WEBVTT

1 00:00:00.000 --> 00:00:02.057 <v Wayne>We introduce Dr. Alex Kaizer.</v>

2 00:00:04.350 --> 00:00:06.976 Dr. Kaizer is an assistant professor

3 00:00:06.976 --> 00:00:10.380 in the Department of Biostatistics and Informatics,

4 00:00:10.380 --> 00:00:11.640 and he's a faculty member

5 00:00:11.640 --> 00:00:14.280 in the Center for Innovative Design and Analysis

6 00:00:14.280 --> 00:00:18.780 at the University of Colorado Medical Campus.

7 00:00:18.780 --> 00:00:21.180 He's passionate about translational research

8 00:00:21.180 --> 00:00:23.817 and the development of normal

9 00:00:23.817 --> 00:00:25.260 and at clinical trial designs

10 00:00:25.260 --> 00:00:26.093 that the more efficient that they

11 00:00:27.458 --> 00:00:30.570 and factually utilize available resources

12 00:00:30.570 --> 00:00:33.660 including past trails and past studies.

13 00:00:33.660 --> 00:00:36.610 And Dr. Kaizer strives to translate

14 00:00:36.610 --> 00:00:40.530 (indistinct) topics into understandable material

15 00:00:40.530 --> 00:00:42.240 that is more than just a mask

16 00:00:42.240 --> 00:00:43.740 and something we can appropriate

17 00:00:43.740 --> 00:00:46.350 and utilize in our daily lives and research.

18 00:00:46.350 --> 00:00:49.593 Now let's welcome Dr. Kaizer.

19 00:00:52.740 --> 00:00:53.573 <v Alex>Thank you Wayne.</v>

20 00:00:53.573 --> 00:00:57.000 So apologies for my own technical difficulties today,

21 00:00:57.000 --> 00:00:58.650 but I'm going to be presenting on

22 00:00:58.650 --> 00:01:01.620 this idea of a sequential basket trial design

23 00:01:01.620 --> 00:01:03.600 based on multi-source exchangeability

24 00:01:03.600 --> 00:01:05.640 with predictive probability monitoring.

25 00:01:05.640 --> 00:01:08.790 And that is admittedly quite the mouthful

26 00:01:08.790 --> 00:01:10.830 and I'm hoping throughout this presentation

27 00:01:10.830 --> 00:01:13.020 to break down each of these concepts

28 00:01:13.020 --> 00:01:16.500 and ideas building upon them sort of until we have this

29 00:01:16.500 --> 00:01:19.383 cumulative effect that represents this title today.

30 00:01:20.760 --> 00:01:22.830 Before jumping into everything though,

31 00:01:22.830 --> 00:01:24.510 I do wanna make a few acknowledgements.

32 00:01:24.510 --> 00:01:26.310 This paper was actually published

33 00:01:26.310 --> 00:01:28.260 just at the end of this past summer in PLOS ONE,

34 00:01:28.260 --> 00:01:31.140 and so if you're interested in more of the technical details

35 00:01:31.140 --> 00:01:32.760 or additional simulation examples

36 00:01:32.760 --> 00:01:34.860 and things beyond what I present today,

37 00:01:34.860 --> 00:01:36.000 I include this paper here

38 00:01:36.000 --> 00:01:38.520 and we'll also have it up again at the very end of my talk

39 00:01:38.520 --> 00:01:40.320 just for reference.

40 00:01:40.320 --> 00:01:42.420 Also acknowledgement to Dr. Nan Chen

41 00:01:42.420 --> 00:01:44.190 who helped with some of the initial coding

42 00:01:44.190 --> 00:01:46.773 of some of these methods and approaches.

43 00:01:49.800 --> 00:01:52.470 So to set the context for my seminar today,

44 00:01:52.470 --> 00:01:54.030 I want to think here about

45 00:01:54.030 --> 00:01:56.040 this move towards precision medicine generally,

46 00:01:56.040 --> 00:01:58.920 but especially in the context of oncology.

47 00:01:58.920 --> 00:02:01.920 And so within oncology, like many other disciplines,

48 00:02:01.920 --> 00:02:03.390 when we design research studies,

49 00:02:03.390 --> 00:02:05.460 we often design these for a particular,

50 00:02:05.460 --> 00:02:06.870 what we might call a histology

51 00:02:06.870 --> 00:02:09.360 or an indication or a disease.

52 00:02:09.360 --> 00:02:10.507 So for example, we might say,

53 00:02:10.507 --> 00:02:12.600 "Well, I have a treatment or intervention

54 00:02:12.600 --> 00:02:15.480 which I hope or think will work in lung cancer,

55 00:02:15.480 --> 00:02:17.580 therefore I'm going to design

56 00:02:17.580 --> 00:02:20.700 and enroll in the study for lung cancer."

57 00:02:20.700 --> 00:02:23.850 Now this represents a very standard way

58 00:02:23.850 --> 00:02:25.920 that we do clinical trial design where we try to

59 00:02:25.920 --> 00:02:27.330 really rigorously define

60 00:02:27.330 --> 00:02:30.960 and limitedly define what our scope is.

61 00:02:30.960 --> 00:02:31.920 Now within oncology,

62 00:02:31.920 --> 00:02:34.140 we've had some exciting scientific developments

63 00:02:34.140 --> 00:02:36.030 over the past few decades.

64 00:02:36.030 --> 00:02:38.070 So now instead of seeing cancer as just

65 00:02:38.070 --> 00:02:40.410 based on the site like you have a lung cancer

66 00:02:40.410 --> 00:02:41.970 or a prostate cancer,

67 00:02:41.970 --> 00:02:44.850 we actually have identified that we can partition cancers

68 00:02:44.850 --> 00:02:47.850 into many small molecular subtypes.

69 00:02:47.850 --> 00:02:49.350 And further, we've actually been able to

70 00:02:49.350 --> 00:02:52.050 leverage this information by being able to say that

71 00:02:52.050 --> 00:02:53.940 what we thought of as a holistic lung cancer

72 00:02:53.940 --> 00:02:57.000 isn't just one type of disease,

73 00:02:57.000 --> 00:02:58.770 we can actually develop therapies that we hope to

74 00:02:58.770 --> 00:03:02.160 target some of these differences in genetic alterations.

75 00:03:02.160 --> 00:03:04.620 And this really gets to that idea of precision medicine that

76 00:03:04.620 --> 00:03:06.900 instead of throwing a treatment at someone

77 00:03:06.900 --> 00:03:08.070 where we think it should work

78 00:03:08.070 --> 00:03:10.350 or it has worked in some people on average,

79 00:03:10.350 --> 00:03:12.570 hopefully we can really target the intervention

80 00:03:12.570 --> 00:03:14.400 based off of some signal

81 00:03:14.400 --> 00:03:17.310 or some indication like a biomarker or a genotype

82 00:03:17.310 --> 00:03:20.370 that we actually hope could respond more ideally

83 00:03:20.370 --> 00:03:22.200 to that intervention.

84 00:03:22.200 --> 00:03:25.500 Now what's really interesting about this as well

85 00:03:25.500 --> 00:03:27.840 is that there could be a potential for heterogeneity

86 00:03:27.840 --> 00:03:30.420 in this treatment benefit by indication.

87 00:03:30.420 --> 00:03:32.580 And what I mean by that is once we've identified

88 00:03:32.580 --> 00:03:34.740 that there's these different genetic alterations,

89 00:03:34.740 --> 00:03:37.230 we've actually discovered that these alterations

90 00:03:37.230 --> 00:03:40.770 aren't necessarily unique to one site of cancer.

91 00:03:40.770 --> 00:03:43.080 For example, we may identify a genetic alteration

92 00:03:43.080 --> 00:03:45.600 in the lung that also is present in the prostate, liver,

93 00:03:45.600 --> 00:03:49.170 and kidney in some of those types of cancer.

94 00:03:49.170 --> 00:03:50.490 Now the challenge here though is that

95 00:03:50.490 --> 00:03:53.340 even though we have the same driver hypothetically

96 00:03:53.340 --> 00:03:56.280 based on our clinical or scientific hypothesis

97 00:03:56.280 --> 00:03:58.560 of that potential benefit for a treatment we've designed

98 00:03:58.560 --> 00:03:59.790 to address it,

99 00:03:59.790 --> 00:04:01.440 there's still may be important differences

100 00:04:01.440 --> 00:04:02.340 that we don't know about

101 00:04:02.340 --> 00:04:04.980 or have yet to account for based off of each site.

102 00:04:04.980 --> 00:04:07.110 So what may have worked actually really well in the lung

103 00:04:07.110 --> 00:04:09.030 for one given mutation,

104 00:04:09.030 --> 00:04:12.090 even for that same mutation, let's say present in the liver,

105 00:04:12.090 --> 00:04:13.170 may not work as well.

106 00:04:13.170 --> 00:04:15.990 And that's that idea of heterogeneity and treatment benefit.

107 00:04:15.990 --> 00:04:18.270 That we can have different levels of response

108 00:04:18.270 --> 00:04:20.793 across different sites or groups of individuals.

109 00:04:22.950 --> 00:04:24.270 Now the cool thing I think here

110 00:04:24.270 --> 00:04:26.880 from the statistical perspective is that the scientific

111 00:04:26.880 --> 00:04:27.990 and clinical advancements

112 00:04:27.990 --> 00:04:30.810 have also led to the revolution and statistical

113 00:04:30.810 --> 00:04:34.170 and clinical design challenges and approaches.

114 00:04:34.170 --> 00:04:36.390 And of course that's the sweet spot that I work at.

115 00:04:36.390 --> 00:04:37.350 I know many of you

116 00:04:37.350 --> 00:04:38.760 and especially students are training

117 00:04:38.760 --> 00:04:40.140 and studying to work in this area

118 00:04:40.140 --> 00:04:42.870 to collaborate with scientific and clinical researchers

119 00:04:42.870 --> 00:04:45.120 and leaders to translate those results

120 00:04:45.120 --> 00:04:46.650 in statistically meaningful ways

121 00:04:46.650 --> 00:04:49.380 and to potentially design trials or studies

122 00:04:49.380 --> 00:04:52.830 that really target these questions and hypotheses.

123 00:04:52.830 --> 00:04:55.950 Now specifically in this talk today,

124 00:04:55.950 --> 00:04:57.720 I'm going to focus on this idea of

125 00:04:57.720 --> 00:05:00.510 a master protocol design or evolution.

126 00:05:00.510 --> 00:05:02.130 And these provide a flexible approach

127 00:05:02.130 --> 00:05:04.680 to the design of trials with multiple indications,

128 00:05:04.680 --> 00:05:06.120 but they do have their own unique challenges

129 00:05:06.120 --> 00:05:08.820 that I'm gonna highlight a few of here in a second.

130 00:05:08.820 --> 00:05:11.160 But there are a variety of master protocols out there

131 00:05:11.160 --> 00:05:12.900 in case you've heard some of these buzzwords.

