WEBVTT
1 00:00:00.470 --> 00:00:01.460 <v ->Lets get started</v>
2 00:00:01.460 --> 00:00:03.310 and thank you everyone for coming today.
3 00:00:03.310 --> 00:00:06.520 And this is will be your final seminar
4 00:00:06.520 --> 00:00:09.330 for this semester for the (indistinct) the house seminar.
5 00:00:09.330 --> 00:00:11.230 And we are very, very pleasant
6 00:00:11.230 --> 00:00:14.683 to have very our own affiliate faculty,
7 00:00:15.740 --> 00:00:18.930 Dr. Josh Warren joining us.
8 00:00:18.930 --> 00:00:21.330 Dr. Warren is a associate professor
9 00:00:21.330 --> 00:00:23.950 at the Biostatistics Department here,
10 00:00:23.950 --> 00:00:27.870 and his research focuses on statistical method
11 00:00:27.870 --> 00:00:30.260 in public health with the emphasis
12 00:00:30.260 --> 00:00:32.397 on environmental health programs,
13 00:00:32.397 --> 00:00:35.690 and much of his work involves introducing spatial
14 00:00:35.690 --> 00:00:38.780 and spatial temporal models in the basin setting
15 00:00:38.780 --> 00:00:40.640 to learn about the association
16 00:00:40.640 --> 00:00:42.490 between environmental exposures,
17 00:00:42.490 --> 00:00:45.640 such as air pollution and various health outcomes,
18 00:00:45.640 --> 00:00:49.550 including the stillbirth that we are here today.
19 00:00:49.550 --> 00:00:52.260 He’s also interested in applying and developing
20 00:00:52.260 --> 00:00:56.230 some spatial temper models in collaborative settings,
21 00:00:56.230 --> 00:00:58.480 such as the infectious disease
22 00:00:58.480 --> 00:01:01.570 we been considered during the COVID pandemic.
23 00:01:01.570 --> 00:01:03.820 So without further ado, Josh,
24 00:01:03.820 --> 00:01:05.320 the floor is yours, thank you.
25 00:01:06.410 --> 00:01:08.470 <v ->Thank thank you Kai for the introduction.</v>
26 00:01:08.470 --> 00:01:10.350 Can everyone hear me?
And thanks to Kai for the invitation and Mulholland for setting all of this up and allowing me to do this virtually. It’s nice to be here talking about something other than COVID. I guess more recently in my past, I’ve been doing a lot of infectious disease work, so it’s kind of nice to be back into something that I’m still passionate about and still working heavily on. And so hopefully some of this today will be a little bit of review of what we’ve done and really current project that we’ve just completed and published, but hopefully there are some elements in here that you can find overlap within your own work. And so if you have, if you see something that brings a bell, just please reach out and we can kind of talk. My goal and all of this work is to kind of develop user friendly methods that are useful for people outside the EPI community and at large usually. So the EPI community and at large usually. So, yeah, just feel free to reach out afterwards, and I can share more information.

Critical window variable selection for mixtures and particularly air pollution and stillbirth.
So we’ll go ahead and jump into it. I think probably most people here will know air pollution, reproductive outcomes. There’s a pretty substantial literature at this point that suggests exposure to ambient air pollution during pregnancies associated with a number of adverse birth outcomes, including preterm pregnancy, low birth weight, congenital heart defects, stillbirth, and others. These are some of the main ones. Stillbirth is a more recently kind of emerging outcome of study. Traditionally, it’s been pre-term birth and low birth weight have gotten a lot of attention, but these associations are stable robust, and have been observed across a number of different study settings, designs, pollutants and there are a number of good review papers. If you’re interested in a lot of the EPI literature on this topic, I would kind of summarize previous a number of the previous EPI studies, but as they like to use pollution exposures that are summarized kind of A priorities, so they wanna focus on a trimester, they wanna focus on the entire pregnancy, like, what is the exposure across the entire pregnancy?
What impact does that have with respect to this outcome?

So these are usually pre-specified averaging periods and they’re explored separately in these different usually kind of traditional statistical models like logistic regression. And so lots of different pollutants are floating around in these analyses, lots of different averaging periods in terms of the exposure, relevance exposure period. Luckily working with pregnancy, we have a relatively stable idea of when exposure potentially affects the fetus. So lots of models floating around lots of pollutants and exposure weeks, but this method is inefficient and doesn’t allow for a joint identification of more kind of specific periods across the entire pregnancy in a continuous manner. So more recently there has been a focus on critical window estimation and identification. So this is where I have done quite a bit of work, I think, in this world. And then even more recently, I would say, and I know a number of people I work with even here.
At Yale, pollution mixers are becoming a really big deal.

So in this talk, we’re trying to combine both of these things, things that we know really well or that my group knows really well.

Critical windows, estimation identification, and pollution mixers are getting more attention.

Starting with critical windows of exposure and exactly what am I talking about when I’m talking about critical windows? Usually we’re thinking about pregnancy, but this can go for any really health outcome.

Understanding like specific timing of exposure with respect to outcome development.

The NIH included this identification of critical windows as a part of its strategic goals back in 2012. And the focus has remained since then.

So understanding like specific timing of exposure with respect to outcome development.
has a number of features but importantly, it could lead to improve mechanistic explanations of disease development, and ultimately focus guidelines for protection of the unborn child.

So we have, like I mentioned, we've done a lot of methods work here, trying to understand variability in these windows essentially, and how to estimate them appropriately. So you'll start to see, I show some pictures, some figures here that the models become really lots of parameters in these models. So you, it really becomes an estimation challenge.

Like how do you, the model makes sense, you can write it down, but can you actually fit these models? So we've done these or consider these models in a number of different settings, including the space temporal settings, survival statistics setting, semi parametric, non-parametric bays with multi-varied outcomes, and then more recently variable selection. And so inferences typically carried out in the Bayesian setting where I do most of my work due to increased computational flexibility and importantly incorporation.
of stabilizing prior structure.

