WEBVTT

1 00:00:00.330 --> 00:00:01.500 <v ->And welcome.</v>

2 00:00:01.500 --> 00:00:02.553 Today, it's my, eh.

3 00:00:04.500 --> 00:00:09.180 Today, it is my pleasure to introduce Professor Abhi Datta

4 00:00:09.180 --> 00:00:13.320 from Johns Hopkins University in Baltimore, Maryland.

5 00:00:13.320 --> 00:00:15.480 Professor Datta earned his BS and MS

6 00:00:15.480 --> 00:00:17.310 from the Indian Statistical Institute

7 00:00:17.310 --> 00:00:20.340 in 2008 and 2010 respectively,

8 00:00:20.340 --> 00:00:24.540 and PhD from the University of Minnesota in 2016.

9 00:00:24.540 --> 00:00:26.640 In addition to being a well-cited researcher

10 00:00:26.640 --> 00:00:29.670 with one publication that's almost 600 citations,

11 00:00:29.670 --> 00:00:30.813 which is pretty nice,

12 00:00:31.860 --> 00:00:34.560 he's also a award-winning educator,

13 00:00:34.560 --> 00:00:37.200 having repeatedly won an excellence in teaching award

14 00:00:37.200 --> 00:00:38.820 from his institution.

15 00:00:38.820 --> 00:00:40.413 So let's welcome Dr. Datta.

16 00:00:44.310 --> 00:00:45.143 <v ->Thank you, Robert,</v>

17 00:00:45.143 --> 00:00:47.940 for the invitation to come here and give the seminar,

18 00:00:47.940 --> 00:00:50.070 and for the very nice introduction.

19 00:00:50.070 --> 00:00:51.570 Thank you everyone for coming.

20 00:00:52.440 --> 00:00:56.310 My talk is about improving cause-specific mortality data

21 00:00:56.310 --> 00:00:58.290 in low and middle-income countries

22 00:00:58.290 --> 00:01:00.090 where the main tool to collect data

23 00:01:00.090 --> 00:01:02.280 is something called verbal autopsies.

24 00:01:02.280 --> 00:01:03.150 And the way I do it

25 00:01:03.150 --> 00:01:06.510 is using a statistical approach called generalized Bayes.

26 00:01:06.510 --> 00:01:07.770 If you have not heard

27 00:01:07.770 --> 00:01:10.710 of verbal autopsies or generalized Bayes,

28 00:01:10.710 --> 00:01:14.130 I can tell you that I hadn't heard of either of those things

29 00:01:14.130 --> 00:01:16.590 when I started working on the project,

30 00:01:16.590 --> 00:01:17.760 so don't worry about that,

31 00:01:17.760 --> 00:01:20.280 I try to give an introduction.

32 00:01:20.280 --> 00:01:23.970 'Cause I mostly work on a spatial and spatial temporal data

33 00:01:23.970 --> 00:01:26.503 and this was a project that came along,

34 00:01:26.503 --> 00:01:28.830 which is very different from what I used to work on.

35 00:01:28.830 --> 00:01:31.410 But over the years, there's been a nice body of work

36 00:01:31.410 --> 00:01:33.033 developed in this project.

37 00:01:35.310 --> 00:01:37.630 So this is a joint work

38 00:01:38.914 --> 00:01:43.710 with many different institutes and collaborators.

39 00:01:43.710 --> 00:01:46.230 The top row is the Hopkins bio stats team,

40 00:01:46.230 --> 00:01:48.300 which included my former students,

41 00:01:48.300 --> 00:01:50.700 Jacob Fiksel and Brian Gilbert,

42 00:01:50.700 --> 00:01:53.310 and my current postdoc, Sandi,

43 00:01:53.310 --> 00:01:56.280 and my colleague, Scott Zeger, and I

44 00:01:56.280 --> 00:01:58.263 lead the bio stats part of the team.

45 00:02:00.073 --> 00:02:03.450 Agbessi is the PI of the project in Mozambique

46 00:02:03.450 --> 00:02:07.440 that's sort of picked up developments for this work.

47 00:02:07.440 --> 00:02:08.670 And there are a lot of colleagues

48 00:02:08.670 --> 00:02:10.260 from the International Health Department

49 00:02:10.260 --> 00:02:12.120 that helped to collaborate.

50 00:02:12.120 --> 00:02:15.536 And then Li is the PI of a new project

51 00:02:15.536 --> 00:02:17.430 who we're going to apply our methodology

52 00:02:17.430 --> 00:02:21.660 for producing mortality estimates for the WHO.

53 00:02:21.660 --> 00:02:24.570 So we're collaborating with Li there as well.

54 00:02:24.570 --> 00:02:27.360 And then a couple of people outside Hopkins,

55 00:02:27.360 --> 00:02:30.930 Dianna at CDC and Emory University,

56 00:02:30.930 --> 00:02:34.530 as the director of the CHAMPS project.

57 00:02:34.530 --> 00:02:38.730 And Ivalda in the government body at Mozambique

58 00:02:38.730 --> 00:02:41.670 has been now currently doing the work in Mozambique.

59 00:02:43.770 --> 00:02:48.770 So this is funded by three grants from the Gates Foundation.

60 00:02:48.840 --> 00:02:51.840 The first one was the grant that kind of started things.

61 00:02:51.840 --> 00:02:55.020 And then we have a grant that is kind of developing more

62 00:02:55.020 --> 00:02:56.620 on the method side of the world.

63 00:02:58.860 --> 00:03:03.640 So, many low and middle-income countries

64 00:03:04.920 --> 00:03:08.400 often lack high-quality data on causes of death.

65 00:03:08.400 --> 00:03:09.630 Often for most deaths,

66 00:03:09.630 --> 00:03:13.380 there is no sort of medical certification

67 00:03:13.380 --> 00:03:16.170 or like an autopsy done.

68 00:03:16.170 --> 00:03:18.600 And without kind of high-quality data

69 00:03:18.600 --> 00:03:20.880 on what people are dying of,

70 00:03:20.880 --> 00:03:22.890 it's kind of hard to estimate the disease burden

71 00:03:22.890 --> 00:03:23.943 in these countries.

72 00:03:24.960 --> 00:03:27.090 And specifically, the quantity of interest

73 00:03:27.090 --> 00:03:29.070 is the cause-specific mortality fraction,

74 00:03:29.070 --> 00:03:33.930 which is basically the percentage of deaths in a age group

75 00:03:33.930 --> 00:03:36.303 that can be attributable to a given cause.

76 00:03:37.740 --> 00:03:39.510 So cause-specific mortality fractions

77 00:03:39.510 --> 00:03:41.940 are key pieces of information

78 00:03:41.940 --> 00:03:44.070 in determining the global burden of disease,

79 00:03:44.070 --> 00:03:46.620 which in turn dictates sovereign policy,

80 00:03:46.620 --> 00:03:49.170 as well as like resource allocations

81 00:03:49.170 --> 00:03:51.273 for programs operating in this country.

82 00:03:54.480 --> 00:03:56.580 So verbal autopsy is an alternate way

83 00:03:56.580 --> 00:03:58.770 to count deaths and attribute causes

84 00:03:58.770 --> 00:04:02.130 without actually doing a clinical autopsy.

85 00:04:02.130 --> 00:04:04.320 So verbal autopsy is basically

86 00:04:04.320 --> 00:04:06.720 a sort of a systematic interview

87 00:04:06.720 --> 00:04:08.340 of the household members of the deceased.

88 00:04:08.340 --> 00:04:11.760 So the government or the program has a set of field workers

89 00:04:11.760 --> 00:04:14.580 who go out and go from household to household

90 00:04:14.580 --> 00:04:16.530 and ask if anyone died in their household

91 00:04:16.530 --> 00:04:18.120 within the last several months.

92 00:04:18.120 --> 00:04:19.920 And if they died, what were the symptoms?

93 00:04:19.920 --> 00:04:22.770 And the set of questions they ask is not standardized

94 00:04:22.770 --> 00:04:24.360 by the WHO.

95 00:04:24.360 --> 00:04:26.610 Some example questions are here.

96 00:04:26.610 --> 00:04:29.190 Most of the questions would have binary answers

97 00:04:29.190 --> 00:04:31.530 like yes, no, but there are some questions

98 00:04:31.530 --> 00:04:35.793 that have more like continuous responses.

99 00:04:38.430 --> 00:04:40.530 So they said the WHO has standardized

100 00:04:40.530 --> 00:04:41.730 the verbal autopsy tool.

101 00:04:42.990 --> 00:04:46.530 The 2016 version has around 200 to 350 questions,

102 00:04:46.530 --> 00:04:48.360 depending on the age group.

103 00:04:48.360 --> 00:04:50.220 There are separate sections of the questionnaire

104 00:04:50.220 --> 00:04:53.880 for neonates, children deaths and adult deaths.

105 00:04:53.880 --> 00:04:55.770 And if you're interested in more information

106 00:04:55.770 --> 00:05:00.063 about verbal autopsy, there's a page in WHO about it.

107 00:05:01.560 --> 00:05:03.720 So a verbal autopsy, of course,

108 00:05:03.720 --> 00:05:05.070 doesn't give you a cause of death,

109 00:05:05.070 --> 00:05:07.620 it just gives you a bunch of yes-no responses

110 00:05:07.620 --> 00:05:10.233 to various questions related to the symptoms.

111 00:05:14.325 --> 00:05:17.187 So a verbal autopsy is basically a survey questionnaire.

112 00:05:17.187 --> 00:05:19.710 So you can pass that survey through a computer software

113 00:05:19.710 --> 00:05:22.740 and that can give a predictive cause of death.

114 00:05:22.740 --> 00:05:23.700 And so there are a bunch

115 00:05:23.700 --> 00:05:26.163 of different computer software available.

116 00:05:27.120 --> 00:05:30.540 InSilicoVA, developed by Tyler McCormick,

117 00:05:30.540 --> 00:05:32.403 Richard Li was a postdoc here,

118 00:05:33.750 --> 00:05:36.240 is published in "JASA" in 2016,

119 00:05:36.240 --> 00:05:37.440 is one of the, I think,

120 00:05:37.440 --> 00:05:39.900 most statistically-principled approaches to do it.

121 00:05:39.900 --> 00:05:42.660 But there are other approaches and then you can,

122 00:05:42.660 --> 00:05:44.700 this is basically a classification problem.

123 00:05:44.700 --> 00:05:47.700 So you're basically given your data on symptoms,

124 00:05:47.700 --> 00:05:50.000 you're kind of classifying the cause of death

125 00:05:50.000 --> 00:05:51.420 as one of several causes.

126 00:05:51.420 --> 00:05:54.420 So you can use standard classifiers

127 00:05:54.420 --> 00:05:56.420 and machine learning approaches as well.

128 00:05:57.606 --> 00:05:59.010 OpenVA is an excellent resource

129 00:05:59.010 --> 00:06:00.480 to learn about verbal autopsies.