132 00:05:12.900 --> 00:05:15.750 I'll be focusing on basket trials today,

133 00:05:15.750 --> 00:05:17.820 but you may have also heard of things like umbrella trials

134 00:05:17.820 --> 00:05:20.433 or even more generally platform trial designs.

135 00:05:23.520 --> 00:05:25.687 And so one example of what this looks like here is

136 00:05:25.687 --> 00:05:28.320 this is a graphic from a paper in the

137 00:05:28.320 --> 00:05:31.050 New England Journal by Dr. Woodcock and LaVange,

138 00:05:31.050 --> 00:05:32.430 Dr. Woodcock being a clinician,

139 00:05:32.430 --> 00:05:34.590 and Dr. Lisa LaVange being a past president of

140 00:05:34.590 --> 00:05:36.870 The American Statistical Association,

141 00:05:36.870 --> 00:05:39.330 where they actually tried to put to rest some of the

142 00:05:39.330 --> 00:05:42.450 confusion surrounding some of these design types

143 00:05:42.450 --> 00:05:43.283 because it turns out,

144 00:05:43.283 --> 00:05:46.200 up until 2017 when we discussed these designs

145 00:05:46.200 --> 00:05:48.180 across even statistical communities

146 00:05:48.180 --> 00:05:50.070 and with clinical researchers,

147 00:05:50.070 --> 00:05:53.460 we tend to use these terms fairly interchangeably

148 00:05:53.460 --> 00:05:54.810 even though we are really getting at

149 00:05:54.810 --> 00:05:57.120 very different concepts.

150 00:05:57.120 --> 00:05:58.020 So for example,

151 00:05:58.020 --> 00:06:01.710 in the top here we have this idea of an umbrella trial

152 00:06:01.710 --> 00:06:04.110 and this is really the context of a single disease

153 00:06:04.110 --> 00:06:05.220 like lung cancer,

154 00:06:05.220 --> 00:06:06.660 but we actually then will screen for

155 00:06:06.660 --> 00:06:07.770 those genetic alterations

156 00:06:07.770 --> 00:06:09.810 and have different therapies that we're trying to

157 00:06:09.810 --> 00:06:13.830 target a different biomarker or genetic alteration for.

158 00:06:13.830 --> 00:06:16.500 This contrasts to what we're focusing on today below

159 00:06:16.500 --> 00:06:18.090 of a basket trial,

160 00:06:18.090 --> 00:06:20.400 we actually have different diseases or indications,

161 00:06:20.400 --> 00:06:23.280 but they share a common target or genetic alteration

162 00:06:23.280 --> 00:06:24.870 which we wish to target.

163 00:06:24.870 --> 00:06:26.430 And in this sense we can think of it potentially

164 00:06:26.430 --> 00:06:27.780 as them sharing a basket

165 00:06:27.780 --> 00:06:31.560 or sharing a sort of that commonality there.

166 00:06:31.560 --> 00:06:35.400 Now, this is a fairly broad general idea of these designs.

167 00:06:35.400 --> 00:06:36.900 And so I think for the sake of

168 00:06:36.900 --> 00:06:38.040 what we're gonna talk about today

169 00:06:38.040 --> 00:06:39.870 and some of the statistical considerations

170 00:06:39.870 --> 00:06:42.030 that can be helpful to do a bit of a

171 00:06:42.030 --> 00:06:45.690 oversimplification of what a design might look like here.

172 00:06:45.690 --> 00:06:47.970 And so on the slide that I've presented,

173 00:06:47.970 --> 00:06:52.596 I have this kind of naive graphic of actual baskets

174 00:06:52.596 --> 00:06:54.150 and we're going to assume that in each column

175 00:06:54.150 --> 00:06:56.400 we have a different indication or site of cancer

176 00:06:56.400 --> 00:06:58.920 that has that common genetic alteration.

177 00:06:58.920 --> 00:07:01.590 So for example, basket one may represent the lung,

178 00:07:01.590 --> 00:07:04.740 basket two may represent the liver and so on.

179 00:07:04.740 --> 00:07:06.870 Now when we're in the case of designing

180 00:07:06.870 --> 00:07:08.520 or the design stage of a study,

181 00:07:08.520 --> 00:07:10.590 we tend to make oversimplifying assumptions

182 00:07:10.590 --> 00:07:13.380 to address these potential calculations for

183 00:07:13.380 --> 00:07:14.730 power, sample size,

184 00:07:14.730 --> 00:07:16.560 and quantities that we're usually interested in

185 00:07:16.560 --> 00:07:17.493 for study design.

186 00:07:18.600 --> 00:07:19.830 So here on this graph,

187 00:07:19.830 --> 00:07:21.720 we are gonna make a assumption that

188 00:07:21.720 --> 00:07:25.230 there's only two possible responses in this planning stage.

189 00:07:25.230 --> 00:07:28.680 One is that the baskets have no response or a null basket,

190 00:07:28.680 --> 00:07:31.890 that's the blue colored solid baskets on the screen.

191 00:07:31.890 --> 00:07:34.710 The other case would be a alternative response

192 00:07:34.710 --> 00:07:37.440 where there is some hopeful benefit to the treatment

193 00:07:37.440 --> 00:07:40.800 and those are the open orange colored baskets

194 00:07:40.800 --> 00:07:42.930 we see on the screen here.

195 00:07:42.930 --> 00:07:46.080 Now, one of the challenges I think with basket trial design

196 00:07:46.080 --> 00:07:48.120 that can be overlooked sometime,

197 00:07:48.120 --> 00:07:49.260 even in this design stage,

198 00:07:49.260 --> 00:07:50.790 is that for a standard two arm trial,

199 00:07:50.790 --> 00:07:52.050 we do have to make this assumption of,

200 00:07:52.050 --> 00:07:54.870 what is our null hypothesis or response?

201 00:07:54.870 --> 00:07:57.420 What's our alternative hypothesis or response?

202 00:07:57.420 --> 00:08:00.150 We really only have to do that for one configuration

203 00:08:00.150 --> 00:08:02.790 or combination because we have two arms.

204 00:08:02.790 --> 00:08:05.070 In the case of a single arm basket trial here,

205 00:08:05.070 --> 00:08:05.903 we actually see that

206 00:08:05.903 --> 00:08:08.010 just by having five baskets in a study

207 00:08:08.010 --> 00:08:10.470 and many actual trials that are implemented at

208 00:08:10.470 --> 00:08:12.060 far more baskets,

209 00:08:12.060 --> 00:08:14.040 we actually see a range of just six possible

210 00:08:14.040 --> 00:08:17.010 binary combinations of the basket works

211 00:08:17.010 --> 00:08:17.910 or it doesn't work,

212 00:08:17.910 --> 00:08:21.480 ranging from at the extremes a global null

213 00:08:21.480 --> 00:08:23.400 where unfortunately the treatment does not work

214 00:08:23.400 --> 00:08:26.400 in any basket down to the sort of dream scenario

215 00:08:26.400 --> 00:08:28.040 where the basket is actually,

216 00:08:28.040 --> 00:08:30.840 or the drug actually works across all baskets.

217 00:08:30.840 --> 00:08:33.840 There is this homogenous actually response

218 00:08:33.840 --> 00:08:38.013 in a positive direction for the sort of clinical outcome.

219 00:08:39.030 --> 00:08:39.900 More realistically,

220 00:08:39.900 --> 00:08:41.550 we actually will probably encounter

221 00:08:41.550 --> 00:08:43.890 something that we see falls in the middle here,

222 00:08:43.890 --> 00:08:45.780 scenarios two through five,

223 00:08:45.780 --> 00:08:47.640 where there's some mixture of baskets

224 00:08:47.640 --> 00:08:48.777 that actually do show a response

225 00:08:48.777 --> 00:08:51.360 and some that for whatever reason we might not know yet,

226 00:08:51.360 --> 00:08:52.767 it just doesn't appear to have any effect

227 00:08:52.767 --> 00:08:55.050 and is a null response.

228 00:08:55.050 --> 00:08:56.310 So this can make it challenging

229 00:08:56.310 --> 00:08:57.720 for some of the considerations of

230 00:08:57.720 --> 00:09:00.873 what analysis strategy you plan to use in practice.

231 00:09:02.670 --> 00:09:05.340 And so to just, at a high level,

232 00:09:05.340 --> 00:09:06.960 highlight some of these challenges

233 00:09:06.960 --> 00:09:10.500 before we jump into the methods for today's talk.

234 00:09:10.500 --> 00:09:12.780 In practice, each of these baskets within trial

235 00:09:12.780 --> 00:09:14.340 often have what we call a small

236 00:09:14.340 --> 00:09:17.520 and or small sample size for each of those indications.

237 00:09:17.520 --> 00:09:18.353 It turns out

238 00:09:18.353 --> 00:09:20.340 once we actually have this idea of precision medicine

239 00:09:20.340 --> 00:09:21.780 and we can be fairly precise

240 00:09:21.780 --> 00:09:22.830 for who counts for a trial,

241 00:09:22.830 --> 00:09:25.650 we actually have a much smaller potential sample

242 00:09:25.650 --> 00:09:27.420 or population to enroll.

243 00:09:27.420 --> 00:09:29.190 This means that even though we might have a treatment

244 00:09:29.190 --> 00:09:30.180 that works really well,

245 00:09:30.180 --> 00:09:32.640 it can be challenging to find individuals who qualify

246 00:09:32.640 --> 00:09:34.470 or are eligible to enroll

247 00:09:34.470 --> 00:09:36.480 or they may have competing trials or demands

248 00:09:36.480 --> 00:09:39.213 for other studies or care to consider.

249 00:09:40.500 --> 00:09:42.540 As I've also alluded to earlier the challenge,

250 00:09:42.540 --> 00:09:44.340 we also have this potential for indication

251 00:09:44.340 --> 00:09:47.160 or subgroup heterogeneity and that may be likely.

252 00:09:47.160 --> 00:09:47.993 In other words,

253 00:09:47.993 --> 00:09:49.590 we might not expect the same response

254 00:09:49.590 --> 00:09:50.457 across all those baskets.

255 00:09:50.457 --> 00:09:52.350 And that gets back to the previous graphic

256 00:09:52.350 --> 00:09:53.220 on that last slide

257 00:09:53.220 --> 00:09:55.680 where we might have something like two null baskets

258 00:09:55.680 --> 00:09:57.030 and three alternative baskets.

259 00:09:57.030 --> 00:09:59.040 And that can make it really challenging in the presence

260 00:09:59.040 --> 00:10:01.530 of a small n to determine how do we

261 00:10:01.530 --> 00:10:03.090 appropriately analyze that data

262 00:10:03.090 --> 00:10:05.970 so we capture the potentially applications baskets

263 00:10:05.970 --> 00:10:08.700 and can move those forward so patients benefit

264 00:10:08.700 --> 00:10:10.950 while not carrying forward null baskets

265 00:10:10.950 --> 00:10:13.593 where there is no response for those patients.

