

Matched designs and causal diagrams

Mohammad A Mansournia,^{1*} Miguel A Hernán^{2,3} and Sander Greenland^{4,5}

¹Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran,

²Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA, USA, ³Harvard-MIT Division of Health Sciences and Technology, Boston, MA, USA and ⁴Department of Epidemiology, UCLA School of Public Health, University of California, Los Angeles, CA, USA, and ⁵Department of Statistics, UCLA College of Letters and Science, University of California, Los Angeles, CA, USA

*Corresponding author. Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, PO Box 14155-6446, Tehran, Iran. E-mail: mansournia_ma@yahoo.com

Accepted 10 April 2013

We use causal diagrams to illustrate the consequences of matching and the appropriate handling of matched variables in cohort and case-control studies. The matching process generally forces certain variables to be independent despite their being connected in the causal diagram, a phenomenon known as unfaithfulness. We show how causal diagrams can be used to visualize many previous results about matched studies. Cohort matching can prevent confounding by the matched variables, but censoring or other missing data and further adjustment may necessitate control of matching variables. Case-control matching generally does not prevent confounding by the matched variables, and control of matching variables may be necessary even if those were not confounders initially. Matching on variables that are affected by the exposure and the outcome, or intermediates between the exposure and the outcome, will ordinarily produce irremediable bias.

Keywords Matching, causal diagram, bias, unfaithfulness

Introduction

Balanced matching forces the distribution of the matching factors to be identical across groups of individuals. These groups are defined by the exposure in cohort studies and by the outcome in case-control studies. The goal of matching differs by study design. In cohort studies, matching is used to prevent confounding due to the matching factors; if there is no source of bias other than confounding by the matching factors, adjustment for these factors may be unnecessary to remove bias. In case-control studies, matching is used to increase statistical efficiency when a subsequent procedure (e.g. stratification) is used to adjust for confounding, but introduces selection bias; thus, adjustment for the matching factors may be necessary to remove this bias even if the factors were not confounders to begin with.¹

In this paper, we use causal diagrams to represent matched studies and describe how appropriate handling of matching variables can be inferred from the diagrams. We provide diagrammatic explanations of well-known statistical results for matched studies (which we have often found some students have difficulty understanding in their algebraic form). We consider cohort studies with and without censoring or other missing data, and case-control studies. We describe matching on factors that would and would not be confounders in the absence of matching. In Appendix 1, we consider matching that produces irremediable bias (e.g. on variables that are affected by the exposure and the outcome, or intermediates between the exposure and the outcome). In Appendix 2, we provide results about the direction of bias without adjustment for the matching variable. For simplicity, we assume throughout that there is no measurement error, and we deal only with expected values, ignoring random error.

Causal diagrams and bias

The theory of causal directed acyclic graphs (DAGs) can now be found in many reviews and books.²⁻⁵ Briefly, a DAG includes nodes (measured and unmeasured variables) linked by directed edges (arrows). A causal DAG is one in which the absence of an arrow between two variables implies the absence of a direct causal effect, and in which it is assumed that all shared causes of any pair of variables are included in the graph.

A path between two variables is any non-crossing, non-repeating sequence of arrows (regardless of their directions) connecting the two variables. A directed path is a path traced out entirely along arrows head to tail. If there is a directed path from X to Y, X is called an ancestor of Y and Y is called a descendant of X. A variable is a collider on a path if two arrowheads on the path converge on it. A path is said to be blocked or closed if it contains a non-collider that has been conditioned on, or if it contains a collider that has not been conditioned on and no descendants of the collider have been conditioned on; otherwise it is unblocked or open. Two variables are d-separated if all paths between them are blocked; otherwise they are d-connected. Two sets of variables are said to be

d-separated if each variable in the first set is d-separated from every variable in the second set.

The directed paths from the exposure to the outcome transmit the effects under study and so are target (causal) paths; non-directed open paths between the exposure and the outcome transmit bias for estimating the effect of the exposure on the outcome and so are biasing paths. Given the background knowledge encoded in a causal DAG, we can use the compatibility assumption to determine whether the causal effect of the exposure on the outcome can be identified given certain variables in the graph, and can find sets of variables that are sufficient to identify that effect. A DAG is compatible with the joint probability distribution of the variables in the graph if, whenever two sets of variables are d-separated given a third set, the two sets are conditionally independent given the third. Compatibility is essential for the DAG to tell us which biases are removed or may be produced by covariate adjustment given the data distribution (adjustment validity). In all our examples we assume that the DAGs are causal and compatible with the underlying distribution.