130 00:06:00.480 --> 00:06:02.943 Again, openVA is,

131 00:06:03.811 --> 00:06:05.520 I think Richard is one of the maintainers

132 00:06:05.520 --> 00:06:06.693 and creators of openVA.

133 00:06:11.400 --> 00:06:14.040 So the COMSA project in Mozambique,

134 00:06:14.040 --> 00:06:16.710 one of the main goals was to generate

135 00:06:16.710 --> 00:06:19.440 this cause-specific mortality fractions

136 00:06:19.440 --> 00:06:21.360 for children's and under,

137 00:06:21.360 --> 00:06:23.160 for neonates and under-five children

138 00:06:24.360 --> 00:06:26.250 for the country of Mozambique.

139 00:06:26.250 --> 00:06:30.300 And the data that we collected was a large dataset

140 00:06:30.300 --> 00:06:32.037 of vocal autopsy record

141 00:06:32.037 --> 00:06:34.080 for different households that were surveyed

142 00:06:34.080 --> 00:06:37.860 and that was a map of Mozambique

143 00:06:37.860 --> 00:06:41.080 and the green region show

144 00:06:41.080 --> 00:06:42.960 where the data was collected

145 00:06:42.960 --> 00:06:44.370 as part of the COMSA project.

146 00:06:44.370 --> 00:06:49.370 So in statistical terms, the data just has the symptoms,

147 00:06:49.380 --> 00:06:50.970 it doesn't have the true cause of death,

148 00:06:50.970 --> 00:06:52.863 so we call it the unlabeled data.

149 00:06:56.970 --> 00:07:00.060 So how to go from an unlabeled data to the labeling

150 00:07:00.060 --> 00:07:01.491 of the causes of death

151 00:07:01.491 --> 00:07:03.720 and then estimate these cause fractions.

152 00:07:03.720 --> 00:07:07.755 This is the standard procedure that is typically done

153 00:07:07.755 --> 00:07:09.870 and this is what we were supposed to do as well,

154 00:07:09.870 --> 00:07:12.300 which is simply take each record,

155 00:07:12.300 --> 00:07:14.430 pass it through the computer software

156 00:07:14.430 --> 00:07:16.050 and get a cause of death.

157 00:07:16.050 --> 00:07:17.580 And once you get a cause of death,

158 00:07:17.580 --> 00:07:19.440 then you can sort of simply aggregate.

159 00:07:19.440 --> 00:07:21.210 So in the story example,

160 00:07:21.210 --> 00:07:24.930 three out of the six cases were assigned to be from HIV.

161 00:07:24.930 --> 00:07:27.390 And so the cause-specific mortality fraction for HIV

162 00:07:27.390 --> 00:07:31.950 would be 50% and similar for malaria and sepsis and so on.

163 00:07:31.950 --> 00:07:35.160 So that's the basic template

164 00:07:35.160 --> 00:07:37.590 of how to get a cause-specific mortality fractions

165 00:07:37.590 --> 00:07:39.060 from verbal autopsies.

166 00:07:39.060 --> 00:07:41.010 The question is can we trust this estimates?

167 00:07:41.010 --> 00:07:42.960 Because these are not true causes of death

168 00:07:42.960 --> 00:07:45.900 as determined by a doctor or by a clinical procedure.

169 00:07:45.900 --> 00:07:48.300 These are cause of death predicted by an algorithm

170 00:07:48.300 --> 00:07:52.140 based on just surveying the household members

171 00:07:52.140 --> 00:07:53.103 of the deceased.

172 00:07:57.295 --> 00:07:59.730 So turns out machine learning has a name

173 00:07:59.730 --> 00:08:01.020 for this type of problems,

174 00:08:01.020 --> 00:08:03.630 it's called quantification learning,

175 00:08:03.630 --> 00:08:06.870 which is basically estimating population prevalence

176 00:08:06.870 --> 00:08:09.900 using predicted levels instead of true levels

177 00:08:09.900 --> 00:08:12.570 and the predictions are coming from a classifier.

178 00:08:12.570 --> 00:08:15.510 And so there has been some work in quantification learning

179 00:08:15.510 --> 00:08:18.900 and in the machine learning literature.

180 00:08:18.900 --> 00:08:20.640 So when we were working on this problem,

181 00:08:20.640 --> 00:08:21.960 we realized that estimating

182 00:08:21.960 --> 00:08:23.760 cause-specific mortality fractions

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.

209 00:09:46.500 --> 00:09:50.070 So even these covariates, x, may or may not be available.

210 00:09:50.070 --> 00:09:53.340 And then we are interested in estimating the probability

211 00:09:53.340 --> 00:09:57.720 that y belongs to one of the C many cause categories,

212 00:09:57.720 --> 00:09:59.913 so that's a quantity of interest.

213 00:10:05.160 --> 00:10:07.470 For some reason, there is a conditional sign

214 00:10:07.470 --> 00:10:09.090 that's missing there.

215 00:10:09.090 --> 00:10:13.080 But you can use the law of total probability

216 00:10:13.080 --> 00:10:16.050 to write the probability of the predicted cause of death,

217 00:10:16.050 --> 00:10:17.610 which is the a,

218 00:10:17.610 --> 00:10:22.020 probability of a as a sum of our probability of a given y

219 00:10:22.020 --> 00:10:24.150 times probability of y.

220 00:10:24.150 --> 00:10:26.190 So there's a conditional sign missing here,

221 00:10:26.190 --> 00:10:28.190 I don't don't know what's going on here.

222 00:10:32.010 --> 00:10:33.180 But the COMSA data,

223 00:10:33.180 --> 00:10:36.090 we only get information on the left-hand side, right?

224 00:10:36.090 --> 00:10:40.770 And we want to input upon the quantity probability of y

225 00:10:40.770 --> 00:10:42.863 which would be the true CSMFs.

226 00:10:44.031 --> 00:10:45.960 So there is only one known quantity

227 00:10:45.960 --> 00:10:48.193 with which you can estimate the left-hand side.

228 00:10:48.193 --> 00:10:50.010 There are two unknown quantities on the right-hand side.

229 00:10:50.010 --> 00:10:53.820 So without making assumptions, you cannot really identify

230 00:10:53.820 --> 00:10:55.950 probability of y, right?

231 00:10:55.950 --> 00:10:58.530 So any quantification learning methods

232 00:10:58.530 --> 00:11:01.620 need to either estimate those conditional probabilities,

233 00:11:01.620 --> 00:11:03.510 probability of a given y,

234 00:11:03.510 --> 00:11:05.133 or make some assumptions on it.

235 00:11:07.680 --> 00:11:12.680 So again, all the conditional signs are missing.

236 00:11:16.410 --> 00:11:18.990 The one of the most common approaches,

237 00:11:18.990 --> 00:11:22.170 and this is what is used in the verbal autopsy world

238 00:11:22.170 --> 00:11:24.540 is called classify and count,

239 00:11:24.540 --> 00:11:27.930 which is you simply predict the cause of death

240 00:11:27.930 --> 00:11:29.220 and then aggregate.

241 00:11:29.220 --> 00:11:33.439 So you're simply claiming that probability of a

242 00:11:33.439 --> 00:11:36.420 is same as probability of y which is equivalent to claiming

243 00:11:36.420 --> 00:11:38.850 that this misclassification rate matrix

244 00:11:38.850 --> 00:11:41.310 is an identity matrix, right?

245 00:11:41.310 --> 00:11:43.740 Because you're saying that the left hand quantity

246 00:11:43.740 --> 00:11:47.530 is the same as the rightmost quantity, which would be true

247 00:11:48.390 --> 00:11:50.760 if there is no misclassification by the algorithm

248 00:11:50.760 --> 00:11:52.680 and if the predicted cause of death

249 00:11:52.680 --> 00:11:54.423 is always the true cause of death.

250 00:11:55.860 --> 00:11:58.110 And that's what is typically done

251 00:11:58.110 --> 00:12:01.890 in this cause-specific mortality fraction estimates.

252 00:12:01.890 --> 00:12:03.630 But it's a very strong assumption, right?

253 00:12:03.630 --> 00:12:07.200 Because it says assuming perfect sensitivity and specificity

254 00:12:07.200 --> 00:12:08.050 of the algorithm.

255 00:12:09.570 --> 00:12:11.880 So let's look at how perfect the algorithms are.

256 00:12:11.880 --> 00:12:13.320 So these are two algorithms,

257 00:12:13.320 --> 00:12:15.510 Tariff and InSilicoVA,

258 00:12:15.510 --> 00:12:19.950 PHMRC data is a benchmark dataset from four countries

259 00:12:19.950 --> 00:12:21.870 that has both the verbal autopsy data

260 00:12:21.870 --> 00:12:26.250 as well as a gold standard cause of death diagnosis.

261 00:12:26.250 --> 00:12:30.000 And you can see the accuracies of either method

262 00:12:30.000 --> 00:12:32.940 is around 30%, so they're far from being

263 00:12:32.940 --> 00:12:34.443 like fully accurate.

264 00:12:35.850 --> 00:12:39.330 So there is large misclassification rates

265 00:12:39.330 --> 00:12:41.790 of these algorithms and if you don't kind of adjust

266 00:12:41.790 --> 00:12:44.430 for these misclassifications,

267 00:12:44.430 --> 00:12:45.540 this is burden estimates

268 00:12:45.540 --> 00:12:48.480 of the cause-specific mortality fractions you get

269 00:12:48.480 --> 00:12:50.230 are likely going to be very biased.

270 00:12:53.610 --> 00:12:57.660 So this is where the CHAMPS project comes into play.

271 00:12:57.660 --> 00:13:00.090 So the CHAMPS is an ongoing project

272 00:13:00.090 --> 00:13:04.650 in like seven or eight countries including Mozambique,

273 00:13:04.650 --> 00:13:07.380 which is collecting data on both verbal autopsy

274 00:13:07.380 --> 00:13:11.310 and a more comprehensive cause of death procedure

275 00:13:11.310 --> 00:13:13.830 called minimally invasive tissue sampling.

276 00:13:13.830 --> 00:13:17.490 So it basically takes a sample of your tissue

277 00:13:17.490 --> 00:13:20.460 of the deceased person and then runs a bunch

278 00:13:20.460 --> 00:13:23.070 of pathological tests and imaging analysis

279 00:13:23.070 --> 00:13:25.410 and then gives a cause of death.

280 00:13:25.410 --> 00:13:29.080 And the MITS cause of death assignments

281 00:13:30.330 --> 00:13:32.790 have been shown to be quite accurate when you compare

282 00:13:32.790 --> 00:13:34.593 to like a full diagnostic autopsy.

283 00:13:36.210 --> 00:13:37.920 So MITS is being done in a bunch

284 00:13:37.920 --> 00:13:40.950 of different countries including Mozambique.

285 00:13:40.950 --> 00:13:43.380 And for the cases where MITS is being done,

286 00:13:43.380 --> 00:13:45.990 the verbal autopsies are also collected.