266 00:10:15.780 --> 00:10:16.860 Statistically speaking,

267 00:10:16.860 --> 00:10:19.440 we also have these ideas of operating characteristics

268 00:10:19.440 --> 00:10:20.670 and in the context of a trial,

269 00:10:20.670 --> 00:10:22.410 what we mean by that is things like power

270 00:10:22.410 --> 00:10:23.670 and type one error

271 00:10:23.670 --> 00:10:26.040 and I just have additional considerations with respect to

272 00:10:26.040 --> 00:10:27.840 how do we summarize these?

273 00:10:27.840 --> 00:10:30.540 Do we summarize them within each basket or each column

274 00:10:30.540 --> 00:10:32.220 on that graphic on the previous slide,

275 00:10:32.220 --> 00:10:34.230 essentially treating it as a bunch of

276 00:10:34.230 --> 00:10:36.360 standalone independent one arm trials

277 00:10:36.360 --> 00:10:39.960 just under one overall study design or idea?

278 00:10:39.960 --> 00:10:42.000 Or do we try to account for the fact that we have

279 00:10:42.000 --> 00:10:45.270 five baskets enrolling like on the graphic before

280 00:10:45.270 --> 00:10:46.710 and we might wanna consider something like a

281 00:10:46.710 --> 00:10:48.450 family wise type one error rate

282 00:10:48.450 --> 00:10:51.510 where any false positive would be a negative outcome

283 00:10:51.510 --> 00:10:53.460 if we're trying to correctly predict

284 00:10:53.460 --> 00:10:55.113 or identify associations?

285 00:10:56.820 --> 00:10:58.260 Now the focus of today's talk,

286 00:10:58.260 --> 00:11:00.360 and I could talk about these other points

287 00:11:00.360 --> 00:11:01.980 till the cows come home,

288 00:11:01.980 --> 00:11:04.200 but I'm gonna focus today on

289 00:11:04.200 --> 00:11:05.790 depending on that research stage we're at,

290 00:11:05.790 --> 00:11:07.800 if it's a phase one, two or three trial,

291 00:11:07.800 --> 00:11:09.900 we may wish to terminate early for some reason like

292 00:11:09.900 --> 00:11:11.610 efficacy or futility.

293 00:11:11.610 --> 00:11:13.080 And specifically for time today,

294 00:11:13.080 --> 00:11:15.810 I'm gonna focus on the idea of stopping for futility

295 00:11:15.810 --> 00:11:17.610 where we don't wanna keep enrolling baskets

296 00:11:17.610 --> 00:11:20.400 that are poorly performing both for ethical reasons.

297 00:11:20.400 --> 00:11:21.420 In other words,

298 00:11:21.420 --> 00:11:23.970 patients may benefit from other trials or treatments

299 00:11:23.970 --> 00:11:25.680 that are out there and we don't wanna subject them to

300 00:11:25.680 --> 00:11:27.570 treatments that have no benefit.

301 00:11:27.570 --> 00:11:30.840 But also from a resource consideration perspective.

302 00:11:30.840 --> 00:11:33.990 You can imagine that running a study or trial is expensive

303 00:11:33.990 --> 00:11:35.580 and can be complicated.

304 00:11:35.580 --> 00:11:37.830 And especially if we're doing something like a basket trial

305 00:11:37.830 --> 00:11:40.410 where we're having to enroll across multiple baskets,

306 00:11:40.410 --> 00:11:43.290 it may be ideal to be able to drop baskets early on

307 00:11:43.290 --> 00:11:44.400 that don't show promise

308 00:11:44.400 --> 00:11:46.320 so we can reallocate those resources to

309 00:11:46.320 --> 00:11:48.600 either different studies, research projects,

310 00:11:48.600 --> 00:11:52.353 or trials that we're trying to implement or run.

311 00:11:54.810 --> 00:11:56.760 So the motivation for today's talk

312 00:11:56.760 --> 00:11:58.290 building off of these ideas is that

313 00:11:58.290 --> 00:12:01.230 I want to demonstrate that a design that's very popular

314 00:12:01.230 --> 00:12:03.720 called Simon's two-stage design is

315 00:12:03.720 --> 00:12:05.400 generally speaking suboptimal

316 00:12:05.400 --> 00:12:08.430 compared to the multitude of alternative methods

317 00:12:08.430 --> 00:12:10.410 and designs that are out there.

318 00:12:10.410 --> 00:12:12.420 And then this is especially true in our context of

319 00:12:12.420 --> 00:12:14.640 a basket trial where within the single study

320 00:12:14.640 --> 00:12:16.710 we actually are simultaneously enrolling

321 00:12:16.710 --> 00:12:20.130 multiple one arm trials in our case today.

322 00:12:20.130 --> 00:12:22.260 Then the second point I'd like to highlight is

323 00:12:22.260 --> 00:12:24.360 we can identify when methods for sharing information

324 00:12:24.360 --> 00:12:27.090 across baskets could be beneficial to further improve

325 00:12:27.090 --> 00:12:29.403 the efficiency of our clinical trials.

326 00:12:30.960 --> 00:12:31.800 And so to highlight this,

327 00:12:31.800 --> 00:12:33.900 I wanna first just build us through

328 00:12:33.900 --> 00:12:35.850 and sort of illustrate or introduce these designs

329 00:12:35.850 --> 00:12:37.440 and the general concepts behind them

330 00:12:37.440 --> 00:12:39.600 because I know if you don't work in this space

331 00:12:39.600 --> 00:12:42.720 it may be sort of just ideas vaguely.

332 00:12:42.720 --> 00:12:44.610 So I wanna start with the Simon two-stage design,

333 00:12:44.610 --> 00:12:47.970 that comparator that people are commonly using.

334 00:12:47.970 --> 00:12:50.790 So Richard Simon, and this is back in 1989,

335 00:12:50.790 --> 00:12:53.550 introduced what he called optimal two-stage designs

336 00:12:53.550 --> 00:12:55.650 for phase two clinical trials.

337 00:12:55.650 --> 00:12:57.150 And this was specifically in the context

338 00:12:57.150 --> 00:12:59.490 that we're focusing on today for a one sample trial

339 00:12:59.490 --> 00:13:02.420 to evaluate the success of a binary outcome.

340 00:13:02.420 --> 00:13:05.100 So for oncology we might think of this as a yes no outcome

341 00:13:05.100 --> 00:13:07.170 for is there a reduction in tumor size

342 00:13:07.170 --> 00:13:11.193 or a survival to some predefined time point.

343 00:13:13.230 --> 00:13:16.350 Now specifically what Dr. Simon was motivated by

344 00:13:16.350 --> 00:13:17.760 was the stage-two trials

345 00:13:17.760 --> 00:13:20.220 as it says in the title of his paper,

346 00:13:20.220 --> 00:13:21.510 and just to kind of

347 00:13:21.510 --> 00:13:23.400 give a common lay of the land for everyone,

348 00:13:23.400 --> 00:13:26.340 the purpose generally speaking of a phase two trial

349 00:13:26.340 --> 00:13:28.200 is to identify if the intervention

350 00:13:28.200 --> 00:13:29.670 warrants further development

351 00:13:29.670 --> 00:13:32.370 while collecting additional safety data.

352 00:13:32.370 --> 00:13:33.300 Generally speaking,

353 00:13:33.300 --> 00:13:34.980 we will have already completed what we call

354 00:13:34.980 --> 00:13:37.980 a phase one trial where we collect preliminary safety data

355 00:13:37.980 --> 00:13:40.410 to make sure that the drug is not toxic

356 00:13:40.410 --> 00:13:43.500 or at least has expected side effects

357 00:13:43.500 --> 00:13:45.090 that we are willing to tolerate for that

358 00:13:45.090 --> 00:13:47.580 potential gain in efficacy.

359 00:13:47.580 --> 00:13:49.627 And then in phase two here we're actually trying to say,

360 00:13:49.627 --> 00:13:51.330 "You know, is there some benefit?

361 00:13:51.330 --> 00:13:53.250 Is it worth potentially moving this drug

362 00:13:53.250 --> 00:13:54.420 on either for approval

363 00:13:54.420 --> 00:13:56.940 or some larger confirmatory study

364 00:13:56.940 --> 00:13:59.097 to identify if it truly works or doesn't?"

365 00:14:01.020 --> 00:14:03.090 Now the motivation for Dr. Simon is that

366 00:14:03.090 --> 00:14:04.650 we would like to terminate studies earlier,

367 00:14:04.650 --> 00:14:05.520 as I mentioned before,

368 00:14:05.520 --> 00:14:07.800 for both ethical and resource considerations

369 00:14:07.800 --> 00:14:09.240 that they appear futile.

370 00:14:09.240 --> 00:14:11.430 In other words, it's not a great use of our resources

371 00:14:11.430 --> 00:14:12.390 and we should try in some

372 00:14:12.390 --> 00:14:14.733 rigorous statistical way to address this.

373 00:14:17.040 --> 00:14:19.860 If you do go back and look at Simon's 1989 paper

374 00:14:19.860 --> 00:14:20.880 or you just Google this

375 00:14:20.880 --> 00:14:22.470 and there's various calculators that people have

376 00:14:22.470 --> 00:14:23.430 put out there,

377 00:14:23.430 --> 00:14:25.590 there are two flavors of this design that exist

378 00:14:25.590 --> 00:14:27.210 from this original paper.

379 00:14:27.210 --> 00:14:28.410 One is an optimal

380 00:14:28.410 --> 00:14:30.840 and one is called a minimax design.

381 00:14:30.840 --> 00:14:31.920 Within clinical trials,

382 00:14:31.920 --> 00:14:35.730 once we introduce this idea of stopping early potentially

383 00:14:35.730 --> 00:14:38.700 or have the chance to stop early based on our data,

384 00:14:38.700 --> 00:14:41.820 we now have this idea that there's this expected sample size

385 00:14:41.820 --> 00:14:43.830 because we could enroll the entire sample size

386 00:14:43.830 --> 00:14:47.370 that we planned for or we could potentially stop early.

387 00:14:47.370 --> 00:14:49.320 And since we could stop early or go the whole way

388 00:14:49.320 --> 00:14:50.760 and we don't know what our choice will be

389 00:14:50.760 --> 00:14:53.220 until we actually collect the data and do the study,

390 00:14:53.220 --> 00:14:55.740 we now have sample size of the random variable,

391 00:14:55.740 --> 00:14:57.690 something that we can calculate an expectation

392 00:14:57.690 --> 00:14:59.070 or an average for.

393 00:14:59.070 --> 00:15:01.080 And so Simon's optimal design tries to

394 00:15:01.080 --> 00:15:05.820 minimize what that average sample size might be in theory.

395 00:15:05.820 --> 00:15:08.190 In contrast, the minimax design

396 00:15:08.190 --> 00:15:11.040 tries to minimize whatever that largest sample size would be

397 00:15:11.040 --> 00:15:12.690 if we didn't stop early.