So not only have these been done on the
method side
where a lot of my time is spent,
but I really like seeing them translated
to actual practice too.
So these methods and kind of variants
has successfully identified these critical win-
dows in a number of outcomes and settings
and different populations,
but pre-term birth, low birth weight,
so across a number of studies now.
So they’re getting good traction in other stud-
ies.
Well, not just in the stat literature,
which is nice to see.
To give you a more kind of practical view
of what I’m talking about,
this is one of the first studies we published on
way back in 2012.
And this is for Harris County Texas,
home of Houston, Texas.
And on the left two panels,
you’ll see output from our newly developed
method
on the right two panels,
you’ll see output from more of a naive ap-
proach
that was that we were considering at the time.
So what we’re talking about these critical windows
190 00:07:36.930 --> 00:07:38.940 are exactly what you’re seeing.
191 00:07:38.940 --> 00:07:40.920 Maybe you can see my mouse here,
192 00:07:40.920 --> 00:07:45.250 but these periods where these risk ratios
193 00:07:45.250 --> 00:07:48.910 in this case kind of exclude zero
194 00:07:48.910 --> 00:07:50.380 or these risk parameters,
195 00:07:50.380 --> 00:07:51.750 they’re not on any particular scale.
196 00:07:51.750 --> 00:07:54.370 That’s easily to interpreted in this case, un-
197 00:07:54.370 --> 00:07:56.510 but this means that elevated exposure
198 00:07:56.510 --> 00:07:59.400 during pregnancy week 10 for example,
199 00:07:59.400 --> 00:08:01.080 leads to an increase in this case,
200 00:08:01.080 --> 00:08:03.730 was preterm birth, a preterm birth risk.
201 00:08:03.730 --> 00:08:06.610 So during your early kind of mid first
202 00:08:06.610 --> 00:08:09.110 and early second trimester pregnancy,
203 00:08:09.110 --> 00:08:13.910 we were noticing some interesting elevated
204 00:08:13.910 --> 00:08:16.460 And what we’ve seen across a number of
205 00:08:16.460 --> 00:08:20.360 is that these windows vary by pollutant by
206 00:08:20.360 --> 00:08:21.580 they’re very different.
207 00:08:21.580 --> 00:08:24.230 There’s lots of variability for ozone for exam-
208 00:08:24.230 --> 00:08:26.773 it seemed to be early on in the first trimester.
209 00:08:26.773 --> 00:08:31.430 So this new methodology allows us to kind of
210 00:08:31.430 --> 00:08:34.930 on the signal and reduce some of this noise.
211 00:08:34.930 --> 00:08:37.970 So if you try to basically imagine your data
212 00:08:37.970 --> 00:08:40.840 you have lots of pregnant women in your
213 00:08:40.840 --> 00:08:43.550 and you have linked with that pollution expo-
214 00:08:43.550 --> 00:08:45.990 for the first 36 weeks of pregnancy.
A really naive thing to do would be, let’s just throw all of those into a multiple regression model, some binary regression model, all at the same time. Clearly there’s going to be correlation across time because exposure week one looks like exposure week two, et cetera. And if you do that, you can expect multicollinearity, which is jumping around of point estimates, increased variability, which is exactly what you see here. So our new methodology, which relied on like Gaussian processes and other smoothing techniques allowed us to in a data driven way, kind of tease out signal that you could almost make out by eye here. So if you look hard enough, you can see kind of a similar shape in both cases, but we were able to see a better shape here. So this is what we’re generally in the past have been talking about with critical window estimation and identification. We mentioned that we worked on the survival outcome, we started to think about preterm birth instead of just a binary outcome yes or no.
242 00:09:40.690 --> 00:09:42.790 We wanted to consider it as a survival outcome.
243 00:09:42.790 --> 00:09:44.330 So what’s the probability you make it
244 00:09:44.330 --> 00:09:46.090 to week 35 of your pregnancy,
245 00:09:46.090 --> 00:09:48.700 given that you’ve made it to 34 for example.
246 00:09:48.700 --> 00:09:50.050 So what this opened up was,
247 00:09:50.050 --> 00:09:52.830 well, maybe there are different exposure windows
248 00:09:52.830 --> 00:09:54.600 given different outcome weeks.
249 00:09:54.600 --> 00:09:57.870 So you can think of outcome week on the X axis
250 00:09:57.870 --> 00:10:00.610 on the Y axis here on an exposure week on the Y axis.
251 00:10:00.610 --> 00:10:04.100 So if you gave birth that week 27,
252 00:10:04.100 --> 00:10:07.100 you only had 27 weeks of exposure, for example.
253 00:10:07.100 --> 00:10:09.950 So people were leaving the set as pregnancy happened.
254 00:10:09.950 --> 00:10:11.780 And so introduced methodology
255 00:10:11.780 --> 00:10:14.790 that not only kind of smoothed in the exposure direction,
256 00:10:14.790 --> 00:10:17.470 but also smooth across the outcome direction.
257 00:10:17.470 --> 00:10:20.790 And so these darker areas indicate weeks
258 00:10:20.790 --> 00:10:23.150 and outcome weeks, exposure weeks and outcome weeks
259 00:10:23.150 --> 00:10:26.510 where elevated exposure more adversely impacts
260 00:10:26.510 --> 00:10:29.330 like the risk of preterm birth in this case.
261 00:10:29.330 --> 00:10:32.380 So there was a distinct difference in this early preterm
262 00:10:32.380 --> 00:10:33.440 and then this late preterm,
263 00:10:33.440 --> 00:10:35.370 which kind of was impacted by exposures later
264 00:10:35.370 --> 00:10:36.363 in the pregnancy.
And so underlying all of these kind of simplified plots

I’m showing you were

ds individual outcome week specific critical window plots

that we kind of are more accustomed to interpreting.

