confounder you can't ignore: you may have to admit defeat*
- *Co-linearity between multiple predictors of interest: if it makes sense, create summary variables or select among them*
- *Co-linearity between control variables: not a problem*

Selection algorithms: univariate screening

- Screen using single-predictor models, retain predictors with $p < 0.2$
- Liberal retention criterion to avoid missing negatively confounded variables
- Initial list should reflect substantive considerations

Selection algorithms: univariate screening

- Keep everything meeting screening criterion in multi-predictor model?
 - can compare unadjusted and adjusted coefficients
 - unimportant, redundant variables may be kept
 - cautious interpretation would be key
- Alternatively: do further selection

Backward selection

- Start from big model including all candidate predictors
- Sequentially delete predictor with biggest p -value and re-fit model
- Stop when all remaining variables meet inclusion criterion (e.g., $p < 0.2$)

Backward selection

- Advantages: should do well with negative confounding
- Disadvantages:
 - initial model could be unstable if too large
 - selections subject to chance, especially between co-linear variables
 - sequential algorithm, so good models may be missed

Forward, stepwise selection

- Start from model including intercept only
- Sequentially add most important omitted predictor
- Stop when no more variables meet inclusion criterion
- Stepwise also allows variables to be deleted

Forward, stepwise selection

- Advantages: avoids problem with too-large initial model
- Disadvantages:
 - may not do well with negative confounding
 - selections subject to chance
 - sequential algorithm

Categorical variables

- Selection procedures should keep or discard all categories
 - use parentheses in STATA to force this:
`xi: sw, pr(.2): outcome x1 (i.categorical) x3 x4`
 - uses omnibus F or χ^2 test to guide selection
- Should we collapse categories *post hoc*?

Allen-Cady procedure

- Modification of backwards selection
 1. designate variables to be included by default
 - those of primary interest
 - for face validity
 2. rank remaining variables in order of importance
 3. delete least important variable from full model if p -value exceeds a cutoff you choose
 4. iterate until p -value for the least important remaining variable is smaller than the cutoff

Allen-Cady procedure

- May retain variables with p -values larger than criterion
- Limits inflation of type-I error by limiting number of models searched over
- Incorporates *a priori* considerations through ranking

Allen-Cady procedure: example

- Model to evaluate mortality risk factors in patients with CHF
- Included by default: age, race/ethnicity, DM, eGFR, NYHA class, AFIB, aspirin, diuretics, β -blockers, ACE-I, CCBs, statins
- Other covariates ranked by descending importance: SBP, tachycardia, smoking, BMI, exercise, alcohol use, Na, BUN, LDL, HDL, TG, Lp(a), left bundle branch block, history of MI, CABG, PTCA
- Retention criterion: $p < 0.1$; procedure stopped after removing LDL; four NS predictors retained

A simpler version of Allen-Cady

- Designate *a priori* variables to be included by default
 - then use standard backwards selection on the rest
- Only slightly more subject to type-I errors than Allen-Cady
- Avoids importance ranking, but only reflects variable importance via inclusion by default

Pitfalls of model selection

- P -values too small, CIs too narrow, but hard to fix
- Type-I error rate inflated by searching over many models
- Only complete cases used in automated procedures (i.e., no missings on any predictors in initial list)
- Poorly motivated, unstable choices between predictors
- Coefficient estimates biased away from null, especially for weak predictors (only retained when coefficient estimate is big by chance)

A priori model selection

- Select complete model *a priori* using DAG
- Advantages
 - avoids pitfalls of model selection
 - *p*-values retain their theoretical meaning

A priori model selection

- Disadvantages
 - mainly suitable for well-studied areas
 - opens door to residual confounding
- Still requires checking model for unmodeled non-linearities, omitted interactions

Sensitivity analysis

- Are substantive results sensitive to
 - model selection procedure (forwards, backwards, stepwise)?
 - model size / retention criterion?
 - *a priori* selections between co-linear covariates?
- Robustness of main findings using different model selection methods strengthens conclusions

Summary recommendations

- Goal 1 – prediction: select model using aggressive search plus cross-validation or AIC/BIC, validate in external sample
- Goal 2 – predictor of primary interest: minimize confounding by using an inclusive model strongly motivated by DAG
- Goal 3 – identifying multiple independent predictors: minimize confounding by using an inclusive model, interpret weak findings cautiously
- Specify models for goals 2 and 3 *a priori* if possible; otherwise use modified backwards procedures

Summary recommendations

- Numbers of predictors: relax the rule of thumb if necessary to rule out confounding, but check for bad behavior
- Co-linearity:
 - between control variables, not a problem
 - between a predictor of primary interest and a non-ignorable confounder, may mean defeat
- Use sensitivity analyses