# Crash Course in Causality Quiz Questions

**Submitted By: Akshit Verma**
**NUID: 002059756**

Here are 15 multi-select MCQs based on the “Crash Course in Causality” written section. Each has multiple correct answers. After each question, I list the correct options and then explain every option (why it’s right or wrong).

**1) Which statements about causality vs correlation are true?**
A. Causal inference aims to estimate effects under hypothetical interventions.
B. Correlation guarantees a causal relationship if it’s statistically significant.
C. Counterfactuals encode “what would happen if we changed X.”
D. Predictive accuracy alone is sufficient for causal conclusions.

**Correct:** A, C
**Explanations:**

- A ✅ Correct: Causal inference targets interventional effects (e.g., do(X=x)).
- B ❌ Incorrect: Significance ≠ causality; confounding/colliders can create spurious correlations.
- C ✅ Correct: Counterfactuals formalize hypothetical outcomes under alternate treatments.
- D ❌ Incorrect: A model can predict well without capturing causal mechanisms.

---

**2) Which preprocessing choices support valid causal estimation?**
A. Adjusting for known confounders that affect both treatment and outcome.
B. Adjusting for post-treatment mediators to “get more precision.”
C. Avoiding adjustment for colliders on the treatment–outcome path.
D. Throwing every variable into the model to be safe.

**Correct:** A, C
**Explanations:**

- A ✅ Correct: Back-door adjustment blocks confounding bias.

- B ❌ Incorrect: Conditioning on mediators blocks part of the true effect (biasing total effects).
- C ✅ Correct: Conditioning on colliders opens spurious paths → bias.
- D ❌ Incorrect: "Kitchen-sink" adjustment can introduce collider bias and overfit.

---

**3) About missing data mechanisms (MCAR/MAR/MNAR), which are correct?**
A. Under MCAR, listwise deletion yields unbiased estimates (but loses power).
B. Under MAR, multiple imputation can be unbiased if the imputation model is correct.
C. Under MNAR, simple mean imputation is unbiased.
D. Using domain knowledge to include predictors of missingness can help meet MAR.

**Correct:** A, B, D
**Explanations:**

- A ✅ Correct: MCAR → deletion unbiased, though less efficient.
- B ✅ Correct: MAR + proper imputation model → unbiased in expectation.
- C ❌ Incorrect: MNAR requires modeling the missingness mechanism; mean imputation is biased.
- D ✅ Correct: Including missingness drivers helps approximate MAR.

---

**4) Regarding categorical encoding in a causal workflow:**
A. One-hot encoding preserves category separation without imposing order.
B. Ordinal encoding is ideal for unordered nominal variables.
C. Target encoding can leak outcome information if not done with proper CV.
D. Dropping one dummy avoids perfect multicollinearity.

**Correct:** A, C, D
**Explanations:**

- A ✅ Correct: One-hot is standard for nominal categories.
- B ❌ Incorrect: Ordinal implies order; wrong for purely nominal data.
- C ✅ Correct: Use nested CV/regularization to avoid target leakage.

- D ✅ Correct: Reference category prevents the dummy variable trap.

---

**5) Lasso (L1) in data preparation for causal analysis—what's true?**
A. Lasso can remove irrelevant/noisy predictors to improve stability.
B. Lasso guarantees selection of the correct causal adjustment set.
C. Lasso coefficients can be shrunk to zero, aiding feature parsimony.
D. Lasso alone can open collider paths if you include the wrong variables.

**Correct:** A, C, D
**Explanations:**

- A ✅ Correct: Useful for regularization and variance control.
- B ❌ Incorrect: Causal sufficiency is a causal/graph problem, not guaranteed by L1.
- C ✅ Correct: Core property of L1 penalty.
- D ✅ Correct: Automated selection without causal knowledge can condition on colliders.

---

**6) Good practices for an end-to-end preprocessing pipeline:**
A. Separate numeric/categorical transformers (impute, scale, encode) via a ColumnTransformer.
B. Fit transformers on the full dataset before the split to stabilize estimates.
C. Encapsulate steps in a Pipeline to avoid leakage and ensure reproducibility.
D. Apply imputation and scaling within CV folds.

**Correct:** A, C, D
**Explanations:**

- A ✅ Correct: Standard pattern to tailor transforms by type.
- B ❌ Incorrect: Fit on train only; fitting on full data leaks test info.
- C ✅ Correct: Pipelines prevent leakage and keep steps consistent.
- D ✅ Correct: All transforms must be fit inside the training fold.

---

**7) About propensity scores (PS) and IPW:**
A. The PS is P(T=1|X) and is used to balance covariates across treatment groups.
B. IPW weights are 1/PS for treated and 1/(1−PS) for controls (often stabilized).
C. If PS are near 0 or 1, weight variance inflates; trimming/clipping can help.
D. Estimating PS on the test set improves generalization.

**Correct:** A, B, C
**Explanations:**

- A ✅ Correct: PS reweights observational data toward randomized balance.
- B ✅ Correct: That's the basic form; stabilized weights use marginal treatment probability.
- C ✅ Correct: Extreme propensities → unstable weights; clipping/overlap checks are essential.
- D ❌ Incorrect: Fit PS on the analysis sample (train); never peek at test labels.