So more recently we got into the spacial world noticing

that, well, we started noticing that

when we applied these methods
to different data sets in different areas,

we were seeing different shapes, different windows,

different pollutants, emerging as important.

And so we begin to think, well, is there spatial variability in even at a local scale?

And so we develop new methodology

that can kind of tease out

not only temporal changes and exposure risk,

but also spatial variability as well.

So there’s spatial correlation component here along

with kind of these critical windows floating around as well.

So this was 11 counties in North Carolina,

including Wake County and the county to House Charlotte,

and this was a low birth weight study.

So there’s methodology around that can do this.

So we were working on these for a number of years
and we got approached basically with a question, how are you actually defining a critical pregnancy window? And it seemed obvious at first, but then we started to really question the assumptions we had been making, but obviously what we had been doing, if I go back a few slides here is just looking when did these individual week or time specific parameters exclude the critical value in zero in this case? And we were calling that a critical window but we started to worry that this might not be getting exactly what we’re hoping is this doing a good job? In particular, we were worried about over smoothing with something like a Gaussian process and specifically with the endpoint. So if you can imagine, I’ll go back one more time, sorry to scroll. Imagine the end points here and here, we begin to worry that the over smoothness could be pulling some of these actually null results into the critical set or vice versa, kind of pulling some important ones down to the null set.
So we were very concerned about the endpoints here when we started working on this more recent work. Our solution to this was critical window variable selection. So we like the smoothness, we like the plots that emerge. We like how we can interpret these things, but a variable selection component would allow us to turn some of these effects off, even if they appear to be significant in the plots. And so what this meant is, we introduced like a bayesian variable selection technique called critical window variable selection, where basically you still have the critical window plots that you know and love, and you know how to interpret, but underlying each effect now, you actually have this binary exclusionary, or inclusion variable that tells you whether this thing should be included. This particular weekly effect should be included in the critical window set. And what we found is that there are a number of times, not in this particular real case study in North Carolina, but through simulation,
we noticed that there were times when exactly what we had worried was happening. had been happening so effects near the border here were being pulled into the set, but luckily they were not being included in the variable selection component. So to be in the variable selection set now, you had to have posterior inclusion probability bigger than point five, so bigger than this line and your individual weekly effects had to be exclude zero with a 95% credible vulnerable. So with these two kind of definitions we were doing, we were getting a much better kind of recovering the true set of critical windows in simulation, at least. So this really outperformed what we had been doing previously. So we’ve been moving forward with this variable selection concept since then. All right, so we like critical window variable selection, we like a lot of these other methods. The problem is that as I know, a number of you are aware, the literature has really moved towards the science has moved towards pollution, mixtures and multiple exposures.
And a lot of these methodologies were developed with one pollutant in mind at the most two to three, but they were not generally meant for pollution mixtures for example. So our goal in this work was to extend what we liked the CWVS, critical and variable selection to accommodate mixtures. And so when we started to thinking about mixtures, when you have time varying exposures and time varying effects, it became relatively conceptually complicated because you have lots of parameters floating around. So we wanted something that could do like a dimension reduction essentially. So what we thought is a nice solution, like in a single pollutant context, or I’m sorry, in a single exposure time period context this weighted quantile sum regression, which I know a lot of you are familiar with, that have discussed weighted quantile sum regression here, but it offers a nice interpretable solution for estimating the impact of a mixture on an outcome. And it has this really nice sum to one constraint on the regression parameters.
And so you get in the end, you have 20 pollutants for example, and you get to see the relative contribution of each of these pollutants in terms of the entire mixture. So you have these little sum to one between zero and one probabilities or proportions that describe the role of individual pollutants. And then you have this global regression parameter that describes the impact of that mixture as defined by those weights on the health outcome. So it does a little two stage process estimate weights and then global regression parameter, not important for this talk. More recently in 2020, this was extended to the lag weighted quantile sum regression. And yeah, it extended WQS to the multiple pollutants setting in a really, I think of it as a relatively ad hoc solution, but basically WQS has fit at each exposure week separately. The weights are estimated, the mixtures are combined based on those weights. And then those kind of package mixtures are thrown into like a distributed live model to estimate similar curves is what I've been showing you so far. So the estimation of the weights
and their relative importance in the mixture are done separately outside of kind of the estimation of the regression parameters as well.

So this more, again, more of a two stage approach.

All right, so we like WQS because of its relative simplicity and its interpretability, we liked critical and variable selection.

So the goals here were to combine the estimation identification ability of CWVS with the interpretability and shrinkage properties of WQS within a unified modeling framework and extending WQS is nice.

It has zero to some to one components that are between zero and one, but you don’t actually get a sense of variable selection when doing this.

So none of the weights can exactly equal zero. We wanted a more sparse solution and so we introduced also a way to make these weights exactly zero.

So you can get a better sense of which pollutants are the main players in the mixture.

And so what we’re calling this is CWVS for mixtures or CWVS mix.

