And welcome.

Today, it's my pleasure to introduce Professor Abhi Datta from Johns Hopkins University in Baltimore, Maryland.

Professor Datta earned his BS and MS from the Indian Statistical Institute in 2008 and 2010 respectively, and PhD from the University of Minnesota in 2016.

In addition to being a well-cited researcher with one publication that's almost 600 citations, he's also a award-winning educator, having repeatedly won an excellence in teaching award from his institution.

My talk is about improving cause-specific mortality data in low and middle-income countries where the main tool to collect data is something called verbal autopsies.

And the way I do it
is using a statistical approach called generalized Bayes.

If you have not heard of verbal autopsies or generalized Bayes, I can tell you that I hadn’t heard of either of those things when I started working on the project, so don’t worry about that, I try to give an introduction. ’Cause I mostly work on a spatial and spatial temporal data and this was a project that came along, which is very different from what I used to work on. But over the years, there’s been a nice body of work developed in this project. So this is a joint work with many different institutes and collaborators. The top row is the Hopkins bio stats team, which included my former students, Jacob Fiksel and Brian Gilbert, and my current postdoc, Sandi, my colleague, Scott Zeger, and I lead the bio stats part of the team. Agbessi is the PI of the project in Mozambique. And there are a lot of colleagues from the International Health Department that helped to collaborate. And then Li is the PI of a new project.
who we’re going to apply our methodology for producing mortality estimates for the WHO. So we’re collaborating with Li there as well. And then a couple of people outside Hopkins, Dianna at CDC and Emory University,  as the director of the CHAMPS project. Ivalda in the government body at Mozambique has been currently doing the work in Mozambique. So this is funded by three grants from the Gates Foundation. The first one was the grant that kind of started things. And then we have a grant that is kind of developing more on the method side of the world. So, many low and middle-income countries often lack high-quality data on causes of death. Often for most deaths, there is no sort of medical certification or like an autopsy done. And without kind of high-quality data on what people are dying of, it’s kind of hard to estimate the disease burden in these countries. And specifically, the quantity of interest is the cause-specific mortality fraction, which is basically the percentage of deaths in a age group that can be attributable to a given cause. So cause-specific mortality fractions are key pieces of information.
in determining the global burden of disease, which in turn dictates sovereign policy, as well as like resource allocations for programs operating in this country. So verbal autopsy is an alternate way to count deaths and attribute causes without actually doing a clinical autopsy. So verbal autopsy is basically a sort of a systematic interview of the household members of the deceased. So the government or the program has a set of field workers who go out and go from household to household and ask if anyone died in their household within the last several months. And if they died, what were the symptoms? The set of questions they ask is not standardized by the WHO. The 2016 version has around 200 to 350 questions, depending on the age group. There are separate sections of the questionnaire for neonates, children deaths and adult deaths.
And if you’re interested in more information about verbal autopsy, there’s a page in WHO about it. So a verbal autopsy, of course, doesn’t give you a cause of death, it just gives you a bunch of yes-no responses to various questions related to the symptoms. So a verbal autopsy is basically a survey questionnaire. So you can pass that survey through a computer software and that can give a predictive cause of death. And so there are a bunch of different computer software available. InSilicoVA, developed by Tyler McCormick, Richard Li was a postdoc here, published in “JASA” in 2016, is one of the, I think, most statistically-principled approaches to do it. But there are other approaches and then you can, this is basically a classification problem. So you’re basically given your data on symptoms, you’re kind of classifying the cause of death as one of several causes. So you can use standard classifiers and machine learning approaches as well. OpenVA is an excellent resource to learn about verbal autopsies. Again, openVA is, I think Richard is one of the maintainers.
and creators of openVA.

So the COMSA project in Mozambique, one of the main goals was to generate this cause-specific mortality fractions for children’s and under,

for neonates and under-five children

And the data that we collected was a large dataset of vocal autopsy record for different households that were surveyed and that was a map of Mozambique and the green region show where the data was collected as part of the COMSA project.

So in statistical terms, the data just has the symptoms, it doesn’t have the true cause of death, so we call it the unlabeled data.

So how to go from an unlabeled data to the labeling and then estimate these cause fractions.