The distribution is faithful to a DAG if two sets of variables are associated conditional on a third set

BOX 1 BASIC CAUSAL DIAGRAM TERMINOLOGY

- Parent/Child: if there is an arrow from X to Y, X is called a parent of Y and Y is called a child of X.
- Path: any non-crossing, non-repeating sequence of arrows (regardless of their directions) connecting two variables.
- Directed path: a path traced out entirely along arrows head to tail.
- Ancestor/Descendant: if there is a directed path from X to Y, X is called an ancestor of Y and Y is called a descendant of X.
- Collider: a variable is a collider on a path if two arrowheads converge on it.
- Blocked (closed) path/Unblocked (open) path: a path is said to be blocked or closed if it contains a non-collider that has been conditioned on, or it contains a collider that has not been conditioned on and no descendants of the collider have been conditioned on; otherwise it is unblocked or open.
- D-separation/D-connectedness: Two variables (or two sets of variables) are said to be d-separated if every path between them is blocked; otherwise they are d-connected.
- Target (causal) paths: the directed paths from the exposure to the outcome which transmit the target effect.
- Biasing paths: non-directed open paths between the exposure and the outcome, which transmit bias for estimating the effect of the exposure on the outcome.
- Compatibility: two sets of variables are independent conditional on a third set whenever the two sets are d-separated given the third.
- Faithfulness: two sets of variables are associated conditional on a third set whenever the two sets are d-connected given the third.

whenever the two sets are d-connected given the third. Faithfulness is not necessary to determine adjustment validity from the DAG; nonetheless, its violation implies that there will be adjustments with no impact on bias despite the DAG suggesting otherwise. Fortunately, unfaithfulness requires very specific cancellation or balancing of associations that is unlikely to occur naturally. One can however impose certain constraints on the data-generating process that produce unfaithfulness. Matching with a constant subject ratio within matched sets (balanced matching) is an example in which the selection process forces certain variables to be independent despite their being d-connected, thus inducing unfaithfulness. Balanced matching is the most common matching strategy, and we will use the term 'matching' to refer to that unless noted otherwise.

To deduce the impacts of adjustment, we will make use of well-known results on noncollapsibility:⁶ adjustment for a binary covariate L will change the risk ratio and difference if and only if L is associated with (and thus d-connected to) exposure unconditionally and with disease given exposure; L adjustment will change the odds ratio if and only if L is associated with exposure given disease and is associated with disease given exposure. These results continue to hold for nonbinary L , apart from artificial exceptions in which no change (collapsibility) occurs despite d-connections, including unfaithful distributions produced by matching.

Cohort studies without censoring or other missing data

Most cohort matching starts with exposed subjects and for each one selects one or more unexposed subjects with the same values for the matching factors. Subjects that are not matched are usually discarded from the analysis. Often cohort matching is balanced in that the ratio of exposed to unexposed is constant across the sets (e.g. two unexposed might be matched to every exposed). Balanced cohort matching forces the baseline distribution of the matching factors to be identical between the exposed and the unexposed. With balanced matching to the exposed, the distribution of matching factors in the sub-cohort formed by all matched sets is the same as the distribution of the matching factors among the exposed subjects in the original cohort. Because balanced matching on a variable L ensures that the distribution of L will be identical between the exposed and unexposed in the matched sub-cohort, an analysis restricted to the sub-cohort does not need adjustment for L to remove bias.

The concepts discussed in this section apply to all forms of balanced matching, including individual and frequency matching, matching the exposed to the unexposed or the unexposed to the exposed,

and matching on variables or on propensity scores. For simplicity, however, we will focus on the consequences of matching to the exposed on a variable L . Our implicit causal parameter is then the average causal effect of exposure in the exposed. We consider three scenarios: (i) L is a confounder, (ii) L is associated with the exposure but is not a confounder, and (iii) L is associated with the outcome but is not a confounder.

(i) L is a confounder

Figure 1 represents a matched cohort study of the effect of statin therapy E (1: yes, 0: no) on cardiovascular disease D (1: yes, 0: no) with confounding by hypercholesterolaemia L (1: yes, 0: no). The variable S indicates whether an individual in the source population is selected for the matched study (1: yes, 0: no). The square around $S=1$ indicates that the analysis is conditional on having been selected into the matched study ($S=1$), i.e. only subjects in the matched sets are included in the analysis.

There are arrows from E and L to S because selection into the sub-cohort depends on the values of E and L . The arrow from E to S indicates that the exposed are more likely to be selected into the matched sub-cohort. The arrow from L to S indicates that, among the unexposed, a subject's value of L will affect selection given that L affects E , for then the exposed and unexposed will have a different distribution of L in the original cohort, and thus the distribution of L in the unexposed must be modified by the selection to match that in the exposed.

The variables L and E are d-connected via two paths: $L-S-E$ and $L-E$. Nonetheless, L and E are independent (by design) in the matched sub-cohort. Matching induces an association via the path $L-S-E$ that is of equal magnitude, but opposite direction, to the association via the path $L-E$, ensuring that L and E are independent in the matched sub-cohort. Thus, because of the matching, the joint distribution of the matched data is unfaithful to the DAG in Figure 1.⁷ The independence of E and L ensures there is no confounding by L in the matched sub-cohort, and that adjustment for the matching factors is not necessary to obtain valid point estimates of the effect of E on D in the exposed.