And so some features before we get
into a little bit of the details of the model, these are like the high, just if you take nothing else away from like what this model does, this is, I think the important slide here is that, we have main effects and first order interactions between the pollutants during each exposure period. So week one of pregnancy, week two of pregnancy, all of these interactions, all of these main effects are included. So there's lots of parameters you can already imagine are floating around here. We still hold onto this sum to one mixture weights at each exposure week separately. But we want to account for the fact that, what's happening in exposure week one may be similar to exposure week two to three to four, with this correlation dying out as you get further apart in exposure time. So we want these weights not to have to be estimated kind of independently at each exposure week. We want to enforce some smoothness, data driven smoothness preferably to estimate these weights. And as I mentioned, we want these weights to have a variable selection component.
So we can actually identify individual elements of the mixture and we still have this global risk parameter, and this is going to follow the CWVS model so that we can estimate these critical windows more accurately. All right, so the goals of this study are to develop CWVS mix. As I mentioned, simulation is really important in this world. I wanna make sure that what we’re doing is not just duplicating other efforts and that it’s actually offering something new, something helpful to the literature that we can point to. I think I know the shortcomings of something like lag, weighted quantile sum regression, but until I see it actually happen in simulation it’s just kind of hypothetical. So finally we wanna investigate the impact of multiple ambient air pollutants on stillbirth risk. And in this case, we’re focusing on New Jersey from 2005 to 2014. And actually we have really nice output from a novel data fusion model. There are lots of data fusion models floating
around right now, but this is a one from 2019, from our collaborator at Georgia tech and at Emory that provided 12 pollutants, 12 kilometer grid cell size across the entire US daily no missing this things like that. So for these particular pollutants. All right, so let’s talk a little bit about the model and what it does and some of the intuitive features that I think it has and why it might work well. So yeah, we’re starting with some outcome, it could be some adverse health outcome like preterm pregnancy or not, or stillbirth or not some be newly outcome where this PI describes kind of the probability that person I experiences this outcome. We model this probability using logistic regression as we normally would, these green I’m kind of trying to different. I’m trying to keep people’s attention to the parameters and how I’m mentally grouping them as well. So these green represent these typical like demographics. We know there are certain risk factors for different health outcomes, particularly pregnancy outcomes being over 35 for example, with preterm pregnancy, alcohol, smoking, et cetera.
So this would go into this ex transpose data. This specter here where a lot of our work came in are on these blue parameters, which are the weights that I’ve been talking about. So these weights, these blue parameters actually sum to one at each exposure week. So each exposure week T we basically have a vector of Lambdas that are weights between zero and one could be actually equal to zero exactly. And they sum to one at each exposure week separately, you notice their index Byte because we’re allowing the possibility that the exposure profile changes across the pregnancy. So it early on in the pregnancy, maybe the risk is primarily driven by pollutant A but later on in the pregnancy, perhaps that shifts. And so the weights would shift well as well, but we expect this shift to be smoother rather than complete choppiness across the exposure weeks. And so what these weights do are they kind of multiply here with the main effects and these first order interactions. And if you think about taking this sum across main effects and interactions,
you have this package of weighted exposure essentially.

And the alpha here tells us whether at exposure period $T$

this package has any impact on

your ultimate probability of developing the outcome.

So we have this nice sense of the weights,

help us describe what’s happening with the mixture profiles.

And, but the alpha keeps us honest

and keeps us able to say,

well, you know, this mixture’s interesting,

but it has no impact on the health outcome of interest here.

So how do we do these mixture weights?

As I mentioned, two features that we’re interested in

the ability to actually equal zero

and smoothness across time.

And so first point is to,

we introduce these latent weight parameters

that I’m calling Lambda star,

not to get too caught up in them.

Basically they’re continuously varying parameters

that as soon as they cross the zero threshold,

they turn on in our model.

So that’s what this maximum is doing.

So they turn on and they give you some weight

and then as soon as they cross into negative territory,

they go to zero.
So this is how we're getting actual zeros in these weights.
So the Lambdas and the Lambda Tilda can actually equal zero based on these underlying latent weight parameters.
All right, so we keep them summing to one by dividing by the sum of the numerator, essentially.
So whatever weights are positive gets summed and we're dividing by, we're basically self kind of correcting here so that the weights always come to one, these weights combined.
For the interactions, we don’t want the case.
We prefer sparse model, particularly as the number of pollutants get really large.
So the number of interactions will grow.
So what we want is our interactions that are only turned on essentially when the main effects are turned on.
So you can see these two indicators I've added basically say if the main effect themselves aren’t both turned on, this interaction effect gets zeroed out already.
So the interaction has a kind of a higher bar clear this strict hierarchy basically where both main effects have to be on and the interaction latent variable has to be on.
So there's the zero component now,
how do we do smoothness across time?
Well, it’s all about this correlation structure. So these latent Lambda star parameters that control the weights are actually modeled as a multi Gaussian process.
And I think the key thing to focus on here is that there’s this underlying correlation structure that tells us as two exposure time points get further apart.
This exponential of a negative number will get closer to zero. So correlation dies out as exposure time gets further apart.
now, as they get closer together, this correlation is gonna be higher.
And the main parameter that controls this level of correlation is this fee parameter.
And we actually put prior distributions on this to allow the data, to drive the inference, rather than like our view of what we expect this smoothness to look like.
So yeah, this is data driven kind of smoothness across exposure time.
All right, so now, so we’ve got the weights handled.
they have both properties that we care about. Now let’s talk about the mixture impact itself.
So this alpha recall tells us whether the mixture that we observe at time point T
or that we estimated exposure time point T
is actually relevant to the health outcome.
So we want, again,
we want this variable selection here
because we’ve noticed the problem with the
end points
that I described earlier.
So to do this, we decompose this effect into
two pieces,
a continuously varying piece.
And then this binary piece
that I mentioned earlier on in the talk,
the binary piece are just independent
but newly random variables.
But we imagine that if you’re in the critical
window set
at time one, then you may be in it at time
two
and may be more likely to be in at time three.
So there may be some sense of correlation
across exposure time here as well.
So while we model these things as independent,
the probabilities that underlie these zero
and one variables are actually smoothly varying
and correlated across time.
So again,
we use this kind of exponential correlation
structure.
We allow for cross correlation between the
continuous
and the binary piece.
Not important to get into here, you can kind of read back over. I can share a paper with you if you want to, or talk more about it offline. But essentially there's some cross correlation, there's correlation across time, but this allows for smoothness in the effects and the kind of the regression parameter effects. And these, both of these things come together to kind of define the critical window variable selection model.

