Does face-to-face compared to online teaching improve student exam performance in functional musculoskeletal anatomy? A causal inference approach

Project members:

Joanna Diong
Darren Reed
Hopin Lee
with thanks to Jan Douglas-Morris

February 8, 2021

Contents

1 Introduction

1.1 Background 15 Nov 2020

The Discipline of Anatomy and Histology (and former Discipline of Biomedical Science) at the University of Sydney delivers musculoskeletal anatomy teaching in face-to-face practical classes using human cadaver specimens and other media in anatomy laboratories. This approach to anatomy education has persisted based on the belief that it is more effective (than online or other modes of teaching) at helping students learn anatomical concepts and relate them to human movement. A key question is: Does face-to-face teaching (compared to online teaching) improve student exam performance in functional musculoskeletal anatomy?

Previous observational studies have attempted to quantify the causal benefit of face-to-face teaching compared to other modes of teaching on student performance (cite examples). Although promising, the findings from these studies are questionnable because investigators could not control for bias due to unmeasured confounding. With new developments in causal inference methods (Hernán and Robins, 2020; Herbert, 2020), it is now possible to determine causal effects under plausible assumptions using observational studies, provided these assumptions are explicitly pre-specified in a causal graph.

In March 2020, greater metropolitan Sydney was placed under lockdown to mitigate the spread of COVID-19. Anatomy units rapidly transitioned from face-to-face to online teaching in Semester 1, week 4. Subsequently, in the remaining 9 weeks of the 13 week semester, all anatomy teaching was conducted online. Both the mid- and end-semester theory exams that determine the final exam mark were delivered online. That is, the majority of anatomy content in Sem 1, 2020 and all formal assessments were delivered online. This is substantially different compared to Sem 1, 2019, where all lecture and practical class teaching and all assessments (mid-semester prac exam, end-semester prac and theory exams) were face-to-face.

In Figure 1, we specify a causal graph to how mode of teaching (face-to-face or online; exposure) causes student performance (exam mark; outcome).

Under this causal graph, the causal effect of mode of teaching on student performance is mediated through student engagement time with Canvas (through the Unit of Study resources), exam difficulty, and exam format. In addition, student Australian Tertiary Admission Rank (ATAR) scores directly and indirectly influence student performance through 'backdoor paths', which can be controlled. Importantly, there are no plausible common causes of mode of teaching and student performance. (A common cause of mode of teaching and student performance might be e.g. malicious intent, where students deliberately started COVID, then enrolled in the Unit to sit the exam; however, this example is implausible and resorts to conspiracy theory).

Based on this model, it would be possible to determine wither face-to-face teaching improves student performance by comparing exam marks from students in Sem 1, 2020 to Sem 1, 2019. The causal graph was specified using DAGgity software. To examine the *total effect* of mode of teaching on exam mark, no adjustment is necessary. To examine the *direct effect* of mode of teaching on exam mark, the minimal sufficient adjustment set is ATAR, Canvas engagement time, Exam difficulty, Exam format.