This is the standard procedure that is typically done.

And this is what we were supposed to do as well, which is simply take each record, pass it through the computer software and get a cause of death.

And once you get a cause of death, then you can sort of simply aggregate.
So in the story example, three out of the six cases were assigned to be from HIV. And so the cause-specific mortality fraction for HIV would be 50% and similar for malaria and sepsis and so on. So that’s the basic template of how to get a cause-specific mortality fractions from verbal autopsies. The question is can we trust this estimates? Because these are not true causes of death as determined by a doctor or by a clinical procedure. These are cause of death predicted by an algorithm. Based on just surveying the household members of the deceased. So turns out machine learning has a name it’s called quantification learning, which is basically estimating population prevalence using predicted levels instead of true levels and the predictions are coming from a classifier. And so there has been some work in quantification learning and in the machine learning literature.
183 00:08:23.760 --> 00:08:26.760 using predicted cause of death data from verbal autopsy
184 00:08:26.760 --> 00:08:28.953 is an example of quantification learning.
185 00:08:30.690 --> 00:08:34.620 So just a sort of an overview of terms that we’ll be using
186 00:08:34.620 --> 00:08:36.570 and the corresponding statistical notation.
187 00:08:36.570 --> 00:08:41.570 So our true cause of death is $y$ which we do not observe.
188 00:08:41.760 --> 00:08:43.310 We want to estimate the probability
189 00:08:43.310 --> 00:08:45.330 of population prevalence of $y$,
190 00:08:45.330 --> 00:08:47.433 so $y$ is a categorical variable.
191 00:08:48.510 --> 00:08:50.640 And so probability of $y$ or $p$
192 00:08:50.640 --> 00:08:52.770 is our cause-specific mortality fraction,
193 00:08:52.770 --> 00:08:54.780 which is the estimand.
194 00:08:54.780 --> 00:08:57.390 We observed the verbal autopsy, which is a,
195 00:08:57.390 --> 00:09:00.180 think of this as a high dimensional
196 00:09:00.180 --> 00:09:01.740 or a long list of yes-no answers
197 00:09:01.740 --> 00:09:05.850 to the verbal autopsy questions, so that is $x$,
198 00:09:05.850 --> 00:09:08.010 and this $x$ is passed through a software
199 00:09:08.010 --> 00:09:11.913 to give a predicted level, which is $a$ of $x$ or simply $a$.
200 00:09:17.070 --> 00:09:21.060 So what we have in the COMSA project
201 00:09:21.060 --> 00:09:24.600 is simply an unlabeled dataset
202 00:09:24.600 --> 00:09:28.350 which uses these verbal autopsy responses,
203 00:09:28.350 --> 00:09:33.350 pass it through a software and get the predicted levels.
204 00:09:33.510 --> 00:09:36.870 We do not observe the true levels, $y$,
205 00:09:36.870 --> 00:09:40.170 we may or may not retain the verbal autopsy responses
206 00:09:40.170 --> 00:09:41.790 because those are identifiable data
207 00:09:41.790 --> 00:09:43.290 and those are often not released,
208 00:09:43.290 --> 00:09:46.500 so often, just the predicted cause of that is available.
So even these covariates, $x$, may or may not be available.

And then we are interested in estimating the probability that $y$ belongs to one of the $C$ many cause categories,

so that’s a quantity of interest.

For some reason, there is a conditional sign missing there.

But you can use the law of total probability to write the probability of the predicted cause of death,

which is the $a$,

probability of $a$ as a sum of our probability of a given $y$
times probability of $y$.

So there’s a conditional sign missing here,

I don’t know what’s going on here.

But the COMSA data,

we only get information on the left-hand side, right?

And we want to input upon the quantity of $y$
which would be the true CSMFs.

So there is only one known quantity
with which you can estimate the left-hand side.

There are two unknown quantities on the right-hand side.

So without making assumptions, you cannot really identify
probability of $y$, right?