398 00:15:12.690 --> 00:15:13.650 So if we kept enrolling

399 00:15:13.650 --> 00:15:15.960 and we never stopped at any of our interim looks,

400 00:15:15.960 --> 00:15:17.970 how much data would we need to collect

401 00:15:17.970 --> 00:15:20.280 until we choose a design that minimizes that

402 00:15:20.280 --> 00:15:22.563 at the expense of potentially stopping early?

403 00:15:24.930 --> 00:15:26.820 I think this is most helpful to see the

404 00:15:26.820 --> 00:15:28.590 sort of elegance of this design

405 00:15:28.590 --> 00:15:30.060 and why it's I think so popular

406 00:15:30.060 --> 00:15:31.260 by just introducing example

407 00:15:31.260 --> 00:15:33.390 that will also motivate our simulations

408 00:15:33.390 --> 00:15:35.610 here that we're gonna talk about in a minute.

409 00:15:35.610 --> 00:15:36.960 We're gonna consider a study where

410 00:15:36.960 --> 00:15:39.883 the null response rate is 10%.

411 00:15:39.883 --> 00:15:41.730 And we're going to consider a target

412 00:15:41.730 --> 00:15:43.800 for an alternative response rate of 30%.

413 00:15:43.800 --> 00:15:45.090 So this isn't a situation where

414 00:15:45.090 --> 00:15:47.820 we're looking for necessarily a curative drug,

415 00:15:47.820 --> 00:15:49.380 but something that does show what we think of

416 00:15:49.380 --> 00:15:52.410 as a clinically meaningful benefit from 10 to 30%,

417 00:15:52.410 --> 00:15:54.333 let's say survival or tumor response.

418 00:15:55.200 --> 00:15:57.480 Now if we have these two parameters

419 00:15:57.480 --> 00:16:00.180 and we wanna do a Simon two-stage minimax design

420 00:16:00.180 --> 00:16:02.970 to minimize that maximum possible sample size

421 00:16:02.970 --> 00:16:04.230 we would enroll,

422 00:16:04.230 --> 00:16:06.060 we would have to also define

423 00:16:06.060 --> 00:16:07.590 the type one error rate or alpha

424 00:16:07.590 --> 00:16:09.600 that cancels a false positive.

425 00:16:09.600 --> 00:16:12.447 Here we're going to set 10% for this phase two design

426 00:16:12.447 --> 00:16:15.330 and we also wish to target a 90% power

427 00:16:15.330 --> 00:16:19.530 to detect that treatment of 30% if it truly exists.

428 00:16:19.530 --> 00:16:21.660 So we put all of this into our calculator

429 00:16:21.660 --> 00:16:24.900 to Simon's framework and we turn that statistical crank.

430 00:16:24.900 --> 00:16:25.980 What we see is that

431 00:16:25.980 --> 00:16:28.710 it gives us this approach where in stage one

432 00:16:28.710 --> 00:16:30.780 we would enroll 16 participants

433 00:16:30.780 --> 00:16:33.600 and we would terminate the trial or this study arm

434 00:16:33.600 --> 00:16:37.410 for futility if one or fewer responses are observed.

435 00:16:37.410 --> 00:16:41.130 Now if we observe two or more responses,

436 00:16:41.130 --> 00:16:42.570 we would continue enrollment

437 00:16:42.570 --> 00:16:45.360 to the overall maximum sample size that we plan for

438 00:16:45.360 --> 00:16:47.580 of 25 in the second stage.

439 00:16:47.580 --> 00:16:50.760 And at this point if four or fewer responses are observed,

440 00:16:50.760 --> 00:16:53.220 no further investigation is warranted

441 00:16:53.220 --> 00:16:54.900 or we can think of this as a situation where

442 00:16:54.900 --> 00:16:58.923 our P value would be larger than our defined alpha 0.1.

443 00:16:59.970 --> 00:17:02.730 Now, the nice thing here is that it is quite simple.

444 00:17:02.730 --> 00:17:04.620 In fact, after we trim that statistical crank

445 00:17:04.620 --> 00:17:06.240 and we have this decision rule,

446 00:17:06.240 --> 00:17:08.490 you in theory don't even need a statistician

447 00:17:08.490 --> 00:17:10.380 because you can count the number of responses

448 00:17:10.380 --> 00:17:12.240 for your binary outcome on your hand

449 00:17:12.240 --> 00:17:15.330 and determine should I stop early, should I continue?

450 00:17:15.330 --> 00:17:16.260 And if I continue,

451 00:17:16.260 --> 00:17:18.090 do I have some benefit potentially

452 00:17:18.090 --> 00:17:20.610 that says it's worth either doing a future study

453 00:17:20.610 --> 00:17:22.860 or I did a statistical test,

454 00:17:22.860 --> 00:17:25.110 would find that the P value meets my threshold

455 00:17:25.110 --> 00:17:26.823 I set for significance.

456 00:17:29.310 --> 00:17:30.690 Now, of course,

457 00:17:30.690 --> 00:17:32.617 it wouldn't be a great talk if I stopped there and said,

458 00:17:32.617 --> 00:17:34.140 "You know, this is everything.

459 00:17:34.140 --> 00:17:35.760 It's perfect. There's nothing to change."

460 00:17:35.760 --> 00:17:37.350 There are some potential limitations

461 00:17:37.350 --> 00:17:39.270 and of course some solutions I think

462 00:17:39.270 --> 00:17:41.850 that we could address in this talk.

463 00:17:41.850 --> 00:17:43.020 The first thing to note is that

464 00:17:43.020 --> 00:17:45.750 this is extremely restrictive in when it could terminate

465 00:17:45.750 --> 00:17:47.940 and it may continue to the maximum sample size

466 00:17:47.940 --> 00:17:49.980 even if a null effect is present.

467 00:17:49.980 --> 00:17:51.840 And we're gonna see this come to fruition

468 00:17:51.840 --> 00:17:53.580 in the simulation studies,

469 00:17:53.580 --> 00:17:55.380 but it's worth noting here it only looks once.

470 00:17:55.380 --> 00:17:57.630 It's a two stage design.

471 00:17:57.630 --> 00:18:00.210 And depending on the criteria you plug in,

472 00:18:00.210 --> 00:18:01.800 it might not look for quite some time.

473 00:18:01.800 --> 00:18:04.920 16 out of 25 total participants enrolled

474 00:18:04.920 --> 00:18:07.440 is still a pretty large sample size

475 00:18:07.440 --> 00:18:09.273 relative to where we expect to be.

476 00:18:10.710 --> 00:18:12.057 One solution that we could look at

477 00:18:12.057 --> 00:18:13.890 and that I'm going to propose today

478 00:18:13.890 --> 00:18:15.960 is that we could use Bayesian methods instead

479 00:18:15.960 --> 00:18:18.150 for more frequent interim monitoring.

480 00:18:18.150 --> 00:18:19.607 And this could use quantities that we think of

481 00:18:19.607 --> 00:18:23.280 as the posterior or the predictive probabilities

482 00:18:23.280 --> 00:18:24.243 of our data.

483 00:18:25.830 --> 00:18:28.260 Another limitation that we wish to address as well is that

484 00:18:28.260 --> 00:18:29.820 in designs like a basket trial

485 00:18:29.820 --> 00:18:30.960 that have multiple indications

486 00:18:30.960 --> 00:18:34.560 or multiple arms that have the same entry criteria,

487 00:18:34.560 --> 00:18:36.300 Simon's two-stage design is going to

488 00:18:36.300 --> 00:18:38.040 fail to take advantage of the potential

489 00:18:38.040 --> 00:18:40.740 what we call exchange ability across baskets.

490 00:18:40.740 --> 00:18:44.550 In other words, if baskets appear to have the same response,

491 00:18:44.550 --> 00:18:45.900 whether it's let's say that null

492 00:18:45.900 --> 00:18:47.850 or that alternative response,

493 00:18:47.850 --> 00:18:49.230 it would be great if we could

494 00:18:49.230 --> 00:18:52.230 informatively pull them together into meta subgroups

495 00:18:52.230 --> 00:18:53.790 so we can increase the sample size

496 00:18:53.790 --> 00:18:56.310 and start to address that challenge of the small n

497 00:18:56.310 --> 00:18:59.250 that I mentioned earlier for these basket trial designs.

498 00:18:59.250 --> 00:19:02.280 And specifically today we're going to examine the use of

499 00:19:02.280 --> 00:19:04.740 what we call multi-source exchangeability models

500 00:19:04.740 --> 00:19:07.950 to share information across baskets when appropriate.

501 00:19:07.950 --> 00:19:10.170 And I'll walk through a very high level sort of

502 00:19:10.170 --> 00:19:12.120 conceptual idea of what these models

503 00:19:12.120 --> 00:19:14.220 and how they work and what they look like.

504 00:19:16.770 --> 00:19:17.700 Before we get into that though,

505 00:19:17.700 --> 00:19:20.247 I wanna just briefly mention the idea of posterior

506 00:19:20.247 --> 00:19:21.600 and predictive probabilities

507 00:19:21.600 --> 00:19:22.950 and give some definitions here

508 00:19:22.950 --> 00:19:25.200 so we can conceptually envision what we mean

509 00:19:25.200 --> 00:19:26.850 and especially if you haven't had the chance

510 00:19:26.850 --> 00:19:28.920 to work with a lot of patient methods,

511 00:19:28.920 --> 00:19:30.720 this can help give us an idea

512 00:19:30.720 --> 00:19:33.150 of some of the analogs to maybe a frequentist approach

513 00:19:33.150 --> 00:19:34.260 or what we're trying to do here

514 00:19:34.260 --> 00:19:36.450 that you may be familiar with.

515 00:19:36.450 --> 00:19:37.380 Now I will mention,

516 00:19:37.380 --> 00:19:39.360 I'm not the first person to propose looking at

517 00:19:39.360 --> 00:19:40.500 Bayesian interim stopping rules.

518 00:19:40.500 --> 00:19:43.170 I have a couple citations here by Dmitrienko

519 00:19:43.170 --> 00:19:44.940 and Wang and Saville et all

520 00:19:44.940 --> 00:19:46.860 and they do a lot of extensive work in addition to

521 00:19:46.860 --> 00:19:48.930 hundreds of other papers considering

522 00:19:48.930 --> 00:19:51.300 Bayesian interim monitoring.

523 00:19:51.300 --> 00:19:53.160 But specifically to motivate this

524 00:19:53.160 --> 00:19:55.590 we have these two concepts that commonly come up

525 00:19:55.590 --> 00:19:56.880 in Bayesian analysis,

526 00:19:56.880 --> 00:20:00.870 a posterior probability or a predictive probability.

527 00:20:00.870 --> 00:20:02.880 The posterior probability

528 00:20:02.880 --> 00:20:05.340 is very much analogous to kinda like a P value

529 00:20:05.340 --> 00:20:06.330 in a frequent significance.