287 00:13:45.990 --> 00:13:48.120 So what you get from this CHAMPS data

288 00:13:48.120 --> 00:13:50.310 is a labeled or paired dataset

289 00:13:50.310 --> 00:13:51.930 where you have both the verbal autopsy

290 00:13:51.930 --> 00:13:54.000 as well as the MITS cause of death

291 00:13:54.000 --> 00:13:57.630 and you can pass the verbal autopsy to the software

292 00:13:57.630 --> 00:14:00.254 to get the verbal autopsy predicted cause of death.

293 00:14:00.254 --> 00:14:01.770 And then you can cross tabulate the two

294 00:14:01.770 --> 00:14:04.470 and get an estimate of the misclassification rates, right?

295 00:14:04.470 --> 00:14:05.917 Like you can say like,

296 00:14:05.917 --> 00:14:08.370 "Oh okay, so there are 10 cases

297 00:14:08.370 --> 00:14:10.830 that the MITS cause of death was HIV,

298 00:14:10.830 --> 00:14:12.180 out of those 10 cases,

299 00:14:12.180 --> 00:14:15.060 seven of them were correctly assigned to HIV

300 00:14:15.060 --> 00:14:16.380 by verbal autopsy.

301 00:14:16.380 --> 00:14:19.980 So then the sensitivity would be 70%

302 00:14:19.980 --> 00:14:22.827 and the false positive would be 30%, so on."

303 00:14:27.060 --> 00:14:29.130 So this is the broad idea of the methodology.

304 00:14:29.130 --> 00:14:32.250 So for the COMSA data, which is the unpaired data,

305 00:14:32.250 --> 00:14:34.440 you get only the verbal autopsy record

306 00:14:34.440 --> 00:14:37.110 so you can get an estimate of the predicted cause of deaths

307 00:14:37.110 --> 00:14:38.880 from the verbal autopsy.

308 00:14:38.880 --> 00:14:41.190 From the CHAMPS data, which is the paired data,

309 00:14:41.190 --> 00:14:44.400 you can get an estimate of the misclassification rates.

310 00:14:44.400 --> 00:14:47.670 And then the only unknown is then the probabilities

311 00:14:47.670 --> 00:14:49.500 of the cause of death

312 00:14:49.500 --> 00:14:54.090 if you were able to do the MITS autopsy for every death.

313 00:14:54.090 --> 00:14:57.859 So then this is an equation with two knowns and one unknown

314 00:14:57.859 --> 00:15:01.320 and you can solve for it and get the calibrating message.

315 00:15:01.320 --> 00:15:04.533 So that's the broad idea and we do it in a model-based way.

316 00:15:08.880 --> 00:15:10.650 So here's the formal model.

317 00:15:10.650 --> 00:15:14.700 So for the CHAMPS dataset with the unlabeled data or the U,

318 00:15:14.700 --> 00:15:17.280 we have the predicted labels, ar,

319 00:15:17.280 --> 00:15:18.483 and then for the,

320 00:15:19.560 --> 00:15:21.000 that's for the COMSA data,

321 00:15:21.000 --> 00:15:22.110 and for the CHAMPS data,

322 00:15:22.110 --> 00:15:25.560 we have both the predicted labels from verbal autopsy, ar,

323 00:15:25.560 --> 00:15:27.783 as well as the MITS determine labels, yr.

324 00:15:28.800 --> 00:15:33.120 And our quantity of interest is the probabilities of yr

325 00:15:34.284 --> 00:15:35.984 belonging to the different causes.

326 00:15:40.740 --> 00:15:43.110 There's a conditional sign missing here.

327 00:15:44.250 --> 00:15:47.730 But if the conditional probabilities

328 00:15:47.730 --> 00:15:52.380 are denoted by Mij, which is if the MITS cause is i,

329 00:15:52.380 --> 00:15:55.563 what is the probability that the via predicted cause is j?

330 00:15:57.090 --> 00:15:59.340 Then you can use a law of total probability

331 00:15:59.340 --> 00:16:01.650 to write down the marginal distribution

332 00:16:01.650 --> 00:16:03.270 of the via predicted cause.

333 00:16:03.270 --> 00:16:06.720 So that would be in terms of the misclassification rates

334 00:16:06.720 --> 00:16:09.680 and the marginal cause distribution of the MITS-COD.

335 00:16:09.680 --> 00:16:11.010 So that's the whole idea.

336 00:16:11.010 --> 00:16:14.880 So you can write this in terms of a matrix vector notation

337 00:16:14.880 --> 00:16:18.030 as probability of a as M transpose p

338 00:16:18.030 --> 00:16:20.760 where M is the misclassification rate matrix,

339 00:16:20.760 --> 00:16:23.640 p is the unknown quantity of interest,

340 00:16:23.640 --> 00:16:26.610 which is probability that the cause of death

341 00:16:26.610 --> 00:16:29.390 is coming from an unknown cause.

342 00:16:31.440 --> 00:16:33.840 So the data model is very simple,

343 00:16:33.840 --> 00:16:36.000 but the unlabeled data,

344 00:16:36.000 --> 00:16:38.220 it follows multinomial with this probability

345 00:16:38.220 --> 00:16:41.400 which is coming from this law of total probability.

346 00:16:41.400 --> 00:16:42.690 And then for the label data,

347 00:16:42.690 --> 00:16:46.320 this is ar given yr equals to i,

348 00:16:46.320 --> 00:16:47.850 it follows multinomial with the i

349 00:16:47.850 --> 00:16:49.410 throughout the misclassification matrix.

350 00:16:49.410 --> 00:16:51.030 So if the MITS-COD is i,

351 00:16:51.030 --> 00:16:53.010 the misclassification rates are given by the i

352 00:16:53.010 --> 00:16:55.350 throughout the misclassification matrix,

353 00:16:55.350 --> 00:16:58.500 so it's multinomial with that probability.

354 00:16:58.500 --> 00:17:00.477 And then we've put priors on M and p

355 00:17:01.349 --> 00:17:03.930 and then we can get estimates of both M and p.

356 00:17:03.930 --> 00:17:06.830 M is a nuisance parameter, p is the parameter of interest.

357 00:17:09.900 --> 00:17:13.380 Just to carefully go over what are the assumptions here.

358 00:17:13.380 --> 00:17:17.610 The main assumption is that the misclassification rates

359 00:17:17.610 --> 00:17:20.040 of verbal autopsy given MITS

360 00:17:20.040 --> 00:17:22.530 are the same in your label data

361 00:17:22.530 --> 00:17:24.750 as they would be in your unlabeled data.

362 00:17:24.750 --> 00:17:27.540 This is not verifiable because we don't have

363 00:17:27.540 --> 00:17:29.760 any true cause of death in the unlabeled data,

364 00:17:29.760 --> 00:17:30.873 so it's an assumption.

365 00:17:33.210 --> 00:17:34.890 Given that the verbal autopsy

366 00:17:34.890 --> 00:17:36.930 is a function of your symptoms,

367 00:17:36.930 --> 00:17:41.133 the assumption is essentially that given a true cause,

368 00:17:42.000 --> 00:17:44.370 the probability of the symptoms are going to be same

369 00:17:44.370 --> 00:17:46.403 in your unlabeled dataset as in your labeled dataset.

370 00:17:49.207 --> 00:17:50.100 And it's a reasonable assumption

371 00:17:50.100 --> 00:17:52.530 as if you have a cause of death,

372 00:17:52.530 --> 00:17:56.430 it's likely that you have certain symptoms will appear

373 00:17:56.430 --> 00:17:58.500 and some certain symptoms will not appear.

374 00:17:58.500 --> 00:18:02.400 And that is true regardless of whether the data is coming

375 00:18:02.400 --> 00:18:03.473 from the labeled set or the unlabeled set.

376 00:18:08.462 --> 00:18:12.240 We do not assume that the marginal distribution

377 00:18:12.240 --> 00:18:15.690 of the CHAMPS data of the causes in the label data

378 00:18:15.690 --> 00:18:17.370 is representative of the population

379 00:18:17.370 --> 00:18:19.920 because they are not, because the CHAMPS state,

380 00:18:19.920 --> 00:18:21.450 so the CHAMPS project is done

381 00:18:21.450 --> 00:18:24.420 at specific hospitals in the country

382 00:18:24.420 --> 00:18:27.540 and distribution of causes in hospitals

383 00:18:27.540 --> 00:18:29.910 are typically not same as distribution

384 00:18:29.910 --> 00:18:31.110 of causes in the community.

385 00:18:31.110 --> 00:18:31.950 And we are interested

386 00:18:31.950 --> 00:18:34.080 in the cause distribution in the population.

387 00:18:34.080 --> 00:18:35.470 So there is no assumption

388 00:18:36.509 --> 00:18:40.170 that the marginal distribution of y in the label data

389 00:18:40.170 --> 00:18:42.960 is same as the marginal distribution of y in unlabeled data,

390 00:18:42.960 --> 00:18:44.970 which is our quantity of interest.

391 00:18:44.970 --> 00:18:47.010 And the reason there is no assumption

392 00:18:47.010 --> 00:18:50.610 is we only model a given y in the label data.

393 00:18:50.610 --> 00:18:53.013 We never model y in the label data.

394 00:18:53.910 --> 00:18:55.560 So we only model the conditional

395 00:18:55.560 --> 00:18:56.910 and the assumption is the condition

396 00:18:56.910 --> 00:18:59.610 of misclassification rates are transportable

397 00:18:59.610 --> 00:19:01.883 from the labeled to the unlabeled side.

398 00:19:05.707 --> 00:19:07.230 So that's the main idea.

399 00:19:07.230 --> 00:19:09.380 And this was the first work we did,

400 00:19:09.380 --> 00:19:13.170 we just used this top cause prediction.

401 00:19:13.170 --> 00:19:14.610 But many of these algorithms

402 00:19:14.610 --> 00:19:16.800 are actually probabilistic in nature in the sense

403 00:19:16.800 --> 00:19:18.090 that if you look at their outputs,

404 00:19:18.090 --> 00:19:20.130 they won't give a single cause of death,

405 00:19:20.130 --> 00:19:22.470 but they will give scores to each cause.

406 00:19:22.470 --> 00:19:23.910 So for example,

407 00:19:23.910 --> 00:19:26.460 this would be a typical output of an algorithm

408 00:19:26.460 --> 00:19:28.380 for like say 6%.

409 00:19:28.380 --> 00:19:30.180 So for the first person, it will say

410 00:19:33.194 --> 00:19:35.344 70% HIV, 20% malaria, 10% sepsis and so on.

411 00:19:38.100 --> 00:19:40.770 And the standard procedure is to take the top cause,

412 00:19:40.770 --> 00:19:43.680 so for the first person, it would be HIV,

413 00:19:43.680 --> 00:19:47.610 for the second person, it will be malaria and so on.

414 00:19:47.610 --> 00:19:49.590 So that's how you get a single cause

415 00:19:49.590 --> 00:19:51.190 from a probabilistic prediction.