So any quantification learning methods
need to either estimate those conditional probabilities, probability of a given $y$, or make some assumptions on it. So again, all the conditional signs are missing. The one of the most common approaches, and this is what is used in the verbal autopsy world is called classify and count, which is you simply predict the cause of death and then aggregate. So you’re simply claiming that probability of $a$ is same as probability of $y$ which is equivalent to claiming this misclassification rate matrix is an identity matrix, right? Because you’re saying that the left hand quantity is the same as the rightmost quantity, which would be true if there is no misclassification by the algorithm and if the predicted cause of death is always the true cause of death. And that’s what is typically done in this cause-specific mortality fraction estimates. But it’s a very strong assumption, right? Because it says assuming perfect sensitivity and specificity of the algorithm. So let’s look at how perfect the algorithms are. So these are two algorithms,
Tariff and InSilicoVA, PHMRC data is a benchmark dataset from four countries that has both the verbal autopsy data as well as a gold standard cause of death diagnosis.

And you can see the accuracies of either method is around 30%, so they’re far from being fully accurate. So there is large misclassification rates of these algorithms and if you don’t kind of adjust for these misclassifications, this is burden estimates of the cause-specific mortality fractions you get are likely going to be very biased. So this is where the CHAMPS project comes into play. So the CHAMPS is an ongoing project in like seven or eight countries including Mozambique, which is collecting data on both verbal autopsy and a more comprehensive cause of death procedure called minimally invasive tissue sampling. It basically takes a sample of your tissue of the deceased person and then runs a bunch of pathological tests and imaging analysis and then gives a cause of death. And the MITS cause of death assignments have been shown to be quite accurate when you compare...
to like a full diagnostic autopsy.

So MITS is being done in a bunch of different countries including Mozambique. And for the cases where MITS is being done, the verbal autopsies are also collected.

So what you get from this CHAMPS data is a labeled or paired dataset where you have both the verbal autopsy as well as the MITS cause of death and you can pass the verbal autopsy to the software to get the verbal autopsy predicted cause of death.

And then you can cross tabulate the two and get an estimate of the misclassification rates, right?

Like you can say like, "Oh okay, so there are 10 cases that the MITS cause of death was HIV, out of those 10 cases, seven of them were correctly assigned to HIV by verbal autopsy.

So then the sensitivity would be 70% and the false positive would be 30%, so on.”

So this is the broad idea of the methodology.

So for the COMSA data, which is the unpaired data, you get only the verbal autopsy record so you can get an estimate of the predicted cause of deaths from the verbal autopsy.

From the CHAMPS data, which is the paired data,
you can get an estimate of the misclassification rates.

And then the only unknown is then the probabilities of the cause of death.

if you were able to do the MITS autopsy for every death.

So then this is an equation with two knowns and one unknown.

and you can solve for it and get the calibrating message.

So that’s the broad idea and we do it in a model-based way.

So here’s the formal model.

So for the CHAMPS dataset with the unlabeled data or the U,

we have the predicted labels, ar,

and then for the,

we have both the predicted labels from verbal autopsy, ar,

as well as the MITS determine labels, yr.

And our quantity of interest is the probabilities of yr

belonging to the different causes.

There’s a conditional sign missing here.

But if the conditional probabilities

are denoted by Mij, which is if the MITS cause is i,

what is the probability that the via predicted cause is j?

Then you can use a law of total probability

to write down the marginal distribution.
So that would be in terms of the misclassification rates and the marginal cause distribution of the MITS-COD. So that’s the whole idea.

So you can write this in terms of a matrix vector notation as probability of a as $M^\top p$ where $M$ is the misclassification rate matrix, $p$ is the unknown quantity of interest, which is probability that the cause of death is coming from an unknown cause.

The data model is very simple, but the unlabeled data, it follows multinomial with this probability. And then for the label data, this is $ar$ given $yr$ equals to $i$, it follows multinomial with the $i$ throughout the misclassification matrix.

So if the MITS-COD is $i$, the misclassification rates are given by the $i$ throughout the misclassification matrix, so it’s multinomial with that probability. And then we’ve put priors on $M$ and $p$. Then we can get estimates of both $M$ and $p$.