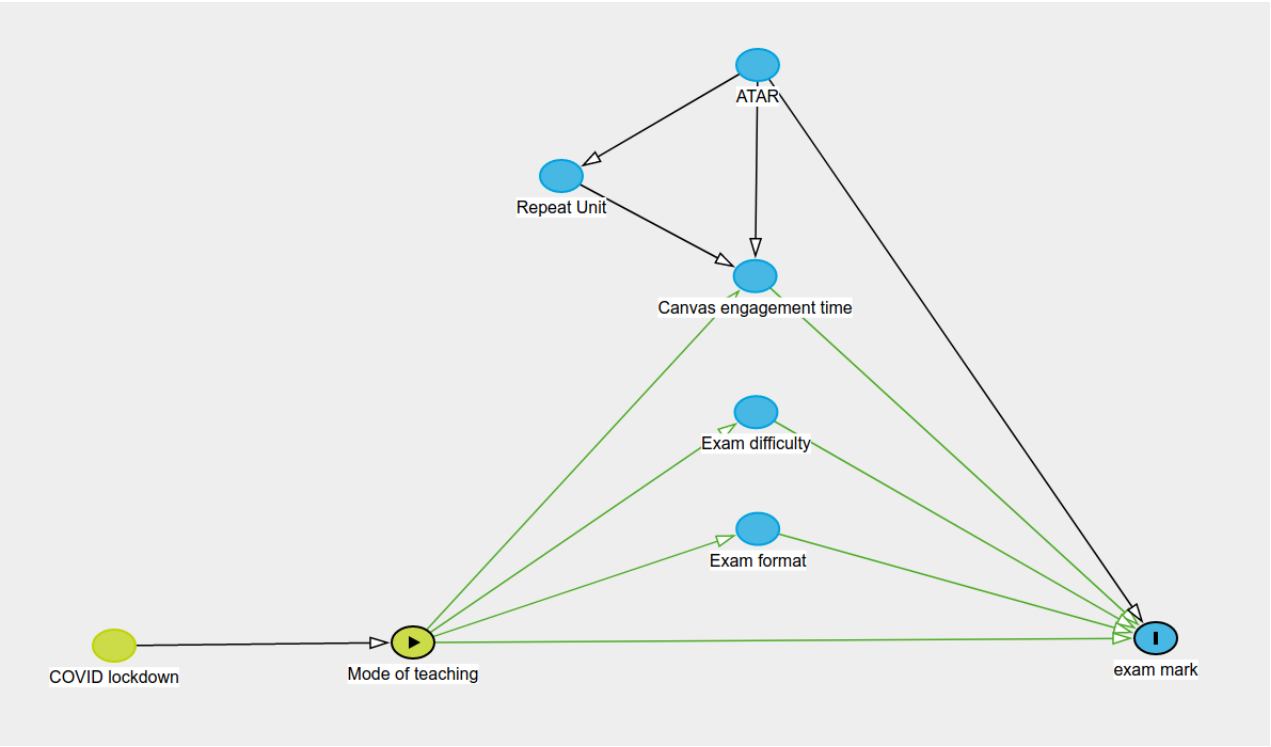

Figure 1: Directed acyclic graph (DAG 1) of mode of teaching on student performance.

1.1.1 Daggity code: dag-1 (get from raw text file)

dag "COVID lockdown" [pos="-1.005,-0.232"] "Canvas engagement time" [pos="-0.304,-0.919"] "Exam difficulty" [pos="-0.303,-0.666"] "Exam format" [pos="-0.301,-0.449"] "Mode of teaching" [exposure,pos="-0.678,-0.238"] "Repeat Unit" [pos="-0.516,-1.105"] "exam mark" [outcome,pos="0.134,-0.24 ATAR [pos="-0.301,-1.311"] "COVID lockdown" -¿ "Mode of teaching" "Canvas engagement time" -¿ "exam mark" "Exam difficulty" -¿ "exam mark" "Exam format" -¿ "exam mark" "Mode of teaching" -¿ "Canvas engagement time" "Mode of teaching" -¿ "Exam difficulty" "Mode of teaching" -¿ "Exam format" "Mode of teaching" -¿ "exam mark" "Repeat Unit" -¿ "Canvas engagement time" ATAR -¿ "Canvas engagement time" ATAR -¿ "Repeat Unit" ATAR -¿ "exam mark"

1.2 Follow up

Greg Londish, Peter Knight: extract de-identified exam marks linked to Canvas engemenet time and ATAR scores for BIOS1168 Sem 1, 2020 and Sem 1, 2019 cohorts. Link Canvas to SRES.

Darren, others: quantify 'exam difficulty'

Code as binary: mode of teaching, exam difficult, exam format

1.3 Email DR: 16 Nov

A couple of comments:

- Was the MSE 2019 a lab prac exam? I'm pretty sure it was but the last 2 years have been a bit of a blur. Were C and L intending to change the MSE this year to theory based?
- Are the ATAR scores enough prove the cohort are the same from one year to the next? Does it matter?
- Do we need to take into account that S1 had 3 weeks of F2F teaching?
- As the exams are different do we need to establish that they are examining similar objectives? Or at least similar numbers of objectives?
- I know you are not so keen on this Jo, but should we consider including other subjects eg BIOS5090, BIOS1169 or Helen's subject?
- Title suggestion: What is the effect on student exam performance of changing from face to face to online teaching in functional musculoskeletal anatomy?