---

**8) Which are assumptions for valid IPW ATE?**
A. No unmeasured confounding (conditional exchangeability).
B. Positivity/overlap: every X has nonzero probability of both treatments.
C. Stable Unit Treatment Value Assumption (no interference, well-defined treatment).
D. The treatment effect must be linear.

**Correct:** A, B, C
**Explanations:**

- A ✅ Correct: Unobserved confounding biases IPW.
- B ✅ Correct: Needed to construct finite weights for all strata.
- C ✅ Correct: SUTVA underlies standard causal estimators.
- D ❌ Incorrect: Linearity is not an assumption of IPW ATE.

---

**9) In EDA for causal workflows, which goals are appropriate?**
A. Inspect outcome distributions across treatment groups.
B. Visualize missingness patterns to inform imputation strategy.

C. Identify variables to condition on purely by highest correlation with outcome.
D. Examine overlap by plotting PS histograms by treatment.

**Correct:** A, B, D
**Explanations:**

- A ✅ Correct: Helps spot imbalances and effect heterogeneity.
- B ✅ Correct: Guides choice of imputation method/mechanism assumption.
- C ❌ Incorrect: Correlation alone is not a safe criterion for adjustment; use causal reasoning.
- D ✅ Correct: Overlap diagnostics are crucial for IPW/PSM.

---

**10) About adjusting variables in causal models:**
A. Adjust for confounders that cause both treatment and outcome.
B. Adjust for mediators to estimate total effects.
C. Don't adjust for colliders to avoid inducing bias.
D. Adjusting for instruments may increase variance without reducing bias.

**Correct:** A, C, D
**Explanations:**

- A ✅ Correct: Blocks back-door paths.
- B ❌ Incorrect: Conditioning on mediators removes part of the total effect.
- C ✅ Correct: Collider bias is introduced by conditioning.
- D ✅ Correct: Instruments don't confound; adjusting can hurt efficiency.

---

**11) Handling categorical variables for causal PS models:**
A. One-hot nominal predictors before logistic PS to avoid fake ordinality.
B. Ordinal-encode nominal predictors to reduce feature count safely.
C. Group rare levels or use `min_frequency` to stabilize PS estimation.
D. Ensure encoders are fit on the training fold only.

**Correct:** A, C, D
**Explanations:**

- A ✅ Correct: Nominal → one-hot is standard.
- B ❌ Incorrect: Ordinal encoding imposes order where none exists → mis-specification.
- C ✅ Correct: Rare levels cause separation/instability; grouping helps.
- D ✅ Correct: Prevents leakage and preserves out-of-fold integrity.

---

**12) About imputation choices in causal pipelines:**
A. Median/most-frequent imputation is a robust baseline for tree/GLM models.
B. Impute using the test set too for consistency.
C. Include missingness indicators if missingness may be informative.
D. Multiple imputation can better reflect uncertainty than single imputation.

**Correct:** A, C, D
**Explanations:**

- A ✅ Correct: Simple, effective baseline inside CV folds.
- B ❌ Incorrect: Never impute using test info; fit imputer on train only.
- C ✅ Correct: "Missingness dummies" can help under MAR-like mechanisms.
- D ✅ Correct: MI propagates imputation uncertainty to estimates.

---

**13) Interpreting Lasso results in the written section context:**
A. Large positive coefficients indicate features increasing the predicted outcome.
B. Zeroed coefficients mean those features were deemed non-useful by the penalized model.
C. Lasso selection always equals the minimal sufficient adjustment set.
D. Standardizing features before Lasso aids fair coefficient shrinkage.

**Correct:** A, B, D
**Explanations:**

- A ✅ Correct: In linear/GLM settings with standardized inputs, sign/size reflect direction/strength.
- B ✅ Correct: L1 penalty shrinks some coefficients to zero.

- C ❌ Incorrect: Causal sufficiency is not guaranteed by Lasso.
- D ✅ Correct: Scaling prevents penalization from being dominated by feature scale.

---

**14) Diagnostics after estimating a PS model:**
A. Check covariate balance (e.g., standardized mean differences) after weighting/matching.
B. Inspect PS overlap (histograms/densities) for treated vs control.
C. If severe non-overlap, consider trimming the non-overlap region.
D. Skip model diagnostics if the AUC of the PS model is high.

**Correct:** A, B, C
**Explanations:**

- A ✅ Correct: Balance, not classifier AUC, validates PS use.
- B ✅ Correct: Overlap is required for credible IPW/PSM.
- C ✅ Correct: Trimming improves internal validity (at cost of generalizability).
- D ❌ Incorrect: High AUC can indicate separation; diagnostics are still required.

---

**15) Reporting a causal analysis in ML context (per the chapter):**
A. State the causal question, treatment, outcome, and the assumed confounders.
B. Declare assumptions (exchangeability, positivity, SUTVA) and limitations (unmeasured confounding).
C. Provide effect estimates with uncertainty (e.g., CIs) and sensitivity considerations.
D. Only report predictive metrics like RMSE/ROC; causal estimates don't need context.

**Correct:** A, B, C
**Explanations:**

- A ✅ Correct: Explicit design clarifies target estimand (e.g., ATE).
- B ✅ Correct: Assumptions drive identification; must be transparent.
- C ✅ Correct: Uncertainty and robustness are essential to interpretability.
- D ❌ Incorrect: Predictive metrics are not substitutes for causal effect reporting.