$M$ is a nuisance parameter, $p$ is the parameter of interest. Just to carefully go over what are the assumptions here.
The main assumption is that the misclassification rates of verbal autopsy given MITS are the same in your label data as they would be in your unlabeled data. This is not verifiable because we don’t have any true cause of death in the unlabeled data, so it’s an assumption. Given that the verbal autopsy is a function of your symptoms, the assumption is essentially that given a true cause, the probability of the symptoms are going to be same in your unlabeled dataset as in your labeled dataset. And it’s a reasonable assumption as if you have a cause of death, it’s likely that you have certain symptoms will appear and some certain symptoms will not appear. And that is true regardless of whether the data is coming from the labeled set or the unlabeled set. We do not assume that the marginal distribution of the CHAMPS data of the causes in the label data is representative of the population because they are not, because the CHAMPS state, so the CHAMPS project is done at specific hospitals in the country and distribution of causes in hospitals.
are typically not same as distribution of causes in the community. And we are interested in the cause distribution in the population. So there is no assumption that the marginal distribution of y in the label data is same as the marginal distribution of y in unlabeled data, which is our quantity of interest. We never model y in the label data. So we only model the conditional and the assumption is the condition of misclassification rates are transportable from the labeled to the unlabeled side. So that’s the main idea. And this was the first work we did, we just used this top cause prediction. But many of these algorithms are actually probabilistic in nature in the sense that if you look at their outputs, they won’t give a single cause of death, but they will give scores to each cause. So for example, this would be a typical output of an algorithm for like say 6%. So for the first person, it will say 70% HIV, 20% malaria, 10% sepsis and so on. And the standard procedure is to take the top cause,
so for the first person, it would be HIV,
for the second person, it will be malaria and so on.

So that’s how you get a single cause from a probabilistic prediction.
So that essentially ignores sort of the scores assigned to the second most likely cause,
the third most likely cause and so on. And you ignore those, you can end up with a biased estimate.

So you can see these are the CSMF estimates using the top cause,
these are the CSM estimates using the exact scores that are assigned and those are different, right?

So when we kind of change this probabilistic output to a single cause output, we discard information.
So we wanted to extend the work to kind of use the full set of scores and the set of scores can be thought of as a compositional data in the sense that the scores sum up to one because it assigns 100% probability across all causes and then they’re each non-negative.
The issue is that for the categorical data, our model is based on multinomial distribution.
And then for compositional data, the models are typically like Dirichlet.
or log ratio based models, which are very different from the multinomial distribution.

So if we have some cases for which we have categorical output, for some, we have compositional output, this would lead to different models for different parts of the dataset. These Dirichlet or log-ratio models also do not allow zeros in the data. So if you have zeros or ones in the composition, they don’t allow that. And then there are very specific models about the data which are subjective model and specification. So the data distribution does not look like a Dirichlet assuming a Dirichlet layer would lead to kind of wrong results.

How do we extend the multinomial framework we had for categorical data to compositional data? Again, there would be a conditional sign here. But the basic assumption that we had for the multinomial case was probability of a given $y$ is the $i$ throughout misclassification matrix, right? And for categorical data, a probability statement is same as an expectation statement, right? So we can equivalently write this as expectation of a given $y$. 