(ii) L is associated with E but is not a confounder

Figure 2 represents a matched cohort study of the effect of oral contraceptive use E on cardiovascular disease D . Suppose religion L (1: religion forbids contraception, 0: other) has a direct effect on E , but not on D . In this setting, there is no confounding by L , and so adjustment or matching on L would be pointless. Again the joint distribution is unfaithful to the DAG, because the associations induced by each of the paths $L-S-E$ and $L-E$ cancel each other.

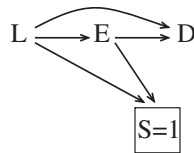


Figure 1 Cohort matching on a confounder

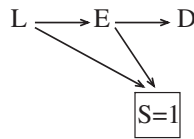


Figure 2 Cohort matching on a variable associated with exposure but that is not a confounder

(iii) *L is associated with D but is not a confounder*

Figure 3 represents a matched cohort study of the effect of ABO blood group E (1: non-O blood groups, 0: blood group O) on cardiovascular disease D. Sex L (1: woman, 0: man) is associated with D but not E in the original population, i.e. L is just a risk factor for D. In this setting, matching on L is pointless, because there was no confounding by L for the effect of E on D in the original population. The E-S-L-D path can produce bias for estimating the E effect on D if we fail to condition on L, except for the unfaithful case of balanced matching, in which selection is done to maintain the null unconditional L-E association.

Suppose now we wish to estimate the causal rate ratio in the exposed. The original balance produced by matching at the start of follow-up will not extend to the person-time available for the analysis if the outcome is associated with the exposure given the matching variable (L), and with the matching variable given exposure (as in Figures 1 and 3). The resulting L-E association in the person-time may deceive one into thinking that adjustment for L is necessary in the balanced-matched cohort study. In fact, the unconditional rate ratio is unconfounded and unbiasedly estimates the effect of the exposure in the exposed, even though it does not equal the common L-specific rate ratio due to non-collapsibility of the rate ratio.⁸ Moreover, the L-standardized rate ratio (the rate ratio standardized to the L-specific person-time distribution in the exposed) does not equal the unconditional rate ratio and thus does not unbiasedly estimate the effect of the exposure in the exposed. This is because both the exposure and the matching variable alter the person-time distribution, and thus adjusting for differences in person-time generally leads to collider bias.^{6,9,10} The same comments apply to the rate difference.

Cohort studies with censoring or other missing data

In the previous section, we have assumed that each sub-cohort member is followed until either the event

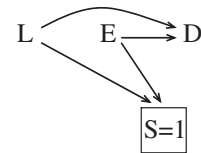


Figure 3 Cohort matching on a variable associated with outcome but that is not a confounder

of interest occurs or the study is terminated. In most real studies, however, this assumption is not met, because some subjects are censored due to losses to follow-up (e.g. emigration) or competing risks (e.g. deaths from other causes), and often subjects are dropped from the analysis due to missing data.

In this section, we use causal diagrams to illustrate how matching does not eliminate the need to adjust for the matching factor in cohort studies with censoring if (i) the matching factor is a risk factor for the outcome, and (ii) censoring is associated with the exposure given the matching factor and with the matching factor given the exposure.

The causal DAG in Figure 4 adds to the DAG in Figure 3 a selection variable C representing censoring (1: censored, 0: uncensored). The arrow from L to C reflects the assumption that women are more likely to drop out than men. ABO blood group is associated with certain cancers which may make subjects drop out of the study, as represented by the arrow from E to C. The square around C=0 indicates that the analysis is necessarily restricted to those who were not censored (C=0). In this setting, although there is no biasing path at the start of the follow-up, a biasing path E-C-L-D is opened due to censoring during the follow-up, necessitating adjustment for L.

Case-control studies

In case-control studies, matching ensures that cases and controls have the same unconditional (marginal) distribution of matching factors. The primary statistical reason for matching in case-control studies is to improve the efficiency (reduce the variance) of the effect estimates upon adjustment for the matching factors. The gains in efficiency are directly related to the strength of the association between matching factor and outcome and inversely related to the strength of the association between exposure and outcome.¹¹ Other reasons for matched case-control designs are (i) to enable adjustment for confounding variables which are difficult to measure, e.g., use of siblings as controls may help to adjust for genetic factors, and (ii) to provide an appropriate or convenient sampling frame for controls (e.g. one may sample controls who are individually matched to cases from the same neighbourhood without enumerating the entire population).

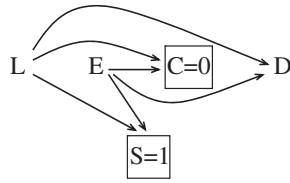


Figure 4 Cohort matching on a variable associated with both outcome and censoring

The concepts discussed in this section apply generally. For simplicity, however, we assume that cases are randomly sampled from those occurring in a closed cohort over a defined time period and controls are sampled from non-cases in that cohort. We will discuss the consequences of case-control matching on a variable L under three scenarios: (i) L is a confounder, (ii) L is associated with the exposure but is not a confounder, and (iii) L is associated with the outcome but is not a confounder. Appendix 3 discusses the counter-matched design.