To finish the model recall everything's in the base setting so really weekly informative prior distributions kind of standard prior distributions when possible, nothing too interesting here. So the model you may be looking at on this previous slide and thinking there's a lot of parameters floating around here. There's a lot of output that you're going to be estimating. So how do you make sense of this as a practitioner, someone who actually wants to know if a mixture is having an impact on your health? Well, luckily we still have relatively nice
and estimable kind of effects here, associations that we can talk about. So for example, for a change in the log odds for a one unit increase in each pollutant during a particular exposure period, this would be the quantity that you would make (indistinct). You would exponentiate this, and you would have like an odd ratio, for example, now recall for any model that includes interactions. The interpretation is always increasingly complicated because it matters where you start when you have interactions. So if you’re already at a high level, so the values themselves of exposure have to come into play, but nonetheless, you can still get nice quantities to estimate in the end. And if you’re only interested in what happens if pollutant A increased during exposure period T you can write down actually what looks like as well. So you can estimate both of these things relatively easily from our output, from our model. Alright, so we have a model that kind of checked all the boxes,
at least in my head when I was writing it down
and we can, I tested it, we can fit it,
it seems to work and that it’s converging
and it’s producing things that look reasonable,
but the simulation study really allows us to dig deeper
and say, is there anything, this it’s obviously new,
but is there anything beneficial to what we’re doing?
Or should we just be doing something simpler
that already exists?
So we wanted particularly to ask,
how does CWVS mix compared
to some of these existing approaches
for three different factors that we’re interested in?
So first identifying the true critical window set,
obviously probably the most important part
of critical window research here is like,
let’s get the critical window set right
when we’re estimating and identifying these parameters.
But obviously when you’re talking about mixtures,
we also care about these weights.
We want to know that the mixture profile we’re looking at
on a certain exposure period actually is,
reflective of the true mixture profile
that makes sense here.
So how well do we do at estimating these Lambdas and Lambdas Tilda parameters that describe the effects of main effects and interactions, and then finally, how well do we do it at estimating the magnitude of risk, these alpha T parameters. We wanna make sure we’re getting these right as well. And as a side issue, I guess, just more of our curiosity, how well does this variable selection process work for the weights that we’ve introduced? So now we need to think about what are competing methods in this space. There aren’t a lot of methods out there that aim to estimate critical windows for time bearing exposures and multiple pollutants. and the ones that are out there give different enough output. that’s hard to compare one model to the next, but here are three approaches that we kind of came up with. One is the most naive kind of give different enough output. as a practitioner with a new data set, this equal weights approach. So maybe just averaging all of the exposures for a person.
on a given exposure week and including that average and the interactions with the other exposure periods in a framework, a distributed lag framework. So yeah, this is called equal weights or EW. A PCA approach also makes sense. So let’s allow the data to determine the correct weights of these Lambdas, but let’s focus it only on the exposure period, only the exposure data. So at each exposure period, fit a PCA to the person specific exposures and generate these weights. That kind of describe the relative contribution of the different interactions and main effects in a mixture, and then weight the mixtures in that way and throw that weighted value into the distributed regression model. For all of these methods, we’re using the original CWVS, so that we’re comparable so that the method results are actually comparable across. And that the only thing that is changing essentially is how we define the weights. And then finally, the most sophisticated approach at that time was this lag, weighted quantal sum regression that we talked a little bit about.
where we applied weighted quantal sum regression separately to each exposure period, let that estimate the weights, create the little package of exposure, and then throw those packages into the regression model using CWVS. So once you have the weights, like once you condition on the weights and you know the weights, you basically have one exposure and that exposure is the package, the mixture package that you’ve made. So the model, the modeling becomes much simpler in that case.

So how did we go about to test these different methods? Well, we started very simply. So these represent the weights cross exposure period. In this case, I’m pretending like there’s only five weeks in the exposure set. So reality, I let that vary for each data set the length and the start time of the exposure window changed but for this case, we assumed it started at pregnancy week one and went to week five. And so in the simplest case, we had just assumed there was one pollutant at play.
and it stayed constant across the exposure period.

This is really simple.

One pollutant is driving the entire risk that we’re seeing.

In another setting, we assumed that there were two,

but there was no changes over time.

They were always static across time

and three, there were three that were coming into play

at four, four, and then five, five of them,

obviously as more come online

and become important players in the mixture.

The weights generally go down

because all of lots of these have to be non zero.

In setting B,

we wanted to allow for some variability

among the important pollutants.

So we still allow for the same important pollutants

to be important at each exposure period,

but we allowed their relative contribution
to change across time.

So early on in pregnancy, this one was important,

but then it’s contribution went down

and it was kind of surpassed by number two here

at pollutant two,

and then they can keep swapping in and out

across the exposure.
And in setting C it was complete chaos essentially
different pollutants could come online and then leave and become important
or not important go to zero.
We don’t anticipate this would ever,
or this would be the case,
but it would be nice to know if our model can somehow collapse and kind of accommodate this reckless,
this wild behavior, I guess.
this is something that kind of testing the extreme
of all these methods is what we were trying to do here.
So we’ll jump right into the results.
Just to give you a sense of what happened when we tested these models
with lots of simulated data sets,
continuously and kind of consistently
was able to get the critical windows set more accurately than the other methods,
which struggled kind of in varying degrees across these different settings,
in terms of estimating the weight parameters.
There’s a generally CWVS mix has a lower means scored error so it’s doing a better job of estimating these parameters,
as you would expect, like with equal weights,
if you assume each weight,
847 00:35:05.220 --> 00:35:08.580 each pollutant and interaction is playing
848 00:35:08.580 --> 00:35:09.930 an equal part in the story,
849 00:35:09.930 --> 00:35:12.540 you can be very bad off a lot of times,
850 00:35:12.540 --> 00:35:15.740 which is given, which is why these weights
851 00:35:15.740 --> 00:35:18.453 these values are so high for some of these
852 00:35:19.500 --> 00:35:20.560 And finally,
853 00:35:20.560 --> 00:35:23.220 with the estimation of the regression param-
854 00:35:23.220 --> 00:35:25.073 eters
855 00:35:26.270 --> 00:35:30.740 Generally, we’re seeing improved performance
856 00:35:30.740 --> 00:35:31.710 with CWVS mix,
857 00:35:31.710 --> 00:35:34.300 at least at the time when we first saw this
858 00:35:34.300 --> 00:35:37.870 is that the equal weights method does a pretty
859 00:35:37.870 --> 00:35:42.800 good job
860 00:35:42.800 --> 00:35:46.060 of estimating these risk magnitude parameters
861 00:35:46.060 --> 00:35:48.100 So if you tell me that every one of your pol-
862 00:35:48.100 --> 00:35:49.220 lutants
863 00:35:49.220 --> 00:35:52.610 are important,
864 00:35:52.610 --> 00:35:55.030 then it’s going to be hard to beat that some-
865 00:35:55.030 --> 00:35:56.650 thing
866 00:35:56.650 --> 00:35:58.210 As more pollutants become important,
867 00:35:58.210 --> 00:36:01.410 giving everything equal weight is not such a
868 00:36:01.410 --> 00:36:04.340 bad ideas,
869 00:36:04.340 --> 00:36:06.590 almost it’s just averaging away some of that
870 00:36:06.590 --> 00:36:08.600 error,
that’s really importantly, 'cause at the time this was the kind of the main method out there that aimed to do the same thing we were doing.