1.4 Email HL: 16 Nov

Sounds like a neat idea. I had some thoughts on your DAG - copied below. I wondered how plausible it would be to assume there are no backdoor pathways via COVID-lockdown and other unmeasured mediators of COVID → Exam mark. e.g student motivation. Some thoughts on violation of positivity too.

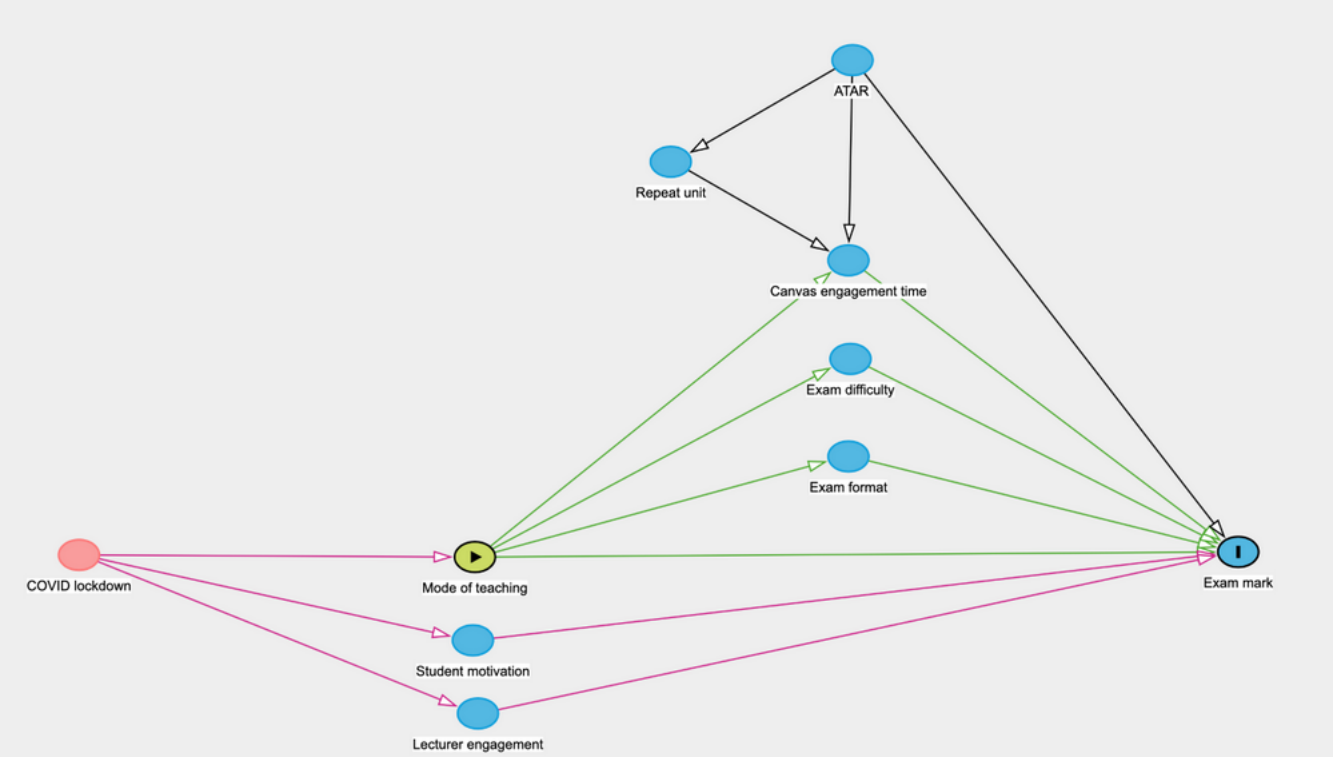

Figure 2: DAG 2

Under this revised DAG, the following adjustment sets would allow unbiased identification of the total effect:

- COVID lockdown
- Student motivation, Lecturer engagement

However, both of these confounding mechanisms may be intractable.