416 00:19:53.430 --> 00:19:56.037 So that essentially ignores sort of the scores

417 00:19:57.390 --> 00:20:00.810 assigned to the second most likely cause,

418 00:20:00.810 --> 00:20:03.630 the third most likely cause and so on.

419 00:20:03.630 --> 00:20:08.630 And you ignore those, you can end up with a biased estimate.

420 00:20:09.030 --> 00:20:11.940 So you can see these are the CSMF estimates

421 00:20:11.940 --> 00:20:13.650 using the top cause,

422 00:20:13.650 --> 00:20:14.940 these are the CSM estimates

423 00:20:14.940 --> 00:20:16.950 using the exact scores that are assigned

424 00:20:16.950 --> 00:20:18.300 and those are different, right?

425 00:20:18.300 --> 00:20:21.600 So when we kind of change this probabilistic output

426 00:20:21.600 --> 00:20:25.863 to a single cause output, we discard information.

427 00:20:29.640 --> 00:20:31.530 So we wanted to extend the work

428 00:20:31.530 --> 00:20:35.790 to kind of use the full set of scores and the set of scores

429 00:20:35.790 --> 00:20:38.100 can be thought of as a compositional data in the sense

430 00:20:38.100 --> 00:20:40.170 that the scores sum up to one

431 00:20:40.170 --> 00:20:44.610 because it assigns 100% probability across all causes

432 00:20:44.610 --> 00:20:47.670 and then they're each non-negative.

433 00:20:47.670 --> 00:20:50.610 The issue is that for the categorical data,

434 00:20:50.610 --> 00:20:53.460 our model is based on multinomial distribution.

435 00:20:53.460 --> 00:20:55.110 And then for compositional data,

436 00:20:55.110 --> 00:20:57.030 the models are typically like Dirichlet

437 00:20:57.030 --> 00:20:58.920 or log ratio based models,

438 00:20:58.920 --> 00:21:01.870 which are very different from the multinomial distribution.

439 00:21:03.450 --> 00:21:05.070 So if we have some cases

440 00:21:05.070 --> 00:21:07.050 for which we have categorical output,

441 00:21:07.050 --> 00:21:09.090 for some, we have compositional output,

442 00:21:09.090 --> 00:21:10.830 this would lead to different models

443 00:21:10.830 --> 00:21:12.580 for different parts of the dataset.

444 00:21:14.760 --> 00:21:16.710 These Dirichlet or log-ratio models

445 00:21:16.710 --> 00:21:19.500 also do not allow zeros in the data.

446 00:21:19.500 --> 00:21:21.810 So if you have zeros or ones in the composition,

447 00:21:21.810 --> 00:21:23.430 they don't allow that.

448 00:21:23.430 --> 00:21:26.820 And then there are very specific models about the data

449 00:21:26.820 --> 00:21:29.100 which are subjective model and specification.

450 00:21:29.100 --> 00:21:32.670 So the data distribution does not look like a Dirichlet

451 00:21:32.670 --> 00:21:33.660 assuming a Dirichlet layer

452 00:21:33.660 --> 00:21:37.713 would lead to kind of wrong results.

453 00:21:40.800 --> 00:21:45.800 So how do we extend the multinomial framework we had

454 00:21:46.110 --> 00:21:49.233 for categorical data to compositional data?

455 00:21:50.790 --> 00:21:55.680 Again, there would be a conditional sign here.

456 00:21:55.680 --> 00:21:57.750 But the basic assumption that we had

457 00:21:57.750 --> 00:22:01.650 for the multinomial case was probability of a given y

458 00:22:01.650 --> 00:22:04.620 is the i throughout misclassification matrix, right?

459 00:22:04.620 --> 00:22:09.620 And for categorical data, a probability statement

460 00:22:09.900 --> 00:22:12.030 is same as an expectation statement, right?

461 00:22:12.030 --> 00:22:13.860 So we can equivalently write this

462 00:22:13.860 --> 00:22:16.170 as expectation of a given y

463 00:22:16.170 --> 00:22:17.470 is the i throughout the M.

464 00:22:18.919 --> 00:22:20.430 The advantage of the expectation statement

465 00:22:20.430 --> 00:22:23.310 is that it's more generally applicable.

466 00:22:23.310 --> 00:22:27.150 It will not be just for categorical data, right?

467 00:22:27.150 --> 00:22:30.150 So for categorical data, there's a equivalent.

468 00:22:30.150 --> 00:22:33.390 For other data types, this statement can be valid

469 00:22:33.390 --> 00:22:36.690 even though the previous statement may not be applicable.

470 00:22:36.690 --> 00:22:40.887 So we kind of write this as our model

471 00:22:40.887 --> 00:22:45.210 for the compositional data and we make no other assumptions

472 00:22:45.210 --> 00:22:46.260 about this distribution.

473 00:22:46.260 --> 00:22:50.920 So only a first moment conditional expectation statement

474 00:22:53.400 --> 00:22:56.313 without any full distributional specification.

475 00:22:58.650 --> 00:23:00.450 So what do we do?

476 00:23:00.450 --> 00:23:02.880 So we have expectation of a given y

477 00:23:02.880 --> 00:23:05.343 is the i throughout the misclassification matrix.

478 00:23:08.040 --> 00:23:09.567 We can use something called

479 00:23:09.567 --> 00:23:11.520 the Kullback Leibler Divergence

480 00:23:11.520 --> 00:23:13.710 or the cross entropy loss

481 00:23:13.710 --> 00:23:16.770 between a and its model expectation.

482 00:23:16.770 --> 00:23:20.013 So these are all the conditional signs are missing here.

483 00:23:22.050 --> 00:23:25.353 So basically a is the data we observe,

484 00:23:26.400 --> 00:23:28.860 this is the modeled expectation,

485 00:23:28.860 --> 00:23:29.693 which is basically the i

486 00:23:29.693 --> 00:23:31.287 through of the misclassification matrix

487 00:23:31.287 --> 00:23:33.630 and we use the cross entropy loss,

488 00:23:33.630 --> 00:23:36.810 the Kullback Leibler loss between the two.

489 00:23:36.810 --> 00:23:37.800 What's the advantage?

490 00:23:37.800 --> 00:23:38.633 So first of all,

491 00:23:38.633 --> 00:23:41.610 the Kullback Leibler loss allows zeroes in the composition.

492 00:23:41.610 --> 00:23:45.330 So it is well-defined even if you have zeroes or ones.

493 00:23:45.330 --> 00:23:47.970 If you take the negative loss and exponentiate it,

494 00:23:47.970 --> 00:23:49.940 it's exactly the multinomial likelihood.

495 00:23:49.940 --> 00:23:52.050 So if your data is indeed multinomial,

496 00:23:52.050 --> 00:23:54.420 you get back your likelihood that you're using

497 00:23:54.420 --> 00:23:57.120 for your single class model.

498 00:23:57.120 --> 00:23:59.550 But if your data is not multinomial,

499 00:23:59.550 --> 00:24:02.100 you get a pseudo likelihood that you can work with.

500 00:24:03.960 --> 00:24:06.660 If you can take the derivative of the loss function

501 00:24:06.660 --> 00:24:10.170 and take the expectation under the two parameter,

502 00:24:10.170 --> 00:24:13.001 you'll see that it's a valid score function

503 00:24:13.001 --> 00:24:15.750 in the sense that you get an unbiased estimating equation

504 00:24:15.750 --> 00:24:18.900 for your misclassification rate matrix, M,

505 00:24:18.900 --> 00:24:21.033 based on just the first moment as option.

506 00:24:22.890 --> 00:24:24.720 And then similarly, you can do the same thing

507 00:24:24.720 --> 00:24:26.730 for the unlabeled data.

508 00:24:26.730 --> 00:24:29.520 The probability statement becomes expectation statement

509 00:24:29.520 --> 00:24:32.400 and then we have the Kullback Leibler loss.

510 00:24:32.400 --> 00:24:36.360 This is an unbiased estimated equation for both M and p.

511 00:24:36.360 --> 00:24:37.500 And again,

512 00:24:37.500 --> 00:24:40.680 if the data is truly multinomial and not compositional,

513 00:24:40.680 --> 00:24:43.410 this becomes exactly the multinomial likelihood.

514 00:24:43.410 --> 00:24:44.760 If the data is compositional,

515 00:24:44.760 --> 00:24:46.310 it becomes a pseudo likelihood.

516 00:24:49.860 --> 00:24:52.170 Okay, so how do we do Bayes analysis

517 00:24:52.170 --> 00:24:54.240 with pseudo likelihoods?

518 00:24:54.240 --> 00:24:56.970 So this is where this idea of generalized Bayes

519 00:24:56.970 --> 00:24:58.920 or model-free Bayesian inference comes in

520 00:24:58.920 --> 00:25:01.200 and there have been parallel developments

521 00:25:01.200 --> 00:25:04.290 in both computer science, econometrics and statistics

522 00:25:04.290 --> 00:25:06.870 without much communication among the three fields

523 00:25:06.870 --> 00:25:10.080 for the last 30, 40 years.

524 00:25:10.080 --> 00:25:12.570 Basically, if you're given a loss function

525 00:25:12.570 --> 00:25:15.480 without a given like a full likelihood for the data,

526 00:25:15.480 --> 00:25:18.330 you can take negative of that loss function

527 00:25:18.330 --> 00:25:20.823 multiplied by some tuning parameter, alpha,

528 00:25:21.870 --> 00:25:25.620 exponentiate it and treat it as a pseudo likelihood

529 00:25:25.620 --> 00:25:27.270 and apply your priors

530 00:25:27.270 --> 00:25:30.000 and then your posterior is going to be proportional to this

531 00:25:30.000 --> 00:25:32.850 as long as the normalization constant exists.

532 00:25:32.850 --> 00:25:35.460 And there has been a lot of work that has shown

533 00:25:35.460 --> 00:25:37.590 that this is a valid posterior,

534 00:25:37.590 --> 00:25:40.500 it is a generalization of the Bayesian posterior,

535 00:25:40.500 --> 00:25:42.360 like if this is an actual likelihood,

536 00:25:42.360 --> 00:25:44.040 this is the Bayesian posterior,

537 00:25:44.040 --> 00:25:46.173 but if it's not a actual likelihood,

538 00:25:47.654 --> 00:25:49.470 this has been shown that it basically minimizes

539 00:25:49.470 --> 00:25:52.503 the Bayes risk for that loss function.

540 00:25:54.120 --> 00:25:56.280 It has nice asymptotic properties

541 00:25:56.280 --> 00:25:59.400 shown by Victor Chernozhukov in this paper

542 00:25:59.400 --> 00:26:03.960 and then in this JSS paper in 2016 I think

543 00:26:03.960 --> 00:26:06.000 it showed that if you're given a loss function

544 00:26:06.000 --> 00:26:07.140 and a prior,

545 00:26:07.140 --> 00:26:10.173 this is the only coherent way you can get a posterior.