So in summary here with a simulation study, we did really well in critical in terms of accuracy, sorry, weight parameter estimation, and even in the risk magnitude parameter estimation.

So models that don’t have, that they don’t actually estimate weights are more efficient when the complexity or the number of important pollutants grow, and a little bit about the variable selection that we introduced with these latent variables. It appeared to do really well again, as the number of important pollutants was relatively small.

So if you have lots of pollutants that are important, and their interactions are important, it was hard for the variable selection process to kind of tease out when something’s included or excluded. It tended to just say everything was included. So something to keep in mind, I guess, as a limitation per perhaps of this approach. All right, so now onto the real data application that we had,
and this is part of a larger kind of climate change,
heat preterm birth study,
we collected lots of state specific data birth records
for all the way back to 1990 for maybe 12, 14 states.
And so this one was set in New Jersey,
but we focused on stillbirth given their really strong stillbirth data collection
kind of methodology that New Jersey was using.
So stillbirth the death or loss of a baby, at least 20 weeks of pregnancy affects about one in 160 births in the US.
There are some known maternal risk factors, black mother, 35 years age or more of age, low SES, smoking, et cetera.
And recent literature review meta analysis suggest that,
PM 2.5 CO2 and O3 are associated with increased risk.
This was really recent,
but that more studies are definitely needed.
There’s not a lot as in comparison to some of the other adverse birth outcomes, there’s not as much done with stillbirth, at least.
However a majority of these previous studies have focused on again, single pollutant approaches,
wide exposure periods like the entire relevant pregnancy.
before the delivery.
So there is a need for kind of multiple pollutant critical window methods in this setting. This is what made us think about developing this methodology, but also applying it in this case study. A little bit about the data we had access to. We had live birth and fetal death records from New Jersey from 2005 to 14. We included singletons with gestational age of at least 20 weeks, no birth defects, conception date in 25 to 2005 to 2013, we ran a case control analysis here where we five link live births were linked with each stillbirth matching only on race, ethnicity. And we actually ended up running these analyses separately for each group non-Hispanic black, non-Hispanic white and Hispanic. And in terms of what our exposures, we included weekly pollution exposures through gestational week 20 were included in this analysis. All right, a little bit about the pollutants. I mentioned we relied on a data fusion model that gave us kind of fine scale spatially and temporally estimates of 12 pollutants across New Jersey. Across the US actually, but focusing here on New Jersey.
So you can see the pollutants listed here and we linked each woman’s residence at delivery with the closest grid where data were available or the estimates and predictions were available and assigned weekly exposures across the first 20 weeks of gestation. I know there’s always a lot of pushback because we don’t have residential mobility, we don’t have sense of how often people move. And we know moving is differentialable by socioeconomic status for example, there are a lot of factors that influence moving during pregnancy, but if maybe this will make you feel somewhat better, but we did a study in 2019, the kind of assess the robustness of these critical window methods more generally to lots of different sources of error, including residential mobility and the results were actually very promising. I thought so the findings are robust generally to kind of this exposure misclassification or exposure error that’s introduced through mobility. All right, so in summary, I guess for the data we had around 1300 non-Hispanic black,
stillbirths in this time 928 Hispanic, and 1100 non-Hispanic white.

our covariates that we included were a year of conception,
season of conception to control for this kind of seasonality
and long term time trends and pollution exposure,
tobacco use indicator, age category, education.
We had this sex of the fetus
and to control for spatial kind of residual correlation.
We actually included latitude, longitude
of the residents had delivery and their interaction term
as a pre-screening
because we had 12 pollutants to work with.
We didn’t wanna introduce a lot of noise if possible,
into the new framework.
So we did a pre-run of the original critical window variable
selection on each pollutant individually,
as most analysis would do anyway,
and identified a subset across all
of the different data sets and by different data sets.
I mean the non-Hispanic black, non-Hispanic white, and Hispanic.
So all of the relevant and kind of significant exposures
that came up and during any exposure period
were included as a subset into this bigger framework.
And so in total, we had PM 2.5 sulfate, nitrogen oxide, ammonium, and nitrate that kind of made this pre-screening period into the final subset.

So here is some of the output that we thought was interesting. There’s a lot of output that can be shown as you already know. I guess now there’s weight at each exposure period, there’s regression parameters, there’s just a lot that can happen here and there’s interactions, there’s main effects, but first let’s focus on the first column here, and this is at least something we can hold onto that we understand from previous work in this space.

So what we can see for the non-Hispanic black population that we were working with in New Jersey during this time, that elevated exposure to some combination of these five pollutants during pregnancy week two, and then later on in the pregnancy, 16, 17, 20 actually led to increased odds. So these are odds or ratios being presented of excuse me, of stillbirth.

And so we can kind of take these in and say,
we get a sense of the critical windows that are identified.