530 00:20:06.330 --> 00:20:09.090 It says, "Based on the posterior distribution

531 00:20:09.090 --> 00:20:11.100 we arrive at through a Bayesian analysis,

532 00:20:11.100 --> 00:20:13.140 we're gonna calculate the probability

533 00:20:13.140 --> 00:20:15.480 that our proportion exceeds the null response rate

534 00:20:15.480 --> 00:20:16.313 we wish to beat."

535 00:20:16.313 --> 00:20:17.617 So in our case, we're basically saying,

536 00:20:17.617 --> 00:20:19.680 "What's the probability based on our data

537 00:20:19.680 --> 00:20:23.577 and a prior we've given that the response is 10% or higher."

538 00:20:24.510 --> 00:20:25.770 So this covers a lot of ground

539 00:20:25.770 --> 00:20:29.160 'cause anything you know from 10.1 up to 100%

540 00:20:29.160 --> 00:20:32.160 would meet this criteria being better than 10%.

541 00:20:32.160 --> 00:20:34.080 But it does quantify,

542 00:20:34.080 --> 00:20:36.720 based on the evidence we've observed so far,

543 00:20:36.720 --> 00:20:40.020 how the data suggests the

544 00:20:40.020 --> 00:20:41.760 benefit may be with respect to that null.

545 00:20:41.760 --> 00:20:43.700 So in the case of let's say

546 00:20:43.700 --> 00:20:46.860 an interim look for futility at the data, we could say,

547 00:20:46.860 --> 00:20:50.520 if we just use Simon's two-stage design as our motivating

548 00:20:50.520 --> 00:20:52.837 ground to consider, we might say,

549 00:20:52.837 --> 00:20:55.320 "Okay, we have 16 people so far,

550 00:20:55.320 --> 00:20:57.660 what's the probability based on these 16 people

551 00:20:57.660 --> 00:20:58.710 that I could actually say

552 00:20:58.710 --> 00:21:00.360 there's no chance or limited chance

553 00:21:00.360 --> 00:21:02.700 I'm going to detect something in the trial here

554 00:21:02.700 --> 00:21:04.830 based on the data I've seen so far?"

555 00:21:04.830 --> 00:21:06.540 Now the challenge here is that

556 00:21:06.540 --> 00:21:09.180 it is based on off the data we've seen so far

557 00:21:09.180 --> 00:21:12.030 and it doesn't take into account the fact that we still have

558 00:21:12.030 --> 00:21:14.820 another nine potential participants to enroll

559 00:21:14.820 --> 00:21:18.090 to get to that maximum sample size of 25.

560 00:21:18.090 --> 00:21:19.560 That's where this idea of what we call a

561 00:21:19.560 --> 00:21:21.990 predictive probability comes in.

562 00:21:21.990 --> 00:21:24.090 We're considering our accumulated data

563 00:21:24.090 --> 00:21:27.120 and the priors we've specified in our Bayesian context,

564 00:21:27.120 --> 00:21:29.790 it's the probability that we will have observed

565 00:21:29.790 --> 00:21:32.400 a significant result if we've met

566 00:21:32.400 --> 00:21:34.550 and enrolled up to our maximum sample size.

567 00:21:35.550 --> 00:21:37.710 In other words, I think it's a very natural place to be

568 00:21:37.710 --> 00:21:38.610 for interim monitoring

569 00:21:38.610 --> 00:21:40.740 because it says based on the data I've seen so far,

570 00:21:40.740 --> 00:21:42.810 i.e the posterior probability,

571 00:21:42.810 --> 00:21:46.200 if I use that to help identify what are likely futures

572 00:21:46.200 --> 00:21:48.163 to observe or likely sample sizes

573 00:21:48.163 --> 00:21:51.450 I will continue enrolling to get to that maximum of 25,

574 00:21:51.450 --> 00:21:53.130 what's the probability at the end of the day

575 00:21:53.130 --> 00:21:55.560 when I do hit that sample size of 25,

576 00:21:55.560 --> 00:21:58.290 I will have a significant conclusion?

577 00:21:58.290 --> 00:22:00.330 And if it's a really low predictive probability,

578 00:22:00.330 --> 00:22:02.070 if I say there's only a 5% chance

579 00:22:02.070 --> 00:22:04.140 of you actually declaring significance if you

580 00:22:04.140 --> 00:22:05.970 keep enrolling participants,

581 00:22:05.970 --> 00:22:08.280 that can be really informative both statistically

582 00:22:08.280 --> 00:22:10.200 and for clinical partners to say

583 00:22:10.200 --> 00:22:13.380 it doesn't seem very likely that we're gonna hit our target.

584 00:22:13.380 --> 00:22:14.700 That being said,

585 00:22:14.700 --> 00:22:17.310 a lot of people are very happy to continue trials going

586 00:22:17.310 --> 00:22:19.080 with low chances or low probability

587 00:22:19.080 --> 00:22:21.030 because you're saying there's still a chance

588 00:22:21.030 --> 00:22:23.010 I may detect something that could be

589 00:22:23.010 --> 00:22:25.170 significant enough worth.

590 00:22:25.170 --> 00:22:27.600 So we'll see that across a range of these thresholds,

591 00:22:27.600 --> 00:22:29.883 the performance of these models may change.

592 00:22:32.040 --> 00:22:34.410 Now this brings us to a brief recap

593 00:22:34.410 --> 00:22:35.400 of sort of our motivation.

594 00:22:35.400 --> 00:22:36.540 I just spent a few minutes

595 00:22:36.540 --> 00:22:38.970 introducing that popular Simon two-stage design,

596 00:22:38.970 --> 00:22:39.900 the idea behind it,

597 00:22:39.900 --> 00:22:41.910 what it might look like in practice,

598 00:22:41.910 --> 00:22:45.060 as well as some alternatives with the Bayesian flare.

599 00:22:45.060 --> 00:22:47.010 The next part I wanna briefly address is that

600 00:22:47.010 --> 00:22:49.620 we can also now look at this idea

601 00:22:49.620 --> 00:22:52.410 of sharing information across baskets

602 00:22:52.410 --> 00:22:54.090 to further improve that trial efficiency

603 00:22:54.090 --> 00:22:56.490 'cause so far both Simon's design

604 00:22:56.490 --> 00:22:59.130 and the just using a posterior predictive probability

605 00:22:59.130 --> 00:23:01.980 for an interim monitoring will still treat each basket

606 00:23:01.980 --> 00:23:04.203 as its own little one arm trial.

607 00:23:07.020 --> 00:23:09.780 Now specifically today I'm gonna focus on this idea

608 00:23:09.780 --> 00:23:13.433 we call multi-source exchangeability models or MEMs.

609 00:23:13.433 --> 00:23:15.450 This is a general Bayesian framework

610 00:23:15.450 --> 00:23:18.060 to enable the incorporation of independent sources

611 00:23:18.060 --> 00:23:20.220 of supplemental information

612 00:23:20.220 --> 00:23:22.140 and its original work that I developed

613 00:23:22.140 --> 00:23:25.170 during my dissertation at the University of Minnesota.

614 00:23:25.170 --> 00:23:26.003 In this case,

615 00:23:26.003 --> 00:23:27.450 the amount of borrowing is determined by

616 00:23:27.450 --> 00:23:29.370 the exchange ability of our data,

617 00:23:29.370 --> 00:23:30.600 which in our context is really,

618 00:23:30.600 --> 00:23:33.000 how equivalent are the response rates?

619 00:23:33.000 --> 00:23:35.490 If two baskets have the exact same response rate,

620 00:23:35.490 --> 00:23:37.920 we may think that there's a higher probability

621 00:23:37.920 --> 00:23:39.780 that the true underlying population

622 00:23:39.780 --> 00:23:41.880 we are trying to estimate are truly exchangeable.

623 00:23:41.880 --> 00:23:46.140 We wish to combine that data as much as we possibly can.

624 00:23:46.140 --> 00:23:48.360 First is if again we see something that is like a

625 00:23:48.360 --> 00:23:50.190 10% response rate for one basket

626 00:23:50.190 --> 00:23:52.800 and a 30% response rate for another basket,

627 00:23:52.800 --> 00:23:55.110 we likely don't want to combine that data because

628 00:23:55.110 --> 00:23:57.360 those are not very equivalent response rates.

629 00:23:57.360 --> 00:23:59.160 In fact, we seem to have identified

630 00:23:59.160 --> 00:24:00.270 two different subgroups

631 00:24:00.270 --> 00:24:03.393 and performances in those two baskets.

632 00:24:04.380 --> 00:24:06.810 One of the advantages of MEMs relative to

633 00:24:06.810 --> 00:24:09.210 a host of other statistical methods that are out there

634 00:24:09.210 --> 00:24:12.150 that include things like power priors, commensurate priors,

635 00:24:12.150 --> 00:24:14.610 meta analytic priors, and so forth,

636 00:24:14.610 --> 00:24:16.050 is that we've been able to demonstrate that

637 00:24:16.050 --> 00:24:18.360 in their most basic iteration without

638 00:24:18.360 --> 00:24:20.310 any extra bells or whistles,

639 00:24:20.310 --> 00:24:23.130 MEMs are able to actually account for this heterogeneity

640 00:24:23.130 --> 00:24:25.650 across different potential response rates

641 00:24:25.650 --> 00:24:28.950 and appropriately down weight non-changeable sources.

642 00:24:28.950 --> 00:24:30.390 Whereas we show through simulation

643 00:24:30.390 --> 00:24:33.300 and earlier work some of these other methods without

644 00:24:33.300 --> 00:24:35.250 newer advancements to them

645 00:24:35.250 --> 00:24:37.560 actually either naively pull everything together

646 00:24:37.560 --> 00:24:40.560 even if there's non-changeable groups

647 00:24:40.560 --> 00:24:43.140 or they're afraid of the sort of presence of

648 00:24:43.140 --> 00:24:45.120 non-change ability and if anything seems amiss,

649 00:24:45.120 --> 00:24:48.240 they quickly go to an independence analysis

650 00:24:48.240 --> 00:24:50.700 that doesn't leverage this potential sharing

651 00:24:50.700 --> 00:24:54.753 of information across meta subgroups that are exchangeable.

652 00:24:56.460 --> 00:24:58.530 Now again, I don't wanna get too much into the weeds

653 00:24:58.530 --> 00:24:59.667 of the math behind the MEMs,

654 00:24:59.667 --> 00:25:02.250 but I will have a few formulas in a couple slides

655 00:25:02.250 --> 00:25:03.180 but I do think it's helpful to

656 00:25:03.180 --> 00:25:05.490 conceptualize it with graphics.

657 00:25:05.490 --> 00:25:08.050 And so here I just want to illustrate a very simplified case

658 00:25:08.050 --> 00:25:11.160 where we're gonna assume that we have a three basket trial

659 00:25:11.160 --> 00:25:13.830 and for the sake of doing an analysis with MEMs,

660 00:25:13.830 --> 00:25:15.720 I think it's helpful to also think of it as

661 00:25:15.720 --> 00:25:18.090 we're looking at the perspective of the analysis

662 00:25:18.090 --> 00:25:20.400 from one particular basket.