546 00:26:11.670 --> 00:26:14.670 So there's now been a lot of work and it's been called

547 00:26:14.670 --> 00:26:17.340 by different names like Gibbs posteriors,

548 00:26:17.340 --> 00:26:19.740 pseudo posterior, Laplace-type estimators

549 00:26:19.740 --> 00:26:23.043 and quasi-Bayesian estimators along with generalized Bayes.

550 00:26:25.470 --> 00:26:28.470 So for our case, we have the pseudo likelihood

551 00:26:28.470 --> 00:26:29.460 for the label data.

552 00:26:29.460 --> 00:26:31.530 We have the pseudo likelihood for the unlabeled data.

553 00:26:31.530 --> 00:26:33.270 We put priors.

554 00:26:33.270 --> 00:26:35.190 If all of our data were categorical,

555 00:26:35.190 --> 00:26:37.560 this reduces to that multinomial model we had

556 00:26:37.560 --> 00:26:39.120 for the categorical data.

557 00:26:39.120 --> 00:26:41.190 But if some of the data is compositional,

558 00:26:41.190 --> 00:26:43.830 then this becomes generalized Bayes,

559 00:26:43.830 --> 00:26:47.160 so we call it generalized Bayes quantification learning.

560 00:26:47.160 --> 00:26:50.190 It allows sparsity of the outputs in the sense

561 00:26:50.190 --> 00:26:53.520 that if some of the data have zeroes and ones in them,

562 00:26:53.520 --> 00:26:55.590 this is well-defined.

563 00:26:55.590 --> 00:26:57.750 It's the same pseudo likelihood

564 00:26:57.750 --> 00:27:00.510 for categorical compositional predictions.

565 00:27:00.510 --> 00:27:01.950 And then it also allows

566 00:27:01.950 --> 00:27:05.013 a nice Gibbs sample using conjugacy.

567 00:27:10.920 --> 00:27:14.820 One final sort of data aspect we had

568 00:27:14.820 --> 00:27:18.420 was that this minimal tissue sampling

569 00:27:18.420 --> 00:27:20.730 was also sometimes inconclusive in the sense

570 00:27:20.730 --> 00:27:22.230 that they gave two causes.

571 00:27:22.230 --> 00:27:27.230 Like often, they were ambiguous between HIV and tuberculosis

572 00:27:28.890 --> 00:27:30.750 and they would give one as the immediate cause

573 00:27:30.750 --> 00:27:32.040 and one as the underlying cause.

574 00:27:32.040 --> 00:27:35.820 So sometimes, even the true cause of death is compositional.

575 00:27:35.820 --> 00:27:38.790 So your predicted cause of death is compositional,

576 00:27:38.790 --> 00:27:40.647 your true cause of death is also compositional

577 00:27:40.647 --> 00:27:45.270 and we call it like b, which represents the belief.

578 00:27:45.270 --> 00:27:49.380 And you can show that if you're only given b

579 00:27:49.380 --> 00:27:51.273 instead of a single cause of death,

580 00:27:52.603 --> 00:27:55.800 your conditional expectation becomes M transpose b

581 00:27:55.800 --> 00:27:59.340 instead of the i through of the M matrix.

582 00:27:59.340 --> 00:28:01.380 And you can do the same thing

583 00:28:01.380 --> 00:28:04.543 using the compositional true cause of death

584 00:28:04.543 --> 00:28:07.620 instead of the actual true cause of death.

585 00:28:07.620 --> 00:28:09.540 And all the conditional signs are missing here

586 00:28:09.540 --> 00:28:13.800 but you can just formulate the Kullback Leibler likelihood

587 00:28:13.800 --> 00:28:16.593 to generate pseudo likelihood.

588 00:28:18.870 --> 00:28:21.570 So this kind of give rise to a digression

589 00:28:21.570 --> 00:28:24.040 where we kind of looked at this is basically

590 00:28:25.152 --> 00:28:28.080 your true cause of death is a compositional covariate

591 00:28:28.080 --> 00:28:31.350 and your predicted cause of death is a compositional output.

592 00:28:31.350 --> 00:28:33.120 So we kind of looked at regression

593 00:28:33.120 --> 00:28:36.270 of a compositional outcome on compositional predictors.

594 00:28:36.270 --> 00:28:39.750 So this was kind of an offshoot paper

595 00:28:39.750 --> 00:28:41.850 where we just developed this piece

596 00:28:41.850 --> 00:28:45.390 and if you look at compositional regression,

597 00:28:45.390 --> 00:28:50.160 most of the work has been done using Dirichlet models

598 00:28:50.160 --> 00:28:52.440 or log ratio transformations.

599 00:28:52.440 --> 00:28:55.343 So this was a different approach to that in the sense

600 00:28:55.343 --> 00:28:57.060 that it's both transformation free

601 00:28:57.060 --> 00:28:58.920 and it doesn't specify a whole distribution

602 00:28:58.920 --> 00:28:59.753 like the Dirichlet,

603 00:28:59.753 --> 00:29:02.040 it just uses a first moment as option.

604 00:29:02.040 --> 00:29:07.040 And we have an R-package to do a regression on composition,

605 00:29:07.470 --> 00:29:10.370 to do composition on composition regression called codalm.

606 00:29:12.150 --> 00:29:14.673 But going back to the verbal autopsy work,

607 00:29:16.050 --> 00:29:17.220 we have the loss functions

608 00:29:17.220 --> 00:29:19.173 for the labeled and unlabeled data,

609 00:29:20.220 --> 00:29:22.500 we do the negative pseudo likelihoods,

610 00:29:22.500 --> 00:29:26.103 put priors on the parameters and we get posterior inference.

611 00:29:27.780 --> 00:29:30.990 One last extension of the methodology

612 00:29:30.990 --> 00:29:33.780 was that there are multiple different

613 00:29:33.780 --> 00:29:35.970 verbal autopsy algorithms and there are papers

614 00:29:35.970 --> 00:29:38.700 where every new algorithm comes out and they say

615 00:29:38.700 --> 00:29:40.620 they're better than all the previous algorithms.

616 00:29:40.620 --> 00:29:44.190 And in practice, you never know which is the best algorithm.

617 00:29:44.190 --> 00:29:48.990 So we developed an ensemble method that takes in predictions

618 00:29:48.990 --> 00:29:53.760 from multiple algorithms, estimates classifier

619 00:29:53.760 --> 00:29:56.550 algorithm-specific misclassification rates

620 00:29:56.550 --> 00:30:00.270 and then they're connected to the unknown estimand.

621 00:30:00.270 --> 00:30:04.140 So we can show that it gives more weight

622 00:30:04.140 --> 00:30:06.900 to the more accurate algorithm in a data-driven way.

623 00:30:06.900 --> 00:30:10.380 And then you're not kind of,

624 00:30:10.380 --> 00:30:11.970 you don't have to make the choice

625 00:30:11.970 --> 00:30:13.950 of which is the best algorithm in advance.

626 00:30:13.950 --> 00:30:15.300 If you have multiple candidates,

627 00:30:15.300 --> 00:30:18.603 you can use multiple algorithms together.

628 00:30:22.560 --> 00:30:26.340 So we looked at some theoretical properties of the method.

629 00:30:26.340 --> 00:30:28.830 We have two log functions, one for the label data,

630 00:30:28.830 --> 00:30:31.080 one for the unlabeled data.

631 00:30:31.080 --> 00:30:31.913 The label data

632 00:30:31.913 --> 00:30:35.610 doesn't even feature the estimand, which is p,

633 00:30:35.610 --> 00:30:38.910 so it will, on its own, it cannot identify p.

634 00:30:38.910 --> 00:30:43.050 The unlabeled data only uses p through this quantity,

635 00:30:43.050 --> 00:30:44.190 M transpose p.

636 00:30:44.190 --> 00:30:47.640 So again, for different combinations of M and p,

637 00:30:47.640 --> 00:30:49.800 as long as this product is the same,

638 00:30:49.800 --> 00:30:52.680 it will never be able to identify p on its own.

639 00:30:52.680 --> 00:30:54.240 So each loss function on its own

640 00:30:54.240 --> 00:30:56.520 cannot identify through parameters.

641 00:30:56.520 --> 00:30:59.070 But using both the loss functions together,

642 00:30:59.070 --> 00:31:02.070 you can identify the estimand, T,

643 00:31:02.070 --> 00:31:06.360 and we were able to show that posterior has nice properties

644 00:31:06.360 --> 00:31:08.400 in terms of asymptotic normality

645 00:31:08.400 --> 00:31:10.500 and well calibrated interval estimate

646 00:31:10.500 --> 00:31:12.990 and near parametric concentration rates.

647 00:31:12.990 --> 00:31:16.320 And the theory also extends to the ensemble method

648 00:31:16.320 --> 00:31:19.470 and we use some approximations and we give sampler

649 00:31:19.470 --> 00:31:21.273 and theory holds for that.

650 00:31:24.150 --> 00:31:25.953 Some empirical validations,

651 00:31:27.450 --> 00:31:32.220 since we're estimating a probability vector,

652 00:31:32.220 --> 00:31:34.380 the common metric that is used is called

653 00:31:34.380 --> 00:31:37.800 this chance-corrected normalized absolute accuracy,

654 00:31:37.800 --> 00:31:40.743 which is basically a scaled L1 error,

655 00:31:41.670 --> 00:31:45.510 centered by the L1 error you would get if you had predicted

656 00:31:45.510 --> 00:31:46.740 the cause of death randomly.

657 00:31:46.740 --> 00:31:49.500 So this is the error if you predict randomly

658 00:31:49.500 --> 00:31:51.900 and then we look at how much improvement we get

659 00:31:51.900 --> 00:31:53.613 over random predictions.

660 00:31:56.790 --> 00:32:00.510 So this is an illustration of what happens if the data

661 00:32:00.510 --> 00:32:03.420 is not Dirichlet and you use Dirichlet distribution.

662 00:32:03.420 --> 00:32:05.070 So on the left-hand side,

663 00:32:05.070 --> 00:32:07.647 the data is generated from Dirichlet

664 00:32:07.647 --> 00:32:11.880 and we use both our method and the Dirichlet-based model

665 00:32:11.880 --> 00:32:13.650 and they both do well.

666 00:32:13.650 --> 00:32:14.670 On the right-hand side,

667 00:32:14.670 --> 00:32:17.490 the data is from an overdispersed Dirichlet

668 00:32:17.490 --> 00:32:19.530 and we use the Dirichlet in our model.

669 00:32:19.530 --> 00:32:22.080 And because our model doesn't specify a distribution,

670 00:32:22.080 --> 00:32:24.690 it just uses a first moment specification,

671 00:32:24.690 --> 00:32:27.820 it's much robust and has much higher accuracy

672 00:32:28.860 --> 00:32:31.657 than for the Dirichlet which becomes misspecified.

673 00:32:35.010 --> 00:32:37.020 And then we also did a bunch of evaluations

674 00:32:37.020 --> 00:32:38.400 using the PHMRC data.