is the i throughout the M.
The advantage of the expectation statement is that it’s more generally applicable. It will not be just for categorical data, right? So for categorical data, there’s a equivalent. For other data types, this statement can be valid even though the previous statement may not be applicable.
So we kind of write this as our model for the compositional data and we make no other assumptions about this distribution. So only a first moment conditional expectation statement without any full distributional specification. So what do we do? So we have expectation of a given y is the i throughout the misclassification matrix. We can use something called the Kullback Leibler Divergence or the cross entropy loss between a and its model expectation. So these are all the conditional signs are missing here. So basically a is the data we observe, this is the modeled expectation, which is basically the i through of the misclassification matrix and we use the cross entropy loss, the Kullback Leibler loss between the two. What’s the advantage?
So first of all, the Kullback Leibler loss allows zeroes in the composition. So it is well-defined even if you have zeroes or ones. If you take the negative loss and exponentiate it, it’s exactly the multinomial likelihood. So if your data is indeed multinomial, you get back your likelihood that you’re using for your single class model. But if your data is not multinomial, you get a pseudo likelihood that you can work with. If you can take the derivative of the loss function and take the expectation under the two parameter, you’ll see that it’s a valid score function in the sense that you get an unbiased estimating equation for your misclassification rate matrix, M, based on just the first moment as option. And then similarly, you can do the same thing for the unlabeled data. The probability statement becomes expectation statement and we have the Kullback Leibler loss. This is an unbiased estimated equation for both M and p. And again, if the data is truly multinomial and not compositional,
this becomes exactly the multinomial likelihood. If the data is compositional,
it becomes a pseudo likelihood. Okay, so how do we do Bayes analysis
with pseudo likelihoods?
So this is where this idea of generalized Bayes
model-free Bayesian inference comes in and there have been parallel developments
in both computer science, econometrics and statistics
without much communication among the three fields
for the last 30, 40 years. Basically, if you’re given a loss function
without a given like a full likelihood for the data,
you can take negative of that loss function multiplied by some tuning parameter, alpha,
exponentiate it and treat it as a pseudo likelihood
and apply your priors
and then your posterior is going to be proportional to this
as long as the normalization constant exists.
And there has been a lot of work that has shown
that this is a valid posterior,
it is a generalization of the Bayesian posterior,
like if this is an actual likelihood,
is the Bayesian posterior,
but if it’s not a actual likelihood,
this has been shown that it basically minimizes
the Bayes risk for that loss function. It has nice asymptotic properties shown by Victor Chernozhukov in this paper and then in this JSS paper in 2016 I think it showed that if you’re given a loss function and a prior, this is the only coherent way you can get a posterior. So there’s now been a lot of work and it’s been called by different names like Gibbs posteriors, pseudo posterior, Laplace-type estimators and quasi-Bayesian estimators along with generalized Bayes. So for our case, we have the pseudo likelihood for the label data. We have the pseudo likelihood for the unlabeled data. We put priors. If all of our data were categorical, this reduces to that multinomial model we had for the categorical data. But if some of the data is compositional, then this becomes generalized Bayes, so we call it generalized Bayes quantification learning. It allows sparsity of the outputs in the sense that if some of the data have zeroes and ones in them, this is well-defined. It’s the same pseudo likelihood for categorical compositional predictions.
And then it also allows a nice Gibbs sample using conjugacy. One final sort of data aspect we had was that this minimal tissue sampling was also sometimes inconclusive in the sense that they gave two causes. Like often, they were ambiguous between HIV and tuberculosis and they would give one as the immediate cause and one as the underlying cause. So sometimes, even the true cause of death is compositional, your predicted cause of death is compositional, your true cause of death is also compositional and we call it like b, which represents the belief. And you can show that if you’re only given b instead of a single cause of death, your conditional expectation becomes $M^\text{transpose} b$ instead of the i through of the M matrix. And you can do the same thing using the compositional true cause of death instead of the actual true cause of death. And all the conditional signs are missing here but you can just formulate the Kullback Leibler likelihood to generate pseudo likelihood. So this kind of give rise to a digression where we kind of looked at this is basically
Your true cause of death is a compositional covariate and your predicted cause of death is a compositional output. So we kind of looked at regression of a compositional outcome on compositional predictors. This was kind of an offshoot paper where we just developed this piece. Most of the work has been done using Dirichlet models or log ratio transformations. This was a different approach to that in the sense that it’s both transformation free and it doesn’t specify a whole distribution like the Dirichlet. It just uses a first moment as option. We have an R-package to do a regression on composition, called codalm. And we have an R-package to do a regression on composition, called codalm. But going back to the verbal autopsy work, we have the loss functions for the labeled and unlabeled data, we do the negative pseudo likelihoods, put priors on the parameters and we get posterior inference. One last extension of the methodology was that there are multiple different verbal autopsy algorithms and there are papers.
where every new algorithm comes out and they say they’re better than all the previous algorithms. And in practice, you never know which is the best algorithm. So we developed an ensemble method that takes in predictions from multiple algorithms, estimates classifier algorithm-specific misclassification rates and then they’re connected to the unknown estimand. So we can show that it gives more weight to the more accurate algorithm in a data-driven way. And then you’re not kind of, you don’t have to make the choice of which is the best algorithm in advance. If you have multiple candidates, you can use multiple algorithms together. So we looked at some theoretical properties of the method. We have two log functions, one for the label data, one for the unlabeled data. The label data doesn’t even feature the estimand, which is p. so it will, on its own, it cannot identify p. The unlabeled data only uses p through this quantity, M transpose p. So again, for different combinations of M and p, as long as this product is the same,
it will never be able to identify \( p \) on its own. So each loss function on its own cannot identify through parameters. But using both the loss functions together, you can identify the estimand, \( T \), and we were able to show that posterior has nice properties in terms of asymptotic normality and well calibrated interval estimate and near parametric concentration rates. And the theory also extends to the ensemble method and we use some approximations and we give sampler and theory holds for that. Some empirical validations, since we’re estimating a probability vector, the common metric that is used is called this chance-corrected normalized absolute accuracy, which is basically a scaled L1 error, centered by the L1 error you would get if you had predicted the cause of death randomly. So this is the error if you predict randomly and then we look at how much improvement we get over random predictions. So this is an illustration of what happens if the data is not Dirichlet and you use Dirichlet distribution. So on the left-hand side,
the data is generated from Dirichlet and we use both our method and the Dirichlet-based model and they both do well.