- An attempt to control for COVID lockdown will violate the positivity assumption. In other words, if we control for COVID lockdown by limiting the sample to pre-lockdown, the probability of a student being assigned to the online mode of teaching will be near zero. Also, if we control for COVID lockdown by limiting the sample to post-lockdown, the probability of a student being assigned to the FTF mode of teaching will also be near zero. Therefore we're pinned to either violate positivity or to let COVID lockdown to open up a back door path.
- Student motivation and lecturer engagment could be affected by COVID lockdown, and this could in turn affect exam marks. There are also other possible backdoors we may not be able to block. eg. procrastination time.
  – Here you could assume that these confounders might have little effect. But this could be a strong assumption. So you might choose to proceed with estimation but follow up with a sensitivity analysis for unmeasured confounding - eg. the E-value.

One other issue is the possible effect of COVID lockdown on CANVAS engagement time. It seems plausible that with students in lockdown, they may have more or less time to work in CANVAS. ie. more time because theyre at home, or less time because they dont have adequate IT equipment, or they are too distracted by the WWW. So if we draw an edge from COVID lockdown to CANVAS, we have two options to remove this bias (control for COVID lockdown OR control for CANVAS engagement). But pursuing each of these options induce other biases. Controlling for COVID violates positivity (as above), and controlling for CANVAS would control for an intermediate that would give us a direct effect, not a total effect.

1.4.1 Daggity code: dag-2

dag "COVID lockdown" [pos="-1.855,1.200"] "Canvas engagement time" [pos="-0.227,0.306"] "Exam difficulty" [pos="-0.223,0.605"] "Exam format" [pos="-0.227,0.901"] "Exam mark" [outcome,pos="0.598,1.191"] "Lecturer engagement" [pos="-1.011,1.680"] "Mode of teaching" [exposure,pos="-1.017,1.206"] "Repeat unit" [pos="-0.603,0.007"] "Student motivation" [pos="-1.022,1.459"] ATAR [pos="-0.219,-0.301"] "COVID lockdown" -¿ "Lecturer engagement" "COVID lockdown" -¿ "Mode of teaching" "COVID lockdown" -¿ "Student motivation" "Canvas engagement time" -¿ "Exam mark" "Exam difficulty" -¿ "Exam mark" "Exam format" -¿ "Exam mark" "Lecturer engagement" -¿ "Exam mark" "Mode of teaching" -¿ "Canvas engagement time" "Mode of teaching" -¿ "Exam difficulty" "Mode of teaching" -¿ "Exam format" "Mode of teaching" -¿ "Exam mark" "Repeat unit" -¿ "Canvas engagement time" "Student motivation" -¿ "Exam mark" ATAR -¿ "Canvas engagement time" ATAR -¿ "Exam mark" ATAR -¿ "Repeat unit"

1.5 Email DR: 17 Nov

Interesting points. As you say the effect of COVID on some people may have been positive with more time to learn/ engage (less travel time and less distractions?) or negative with less time for various reasons. IT access may have limited time to engage but I found more so

family circumstances, anxiety and lack of connectedness with other students (study groups) and lecturers, which as you say affects motivation. Certainly they were factors for my S1 cohort but the cohort did better than previous years.

Anxiety is always an issue, more so for some students. But personality type (introvert/ extrovert) and past experiences can also affect engagement levels. I would say some introverts may have engaged better in the lockdown state or have been less anxious that others were not louder than them. My son was one of them who thought it was more of an even playing field without one or two voices sucking all the airtime F2F. Anxiety was palpably less with some students feeling 'safe' in isolation and more relaxed.

Lecturer engagement is also an interesting one. I found it harder teaching online and so put in a lot more effort to try and stay connected with students. That may diminish with the novelty of online learning fading.

Will be good to explore this further.