675 00:32:38.400 --> 00:32:41.580 So what we did was we trained the classifiers

676 00:32:41.580 --> 00:32:44.370 on three of the countries leaving one country out

677 00:32:44.370 --> 00:32:47.460 and then used a slice of data from that left out country

678 00:32:47.460 --> 00:32:49.710 to estimate the misclassification rates,

679 00:32:49.710 --> 00:32:51.723 and then we apply our method.

680 00:32:54.600 --> 00:32:56.400 The green one is our method

681 00:32:56.400 --> 00:33:01.400 and the x axis is the sample size of the dataset

682 00:33:02.220 --> 00:33:04.154 used from the left out country

683 00:33:04.154 --> 00:33:06.930 to estimate the misclassification rates.

684 00:33:06.930 --> 00:33:10.650 The blue one is sort of the uncalibrated one,

685 00:33:10.650 --> 00:33:12.750 the red one is the one that is calibrated

686 00:33:12.750 --> 00:33:14.250 using the training data.

687 00:33:14.250 --> 00:33:17.760 So you can see that our method does better than both of them

688 00:33:17.760 --> 00:33:20.220 and the higher the sample size we use

689 00:33:20.220 --> 00:33:22.890 from the left out country of interest

690 00:33:22.890 --> 00:33:25.973 to estimate the misclassifications, the more accurate it is.

691 00:33:29.637 --> 00:33:31.440 And also one interesting aspect

692 00:33:31.440 --> 00:33:33.300 was that we looked at calibration

693 00:33:33.300 --> 00:33:35.700 using individual algorithms and the calibration

694 00:33:35.700 --> 00:33:37.440 using the ensemble one.

695 00:33:37.440 --> 00:33:40.380 And more often than not, the ensemble one,

696 00:33:40.380 --> 00:33:41.970 which is the orange one,

697 00:33:41.970 --> 00:33:45.570 tends to perform similar to the best performing algorithm,

698 00:33:45.570 --> 00:33:48.450 and the best performing algorithm can be very different

699 00:33:48.450 --> 00:33:49.530 across different countries.

700 00:33:49.530 --> 00:33:51.450 For example, in Mexico,

701 00:33:51.450 --> 00:33:54.120 InSilicoVA is one of the best performing algorithms,

702 00:33:54.120 --> 00:33:57.390 but in Tanzania, InSilicoVA was doing very poorly

703 00:33:57.390 --> 00:33:58.660 and then InterVA was one

704 00:33:59.499 --> 00:34:00.332 of the better performing algorithms.

705 00:34:00.332 --> 00:34:02.970 So the ensemble always tend to give more weights

706 00:34:02.970 --> 00:34:04.773 to more accurate algorithms.

707 00:34:07.380 --> 00:34:10.020 So this is an overview of what we did for Mozambique.

708 00:34:10.020 --> 00:34:13.920 So we had the unlabeled data with only verbal autopsies.

709 00:34:13.920 --> 00:34:16.230 We've passed it through two algorithms,

710 00:34:16.230 --> 00:34:20.520 InSilicoVA and Expert VA, to get the uncalibrated estimates.

711 00:34:20.520 --> 00:34:23.070 Then we had the label data with the MITS cause of death

712 00:34:23.070 --> 00:34:25.350 with which we estimated the misclassifications

713 00:34:25.350 --> 00:34:27.660 of those two algorithms

714 00:34:27.660 --> 00:34:30.450 and then we combine them in the ensemble method

715 00:34:30.450 --> 00:34:32.100 and getting calibrated estimates.

716 00:34:37.680 --> 00:34:39.900 Some results from Mozambique.

717 00:34:39.900 --> 00:34:41.520 We have two age groups,

718 00:34:41.520 --> 00:34:44.850 neonatal deaths, first four weeks,

719 00:34:44.850 --> 00:34:47.820 and children that's under five years.

720 00:34:47.820 --> 00:34:51.720 Two algorithms, seven causes of death for children,

721 00:34:51.720 --> 00:34:53.613 five causes of death for neonates.

722 00:34:54.690 --> 00:34:56.880 I'm going to just show the neonatal results here.

723 00:34:56.880 --> 00:35:00.540 So these are the misclassification matrices for neonates.

724 00:35:00.540 --> 00:35:03.060 And ideally, you would want the matrices

725 00:35:03.060 --> 00:35:04.890 to have large numbers on the diagonals

726 00:35:04.890 --> 00:35:06.840 because those are the correct matches

727 00:35:06.840 --> 00:35:08.910 and then small numbers on the off diagonals.

728 00:35:08.910 --> 00:35:09.930 But you don't see that,

729 00:35:09.930 --> 00:35:14.040 you see quite a bit of large numbers on the off diagonals.

730 00:35:14.040 --> 00:35:16.740 One thing that stands out is that

731 00:35:16.740 --> 00:35:20.490 if you look at prematurity, it has a very high sensitivity,

732 00:35:20.490 --> 00:35:21.750 close to 90%,

733 00:35:21.750 --> 00:35:25.110 which means that if the true cause is prematurity,

734 00:35:25.110 --> 00:35:28.050 the verbal autopsy correctly diagnoses it.

735 00:35:28.050 --> 00:35:30.960 But then it also has high false positives

736 00:35:30.960 --> 00:35:34.050 in the sense that if the true cause is infection,

737 00:35:34.050 --> 00:35:37.020 20% of time, it is assigned as prematurity.

738 00:35:37.020 --> 00:35:40.149 If the true cause is intrapartum related events,

739 00:35:40.149 --> 00:35:40.982 almost 30% of time,

740 00:35:40.982 --> 00:35:43.020 it's assigned to be prematurity and so on.

741 00:35:43.020 --> 00:35:46.170 So it tends to over count a lot of deaths

742 00:35:46.170 --> 00:35:48.480 from different causes as prematurity.

743 00:35:48.480 --> 00:35:51.540 So what would be the result after calibration

744 00:35:51.540 --> 00:35:54.240 is that the percentage of prematurity comes down.

745 00:35:54.240 --> 00:35:58.380 So this is the uncalibrated estimate of prematurity.

746 00:35:58.380 --> 00:36:00.780 This is the calibrated estimate of prematurity.

747 00:36:00.780 --> 00:36:02.130 You can see that it comes down

748 00:36:02.130 --> 00:36:04.980 because we can see in the data that there is a lot

749 00:36:04.980 --> 00:36:07.353 of over counting of prematurity deaths.

750 00:36:08.820 --> 00:36:12.093 So after calibration, it tends to come down quite a bit.

751 00:36:16.950 --> 00:36:21.510 And also, we looked at the model estimated sensitivities

752 00:36:21.510 --> 00:36:23.550 using both the single cause

753 00:36:23.550 --> 00:36:26.043 and the compositional cause of the data.

754 00:36:27.180 --> 00:36:29.460 So this is the difference in the sensitivities

755 00:36:29.460 --> 00:36:32.550 and you can see that using the compositional cause of death,

756 00:36:32.550 --> 00:36:36.330 you'll always get a higher match because it kind of uses

757 00:36:36.330 --> 00:36:38.580 information for multiple causes and stuff

758 00:36:38.580 --> 00:36:40.530 just considering the top cause.

759 00:36:40.530 --> 00:36:42.660 And so it generally leads to better matching

760 00:36:42.660 --> 00:36:46.263 between the verbal autopsy and the minimal tissue sampling.

761 00:36:49.440 --> 00:36:50.730 Some ongoing work.

762 00:36:50.730 --> 00:36:53.010 So when we did this for Mozambique,

763 00:36:53.010 --> 00:36:56.820 there was very little amount of payer data.

764 00:36:56.820 --> 00:36:59.070 So even though the data was for seven countries,

765 00:36:59.070 --> 00:37:00.990 we kind of merged them together

766 00:37:00.990 --> 00:37:03.900 and estimated the misclassification rates.

767 00:37:03.900 --> 00:37:06.600 Now we have more data coming in for those countries

768 00:37:06.600 --> 00:37:07.920 so we have a chance to assess

769 00:37:07.920 --> 00:37:11.610 whether the misclassification rates vary by country

770 00:37:11.610 --> 00:37:12.450 because if they do,

771 00:37:12.450 --> 00:37:14.920 we should model the misclassification rates

772 00:37:16.980 --> 00:37:19.173 in a way that's specific to each country.

773 00:37:21.420 --> 00:37:25.890 So these are the misclassification rates now

774 00:37:25.890 --> 00:37:27.270 resolved by country.

775 00:37:27.270 --> 00:37:30.030 So there are six countries, Bangladesh, Ethiopia,

776 00:37:30.030 --> 00:37:32.283 Kenya, Mali, Mozambique and Sierra Leone.

777 00:37:34.560 --> 00:37:35.760 You can see the estimates.

778 00:37:35.760 --> 00:37:37.260 These are the empirical estimates

779 00:37:37.260 --> 00:37:40.020 and the confidence intervals for each country.

780 00:37:40.020 --> 00:37:42.090 And the horizontal black line

781 00:37:42.090 --> 00:37:43.980 is what the pooled estimate looks like.

782 00:37:43.980 --> 00:37:48.660 So you can see that there is for some causes like here,

783 00:37:48.660 --> 00:37:51.240 there is not a variability across countries.

784 00:37:51.240 --> 00:37:55.353 But then for some other cause payers like say here,

785 00:37:56.250 --> 00:37:59.641 there's quite a bit of variability across countries.

786 00:37:59.641 --> 00:38:03.390 And so now that we are getting more data,

787 00:38:03.390 --> 00:38:05.400 the next step for the project

788 00:38:05.400 --> 00:38:08.790 is to estimate country-specific misclassification rates.

789 00:38:08.790 --> 00:38:12.450 The issue however is that even with more data,

790 00:38:12.450 --> 00:38:16.530 there is, I think, around 600 cases here for six countries,

791 00:38:16.530 --> 00:38:19.560 which is approximately 100 case per country.

792 00:38:19.560 --> 00:38:22.680 And there are 25 cells of the misclassification matrix.

793 00:38:22.680 --> 00:38:24.720 So that's like four cases per cell,

794 00:38:24.720 --> 00:38:27.450 so that's clearly not enough to do separate

795 00:38:27.450 --> 00:38:29.670 country specific models.

796 00:38:29.670 --> 00:38:32.220 So we'd have to kind of do

797 00:38:32.220 --> 00:38:34.950 a sort of a borrowing of information

798 00:38:34.950 --> 00:38:37.920 both across the rows and columns of the matrix

799 00:38:37.920 --> 00:38:40.083 but also across different countries.

800 00:38:42.000 --> 00:38:45.480 So what we do first is first, we kind of borrow information

801 00:38:45.480 --> 00:38:48.540 across the rows and columns of the matrix.

802 00:38:48.540 --> 00:38:52.200 And to do this, we start with a,

803 00:38:52.200 --> 00:38:54.510 instead of an unstructured misclassification matrix

804 00:38:54.510 --> 00:38:56.910 where we estimated each cell separately,

805 00:38:56.910 --> 00:39:00.120 we start with a structured misclassification matrix

806 00:39:00.120 --> 00:39:01.680 using two basic mechanisms.