663 00:25:20.400 --> 00:25:23.580 So here on this slide here we see that we have this

664 00:25:23.580 --> 00:25:25.140 theta P circle in the middle

665 00:25:25.140 --> 00:25:27.540 and that's the parameter or parameters of interest

666 00:25:27.540 --> 00:25:29.340 we wish to estimate.

667 00:25:29.340 --> 00:25:30.750 In our case, that would be that

668 00:25:30.750 --> 00:25:32.793 binary outcome in each basket.

669 00:25:33.630 --> 00:25:37.110 Now, for this graphic we're using each of these circles here

670 00:25:37.110 --> 00:25:39.690 to represent a different data source.

671 00:25:39.690 --> 00:25:42.270 We're gonna say Y sub P is that primary basket

672 00:25:42.270 --> 00:25:43.860 that we're interested in or the perspective

673 00:25:43.860 --> 00:25:45.630 we're looking at for this example

674 00:25:45.630 --> 00:25:47.550 and Y sub one and Y sub two

675 00:25:47.550 --> 00:25:50.940 are two of the other baskets enrolled within the trial.

676 00:25:50.940 --> 00:25:52.500 Now a standard analysis

677 00:25:52.500 --> 00:25:55.440 without any information sharing across baskets

678 00:25:55.440 --> 00:25:59.550 would only have a data pooled from the observed data.

679 00:25:59.550 --> 00:26:01.380 I mean this is sort of the unexciting

680 00:26:01.380 --> 00:26:02.640 or unsurprising analysis

681 00:26:02.640 --> 00:26:04.980 where we basically are analyzing the data we have

682 00:26:04.980 --> 00:26:08.253 for the one basket that actually represents that group.

683 00:26:09.630 --> 00:26:11.400 However, we could imagine if we wish

684 00:26:11.400 --> 00:26:13.830 to pool together data from these other sources,

685 00:26:13.830 --> 00:26:16.530 we have different ways we could add arrows to this figure

686 00:26:16.530 --> 00:26:19.803 to represent different combinations of these groups.

687 00:26:20.820 --> 00:26:21.720 And this brings us to

688 00:26:21.720 --> 00:26:24.540 that multi-source exchangeability framework.

689 00:26:24.540 --> 00:26:26.490 So we see here on this slide,

690 00:26:26.490 --> 00:26:29.220 I now of a graphic showing four different combinations

691 00:26:29.220 --> 00:26:32.100 of exchangeability when we have these two other baskets

692 00:26:32.100 --> 00:26:34.750 that compare to our one basket of interest right now.

693 00:26:35.610 --> 00:26:38.100 And from top left to the bottom left

694 00:26:38.100 --> 00:26:39.480 in sort of a clockwise fashion,

695 00:26:39.480 --> 00:26:41.760 we see that making different assumptions from

696 00:26:41.760 --> 00:26:43.580 that standard analysis with no borrowing

697 00:26:43.580 --> 00:26:46.020 in the top right here where I'm drawing that arrow.

698 00:26:46.020 --> 00:26:47.820 So it is possible that

699 00:26:47.820 --> 00:26:49.257 none of our data sources are exchangeable

700 00:26:49.257 --> 00:26:51.150 and we should be doing an analysis that

701 00:26:51.150 --> 00:26:53.160 doesn't share information.

702 00:26:53.160 --> 00:26:55.050 On the right hand side that we might envision that

703 00:26:55.050 --> 00:26:58.320 well maybe the first basket or Y1 is exchangeable.

704 00:26:58.320 --> 00:27:01.050 So we wanna pull that with Y2 or excuse me with Yp,

705 00:27:01.050 --> 00:27:02.670 but Y2 is not.

706 00:27:02.670 --> 00:27:04.830 In the bottom right, this capital omega two,

707 00:27:04.830 --> 00:27:06.720 we actually assume that Y2 is exchangeable

708 00:27:06.720 --> 00:27:08.550 but Y1 is not.

709 00:27:08.550 --> 00:27:10.293 And in the bottom left we assume in this case

710 00:27:10.293 --> 00:27:11.637 that all the data is exchangeable

711 00:27:11.637 --> 00:27:13.653 and we should just pool it all together.

712 00:27:15.300 --> 00:27:16.770 So at this stage we've actually

713 00:27:16.770 --> 00:27:20.040 proposed all the configurations we can pair-wise

714 00:27:20.040 --> 00:27:23.400 of combining these different data sources with Y sub P.

715 00:27:23.400 --> 00:27:25.890 And we know that these are fitting four now different models

716 00:27:25.890 --> 00:27:27.390 based off of the data

717 00:27:27.390 --> 00:27:30.240 because for example in the top left, that standard analysis,

718 00:27:30.240 --> 00:27:33.120 there is no extra information from those other baskets

719 00:27:33.120 --> 00:27:34.800 versus like in the bottom left,

720 00:27:34.800 --> 00:27:36.180 we basically have combined everything

721 00:27:36.180 --> 00:27:38.670 and we think there's some common effect.

722 00:27:38.670 --> 00:27:40.440 Now this leads to two challenges on its own

723 00:27:40.440 --> 00:27:42.510 if we just stopped here with the framework.

724 00:27:42.510 --> 00:27:44.280 One would be that we'd have this idea of maybe

725 00:27:44.280 --> 00:27:46.770 cherry picking or trying to pick whichever combination

726 00:27:46.770 --> 00:27:50.160 best suits your prior hypotheses clinically.

727 00:27:50.160 --> 00:27:51.360 And so that would be a big no-go.

728 00:27:51.360 --> 00:27:52.410 We don't like cherry picking

729 00:27:52.410 --> 00:27:53.970 or fishing for things like P values

730 00:27:53.970 --> 00:27:56.493 or significance in our statistical analyses.

731 00:27:57.330 --> 00:27:59.100 The other challenge also is that

732 00:27:59.100 --> 00:28:01.380 all of these configurations are just assumptions

733 00:28:01.380 --> 00:28:02.520 of how we could combine data

734 00:28:02.520 --> 00:28:05.220 but we know underlying everything in the population is that

735 00:28:05.220 --> 00:28:07.140 true assumption of exchange ability of

736 00:28:07.140 --> 00:28:10.500 are these baskets or groups truly combinable or not?

737 00:28:10.500 --> 00:28:13.320 And we're just approximating that with our sample.

738 00:28:13.320 --> 00:28:15.450 And so right now if we have four separate models

739 00:28:15.450 --> 00:28:17.670 and potentially four separate conclusions,

740 00:28:17.670 --> 00:28:20.010 we need some way of combining these models

741 00:28:20.010 --> 00:28:21.390 to make inference.

742 00:28:21.390 --> 00:28:23.160 And in this case we propose

743 00:28:23.160 --> 00:28:25.980 leveraging a Bayesian model averaging framework

744 00:28:25.980 --> 00:28:27.960 where we calculate in this case

745 00:28:27.960 --> 00:28:28.793 and in our formulas here,

746 00:28:28.793 --> 00:28:31.830 the queues represent a posterior distribution

747 00:28:31.830 --> 00:28:33.180 where I've drawn this little arrow

748 00:28:33.180 --> 00:28:35.190 and I'm underlining right now,

749 00:28:35.190 --> 00:28:38.850 that reflects each square's configuration of

750 00:28:38.850 --> 00:28:41.220 exchange ability for our estimates.

751 00:28:41.220 --> 00:28:42.150 And through this process

752 00:28:42.150 --> 00:28:44.820 we estimate these lower case omega model weights

753 00:28:44.820 --> 00:28:46.860 that tries to estimate the appropriateness

754 00:28:46.860 --> 00:28:49.830 of exchangeability with the ultimate goal of

755 00:28:49.830 --> 00:28:52.860 having a average posterior that we can use

756 00:28:52.860 --> 00:28:54.210 for statistical inference

757 00:28:54.210 --> 00:28:57.060 to draw a conclusion about the potential efficacy

758 00:28:57.060 --> 00:28:58.653 or lack thereof of a treatment.

759 00:29:01.530 --> 00:29:02.640 Now very briefly,

760 00:29:02.640 --> 00:29:05.850 because this is a Bayesian model averaging framework,

761 00:29:05.850 --> 00:29:08.400 just one of the few formulas I have in the presentation,

762 00:29:08.400 --> 00:29:10.350 we just see over here that we have

763 00:29:10.350 --> 00:29:12.720 the way we calculate these posterior model weights

764 00:29:12.720 --> 00:29:14.970 as the prior on each model

765 00:29:14.970 --> 00:29:18.090 multiplied by an integrated marginal likelihood.

766 00:29:18.090 --> 00:29:19.530 Essentially, we can think of that as saying

767 00:29:19.530 --> 00:29:22.020 based off of that square we saw on the previous slide

768 00:29:22.020 --> 00:29:24.630 and combining those different data sources,

769 00:29:24.630 --> 00:29:26.430 what is that estimate of the effect

770 00:29:26.430 --> 00:29:28.890 with those different combinations?

771 00:29:28.890 --> 00:29:31.080 One unique thing about the MEM framework

772 00:29:31.080 --> 00:29:33.540 that differs from Bayesian model averaging though is that

773 00:29:33.540 --> 00:29:37.290 we actually specify priors with respect to these sources.

774 00:29:37.290 --> 00:29:39.030 And in the case of this example

775 00:29:39.030 --> 00:29:42.390 with only two supplemental like sources for our graphic,

776 00:29:42.390 --> 00:29:44.520 it's not a great cost savings,

777 00:29:44.520 --> 00:29:46.710 but we can imagine that if we have more and more sources,

778 00:29:46.710 --> 00:29:50.250 there's actually two to the P if P's the number of sources,

779 00:29:50.250 --> 00:29:51.720 combinations of exchange ability

780 00:29:51.720 --> 00:29:53.610 that we have to consider and model.

781 00:29:53.610 --> 00:29:55.800 And that quickly can become overwhelming if we have

782 00:29:55.800 --> 00:29:57.570 multiple sources that we have to define

783 00:29:57.570 --> 00:29:58.680 for each one of those squares,

784 00:29:58.680 --> 00:30:02.040 what's my prior that each combination of exchangeability

785 00:30:02.040 --> 00:30:03.840 is potentially true.

786 00:30:03.840 --> 00:30:06.210 Versus if we define it with respect to the source,

787 00:30:06.210 --> 00:30:09.480 we now go from two to the P priors to just P priors

788 00:30:09.480 --> 00:30:11.987 we have to specify for exchangeability.

789 00:30:14.340 --> 00:30:17.370 A few more notes about this idea here

790 00:30:17.370 --> 00:30:18.960 and just really zooming in on

791 00:30:18.960 --> 00:30:21.420 what we're gonna focus on for today's presentation.