On the right-hand side, the data is from an overdispersed Dirichlet and we use the Dirichlet in our model. And because our model doesn’t specify a distribution, it just uses a first moment specification, it’s much robust and has much higher accuracy than for the Dirichlet which becomes misspecified.

And then we also did a bunch of evaluations using the PHMRC data. So what we did was we trained the classifiers on three of the countries leaving one country out and then used a slice of data from that left out country to estimate the misclassification rates, and then we apply our method.

The green one is our method and the blue one is sort of the uncalibrated one, the red one is the one that is calibrated using the training data. So you can see that our method does better than both of them and the higher the sample size we use from the left out country of interest.
to estimate the misclassifications, the more accurate it is.
And also one interesting aspect was that we looked at calibration
using individual algorithms and the calibration
using the ensemble one.
And more often than not, the ensemble one, which is the orange one,
tends to perform similar to the best performing algorithm,
and the best performing algorithm can be very different
across different countries.
For example, in Mexico,
InSilicoVA is one of the best performing algorithms,
but in Tanzania, InSilicoVA was doing very poorly
and then InterVA was one of the better performing algorithms.
So the ensemble always tend to give more weights
to more accurate algorithms.
So this is an overview of what we did for Mozambique.
We had the unlabeled data with only verbal autopsies.
We’ve passed it through two algorithms,
InSilicoVA and Expert VA, to get the uncalibrated estimates.
Then we had the label data with the MITS cause of death
with which we estimated the misclassifications
of those two algorithms and then we combine them in the ensemble method and getting calibrated estimates. Some results from Mozambique. We have two age groups, neonatal deaths, first four weeks, and children that’s under five years. Two algorithms, seven causes of death for children, five causes of death for neonates. I’m going to just show the neonatal results here. So these are the misclassification matrices for neonates. And ideally, you would want the matrices to have large numbers on the diagonals because those are the correct matches and small numbers on the off diagonals. But you don’t see that, you see quite a bit of large numbers on the off diagonals. One thing that stands out is that if you look at prematurity, it has a very high sensitivity, close to 90%, which means that if the true cause is prematurity, the verbal autopsy correctly diagnoses it. But then it also has high false positives in the sense that if the true cause is infection, it is assigned as prematurity. If the true cause is intrapartum related events,
almost 30% of time, it’s assigned to be prematurity and so on. So it tends to over count a lot of deaths from different causes as prematurity. So what would be the result after calibration is that the percentage of prematurity comes down. So this is the uncalibrated estimate of prematurity. This is the calibrated estimate of prematurity. You can see that it comes down because we can see in the data that there is a lot of over counting of prematurity deaths. So after calibration, it tends to come down quite a bit. And also, we looked at the model estimated sensitivities using both the single cause and the compositional cause of death, you’ll always get a higher match because it kind of uses information for multiple causes and stuff just considering the top cause. And so it generally leads to better matching between the verbal autopsy and the minimal tissue sampling. Some ongoing work. So when we did this for Mozambique, there was very little amount of payer data.
So even though the data was for seven countries, we kind of merged them together and estimated the misclassification rates. Now we have more data coming in for those countries so we have a chance to assess whether the misclassification rates vary by country because if they do, we should model the misclassification rates in a way that’s specific to each country. So these are the misclassification rates now resolved by country. So there are six countries, Bangladesh, Ethiopia, Kenya, Mali, Mozambique and Sierra Leone. You can see the estimates. These are the empirical estimates and the confidence intervals for each country. And the horizontal black line is what the pooled estimate looks like. So you can see that there is for some causes like here, there is not a variability across countries. But then for some other cause payers like say here, there’s quite a bit of variability across countries. And so now that we are getting more data, the next step for the project is to estimate country-specific misclassification rates.
The issue however is that even with more data, there is, I think, around 600 cases here for six countries, which is approximately 100 case per country. And there are 25 cells of the misclassification matrix. So that's like four cases per cell, so that's clearly not enough to do separate country specific models. So we'd have to kind of do a sort of a borrowing of information both across the rows and columns of the matrix. So we do first, we kind of borrow information across the rows and columns of the matrix. And to do this, we start with a structured misclassification matrix using two basic mechanisms. So we say that a classifier operates using two mechanisms, for a given cause, it can either match that cause and we call that an intrinsic accuracy and that matching probability will be different for different causes, so there are three causes here, and you can see...
that the matching probability can be different.