We also get a sense of the variable selection component that I mentioned and in this case, they line up pretty perfectly. These are consistently in the model actually included in the Bayesian variable selection model, but also they’re when they are in the model they’re positive. So there this risk is in the right direction. So more pollution during these pregnancy windows, more risk of stillbirth in this population.

Now the question becomes, what are you talking about when you talk about the exposure? Like, what is the mixture that you’re talking about in week two, for example? Because we have five pollutants and their interactions floating around.

So focusing first, so now let’s move to the second column. This represents the interactions, this top part and the bottom part represents, I’m sorry, this is main effects. And the bottom part represents interactions. So you can see ammonium is playing a big role throughout.
which is dominated sharply by nitrogen oxides.

And then ammonium comes back into play here in terms of the interactions that are important, it looks like PM 2.5 and ammonium early on. And then later on it’s nitrogen oxides and ammonium kind of come into play. So a lot of this is noise. I did not show you the variable section component, but it probably would be nice to kind of gray these out if they’re not selected in the model. But a lot of these actually are selected in the model with our variable selection. So while these look to be non zero weights, some of them are actually exactly zero essentially because of the variable selection component. But there’s so much output, it’s hard to figure out what exactly to show in a digestible way. So this is where we landed. So, interesting results you get to see how the exposure kind of the mixture transitions across exposure time, you get to see what impact that has on the actual risk of the outcome that you’re talking about.
So a nice, I think coherent story can come, can be told, if you’re picturing your own analysis here, you get to talk about the risk overall to the mixture kind of combination or profile, but also then dig deeper into individual weeks and talk about which ones are important, which interactions are reporting for example. For the non-Hispanic white, there was very little indication that these pollutants were planning a role, I guess, in the kind of development of stillbirth or the risk of stillbirth in this population and for the Hispanic population, it looked like there potentially was some uptick here at the end, but nothing significantly jumped out either. And so at this point, it’s almost... You don’t start to investigate over interpret these white parameters, given that you’re not seeing anything here. So I kind of consider this to be noise essentially for the Hispanic and non-Hispanic white results for example.

So a little brief kind of wrapping up here, summary of our findings is that, for the non-Hispanic black data set and variable selection results PM 2.5 and its chemical constituents are primary drivers of risk.
And this was actually changing across exposure week. So driven in week two by a lot of interactions and kind of individual pieces. Week 16, mainly heavily driven by nitrogen oxides and then week 17, one or two pollutants and their interactions. So all the other interactions that are not listed here among the five variables were actually not significantly important here. So no nothing kind of nothing seen for the non-Hispanic white and Hispanic populations. And I guess in conclusion, we introduce CWVS mix with which combines smooth variable Bayesian variable selection in the weights and the regression parameters with interpretable weighted quantile sum shrinkage to identify critical windows, but also kind of understand and kind of dig deeper into the mixture itself. And importantly, at least from our perspective is that CWVS mix seemed to offer something that the existing methods didn’t, which so consistently outperforming these other methods for identifying the true critical window set, estimating weight parameters,
which is really important for interpreting the mixtures

and then estimating the risk magnitude parameters as well.

And our stillbirth results from New Jersey were in qualitative agreement with those in the literature,

in that PM 2.5 consistent signal across many studies

while developing kind of gaining new insights regarding the exposure timing in this particular study,

obviously more work is needed.

And so I guess before jumping to this,

we were working on extending this framework.

So I’m working with the group at Emory here

on extending this framework to allow the windows themselves

to vary by something like socioeconomic status

or race ethnicity, or other individual level factors.

So there’s this effect modification floating around now

plus the mixtures.

So it’s becoming a really big task to kind of do all of this

in a single framework,

but we’re trying to take baby steps, essentially.

We like where we’re at now, we think it works well,

it’s robust, it fits well
and can we extend it next to the questions that are being asked?

So again, if you're someone who is asking similar questions,
please, we can talk.

And I really like enjoy sitting down with collaborators and trying to figure out,
develop new methods that can answer the questions that they have.

But if you find that,
your setting can already be answered by some of these methods that I've discussed today
on my website and on my GitHub site,
I keep a lot of these packages that I've created with help documentation
and then you are always free to reach out to me as well.

But if you’re looking to do this original Gaussian process,
critical window estimation,
we have a package for that.

Howard Chang at Emory, go through my website again,
you’ll find this his survival version
of the model up there as well.