(i) L is a confounder

Figure 5 represents a matched case-control study of the effect of statin therapy E on cardiovascular disease D with confounding by hypercholesterolaemia L . The variable S indicates whether an individual from the original cohort is selected into the matched case-control study (1: yes, 0: no). The square around $S=1$ indicates that the analysis is conditional on having been selected ($S=1$). There is an arrow from D to S because, by definition of a case-control study, disease status affects selection. The arrow from L to S indicates that, among the controls, a subject's value of L will affect selection given that L affects D (both directly and indirectly).

S is a descendant of the collider D on the path L - D - E , so conditioning on S generally opens this path from L to E . The variables L and D are d-connected via three paths: L - D , L - E - D and L - S - D . Matching induces an association between L and D , via the path L - S - D , that is of equal magnitude but opposite direction to the net association via the paths L - E - D and L - D , ensuring that L and D are unconditionally independent in the matched distribution. Thus, as a result of the matching, the joint distribution of the matched data is unfaithful to the DAG in Figure 5. The cancellation over the three paths L - E - D , L - S - D and L - D implies that the net association over the two paths L - S - D and L - D is not zero, implying that L and D are associated conditional on E . In other words, matching does not break the biasing path E - L - D , i.e. case-control matching on a confounder L does not adjust for confounding due to L .

Moreover, case-control matching on a confounder L introduces selection bias, because it creates the biasing path E - L - S - D . Thus, adjustment for L will usually be necessary. The bias that results from not adjusting

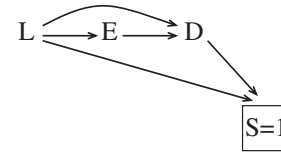


Figure 5 Case-control matching on a confounder

for L is towards the null, because matching makes cases and controls more alike with respect to exposure than they are in the source population (see Appendix 2 for a detailed explanation of direction of bias).

Because E and S are independent conditional on D and L , the L -conditional odds ratios (ORs) are collapsible over S , and thus the L -conditional ORs in the case-control study equal the L -conditional ORs in the original cohort,^{6,12} which have a causal interpretation as the causal ORs within levels of L . With small numbers of cases and controls in each matched set, sparse-data methods and modelling assumptions (e.g. homogeneity of the ORs across L) are required to estimate the L -conditional ORs.^{13,14}

An exception to the above is the scenario depicted in Figure 6 in which E has no effect on D . The variables L and D are d-connected via the paths L - S - D and L - D , but in the matched case-control study these paths cancel each other exactly. Therefore E and D are independent and no adjustment for L is necessary.

(ii) L is associated with E but is not a confounder

Figure 7 represents a matched case-control study of the effect of oral contraceptive use E on cardiovascular disease D . Religion L has a direct effect on E , but not on D . The arrow from L to S indicates that among the controls, L will affect selection given that L affects D through E .

The variables L and D are d-connected via two paths: L - S - D and L - E - D . However, L and D are unconditionally independent (by design) in the matched study distribution. Again the distribution is unfaithful to the DAG, because the associations induced by each of the paths L - S - D and L - E - D cancel each other. Nonetheless, because E is associated with D given L and with L given D , the L - D odds ratio is not collapsible over E , implying that L and D are associated conditional on E .

The open path E - L - S - D in Figure 7 is a biasing path for the effect of E on D and so adjustment for L is necessary. Matching on a variable associated only with the exposure will usually harm efficiency and is considered a type of overmatching.¹ The efficiency loss increases as the effect of the exposure (on the outcome) and its association with the matching variable increase.¹⁵

In the original cohort, both the unconditional and the L -conditional ORs are unbiased and have a causal interpretation. We can see from Figure 7 that E and S are associated conditional on D and so the unconditional OR for the study is not collapsible over S : the

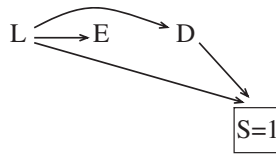


Figure 6 Case-control matching on a confounder (under the causal null hypothesis of no exposure effect)

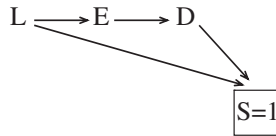


Figure 7 Case-control matching on a variable associated with exposure but that is not a confounder

unconditional OR for the study is not equal to the unconditional OR in the original cohort. But E and S are independent conditional on D and L and so the L-conditional ORs for the matched study are collapsible over S; thus the L-conditional ORs for the study are equal to the L-conditional ORs for the original cohort.^{6,12}

In a case-control study with matching on time (i.e. risk-set sampling), time should be adjusted for in the analysis if exposure prevalence varies over the study interval.^{16,17} Similarly for case-control studies with sibling or friend controls, matching should be taken into account in the analysis, if sibship or friendship is related to exposure (although for friend matching, bias may remain due to overlap of friend sets).¹⁸

Figure 8 is the same as Figure 7 except E has no effect on D. The E-L-S-D path can produce bias for estimating the E effect on D if we fail to condition on L, except for the unfaithful case of balanced matching, in which selection is done to maintain the null unconditional L-D association. With no E-D effect, the L-D association is collapsible over E and thus the conditional L-D association is also null, showing in turn that the null E-D association is collapsible over L and the unadjusted null-hypothesis test is valid. We caution however that estimation and non-null tests involve Figure 7, where L adjustment is needed.