If it doesn’t match the true cause, then it randomly distributes its prediction to the other causes and that random distribution will also have some weights, and those we call the systematic bias or the pool of the classifier. So if it’s not matching, we saw that it’ll often assign a cause to prematurity regardless of what the true cause is.

So that’s kind of the basis for this model. And if you have this model, we kind of rearrange these three bars here and then we put in the circle from there. And these will give you the misclassification priorities.

So we can write each of the misclassification probabilities in terms of just these six parameters and we can do the same for the green cause and for the blue cause. And so basically, these are the nine misclassification rates written in terms of the six parameters.

So this is not that much of a dimension reduction if there are three causes, but if there are in general C causes, this model for misclassification matrix will only have 2C - 1 parameters as opposed to C squared parameters.
So in practice, we use seven causes for children and five causes for neonates, so this leads to a lot of dimension reduction. And one of the justification is that if this model is true then the misclassification into different causes, the odds of misclassification into two causes, \( j \) and \( k \), will not depend on what the true cause is. And we do see that in the data.

So these are different cause payers, \( j \) and \( k \), and these are the odds for what the true cause is. So we are plotting the misclassification rates, \( \frac{m_{ij}}{m_{ik}} \). So this is \( j \) and \( k \) and the colors here give you \( i \). So you do see that they do not vary for different choices of \( i \), it only is specific to \( j \) and \( k \), and that’s an equivalent characterization of that systematic preference. So we allow some diversion or shrinkage towards it and there’s a tuning parameter.
So then we get the homogeneous model and then we have a diversion from the homogeneous model to get country specific model. That’s the broad idea, I won’t go into the modeling details. These are the predictions using the country specific model. I won’t go into details here, but there are many cases, for example, take it here, star is the empirical rate, angle is the heterogeneous model. And you can see it does much better than the horizontal line, which is the homogeneous model. And we do see it throughout the classification rates. These are the estimates for Bangladesh. So the red density is the pooled estimate of the homogeneous estimate. The blue density is the Bangladesh specific estimate. The dotted vertical line is the empirical estimate for Bangladesh and the solid vertical line is the pooled empirical estimate. So you can see that as we get more and more data from Bangladesh, the country specific estimate moves away from the pooled estimate towards the country specific estimate.
So that’s basically the hope is going forward, we will have much more data within each country and we’ll have estimates that are much closer to the dotted lines than the solid lines. So that’s the summary.

So in general, these cause of death classifiers are super inaccurate. So we need to calibrate for that and we have limited data to estimate their inaccuracy, so we calibrate them innovation way. The methods give probabilistic cause of death instead of categorical cause of death. So we develop a generalized Bayes approach that is equivalent to a multinomial model if the data is categorical. But if it’s not categorical, it becomes a pseudo likelihood Bayesian approach for compositional data and that allows zeroes and ones in the data and is not kind of dependent on the model specification. And then it kind of led to this independent development of the composition on composition regression. Some papers and software.