792 00:30:21.420 --> 00:30:22.740 We have developed both fully

793 00:30:22.740 --> 00:30:24.840 and empirically Bayesian prior approaches here,

794 00:30:24.840 --> 00:30:28.740 fully Bayesian meaning that it is defined a priori

795 00:30:28.740 --> 00:30:30.870 and is agnostic to the data you've collected,

796 00:30:30.870 --> 00:30:32.040 empirically Bayesian meaning

797 00:30:32.040 --> 00:30:33.540 we leverage the data we've collected

798 00:30:33.540 --> 00:30:36.843 to help inform that prior for what we've observed.

799 00:30:38.010 --> 00:30:39.930 Specifically there is a what we call a

800 00:30:39.930 --> 00:30:41.580 non constrained, or naive,

801 00:30:41.580 --> 00:30:43.230 empirically based prior

802 00:30:43.230 --> 00:30:45.300 where we would look through all of those growths we had

803 00:30:45.300 --> 00:30:46.590 and we would say, "Whichever one of these

804 00:30:46.590 --> 00:30:49.170 maximizes the integrated marginal likelihood

805 00:30:49.170 --> 00:30:50.790 that's the correct configuration

806 00:30:50.790 --> 00:30:52.890 and we're gonna put all of our eggs into that basket."

807 00:30:52.890 --> 00:30:55.470 Or 100% of the probability there

808 00:30:55.470 --> 00:30:57.813 and that's the only model we use for analysis.

809 00:30:58.830 --> 00:30:59.940 We know, generally speaking,

810 00:30:59.940 --> 00:31:01.920 since we went to all the work to defining

811 00:31:01.920 --> 00:31:04.080 all of these different combinations of exchangeability

812 00:31:04.080 --> 00:31:05.580 and that it's based off of samples,

813 00:31:05.580 --> 00:31:07.350 potentially small samples,

814 00:31:07.350 --> 00:31:10.080 that this can be a very strong assumption.

815 00:31:10.080 --> 00:31:12.120 And so we can also modify this prior

816 00:31:12.120 --> 00:31:14.880 to what we call a constrained EB prior,

817 00:31:14.880 --> 00:31:18.150 where instead of just giving everyone of those model

818 00:31:18.150 --> 00:31:20.010 sources in that MEM that

819 00:31:20.010 --> 00:31:22.710 maximizes the likelihood 100% weight,

820 00:31:22.710 --> 00:31:25.470 we instead give it a weight of what we're calling just B.

821 00:31:25.470 --> 00:31:28.080 This is our hyper prior value here

822 00:31:28.080 --> 00:31:30.870 where if it's a value of zero or up to one,

823 00:31:30.870 --> 00:31:32.430 it'll control the amount of borrowing

824 00:31:32.430 --> 00:31:35.760 and allow other nested models of exchangeability

825 00:31:35.760 --> 00:31:39.300 to also be potentially considered for analysis.

826 00:31:39.300 --> 00:31:40.260 So for example,

827 00:31:40.260 --> 00:31:41.580 if we do set a value of one

828 00:31:41.580 --> 00:31:44.280 that actually replicates the non constrained EB prior

829 00:31:44.280 --> 00:31:47.520 and really aggressively borrows from one specific model.

830 00:31:47.520 --> 00:31:50.490 At the other extreme here, if we set a value of zero,

831 00:31:50.490 --> 00:31:53.070 we essentially recreate an independent analysis

832 00:31:53.070 --> 00:31:55.290 like assign a two stage design or just using those

833 00:31:55.290 --> 00:31:56.940 Bayesian methods for futility monitoring

834 00:31:56.940 --> 00:31:58.740 that doesn't share information.

835 00:31:58.740 --> 00:32:00.180 And then any value in between

836 00:32:00.180 --> 00:32:02.520 gives a little more granularity or control

837 00:32:02.520 --> 00:32:03.993 over the amount of borrowing.

838 00:32:06.257 --> 00:32:08.313 So with that background behind us,

839 00:32:09.276 --> 00:32:10.830 I'm gonna introduce the simulation stuff

840 00:32:10.830 --> 00:32:12.780 and then present results for a couple

841 00:32:12.780 --> 00:32:15.033 key operating characteristics for our trial.

842 00:32:15.870 --> 00:32:18.240 In this case, we're going to assume for our simulations

843 00:32:18.240 --> 00:32:19.380 that we have a basket trial

844 00:32:19.380 --> 00:32:21.300 with 10 different baskets or indications.

845 00:32:21.300 --> 00:32:23.370 So again, that's 10 different types of cancer

846 00:32:23.370 --> 00:32:25.260 that we have enrolled that all have

847 00:32:25.260 --> 00:32:28.290 the same genetic mutation that we think is targeted

848 00:32:28.290 --> 00:32:31.080 by the therapy of interest.

849 00:32:31.080 --> 00:32:31.950 Like we had before,

850 00:32:31.950 --> 00:32:36.420 we're going to assume a null response P knot of 0.1 or 10%.

851 00:32:36.420 --> 00:32:40.053 And an alternative response rate of 30% or P1 here.

852 00:32:41.040 --> 00:32:43.260 We are gonna compare then three different designs

853 00:32:43.260 --> 00:32:46.110 that we just spent some time introducing and outlining.

854 00:32:46.110 --> 00:32:49.110 The first is a Simon minimax two-stage design

855 00:32:49.110 --> 00:32:52.200 using that exact set up that we had before

856 00:32:52.200 --> 00:32:53.700 where we will enroll 16 people,

857 00:32:53.700 --> 00:32:56.490 determine if we have one or fewer observations of success.

858 00:32:56.490 --> 00:32:58.020 If so, stop the trial.

859 00:32:58.020 --> 00:32:59.343 If not, continue on.

860 00:33:00.210 --> 00:33:01.110 In the second case,

861 00:33:01.110 --> 00:33:02.940 we're going to implement a Bayesian design

862 00:33:02.940 --> 00:33:05.100 that uses predictive probability monitoring

863 00:33:05.100 --> 00:33:07.050 but we don't use any information sharing

864 00:33:07.050 --> 00:33:08.760 just to illustrate that we can at least

865 00:33:08.760 --> 00:33:11.670 potentially improve upon the frequency

866 00:33:11.670 --> 00:33:14.400 in use of a interim monitoring above a single look

867 00:33:14.400 --> 00:33:16.590 from the Simon minimax design.

868 00:33:16.590 --> 00:33:17.850 And then the third design

869 00:33:17.850 --> 00:33:19.920 will add another layer of complexity

870 00:33:19.920 --> 00:33:22.830 where we will try to share information across baskets

871 00:33:22.830 --> 00:33:26.613 that have what we estimate to be exchangeable subgroups.

872 00:33:27.510 --> 00:33:28.620 One thing to note here is that

873 00:33:28.620 --> 00:33:32.460 we are setting this hyper parameter value B at 0.1.

874 00:33:32.460 --> 00:33:34.020 This is a fairly conservative value

875 00:33:34.020 --> 00:33:36.240 and admittedly for this design

876 00:33:36.240 --> 00:33:38.520 we actually did not calibrate specifically

877 00:33:38.520 --> 00:33:40.200 for the amount of borrowing to be 0.1.

878 00:33:40.200 --> 00:33:41.130 This is actually based off of

879 00:33:41.130 --> 00:33:42.630 some other prior work we've done

880 00:33:42.630 --> 00:33:44.970 and published on basket trials that just showed that

881 00:33:44.970 --> 00:33:48.750 in the case of an empirically Bayesian prior for MEMs,

882 00:33:48.750 --> 00:33:50.970 this actually allows information sharing

883 00:33:50.970 --> 00:33:53.310 in cases where there's a high degree of exchangeability

884 00:33:53.310 --> 00:33:54.990 and low heterogeneity

885 00:33:54.990 --> 00:33:56.850 and down leap it in cases where we might be

886 00:33:56.850 --> 00:33:57.840 a little more uncertain,

887 00:33:57.840 --> 00:33:59.130 so it's a little more conservative

888 00:33:59.130 --> 00:34:01.290 but we'll see in the simulation results

889 00:34:01.290 --> 00:34:03.063 there are some potential benefits.

890 00:34:05.040 --> 00:34:08.070 For each of the scenarios we're gonna look at today,

891 00:34:08.070 --> 00:34:10.140 we will generate a thousand trials

892 00:34:10.140 --> 00:34:13.950 with a maximum sample size of 25 per basket.

893 00:34:13.950 --> 00:34:15.900 We're gonna look at two cases,

894 00:34:15.900 --> 00:34:17.100 there's a few other in the paper

895 00:34:17.100 --> 00:34:19.590 but we're gonna focus on first the global scenario

896 00:34:19.590 --> 00:34:21.360 where all the baskets are either null

897 00:34:21.360 --> 00:34:24.420 or all 10 baskets have some meaningful effect.

898 00:34:24.420 --> 00:34:25.440 And this is the setting where

899 00:34:25.440 --> 00:34:27.180 information sharing methods like meds

900 00:34:27.180 --> 00:34:29.220 really should outperform anything else

901 00:34:29.220 --> 00:34:31.440 because everything is truly exchangeable

902 00:34:31.440 --> 00:34:33.960 and everything could naively be pooled together

903 00:34:33.960 --> 00:34:37.110 because we're simulating them to have the same response.

904 00:34:37.110 --> 00:34:38.970 We'll then look at what happens if we actually have

905 00:34:38.970 --> 00:34:40.200 a mixed scenario,

906 00:34:40.200 --> 00:34:41.970 which I think is actually more indicative

907 00:34:41.970 --> 00:34:43.170 of what's happened in practice

908 00:34:43.170 --> 00:34:45.300 with some of the published basket trials

909 00:34:45.300 --> 00:34:47.460 and clinically what we've seen from applications

910 00:34:47.460 --> 00:34:49.200 of these types of designs.

911 00:34:49.200 --> 00:34:51.510 Specifically here, we're gonna look at the case where

912 00:34:51.510 --> 00:34:54.813 there are eight null baskets and two alternative baskets.

913 00:34:56.940 --> 00:34:59.220 A few other points just to highlight here.

914 00:34:59.220 --> 00:35:02.070 We're going to assume a beta 0.5 0.5 prior

915 00:35:02.070 --> 00:35:03.540 for our Bayesian models.

916 00:35:03.540 --> 00:35:06.360 This essentially for a binary outcome can be thought of as

917 00:35:06.360 --> 00:35:07.860 adding half of a response

918 00:35:07.860 --> 00:35:12.270 and half of a lack of a response to our observed data.

919 00:35:12.270 --> 00:35:15.810 We're going to look at the most extreme dream Bayesian case

920 00:35:15.810 --> 00:35:17.850 of doing utility monitoring

921 00:35:17.850 --> 00:35:20.340 or any type of interim monitoring continually.

922 00:35:20.340 --> 00:35:22.410 So after every single participant's enrolled

923 00:35:22.410 --> 00:35:23.940 we will do a calculation

924 00:35:23.940 --> 00:35:26.880 and determine if we should stop the trial.