Assuming faithfulness, Figures 1 and 3 (for matched cohort studies) and Figures 5 and 7 (for matched case-control studies) imply that in the matched distribution, L is associated with D conditional on E and L is associated with E conditional on D, so that the OR is not collapsible over L. Figures 1 and 3 illustrate noncollapsibility without confounding or selection bias, whereas Figures 5 and 7 illustrate noncollapsibility with selection bias, with and without confounding.¹⁹

(iii) L is associated with D but is not a confounder

Figure 9 represents a matched case-control study of the effect of ABO blood group E on cardiovascular

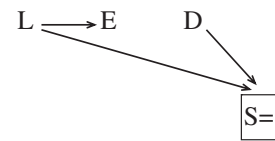


Figure 8 Case-control matching on a variable associated with exposure but that is not a confounder (under the causal null hypothesis of no exposure effect)

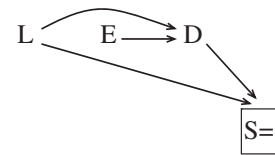


Figure 9 Case-control matching on a variable associated with outcome but that is not a confounder

disease D. Sex L is associated with D but not E in the source population. The arrow from L to S indicates that among the controls, a subject's value of L will affect selection given that L affects D.

Conditioning on S, a descendant of the collider D on the path L-D-E, creates an association between L and E. Thus the variables L and D are d-connected via three paths: L-D, L-E-D (because conditional on S, L is associated with E through D, and E affects D) and L-S-D. Matching induces an association via the path L-S-D that is of equal magnitude, but opposite direction, to the net result of the associations via the path L-E-D and L-D, ensuring that L and D are independent in the matched study distribution. This means the sum of two paths L-S-D and L-D is not zero and L and D are associated conditional on E.

The open paths E-L-D and E-L-S-D in Figure 9 are biasing paths for the effect of E on D and so adjustment for L is necessary.¹² Matching on a risk factor that is not associated with the exposure does not remove bias and could harm the cost efficiency.¹⁰ In the source population, both the unconditional and the L-conditional ORs have causal interpretations, although the unconditional OR is closer to 1 than the common L-specific OR (an example of noncollapsibility without confounding).¹⁹ E and S are associated conditional on D and so the OR in the study is not collapsible over S: the unconditional OR for the study is not equal to the unconditional causal OR for the source population. But E and S are independent conditional on D and L and so the L-conditional ORs for the study are collapsible over S: the L-conditional ORs for the study equal the causal L-conditional ORs for the source population.^{6,12}

The magnitude of bias resulting from not adjusting for L may be less than in the two previous scenarios (Figures 5 and 7). To see why, we can use Cornfield's inequality.²⁰ First, conditional on D, the L-E association should be smaller than the L-D and E-D

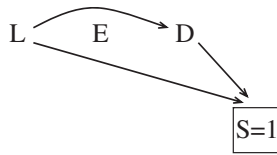


Figure 10 Case-control matching on a variable associated with outcome but that is not a confounder (under the causal null hypothesis of no exposure effect)

associations. This is not the case for scenarios represented in Figures 5 and 7 where L and E are associated in the source population. Second, conditional on D, the E-S association should be smaller than the L-E and L-S associations. These relations suggest that conditional on D, the association E-S should be much smaller than the L-D and E-D associations. A weak association between E and S conditional on D implies that the unconditional OR for the study is not very different from the unconditional causal OR for the source population.

With case-cohort sampling and homogeneity of the risk ratio, L and E are independent conditional on $D=1$ and $D=0$; ¹⁵ thus, E and S are independent conditional on D and so the unconditional study OR equals the unconditional causal risk ratio for the source population. With density sampling and homogeneity of the rate ratio, L and E are not independent conditional on $D=1$ and $D=0$, because the independence of L and E at the start of follow-up will not extend to the person-time experience for the source population. Thus, unconditional OR for the study does not equal the unconditional causal rate ratio. Assuming disease and competing risks are uncommon in all L-E categories and approximate homogeneity of the risk, rate or odds ratio, L and E are approximately independent conditional on $D=1$ and $D=0$. Thus, E and S are approximately independent conditional on D and the unconditional OR for the study is approximately equal to the unconditional causal OR for the source population.

Figure 10 represents the scenario in which E has no effect on D. L and D are d-connected via two paths: L-S-D and L-D, even though they are independent (by design) in the matched study distribution. There is no biasing path from E to D so adjustment for L is unnecessary.

In sum, in case-cohort studies, matching on a variable which is a risk factor for outcome but is not associated with exposure does not introduce selection bias. In contrast, in density or cumulative case-control studies, matching on such a variable may lead to selection bias if exposure affects the outcome, because matching variable and exposure are associated in the matched study distribution. However, the bias resulting from ignoring the matching variable is negligible when the outcome is uncommon over the risk interval.