1.6 Responses JD: 17 Nov

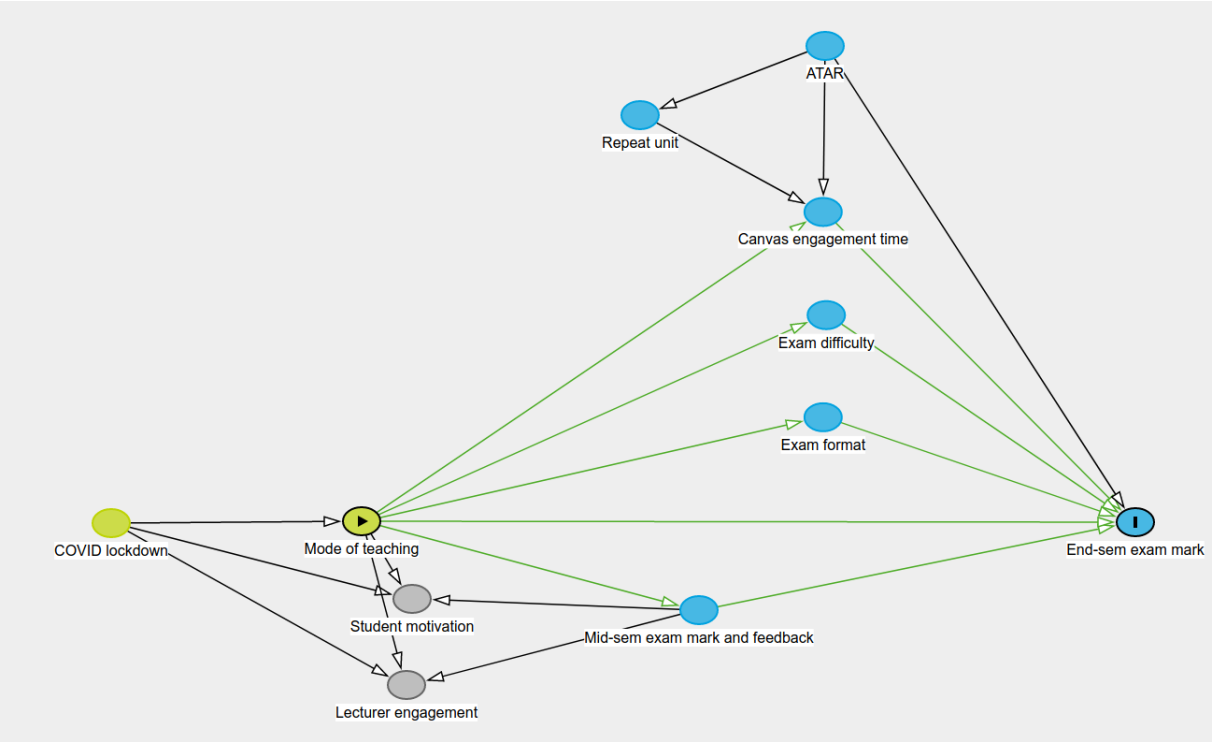

Figure 3: DAG 3

Suppose we only consider the final, End-semester exam component of student performance as the outcome, instead of both Mid- and End-semester exam components. It is plausible that the Mid-sem exam mark serves as feedback to students and lecturers, so that students change their motivation (e.g. study harder and procrastinate less if at risk of failing) and lecturers change their engagement (e.g. develop more content, or relate more with students) based on the marks as feedback. Subsequently, feedback from the Mid-sem exam mark blocks all backdoor paths from COVID lockdown to the End-sem exam mark. This would allow us to determine the causal effect of mode of teaching unconfounded by COVID lockdown on

student performance. Achieve this by comparing only the End-sem exam marks from the 2019 and 2020 Sem 1 cohorts.

Using only the End-sem exam mark also removes 'contamination' of the first 3 weeks of the 2020 Sem 1 cohort by F2F teaching; the End-sem exam tests content taught fully online from weeks 7-13. The exams across both cohorts are weighted differently. For a fairer comparison, compare only the percentage marks of the 2019 ESE paper-based theory exam (worth 40%) to the 2020 ESE online theory exam (worth 60%).