CWVS in this original form is there for download
the spatial version,
which hopefully we’re thinking about extending in
soon to account for something like oxidative potential of these pollutants that’s also there. And then the newly developed methodology is also there for download and for use as well. And this obviously could not have happened without collaborators, including Howard at Chang at Emory, Lauren at RTI did a lot of data management, Matthew Strickland, and Lindsey at University of Nevada Reno, and then James for providing the, or helping with the data fusion output as well. And here, this grant support here that I mentioned in extreme heat duration, and then data integration methods for environmental exposures. So yeah, please feel free to reach out if you have any questions. This work that I went over today is in press at Annals of Applied Statistics, is not on their website yet, but should be really soon. But I think there’s a version on archive if you’re interested or if you want the most up to date version. I actually think I sent it tomorrow who may have passed it out to the class, but yeah, definitely feel free to reach out if there are any questions or anything I can help with.
1199 00:50:26.855 --> 00:50:28.083 Yeah, that’s it.
1200 00:50:29.974 --> 00:50:30.807 <v->Thank you so much.</v>
1201 00:50:30.807 --> 00:50:33.057 (applause)
1202 00:50:35.310 --> 00:50:38.780 Our students were impressed with this
1203 00:50:38.780 --> 00:50:41.613 heavy quantitative focused lecture.
1204 00:50:43.925 --> 00:50:45.240 We already collected some questions
1205 00:50:45.240 --> 00:50:47.030 from our students already,
1206 00:50:47.030 --> 00:50:49.500 but for folks who are joining online,
1207 00:50:49.500 --> 00:50:51.240 if you do have questions,
1208 00:50:51.240 --> 00:50:53.850 please feel free to put in the chat box.
1209 00:50:53.850 --> 00:50:55.530 So the first question,
1210 00:50:55.530 --> 00:50:58.070 one of the students is observing that
1211 00:50:58.070 --> 00:50:61.440 in your study, you found the elevator risk
1212 00:51:01.440 --> 00:51:04.640 was found in week two of the pregnancy,
1213 00:51:04.640 --> 00:51:05.630 which is very early.
1214 00:51:05.630 --> 00:51:06.630 So perhaps many pregnant women are not
1215 00:51:10.820 --> 00:51:12.700 aware of the pregnancy at that time.
1216 00:51:12.700 --> 00:51:15.740 So in terms of the intervention
1217 00:51:15.740 --> 00:51:18.810 at this early stage of pregnancy,
1218 00:51:18.810 --> 00:51:22.440 what’s the kind of policy implications that
1219 00:51:22.440 --> 00:51:24.130 we’ll find?
1220 00:51:24.130 --> 00:51:27.110 Now that’s a really great point.</v>
1221 00:51:27.110 --> 00:51:29.390 And this is something we’ve tried to,
1222 00:51:29.390 --> 00:51:32.877 we haven’t figured out how to deal with either,
1223 00:51:34.930 --> 00:51:37.510 but has we’ve run into a number of interesting
1224 00:51:37.510 --> 00:51:38.490 results we’ve seen early in the pregnancy.
1225 00:51:38.490 --> 00:51:41.700 we’ve seen protective effects at some points
for like PM 2.5 exposure and pre-term pregnancy
very early on in the pregnancy.
And we believe it could be due to the exactly
what we’re talking about.
People who don’t actually know they’re pregnant
at that point.
And so miscarriage is an issue
that isn’t well kind of documented by a lot
of these states.
There could be just fetal loss in general,
that we’re not capturing in the birth records.
And so there’s this population
that we’re not even including in a lot of our
analysis
that are lurking around
and kind of could be biasing
some of these early week results.
In terms of policy implications
it’s a really good question.
I don’t know other than if I guess it really,
if you’re trying to get pregnant,
if you know you’re on that, in that stage,
I mean, maybe it’s helpful for you,
but if you’re someone who doesn’t know
unanticipated
there’s only so much that can go into outside
of just cleaner air altogether.
Which is something everyone can kind of
agree on.
But I think it may only affect a subset of
people
who are either attempting to get pregnant or kind of really regimented and like, know their schedule for example. But there’s this whole other issue about people who aren’t in our data set. That’s a really great point and we have not figured out how to solve that yet. Yet, tough question. Y et, tough question. Yet, tough question.

We do have another question from actually two students read this. They really appreciate your talk about this new metrics. And we realize this is the package.

Our package is available from your GitHub website.

So anyone who’s interested in applying that you can download the app package and run, but the students are wondering like beyond this time wearing air pollution mixtures a lot other mixtures in terms of (indistinct) like temperature, green space, other things. So how does your approach this the CWVS mix apply to a broader setting of environment exposures? I think, my push, and if you read the paper, I think, my push, and if you read the paper, you’ll notice that I really push for people to think about that in their own setting. Cause I think it’s generally applicable to any,
it doesn’t have to be a pregnancy outcome. It doesn’t have to be air pollution. What it does have to be is consistently measured across some exposure period. I’ll often get questions that, I have two time periods measured, in the first trimester and then in the third trimester. can I fit your methodology? Well, we need more fine grained exposure information. That’s consistent across the individuals in order to estimate these critical windows. So I think the only barrier for entry is that you have consistently estimated kind of exposures for the population of interest. It doesn’t matter so much what the exposure is now. I say that, but if you’re bringing binary exposures and you have limit of detection issues, there are obviously some issues that will need to be sorted out, but the framework itself should work really well. The other covariate is, you’ll notice that a lot of my work has been focused on pregnancy outcomes and that’s because the exposure period is so well defined if you’re working with something like cancer for example,
well, how far do you extend back in time, the exposures like how you could go years and years back.

So there’s this cumulative idea as well. That’s really hard to understand and these distributed lag models are great. As long as you can a priority tell me what the relevant exposure period is. I can tell you if any of the interior parts of that exposure period are important, but if you’re telling me you don’t know when the exposure period potentially started or it’s a completely different conversation. So your outcome has to have, or preferably would have some type of relevant exposure period. It’s actually even better for something like cardiac heart defects, which we know the heart forms between like weeks three and eight of pregnancy. So you can really focus in on something like daily or even sub daily if you had that type of exposure information. So yeah, those are the two, generally it should work, but just make sure you have a good sense of the exposure period.

Very good point, thanks Josh. And we do have one comment from our on artist.
So I read Dr. Warren could you please share your thought on applying the critical window analysis? (mutters) <v ->Sorry, with what?</v> That’s a really great point. So over, so I’m actually on sabbatical right now, which is why I couldn’t be there in person with you guys, but over the sabbatical I’ve developed the framework and the code to account for binary outcomes, continuous outcomes and count outcomes as well. Luckily if you’ve taken my (indistinct) course or you’re gonna take it next fall, you’ll see how all of these connect and lend themselves really nicely to kind of full conditional distribution updates that make the model fitting process really kind of slick and nice. So you can have a negative binomial regression, for example, that can do the same thing. You just have count out outcome data, if you have a continuous measure, for example, so I’m really aiming this. I hope this method doesn’t just pop up
and then disappear, I want people to use it, I want it to be useful. And so that’s why I’m trying to extend it and trying to get people to use it in different contexts. So, yeah, definitely I love those types of questions.

Because we actually have another (speech distorted) So we have to end early and we do have a lot of students questions and I’m sure contact you for just once. So thanks again, Josh for wonderful talk. No, yeah thanks for being here.