Discussion

We have used causal diagrams to represent matched studies and have described how the appropriate handling of the matching variables can be inferred from the diagrams. In general, matching produces unfaithfulness, in that certain variables will be independent in the matched-data distribution despite their being d-connected in the causal diagram generating the matched data.

In cohort studies without censoring, balanced matching prevents confounding by the matching factors; thus adjustment for the matching factors is not necessary to obtain valid effect estimates in the matched sub-cohort. Nonetheless, ignoring the matching factors may lead to biased estimates and invalid tests if other adjustments are made (see Appendix 4). Furthermore, adjustment for matching factors is needed if there is censoring associated with the exposure given the factor and with the factor given exposure. Even in cohort studies without censoring, adjustment for matching factors is generally needed to obtain valid variance estimates. ¹

In case-control studies, balanced matching on the variables associated with both exposure and outcome results in selection bias toward the null, and adjustment for the matching factors is necessary to control both the selection bias introduced by matching and the original confounding. Matching on a non-confounder associated with the exposure leads to selection bias. Adjusting for such a variable is necessary to control the induced selection bias, which usually results in reduced efficiency relative to an unmatched study in which no adjustment for the variable would have been needed. ^{1,11} Worse, irremediable bias can be produced by case-control matching on a variable affected by the exposure (such as an intermediate) or the outcome. ^{1,9,10,21}

Keeping within the limits of causal diagrams, we have focused on the qualitative impacts of ignoring matching. Nonetheless, the purpose of matching is the potential gain in efficiency, either in practical terms (e.g. using neighbourhood matching to minimize travel by interviewers) or statistical terms (variance reduction). ¹ For quantitative studies of the impact of case-control matching on efficiency, see the citations on pp. 179–80 of Rothman *et al.*; ¹ for the impact of observational-cohort matching, see Greenland and Morgenstern. ²²

Funding

This work was partly funded by NIH grant R01 HL080644.

Acknowledgements

We thank Tyler VanderWeele for his helpful comments on an earlier draft of this paper.

Conflict of interest: None declared.

KEY MESSAGES

- In cohort studies, matching procedures select individuals based on the values of their covariates and exposure, and thus covariates and exposure are connected in the causal diagram through the selection node. However, matching ensures that the connected variables are independent in the matched subpopulation despite being connected (a phenomenon known as unfaithfulness).
- Cohort matching can prevent confounding by the matched variables; censoring or other missing data and further adjustment may however necessitate control of matching variables.
- Case-control matching generally does not prevent confounding by the matched variables; control of matching variables may be necessary even if those were not confounders initially.
- Matching on variables that are affected by the exposure and the outcome, or intermediates between the exposure and the outcome, will ordinarily produce irremediable bias.

References

- Rothman KJ, Greenland S, Lash TL. Design strategies to improve study accuracy. In Rothman KJ, Greenland S, Lash T (eds). *Modern Epidemiology*. 3rd edn. Philadelphia, PA: Lippincott Williams & Wilkins, 2008, pp. 168–82.
- Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10**:37–48.
- Pearl J. *Causality*. 2nd edn. New York: Cambridge University Press, 2009.
- Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*. 2nd edn. Cambridge, MA: MIT Press, 2001.
- Glymour MM, Greenland S. Causal diagrams. In Rothman KJ, Greenland S, Lash T (eds). *Modern Epidemiology*. 3rd edn. Philadelphia, PA: Lippincott Williams & Wilkins, 2008, pp. 183–209.
- Greenland S, Pearl J. Adjustments and their consequences - collapsibility analysis using graphical models. *Int Stat Rev* 2011;**79**:401–26.
- Hernán MA, Robins JM. *Causal Inference*. London: Chapman & Hall/CRC, 2013.
- Greenland S. Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology* 1996;**7**:498–501.
- Greenland S. Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology* 2003;**14**:300–06.
- Hernán MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;**15**:615–25.
- Thomas DC, Greenland S. The relative efficiencies of matched and independent sample designs for case-control studies. *J Chronic Dis* 1983;**36**:685–97.
- Didelez V, Kreiner S, Keiding N. On the use of graphical models for inference under outcome dependent sampling. *Stat Sci* 2010;**25**:368–87.
- Greenland S. Applications of stratified analysis methods. In Rothman KJ, Greenland S, Lash T (eds). *Modern Epidemiology*. 3rd edn. Philadelphia, PA: Lippincott Williams & Wilkins, 2008, pp. 283–302.
- Greenland S. Introduction to regression modeling. In Rothman KJ, Greenland S, Lash T (eds). *Modern Epidemiology*. 3rd edn. Philadelphia, PA: Lippincott Williams & Wilkins, 2008, pp. 418–55.
- Breslow N. Design and analysis of case-control studies. *Annu Rev Public Health* 1982;**3**:29–54.
- Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol* 1982;**116**:547–53.
- Vandenbroucke JP, Pearce N. Case-control studies: basic concepts. *Int J Epidemiol* 2012;**41**:1480–09.
- Flanders WD, Austin H. Possibility of selection bias in matched case-control studies using friend controls. *Am J Epidemiol* 1986;**124**:150–53.
- Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999;**14**:29–46.
- Cornfield J, Haenszel WH, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. *J Natl Cancer Inst* 1959;**22**:173–203, appendix.
- Schisterman E, Cole S, Platt R. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* 2009;**20**:488–95.
- Greenland S, Morgenstern H. Matching and efficiency in cohort studies. *Am J Epidemiol* 1990;**131**:151–59.
- Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 2009;**20**:512–22.
- Gail MH. Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In Moolgavkar SH, Prentice RL (eds). *Modern Statistical Methods in Chronic Disease Epidemiology*. New York: John Wiley & Sons, 1986, pp. 3–18.
- Langholz B, Clayton D. Sampling strategies in nested case-control studies. *Environ Health Perspect* 1994;**102**(Suppl 8):47–51.