Q. What is the E-value, and how is it calculated?

1.6.1 Daggity code: dag-3

dag "COVID lockdown" [pos="-1.126,1.208"] "Canvas engagement time" [pos="-0.227,0.306"] "End-sem exam mark" [outcome,pos="0.167,1.205"] "Exam difficulty" [pos="-0.223,0.605"] "Exam format" [pos="-0.227,0.901"] "Lecturer engagement" [pos="-0.753,1.677"] "Mid-sem exam mark and feedback" [pos="-0.384,1.460"] "Mode of teaching" [exposure,pos="-0.810,1.202"] "Repeat unit" [pos="-0.458,0.025"] "Student motivation" [pos="-0.746,1.427"] ATAR [pos="-0.224,-0.175"] "COVID lockdown" -¿ "Lecturer engagement" "COVID lockdown" -¿ "Mode of teaching" "COVID lockdown" -¿ "Student motivation" "Canvas engagement time" -¿ "End-sem exam mark" "Exam difficulty" -¿ "End-sem exam mark" "Exam format" -¿ "End-sem exam mark" "Mid-sem exam mark and feedback" -¿ "End-sem exam mark" "Mid-sem exam mark and feedback" -¿ "Lecturer engagement" "Mid-sem exam mark and feedback" -¿ "Student motivation" "Mode of teaching" -¿ "Canvas engagement time" "Mode of teaching" -¿ "End-sem exam mark" "Mode of teaching" -¿ "Exam difficulty" "Mode of teaching" -¿ "Exam format" "Mode of teaching" -¿ "Lecturer engagement" "Mode of teaching" -¿ "Mid-sem exam mark and feedback" "Mode of teaching" -¿ "Student motivation" "Repeat unit" -¿ "Canvas engagement time" ATAR -¿ "Canvas engagement time" ATAR -¿ "End-sem exam mark" ATAR -¿ "Repeat unit"

1.7 Meeting 17 Nov: JD, DR, JDM

Consider other factors in Discussion:

- Amount of class time/ hours of tutor contact time: should be included in Canvas engagement time?
- Types of contact: student-content, student-student, student-tutor
- Lecturer engagement: creating content, relationships to students

1.8 Email HL: 19 Nov

Your response makes plausible sense. But I guess there may also be alternative explanations. For example, can we assume that student motivation and lecturer engagement (affected by mid sem marks) do not affect end-sem exam marks? I would think that whatever motivational or engagement status achieved (as a response to mid-semester feedback) will also influence end-sem marks. See attached DAG. Under this scenario, the only option is to adjust for COVID lockdown - which leaves us violating positivity.

But I guess the upside (preventing contamination from first 3 weeks) is a bonus, so I think that approach is better regardless of the above.

As you know, no DAG is perfect. But awareness of alternative DAG structures might help frame the discussion or to guide any sensitivity analyses.

Here are some papers on the e-value.

- Main paper: https://pubmed.ncbi.nlm.nih.gov/28693043/
- Commentary on usage: https://academic.oup.com/ije/advance-article/doi/10.1093/ije/dyaa094/5879832
- Software: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6066405/
- And a critique form Ioannidis. https://pubmed.ncbi.nlm.nih.gov/30597486/