So the single cause paper was the first one, then we extend it to compositional data and develop the theory for it. The package for calibration is available on GitHub.
and then the composition on composition regression

were the separate piece

and we have the coda linear model package for it on CRAN.

And then we use this approach to produce calibration estimates for neonate and children deaths in Mozambique which were published in the last three papers. Thank you.

Questions? Yes.

So I just had a quick question 'cause you were saying:

the model basically looks at the symptoms that’ll be able to predict which it would be.

Does it also factor in what diseases and stuff are most common in those areas or does it kind of just-

Oh, very good question.

It does factor it in but in a very crude way in the sense that the models have some settings called like high malaria, low malaria or high HIV, low HIV.

So depending on which country you’re running it, you will set the setting to like high HIV country or low HIV country, the same for malaria, but it doesn’t do anything beyond that.

so only at a very close level.

Causes of death or.

So the ICD-10 classification
will have around 30 plus causes of death for children’s and neonates,
I think much more for adults. There are no MITS for adults. MITS was only done for children’s and neonates, only now adult MITS are being started, but we have to kind of group them into broader categories because if you have 30 causes, your misclassification matrix will be 30 times 30. So we don’t have the data to do estimation at that fine resolution. So we group them into broader categories. So seven for children, five for new neonates. Is one of the categories, I have no idea. It is totally unknown. And if so, is that different from the uniform distribution across causes of death? That would be the uniform distribution. There is no category which is, I have no idea, but it’ll be probably reflected in a score that is very flat across the causes. If you think there are seven causes of death, and I’m working with the same dataset and I think there are 100 causes of death, will there be substantial differences in our marginal
estimates of probability? Because our uniform posteriors place such different amounts of mass across the say 30 versus 100 causes of death. <v>Yes, there will be differences</v> and even when we are aggregating from the 30 causes 00:46:58.380 -- 00:47:01.860 to seven causes, the assumption is that within each category the misclassification rates are homogeneous within the finer category. So that is an assumption that we're working with. So definitely, there will be differences. <v>Thank you.</v> I have one more question.<v>I don't know, about halfway through, about how statisticians are working on a thing. Many of us are within the data science track of biostatistics. And nobody talks to each other. Now, many of us are, many of the students here are within the data science track of biostatistics. By the way, love your Twitter handle. But yeah, so how do we bridge those things that we take advantage of these things
and it’s not three separate versions of the same thing?

I don’t know if there’s a systematic way.

Honestly, I came to know about much of the literature going through the revisions and one of the reviewer associate editors said there is a lot of work here in the econometrics literature, you should take a look.

And that’s kind of the value of the peer review system I guess.

And so we looked at it and yes, there was a lot of work and they just called it different things and so I had no idea when I was searching for that in the literature.

And we did see the Victor Chernozhukov paper in "Journal of Economics,"

but it’s basically an asymptotic statistics paper.

It kind of shows that these generalized Bayes stuff,

which they call as Laplace-type estimators, has all these nice properties that a standard vision posterior will have.

But yeah, I think talking to more people like interacting and telling about your work will kind of,
and someone will say that, oh yeah, I do something similar. You should look at this paper, it’s probably. Hopefully Twitter helps. Sorry? Hopefully Twitter helps. Yeah, yeah, definitely. Engagement through any like in-person or social media platform would be useful, yeah.

All right, well thanks so much. I think we’re out of time so we’ll stop it there.

(t attendant muttering indistinctly)

Hope everybody has a wonderful fall break. See you next week.

(t attendants chattering indistinctly)

The other organizer.

(t learner muttering indistinctly)

(t attendants chattering indistinctly)

Or maybe because they’re susceptible.

(t attendants chattering indistinctly)

Thank you. Anyone else need to sign in?

(t attendants chattering indistinctly)

Infection but they’re also premature babies.

(t attendants chattering indistinctly)

Premature, but also it’s that.

(t attendants chattering indistinctly)

It’s not a distinct.
Cause of death is very blurry in this day.

Is that part of why like.

"Cause a symptom given cause session"

Reporting depends on who is answering.

You need one of us to let you.

It might be a short answer.

Yeah, and it’s short answer.

I don’t have to, will you?