Appendix 1 Irremediable bias induced by matching

We now discuss the consequences of matching on the variables that are colliders on a path between the exposure and the outcome, or intermediates between the exposure and the outcome. For brevity, we present here only the results for matched case-control studies. Similar results can be obtained for matched cohort studies. For example, in historical (record-based) cohort studies, analysts might inadvertently match on post-exposure variables; more subtly, missing data might be related to exposure and disease (as with informative censoring), leading to nonrandom selection of matched subjects with complete

data. Both problems are hazards of propensity-score matching in database studies.²³

(i) *L is a collider*

Figure A1 represents a matched case-control study of oral contraceptive use *E* (1: yes, 0: no) on endometrial cancer *D* (1: yes, 0: no). The variable *L* represents presence of vaginal bleeding in the month before the cancer diagnosis (1: yes, 0: no). *L* is affected by both *E* and *D*, and is therefore a collider in the path *E*-*L*-*D*. (For simplicity of presentation, we have not represented ascertainment bias in the causal DAG in Figure A1: vaginal bleeding leads to endometrial cancer being diagnosed.)

The variables *L* and *D* are d-connected via three paths: *D*-*L*, *D*-*E*-*L* and (after conditioning on *S*) *D*-*S*-*L*, but *L* and *D* are independent (by design) in the matched study distribution. *S* is a descendant of the collider *L* on the path *E*-*L*-*D*; thus, conditioning on *S* opens this path and alters the net association between *E* and *D*. Matching induces an association via the path *D*-*S*-*L* that is equal in magnitude, but opposite in direction, to the net association via the paths *D*-*E*-*L* and *D*-*L*. This implies that the net association from the two paths *D*-*S*-*L* and *D*-*L* is not zero.

The open path *E*-*L*-*S*-*D* in Figure A1 is a biasing path for the effect of *E* on *D* and so adjustment for *L* is necessary. On the other hand, conditioning on *L* opens the biasing path *E*-*L*-*D* leading to collider bias.^{6,9,10} In this setting, matching would bias both the unadjusted and the adjusted estimates.¹

Under Figure A2, *E* has no effect on *D*. Conditioning on variable *S*, a descendant of the collider *L* on the path *E*-*L*-*D*, creates an association between *E* and *D*. The causal structure is similar to that of Figure A1, and again matching would bias both the unadjusted and the adjusted estimates. Thus, both adjusted and unadjusted null-hypothesis tests are invalid, with probability of rejection approaching 1 in sufficiently large samples even under the null. The above results also apply when *L* is a descendant of a collider rather than a collider.

(ii) *L is intermediate between E and D*

Figure A3 represents a matched case-control study of oral contraceptive use *E* on endometrial cancer *D*. The variable *L* represents endometrial hyperplasia (1: yes, 0: no). *L* is an intermediate between *E* and *D*.

L and *D* are d-connected via three paths: *L*-*D*, *L*-*E*-*D* and *L*-*S*-*D*. Again, however, *L* and *D* are independent (by design) in the matched study distribution. The open path *E*-*L*-*S*-*D* in Figure A3 is a biasing path for the effect of *E* on *D* and so adjustment for *L* is necessary. On the other hand, conditioning on *L* blocks the causal path *E*-*L*-*D* and so introduces over-adjustment bias.^{6,21} Thus, in this setting, matching biases both the unadjusted and the adjusted estimates.¹

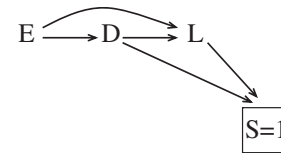


Figure A1 Case-control matching on a variable affected by exposure and outcome

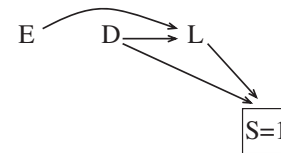


Figure A2 Case-control matching on a variable affected by exposure and outcome (under the causal null hypothesis of no exposure effect)

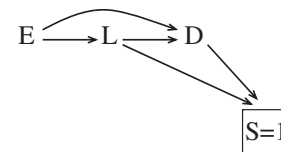


Figure A3 Case-control matching on an intermediate variable between exposure and outcome

Under Figure A4, *E* has no effect on *D*, so the same conclusions apply as with Figure 8 (balanced case-control matching on *L* will lead to a valid test of the *E*-*D* null with no conditioning on *L*, but *L* adjustment is still needed for estimation).