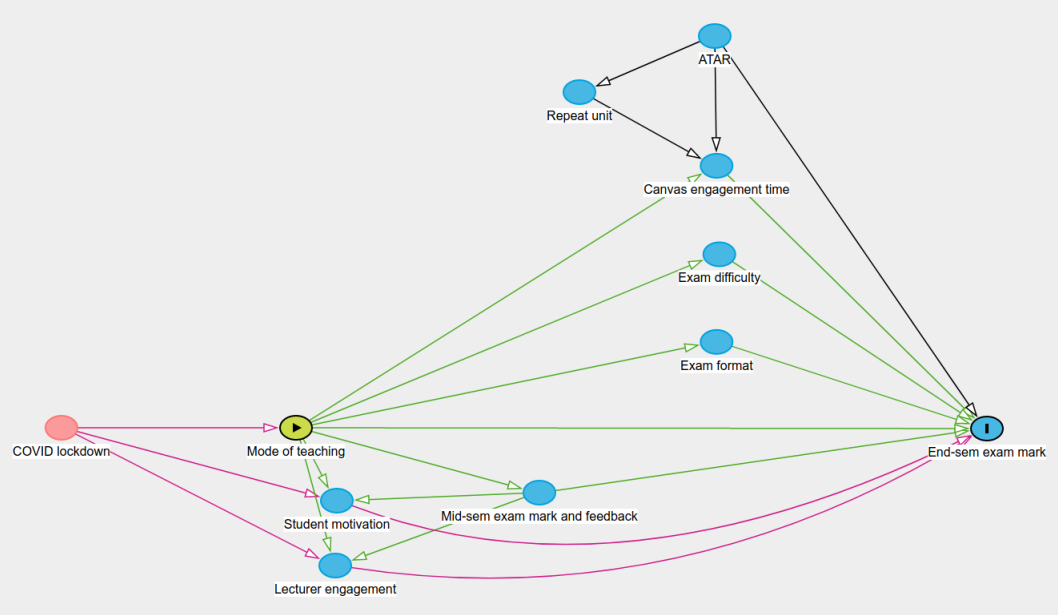

Figure 4: DAG 4

1.8.1 Daggity code: dag-4

dag "COVID lockdown" [pos="-1.187,1.191"] "Canvas engagement time" [pos="-0.227,0.306"] "End-sem exam mark" [outcome,pos="0.169,1.194"] "Exam difficulty" [pos="-0.223,0.605"] "Exam format" [pos="-0.227,0.901"] "Lecturer engagement" [pos="-0.786,1.656"] "Mid-sem exam mark and feedback" [pos="-0.487,1.410"] "Mode of teaching" [exposure,pos="-0.843,1.191"] "Repeat unit" [pos="-0.428,0.057"] "Student motivation" [pos="-0.784,1.437"] ATAR [pos="-0.230,-0.132"] "COVID lockdown" -¿ "Lecturer engagement" "COVID lockdown" -¿ "Mode of teaching" "COVID lockdown" -¿ "Student motivation" "Canvas engagement time" -¿ "End-sem exam mark" "Exam difficulty" -¿ "End-sem exam mark" "Exam format" -¿ "End-sem exam mark" "Lecturer engagement" -¿ "End-sem exam mark" [pos="-0.287,1.830"] "Mid-sem exam mark and feedback" -¿ "End-sem exam mark" "Mid-sem exam mark and feedback" -¿ "Lecturer engagement" "Mid-sem exam mark and feedback" -¿ "Student motivation" "Mode of teaching" -¿ "Canvas engagement time" "Mode of teaching" -¿ "End-sem exam mark" "Mode of teaching" -¿ "Exam difficulty" "Mode of teaching" -¿ "Exam format" "Mode of teaching" -¿ "Lecturer

engagement" "Mode of teaching" -¿ "Mid-sem exam mark and feedback" "Mode of teaching" -¿ "Student motivation" "Repeat unit" -¿ "Canvas engagement time" "Student motivation" -¿ "End-sem exam mark" [pos="-0.364,1.804"] ATAR -¿ "Canvas engagement time" ATAR -¿ "End-sem exam mark" ATAR -¿ "Repeat unit"

1.9 Responses JD: 20 Nov

Points taken – we should add these considerations to the Discussion. Overall, my impression is we can proceed with the study but make conclusions on causal effects (of F2F v online learning) under caveats of how plausible the DAGs are. I hope you agree as well? Some further thoughts:

Violating positivity. If COVID lockdown is adjusted for to close backdoor paths, then I don't think we can answer the question based on this model and we're back to square one. It seems useful to look for published findings where F2F and online learning were randomised in other areas to guide plausible conclusions of effectiveness.

HL response. Yes, I agree - I think as along as we can acknowledge the limitations of our DAG, then I think we can still aim to estimate a causal effect. But based on the plausibility of assumptions, sensitivity analyses, we may or may not decide to give the estimate a causal interpretation.