807 00:39:01.680 --> 00:39:06.680 So we say that a classifier operates using two mechanisms,

808 00:39:07.260 --> 00:39:11.520 for a given cause, it can either match that cause

809 00:39:11.520 --> 00:39:14.760 and we call that an intrinsic accuracy

810 00:39:14.760 --> 00:39:17.550 and that matching probability will be different

811 00:39:17.550 --> 00:39:20.250 for different causes, so there are three causes here,

812 00:39:20.250 --> 00:39:21.330 and you can see

813 00:39:21.330 --> 00:39:23.940 that the matching probability can be different.

814 00:39:23.940 --> 00:39:25.950 If it doesn't match the true cause,

815 00:39:25.950 --> 00:39:28.860 then it randomly distributes its prediction

816 00:39:28.860 --> 00:39:30.750 to the other causes

817 00:39:30.750 --> 00:39:35.750 and that random distribution will also have some weights,

818 00:39:35.970 --> 00:39:38.190 and those we call the systematic bias

819 00:39:38.190 --> 00:39:39.570 or the pool of the classifier.

820 00:39:39.570 --> 00:39:41.550 So if it's not matching,

821 00:39:41.550 --> 00:39:45.780 we saw that it'll often assign a cause to prematurity

822 00:39:45.780 --> 00:39:47.730 regardless of what the true cause is.

823 00:39:47.730 --> 00:39:50.550 So that's kind of the basis for this model.

824 00:39:50.550 --> 00:39:51.810 And if you have this model,

825 00:39:51.810 --> 00:39:56.230 we kind of rearrange these three bars here

826 00:39:57.420 --> 00:39:59.370 and then we put in the circle from there.

827 00:39:59.370 --> 00:40:03.120 And these will give you the misclassification priorities.

828 00:40:03.120 --> 00:40:08.120 So we can write each of the misclassification probabilities

829 00:40:08.340 --> 00:40:12.630 in terms of just these six parameters and we can do the same

830 00:40:12.630 --> 00:40:16.890 for the green cause and for the blue cause.

831 00:40:16.890 --> 00:40:21.570 And so basically, these are the nine misclassification rates

832 00:40:21.570 --> 00:40:23.300 written in terms of the six parameters.

833 00:40:23.300 --> 00:40:25.680 So this is not that much of a dimension reduction

834 00:40:25.680 --> 00:40:27.300 if there are three causes,

835 00:40:27.300 --> 00:40:30.213 but if there are in general C causes,

836 00:40:31.710 --> 00:40:34.470 this model for misclassification matrix will only have

837 00:40:34.470 --> 00:40:38.640 2C - 1 parameters as opposed to C square parameters.

838 00:40:38.640 --> 00:40:43.190 So in practice, we use seven causes for children

839 00:40:43.190 --> 00:40:44.023 and five causes for neonates,

840 00:40:44.023 --> 00:40:46.310 so this leads to a lot of dimension reduction.

841 00:40:48.690 --> 00:40:52.500 And one of the justification

842 00:40:52.500 --> 00:40:54.360 for this dimension reduced model

843 00:40:54.360 --> 00:40:59.070 is that if this model is true then the misclassification

844 00:40:59.070 --> 00:41:01.380 into different causes,

845 00:41:01.380 --> 00:41:05.220 the odds of misclassification into two causes, j and k,

846 00:41:05.220 --> 00:41:08.040 will not depend on what the true cause is.

847 00:41:08.040 --> 00:41:09.720 And we do see that in the data.

848 00:41:09.720 --> 00:41:13.470 So these are different cause payers, j and k,

849 00:41:13.470 --> 00:41:16.920 and these are the odds for what the true cause is.

850 00:41:16.920 --> 00:41:19.890 So we are plotting the misclassification rates,

851 00:41:19.890 --> 00:41:22.290 mij over mik.

852 00:41:22.290 --> 00:41:23.550 So this is j and k

853 00:41:23.550 --> 00:41:25.680 and the colors here give you i.

854 00:41:25.680 --> 00:41:28.470 So you do see that they do not vary

855 00:41:28.470 --> 00:41:30.030 for different choices of i,

856 00:41:30.030 --> 00:41:32.037 it only is specific to j and k,

857 00:41:32.037 --> 00:41:35.730 and that's an equivalent characterization

858 00:41:35.730 --> 00:41:38.970 of that systematic preference

859 00:41:38.970 --> 00:41:41.070 and intrinsic accuracy model that we have,

860 00:41:41.070 --> 00:41:43.203 so we do see that reflected in the data.

861 00:41:44.040 --> 00:41:49.040 But we don't have that as the fixed model we have.

862 00:41:49.230 --> 00:41:50.520 So this is the best model.

863 00:41:50.520 --> 00:41:53.997 We allow some diversion or shrinkage towards it

864 00:41:53.997 --> 00:41:55.800 and there's a tuning parameter.

865 00:41:55.800 --> 00:41:58.230 So then we get the homogeneous model

866 00:41:58.230 --> 00:42:01.260 and then we have a diversion from the homogeneous model

867 00:42:01.260 --> 00:42:02.730 to get country specific model.

868 00:42:02.730 --> 00:42:04.380 So that's the broad idea,

869 00:42:04.380 --> 00:42:06.810 I won't go into the modeling details.

870 00:42:06.810 --> 00:42:08.760 And these are the predictions

871 00:42:08.760 --> 00:42:10.563 using the country specific model.

872 00:42:12.750 --> 00:42:15.270 I won't go into details here, but there are many cases,

873 00:42:15.270 --> 00:42:16.620 for example, take it here,

874 00:42:16.620 --> 00:42:18.393 star is the empirical rate,

875 00:42:19.440 --> 00:42:24.180 angle is the heterogeneous model.

876 00:42:24.180 --> 00:42:25.650 And you can see it does much better

877 00:42:25.650 --> 00:42:29.524 than the horizontal line, which is the homogeneous model.

878 00:42:29.524 --> 00:42:34.163 And we do see it throughout the classification rates.

879 00:42:35.850 --> 00:42:37.620 These are the estimates for Bangladesh.

880 00:42:37.620 --> 00:42:41.030 So the red density is the pooled estimate

881 00:42:41.030 --> 00:42:42.780 of the homogeneous estimate.

882 00:42:42.780 --> 00:42:45.543 The blue density is the Bangladesh specific estimate.

883 00:42:48.090 --> 00:42:49.590 The dotted vertical line

884 00:42:49.590 --> 00:42:51.657 is the empirical estimate for Bangladesh

885 00:42:51.657 --> 00:42:53.430 and the solid vertical line

886 00:42:53.430 --> 00:42:56.250 is the pooled empirical estimate.

887 00:42:56.250 --> 00:42:58.620 So you can see that as we get

888 00:42:58.620 --> 00:43:00.600 more and more data from Bangladesh,

889 00:43:00.600 --> 00:43:02.670 the country specific estimate moves away

890 00:43:02.670 --> 00:43:03.780 from the pooled estimate

891 00:43:03.780 --> 00:43:06.090 towards the country specific estimate.

892 00:43:06.090 --> 00:43:11.090 So that's basically the hope is going forward,

893 00:43:11.790 --> 00:43:14.220 we will have much more data within each country

894 00:43:14.220 --> 00:43:16.410 and we'll have estimates that are much closer

895 00:43:16.410 --> 00:43:20.013 to the dotted lines than the solid lines.

896 00:43:21.810 --> 00:43:22.950 So that's the summary.

897 00:43:22.950 --> 00:43:26.310 So in general, these cause of death classifiers

898 00:43:26.310 --> 00:43:27.810 are super inaccurate.

899 00:43:27.810 --> 00:43:30.840 So we need to calibrate for that and we have limited data

900 00:43:30.840 --> 00:43:32.490 to estimate their inaccuracy,

901 00:43:32.490 --> 00:43:34.773 so we calibrate them innovation way.

902 00:43:36.240 --> 00:43:38.790 The methods give probabilistic cause of death

903 00:43:38.790 --> 00:43:40.350 instead of categorical cause of death.

904 00:43:40.350 --> 00:43:42.960 So we develop a generalized Bayes approach

905 00:43:42.960 --> 00:43:45.060 that is equivalent to a multinomial model

906 00:43:45.060 --> 00:43:47.040 if the data is categorical.

907 00:43:47.040 --> 00:43:50.370 But if it's not categorical, it becomes a pseudo likelihood

908 00:43:50.370 --> 00:43:53.550 Bayesian approach for compositional data

909 00:43:53.550 --> 00:43:57.000 and that allows zeroes and ones in the data

910 00:43:57.000 --> 00:44:01.023 and is not kind of dependent on the model specification.

911 00:44:02.490 --> 00:44:04.830 And then it kind of led to this independent development

912 00:44:04.830 --> 00:44:09.020 of the composition on composition regression.

913 00:44:09.020 --> 00:44:10.216 Some papers and software.

914 00:44:10.216 --> 00:44:13.100 So the single cause paper was the first one,

915 00:44:13.100 --> 00:44:16.934 then we extend it to compositional data

916 00:44:16.934 --> 00:44:18.991 and develop the theory for it.

917 00:44:18.991 --> 00:44:22.394 The package for calibration is available on GitHub

918 00:44:22.394 --> 00:44:24.720 and then the composition on composition regression

919 00:44:24.720 --> 00:44:25.980 were the separate piece

920 00:44:25.980 --> 00:44:30.360 and we have the coda linear model package for it on CRAN.

921 00:44:30.360 --> 00:44:32.460 And then we use this approach

922 00:44:32.460 --> 00:44:34.840 to produce calibration estimates

923 00:44:36.372 --> 00:44:38.970 for neonate and children deaths in Mozambique

924 00:44:38.970 --> 00:44:41.490 which were published in the last three papers.

925 00:44:41.490 --> 00:44:42.323 Thank you.

926 00:44:51.390 --> 00:44:52.950 <v ->Questions? Yes.</v>

927 00:44:52.950 --> 00:44:54.990 <v ->So I just had a quick question 'cause you were saying</v>

928 00:44:54.990 --> 00:44:58.110 the model basically looks at the symptoms

929 00:44:58.110 --> 00:45:00.000 that'll be able to predict which it would be.

930 00:45:00.000 --> 00:45:03.660 Does it also factor in what diseases and stuff

931 00:45:03.660 --> 00:45:07.140 are most common in those areas or does it kind of just-

932 00:45:07.140 --> 00:45:09.360 <v ->Oh, very good question.</v>

933 00:45:09.360 --> 00:45:12.210 It does factor it in but in a very crude way

934 00:45:12.210 --> 00:45:14.280 in the sense that the models have some settings

935 00:45:14.280 --> 00:45:18.360 called like high malaria, low malaria or high HIV, low HIV.