Appendix 2 Direction of bias without adjustment for the matching variable

Suppose that *E*, *L*, and *D* are binary variables and the *L*-conditional *E*-*D* odds ratio is homogeneous (constant) across *L* and not equal to 1. Then if *L* is associated with *D* conditional on *E* and unconditionally unassociated with *E*, the *L*-conditional ORs must be farther from 1 than the unconditional OR.²⁴ The same holds if we exchange *E* and *D*, because of symmetry of OR with respect to exposure and outcome.

Figures 5, 7, 9, A1, A2, and A3 imply that in the matched distribution, *L* is associated with *E* conditional on *D* but unconditionally unassociated with *D* (by design). Thus, we can conclude that in the matched population the conditional OR must be farther from 1 than the unconditional OR.

For Figures 5, 7, 9, the above result suggests that the adjusted OR (which is the conditional causal OR) must be farther from 1 than the unadjusted OR (which is biased). For Figures A1, A2, and A3, we

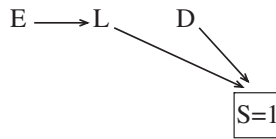


Figure A4 Case-control matching on an intermediate variable between exposure and outcome (under the causal null hypothesis of no exposure effect)

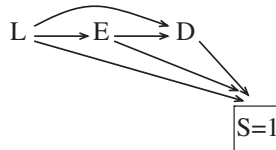


Figure A5 Counter-matching on E (with matching on L)

can only say that the non-causal adjusted OR will be farther from 1 than the non-causal unadjusted OR.

Appendix 3 Counter-matched design

A counter-matched design is a matched case-control design in which information on exposure or a proxy is used to improve statistical efficiency by maximizing the exposure variation within matched sets.²⁵ For a 1:1 counter-matched study, an exposed case will be paired with an unexposed control and an unexposed case with an exposed control, hence the name counter-matching. For a 1:3 counter-matched study, an exposed case will be paired with two unexposed controls and one exposed control, and an unexposed case with two exposed controls and one unexposed control.

The causal DAG in Figure A5 represents a 1:3 counter-matched study of the effect of statin therapy E (1: yes, 0: no) on cardiovascular disease D (1: yes, 0: no) with confounding by hypercholesterolaemia L (1: yes, 0: no). The causal DAG in Figure A5 adds to the DAG in Figure 5 an arrow from E to S representing counter-matching.

The variables L and D are d-connected via L-D, L-E-D, L-S-D and L-S-E-D, but they are unconditionally independent due to matching on L. Similarly, the variables L and E are d-connected via L-E, L-S-E, L-D-E and L-D-S-E, but they are unconditionally independent due to counter-matching in this example (the exposed:unexposed ratio equals 1 across the strata of L). The biasing paths from E to D in Figure A5 are E-L-D, E-L-S-D and E-S-D, represent

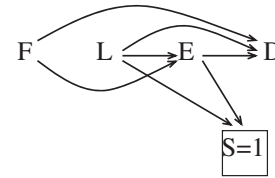


Figure A6 Cohort matching on one of two confounders (L)

confounding in the original cohort, selection bias introduced by matching on L and selection bias due to counter-matching, respectively. Adjustment for L is necessary to control the selection bias introduced by matching and the original confounding. One must adjust for the selection bias induced by counter-matching, for example by a weighting procedure based on the L-E-specific sampling fractions,²⁵ because there is no variable one can condition on to block the biasing path E-S-D.

Appendix 4 Breaking matching with adjustment for unmatched confounders

Matched cohort studies may lead to biased effect estimates if the analysis is adjusted for an unmatched covariate F associated with, but not affected by, the exposure conditional on the matching variable. Figure A6 is a modification of Figure 1 that includes the unmatched confounder F.

L and E are independent in the matched population so the three paths L-S-E, L-E, and L-F-E (because conditional on S, L is associated with F through E, and L affects E) should cancel each other. Adjustment for F removes the association between L and E through F in the matched distribution. This means the sum of the two paths L-S-E and L-E is not zero and thus L and E are associated after adjustment for F. That is, adjustment for F 'breaks the matching' because it makes L associated with E in the matched population. L needs to be adjusted for in the analysis or else the effect estimate will be biased, even under the causal null hypothesis of no effect of E on D. Adjustment for F may be necessary if this variable is associated with, but not affected by, both E and D conditional on L, as is the case in Figure A6. Similarly, breaking the matching in a matched case-control study in which the matching variable is a confounder will lead to an invalid test, if the analysis is adjusted for an unmatched covariate that is associated with the outcome conditional on the matching variable.