Collider blocking. My impression from trying to read Pearl and Hernan is that causal effects are fully blocked at a collider variable. But is partial blocking possible? For example, supposing the strength of the causal effect in one direction is much greater (i.e. bigger effect size) than the causal effect in the other direction, might some causal effects from the stronger variable "trickle through"?

HL response. Agreed - we cant adjust for COVID lockdown. Well violate positivity and well be stuck. Best thing here is to attempt to block this back door through descendants of COVID.

Direct and total effects. Should we compute the total effect only, or both total and direct effects? I'm not sure what best practice is in this area, or what might be the interest of the direct effect alone.

HL response. Im not exactly clear on the question in context to the proposed study. But some random thoughts consider this DAG:

$Z \leftarrow A \rightarrow B \leftarrow C$

Here B is the collider of A and C. I think youre asking whether C can cause Z. IF there is a plausible to draw an edge from $B \rightarrow Z$, then I think it is possible for C to cause Z via the collider B.

Im not sure this gets at your original question though

1.10 Responses JD: 22 Nov

Re. Direct and total effects. I agree and would prefer to just estimate the total effect alone. In our case, for DAG 3 (attached), the total effect requires no adjustment. I'm worried that readers might think that the effect we find could be because the mediators were different (e.g. different ATAR, Canvas engagement time, Exam difficulty, Exam format, Mid-sem exam

mark and feedback). I think this reflects they could be thinking of these as confounders, not mediators, and the mid-sem exam marks is the intermediate. So as you say, it is simpler to just estimate the total effect and interpret it under the DAG. Is my understanding correct?

HL response. Yes, much simpler (and perhaps more meaningful) to estimate the total effect. If we wanted direct effects, we would have to estimate several effects while holding each of the mediators (canvas, difficulty, format, mid sem mark) or some combination of them at once. By default, doing so would also invoke a new set of causal assumptions about confounding of the mediator-outcome effects. A rabbit-hole we may not want to go down! Or at least avoid if we can :)

Think this will be an interesting study, and I think this is a positive way of putting this unique and difficult situation were all in to good use.

2 Meeting 8 Feb 2021: HL, JD

Two approaches:

1) Primary analysis: under DAG 3, perform a linear regression of only the final exam marks between the two cohorts. Unverifiable assumption: no unmeasured confounding.

2) Secondary analysis: Consider COVID lockdown as an instrument. IV regression does not require conditional exchangeability assumption (i.e. there need not be no unmeasured confounding) assumption BUT instrument requires these assumptions:

1. COVID lockdown is associated with mode of teaching
2. COVID lockdown affects marks only through mode of teaching
3. COVID lockdown does not share common causes with marks
4. COVID lockdown does not modify the effect on marks (+ monotonicity)

(1) is probably true.

(3) is probably true.

But (2) is dubious and likely to be a strong assumption. If COVID lockdown does cause marks through paths other than modes (e.g. student motivation, lecturer engagement) we should expect the IV regression 2SLS effect to be *smaller* than the linear regression effect.

The primary analysis on it's own would stand. But it would be interesting to do the secondary analysis also, but we would want to pre-specify the interpretation before doing the analysis otherwise we're fishing.

E.g. suppose the IV regression effect is *bigger* than the primary analysis effect? This is unlikely: confounding bias is in one direction for IV regression (a weak instrument would always make the IV regression effect smaller), but could be in either direction for the primary analysis (if there really is confounding, our effect could be bigger or smaller under the 'no unmeasured confounding' assumption). So, if the IV regression effect appeared bigger, it is really because the primary analysis effect should have been smaller: if so, then perhaps DAG 4, not 3, is the correct DAG

The assumptions trade-off, and assumption sets for both approaches are unverifiable. But could this example be used to suggest which DAG might be correct?

Perhaps do both approaches but pre-specify the interpretation:

If primary analysis effect and IV regress effect are the same: best outcome

If primary analysis effect $>$ IV regress effect: COVID lockdown is a weak instrument, believe the primary analysis effect

If primary analysis effect $<$ IV regress effect: DAG 4 is likely more correct than DAG 3

References

Herbert, RD. Research Note: Causal inference, 2020.

Hernán, M, **Robins, J**. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.