936 00:45:18.360 --> 00:45:20.850 So depending on which country you're running it,

937 00:45:20.850 --> 00:45:24.120 you will set the setting to like high HIV country

938 00:45:24.120 --> 00:45:26.550 or low HIV country, the same for malaria,

939 00:45:26.550 --> 00:45:29.640 but it doesn't do anything beyond that,

940 00:45:29.640 --> 00:45:31.473 so only at a very close level.

941 00:45:34.350 --> 00:45:35.400 <v ->Causes of death or.</v>

942 00:45:36.870 --> 00:45:39.720 <v ->So the ICD-10 classification</v>

943 00:45:39.720 --> 00:45:42.480 will have around 30 plus causes of death

944 00:45:42.480 --> 00:45:44.070 for children's and neonates,

945 00:45:44.070 --> 00:45:45.753 I think much more for adults.

946 00:45:46.620 --> 00:45:48.420 There are no MITS for adults.

947 00:45:48.420 --> 00:45:50.700 MITS was only done for children's and neonates,

948 00:45:50.700 --> 00:45:53.343 only now adult MITS are being started,

949 00:45:54.330 --> 00:45:57.330 but we have to kind of group them into broader categories

950 00:45:57.330 --> 00:45:58.980 because if you have 30 causes,

951 00:45:58.980 --> 00:46:01.500 your misclassification matrix will be 30 times 30.

952 00:46:01.500 --> 00:46:05.040 So we don't have the data to do estimation

953 00:46:05.040 --> 00:46:06.300 at that fine resolution.

954 00:46:06.300 --> 00:46:08.220 So we group them into broader categories.

955 00:46:08.220 --> 00:46:10.950 So seven for children, five for new neonates.

956 00:46:10.950 --> 00:46:13.770 <v ->Is one of the categories, I have no idea,</v>

957 00:46:13.770 --> 00:46:15.210 it is totally unknown.

958 00:46:15.210 --> 00:46:18.450 And if so, is that different from the uniform distribution

959 00:46:18.450 --> 00:46:20.373 across causes of death?

960 00:46:21.240 --> 00:46:22.680 <v ->That would be the uniform distribution.</v>

961 00:46:22.680 --> 00:46:24.810 There is no category which is, I have no idea,

962 00:46:24.810 --> 00:46:27.720 but it'll be probably reflected in a score that is very flat

963 00:46:27.720 --> 00:46:29.550 across the causes.

964 00:46:29.550 --> 00:46:32.040 <v ->If you think there are seven causes of death</v>

965 00:46:32.040 --> 00:46:33.540 and I'm working with the same dataset

966 00:46:33.540 --> 00:46:36.180 and I think there are 100 causes of death,

967 00:46:36.180 --> 00:46:39.420 will there be substantial differences in our marginal

968 00:46:39.420 --> 00:46:41.340 estimates of probability?

969 00:46:41.340 --> 00:46:44.820 Because our uniform posteriors

970 00:46:44.820 --> 00:46:48.030 place such different amounts of mass across the say

971 00:46:48.030 --> 00:46:50.820 30 versus 100 causes of death.

972 00:46:50.820 --> 00:46:52.540 <v ->Yes, there will be differences</v>

973 00:46:54.150 --> 00:46:58.380 and even when we are aggregating from the 30 causes

974 00:46:58.380 --> 00:47:01.860 to seven causes, the assumption is that within each category

975 00:47:01.860 --> 00:47:03.930 the misclassification rates are homogeneous

976 00:47:03.930 --> 00:47:05.130 within the finer category.

977 00:47:05.130 --> 00:47:07.860 So that is an assumption that we're working with.

978 00:47:07.860 --> 00:47:09.910 So definitely, there will be differences.

979 00:47:10.890 --> 00:47:11.723 <v ->Thank you.</v>

980 00:47:16.380 --> 00:47:18.630 <v ->I have one more question.</v>

981 00:47:21.690 --> 00:47:23.100 I'll ask a philosophical question

982 00:47:23.100 --> 00:47:23.933 if I may. <v ->Sure, yeah.</v>

983 00:47:23.933 --> 00:47:24.957 <v ->You commented,</v>

984 00:47:26.010 --> 00:47:27.180 I don't know, about halfway through,

985 00:47:27.180 --> 00:47:31.500 about how statisticians are working on a thing.

986 00:47:31.500 --> 00:47:34.020 Computer scientists are working on the same thing.

987 00:47:34.020 --> 00:47:35.570 There's a third group I forget.

988 00:47:37.320 --> 00:47:38.870 And nobody talks to each other.

989 00:47:39.930 --> 00:47:41.463 Now, many of us are,

990 00:47:42.330 --> 00:47:43.580 many of the students here

991 00:47:44.482 --> 00:47:47.032 are within the data science track of biostatistics.

992 00:47:48.660 --> 00:47:50.523 By the way, love your Twitter handle.

993 00:47:52.230 --> 00:47:55.890 But yeah, so how do we bridge those things

994 00:47:55.890 --> 00:47:57.450 that we take advantage of these things

995 00:47:57.450 --> 00:48:00.213 and it's not three separate versions of the same thing?

996 00:48:01.170 --> 00:48:04.320 <v ->I don't know if there's a systematic way.</v>

997 00:48:04.320 --> 00:48:07.530 Honestly, I came to know about much of the literature

998 00:48:07.530 --> 00:48:08.550 going through the revisions

999 00:48:08.550 --> 00:48:10.680 and one of the reviewer associate editors said

1000 00:48:10.680 --> 00:48:13.620 there is a lot of work here in the econometrics literature,

1001 00:48:13.620 --> 00:48:14.760 you should take a look.

1002 00:48:14.760 --> 00:48:15.720 And that's kind of the value

1003 00:48:15.720 --> 00:48:17.490 of the peer review system I guess.

1004 00:48:17.490 --> 00:48:20.340 And so we looked at it and yes, there was a lot of work

1005 00:48:20.340 --> 00:48:22.260 and they just called it different things

1006 00:48:22.260 --> 00:48:23.250 and so I had no idea

1007 00:48:23.250 --> 00:48:25.800 when I was searching for that in the literature.

1008 00:48:25.800 --> 00:48:28.560 And we did see the Victor Chernozhukov paper

1009 00:48:28.560 --> 00:48:30.150 I think is in "Journal of Economics,"

1010 00:48:30.150 --> 00:48:32.610 but it's basically an asymptotic statistics paper.

1011 00:48:32.610 --> 00:48:35.640 It kind of shows that these generalized Bayes stuff,

1012 00:48:35.640 --> 00:48:38.400 which they call as Laplace-type estimators,

1013 00:48:38.400 --> 00:48:39.990 has all these nice properties

1014 00:48:39.990 --> 00:48:42.140 that a standard vision posterior will have.

1015 00:48:43.200 --> 00:48:46.410 But yeah, I think talking to more people

1016 00:48:46.410 --> 00:48:48.930 and like interacting and telling about your work

1017 00:48:48.930 --> 00:48:49.763 will kind of, 1018 00:48:49.763 --> 00:48:52.320 and someone will say that, oh yeah, I do something similar.

1019 00:48:52.320 --> 00:48:54.600 You should look at this paper,

1020 00:48:54.600 --> 00:48:56.754 it's probably. <v ->Hopefully Twitter helps.</v>

1021 00:48:56.754 --> 00:48:57.587 <v ->Sorry?</v>

1022 00:48:57.587 --> 00:48:58.420 <v ->Hopefully Twitter helps.</v>

1023 00:48:58.420 --> 00:49:00.300 <v ->Yeah, yeah, definitely.</v>

1024 00:49:00.300 --> 00:49:02.340 Engagement through any like in-person

1025 00:49:02.340 --> 00:49:05.463 or social media platform would be useful, yeah.

1026 00:49:07.530 --> 00:49:08.460 <v ->All right, well thanks so much.</v>

1027 00:49:08.460 --> 00:49:11.679 I think we're out of time so we'll stop it there.

1028 00:49:11.679 --> 00:49:14.790 (attendant muttering indistinctly)

1029 00:49:14.790 --> 00:49:16.590 Hope everybody has a wonderful fall break.

1030 00:49:16.590 --> 00:49:17.640 See you next week.

1031 00:49:19.167 --> 00:49:23.584 (attendants chattering indistinctly)

1032 00:49:36.727 --> 00:49:37.923 <v Learner>The other organizer.</v>

1033 00:49:37.923 --> 00:49:39.398 (learner muttering indistinctly)

1034 00:49:39.398 --> 00:49:43.815 (attendants chattering indistinctly)

1035 00:49:52.604 --> 00:49:54.760 <v ->Or maybe because they're susceptible.</v>

1036 00:49:54.760 --> 00:49:59.177 (attendants chattering indistinctly)

1037 00:50:04.226 --> 00:50:06.327 <v ->Thank you. Anyone else need to sign in?</v>

1038 00:50:06.327 --> 00:50:10.744 (attendants chattering indistinctly)

1039 00:50:19.104 --> 00:50:20.966 <v ->Infection but they're also premature babies.</v>

1040 00:50:20.966 --> 00:50:25.383 (attendants chattering indistinctly)

1041 00:50:30.104 --> 00:50:31.890 <v ->Premature, but also it's that</v>

1042 00:50:31.890 --> 00:50:33.506 it's not a distinct.

1043 00:50:33.506 --> 00:50:35.160 (attendants chattering indistinctly)

1044 00:50:35.160 --> 00:50:38.070 <v ->Cause of death is very blurry in this day.</v>

1045 00:50:38.070 --> 00:50:40.414 <v ->Is that part of why like.</v>

1046 00:50:40.414 --> 00:50:44.831 (attendants chattering indistinctly)

1047 00:50:46.157 --> 00:50:48.057 <v ->'Cause a symptom given cause session</v>

1048 00:50:49.011 --> 00:50:50.520 with that much of variation across country.

1049 00:50:50.520 --> 00:50:51.548 <v Learner>Cause.</v>

1050 00:50:51.548 --> 00:50:52.645 (learner muttering indistinctly)

1051 00:50:52.645 --> 00:50:53.790 Cause.

1052 00:50:53.790 --> 00:50:57.614 <v ->Reporting depends on who is answering.</v>

1053 00:50:57.614 --> 00:51:02.031 (attendants chattering indistinctly)

1054 00:51:03.810 --> 00:51:05.400 <v ->You need to go next.</v>

1055 00:51:05.400 --> 00:51:06.233 <v ->Back to.</v>

1056 00:51:09.026 --> 00:51:10.347 <v ->I guess, yeah.</v>

1057 00:51:10.347 --> 00:51:11.795 You need one of us to let you.

1058 00:51:11.795 --> 00:51:14.062 (lecturer muttering indistinctly)

1059 00:51:14.062 --> 00:51:15.929 <v ->It might be a short answer.</v>

1060 00:51:15.929 --> 00:51:16.762 Yeah, and it's short answer.

1061 00:51:16.762 --> 00:51:20.030 (attendants chattering indistinctly)

1062 00:51:20.030 --> 00:51:22.216 <v ->I don't have to, will you? (laughs)</v>