925 00:35:26.880 --> 00:35:29.340 We will then look at the effect of this choice

926 00:35:29.340 --> 00:35:33.660 across a range of predictive probability thresholds

927 00:35:33.660 --> 00:35:35.070 ranging from 0%,

928 00:35:35.070 --> 00:35:36.640 meaning we wouldn't stop early at all,

929 00:35:36.640 --> 00:35:39.360 up to 50% saying if there's anything less

930 00:35:39.360 --> 00:35:41.250 than a 50% chance I'll find success,

931 00:35:41.250 --> 00:35:42.723 I wanna stop that trial.

932 00:35:43.800 --> 00:35:45.180 And then finally it's worth noting

933 00:35:45.180 --> 00:35:49.020 we're actually also completely disregarding calibration

934 00:35:49.020 --> 00:35:50.910 for this interim monitoring.

935 00:35:50.910 --> 00:35:51.930 And so what we're gonna do is

936 00:35:51.930 --> 00:35:53.820 we're gonna calibrate our decision rules

937 00:35:53.820 --> 00:35:57.120 for the posterior probability at the end of the trial

938 00:35:57.120 --> 00:35:58.920 based off of a global scenario where

939 00:35:58.920 --> 00:36:01.560 we think it's ideal to share information

940 00:36:01.560 --> 00:36:03.270 and we're all not gonna account for the fact that

941 00:36:03.270 --> 00:36:05.400 we're doing interim looks at the data.

942 00:36:05.400 --> 00:36:06.810 Part of the question here was

943 00:36:06.810 --> 00:36:08.490 if we truly do all these assumptions

944 00:36:08.490 --> 00:36:11.070 and we do sort of the most naive thing,

945 00:36:11.070 --> 00:36:12.750 how badly do we actually do?

946 00:36:12.750 --> 00:36:15.570 Like is there enough reason to fear the results

947 00:36:15.570 --> 00:36:18.333 if we don't correctly calibrate for everything here?

948 00:36:20.970 --> 00:36:24.090 So I'm gonna paint some pictures here building from the

949 00:36:24.090 --> 00:36:27.210 simpler Simon design to our more complex Bayesian designs

950 00:36:27.210 --> 00:36:28.500 and then with information sharing

951 00:36:28.500 --> 00:36:31.260 just to illustrate three different properties.

952 00:36:31.260 --> 00:36:32.790 I'm gonna go fairly quickly

953 00:36:32.790 --> 00:36:35.490 'cause I know that you all have to vacate the classroom

954 00:36:35.490 --> 00:36:37.200 in about 10 minutes.

955 00:36:37.200 --> 00:36:40.500 So for the global scenario that we're looking at here,

956 00:36:40.500 --> 00:36:43.860 the like rate lines are going to represent

957 00:36:43.860 --> 00:36:45.870 the alternative basket scenario.

958 00:36:45.870 --> 00:36:48.900 So all, in this case, all 10 null baskets.

959 00:36:48.900 --> 00:36:51.030 Here we see we plan for 90% power

960 00:36:51.030 --> 00:36:52.860 Simon's design appropriately achieved

961 00:36:52.860 --> 00:36:55.230 that rejection rate of 90%.

962 00:36:55.230 --> 00:36:57.750 Likewise, the lines at the bottom here,

963 00:36:57.750 --> 00:36:58.950 these black lines,

964 00:36:58.950 --> 00:37:01.200 are going to represent the results of null baskets.

965 00:37:01.200 --> 00:37:02.760 Here are the global null scenario

966 00:37:02.760 --> 00:37:05.433 and we see that it achieves a 10% rejection rate.

967 00:37:06.330 --> 00:37:08.130 Now, this is a flat line here

968 00:37:08.130 --> 00:37:10.830 because again Simon's design is agnostic to things like

969 00:37:10.830 --> 00:37:12.273 the predictive probability.

970 00:37:13.110 --> 00:37:16.320 Now if we do frequent Bayesian monitoring,

971 00:37:16.320 --> 00:37:18.870 we see two interesting things here with these new lines.

972 00:37:18.870 --> 00:37:20.820 We see that at the top

973 00:37:20.820 --> 00:37:22.800 and the bottom, here I add these circles

974 00:37:22.800 --> 00:37:25.320 where the predictive probability threshold is 0%.

975 00:37:25.320 --> 00:37:27.090 This does represent the actual design

976 00:37:27.090 --> 00:37:29.490 that would correspond to the actual calibration we did

977 00:37:29.490 --> 00:37:31.260 without interim monitoring.

978 00:37:31.260 --> 00:37:33.690 And we see that it is possible with Bayesian approaches

979 00:37:33.690 --> 00:37:37.050 to achieve the same frequent operating characteristics

980 00:37:37.050 --> 00:37:40.200 that we would achieve with something like the Simon design.

981 00:37:40.200 --> 00:37:42.720 We can see though that if we want to do interim monitoring

982 00:37:42.720 --> 00:37:43.830 but we didn't calibrate

983 00:37:43.830 --> 00:37:45.840 or think of that in our calculations,

984 00:37:45.840 --> 00:37:48.660 we do see this trade off where we have our

985 00:37:48.660 --> 00:37:50.970 alternative baskets having a decreasing power

986 00:37:50.970 --> 00:37:53.310 or rejection rate as the aggressiveness of the

987 00:37:53.310 --> 00:37:55.920 predictive probability threshold increases.

988 00:37:55.920 --> 00:37:58.620 And likewise the type one error rate or the

989 00:37:58.620 --> 00:38:01.533 rejection rate of the marginal baskets also decreases.

990 00:38:02.370 --> 00:38:05.850 Now if we add information sharing to this design,

991 00:38:05.850 --> 00:38:07.740 we actually see some encouraging results

992 00:38:07.740 --> 00:38:09.360 in this global scenario.

993 00:38:09.360 --> 00:38:11.070 First, it's worth noting that in the case

994 00:38:11.070 --> 00:38:12.360 where we actually calibrated for,

995 00:38:12.360 --> 00:38:17.010 we actually see an increase in power from 90% to about 97%.

996 00:38:17.010 --> 00:38:18.690 And even when we actually have a

997 00:38:18.690 --> 00:38:23.400 10% predictive probability threshold for interim monitoring,

998 00:38:23.400 --> 00:38:26.070 we see that we actually still achieve 90% power

999 00:38:26.070 --> 00:38:30.450 with a corresponding reduction in that type one error rate.

1000 00:38:30.450 --> 00:38:32.100 Of course, this is with the caveat that

1001 00:38:32.100 --> 00:38:34.650 this is the ideal setting for sharing information

1002 00:38:34.650 --> 00:38:37.383 because all of the baskets are truly exchangeable.

1003 00:38:38.220 --> 00:38:40.620 Now the rejection rate correlates to something we call

1004 00:38:40.620 --> 00:38:41.760 that expected sample size.

1005 00:38:41.760 --> 00:38:43.830 What is the average sample size we might enroll

1006 00:38:43.830 --> 00:38:47.010 for each basket of our 10 baskets in the trial?

1007 00:38:47.010 --> 00:38:49.590 We see here that in the case of a null basket

1008 00:38:49.590 --> 00:38:51.363 the Simon design is about 20.

1009 00:38:53.250 --> 00:38:55.560 If we do interim monitoring with Bayesian approaches

1010 00:38:55.560 --> 00:38:57.450 and no information sharing,

1011 00:38:57.450 --> 00:38:59.137 obviously if we don't do any interim looks at the data,

1012 00:38:59.137 --> 00:39:01.380 we have a 0% threshold,

1013 00:39:01.380 --> 00:39:05.040 we're gonna have a sample size of 25 every single time.

1014 00:39:05.040 --> 00:39:06.330 I think what's encouraging though is that

1015 00:39:06.330 --> 00:39:09.450 by looking fairly aggressively we see that our sample size,

1016 00:39:09.450 --> 00:39:10.920 even with a very marginal

1017 00:39:10.920 --> 00:39:13.890 or low 5% threshold for futility monitoring,

1018 00:39:13.890 --> 00:39:17.910 drops from 20 in the assignment design to about 15

1019 00:39:17.910 --> 00:39:19.380 in the Bayesian design,

1020 00:39:19.380 --> 00:39:21.660 the trade-off of course being because we didn't calibrate.

1021 00:39:21.660 --> 00:39:24.300 We also see a reduction in the sample size

1022 00:39:24.300 --> 00:39:25.863 for the alternative baskets.

1023 00:39:27.630 --> 00:39:30.030 And if we add that layer of information sharing,

1024 00:39:30.030 --> 00:39:32.070 we actually see that we do slightly better than

1025 00:39:32.070 --> 00:39:33.840 the design without information sharing

1026 00:39:33.840 --> 00:39:36.870 while attenuating at the top here the effect

1027 00:39:36.870 --> 00:39:39.933 our solid gray line has for the alternative baskets.

1028 00:39:41.580 --> 00:39:44.550 Now, briefly tying this together then to the stopping rate,

1029 00:39:44.550 --> 00:39:47.190 which we can kind of infer from those past results,

1030 00:39:47.190 --> 00:39:50.400 we do see that on average the Simon two-stage design

1031 00:39:50.400 --> 00:39:52.380 for the null baskets stopping for futility

1032 00:39:52.380 --> 00:39:55.170 is only taking place a little over 50% of the time

1033 00:39:55.170 --> 00:39:56.580 in this simulation.

1034 00:39:56.580 --> 00:39:58.290 The advantage here though is that it is

1035 00:39:58.290 --> 00:40:01.233 very rarely stopping for the alternative baskets.

1036 00:40:02.340 --> 00:40:03.480 In our Bayesian approaches,

1037 00:40:03.480 --> 00:40:06.050 we see that there is an over 80%

1038 00:40:06.050 --> 00:40:08.790 of these low thresholds probability of stopping

1039 00:40:08.790 --> 00:40:10.200 if it's a null effect.

1040 00:40:10.200 --> 00:40:12.270 And this is ideal because we have 10 baskets.

1041 00:40:12.270 --> 00:40:14.130 And so these potential savings or effects

1042 00:40:14.130 --> 00:40:16.830 can compound themselves across these multiple baskets.

1043 00:40:18.300 --> 00:40:20.910 We then see that the design adding these solid lines

1044 00:40:20.910 --> 00:40:23.010 for information sharing do very similarly

1045 00:40:23.010 --> 00:40:25.860 where again the the consequence of not calibrating

1046 00:40:25.860 --> 00:40:28.473 are attenuated in this circumstance.

1047 00:40:29.550 --> 00:40:31.230 Now the thing to note here that

1048 00:40:31.230 --> 00:40:33.840 everything I presented on these few graphics

1049 00:40:33.840 --> 00:40:36.210 were with respect to the global scenario,

1050 00:40:36.210 --> 00:40:38.190 that ideal scenario that I actually don't think

1051 00:40:38.190 --> 00:40:40.710 is super realistic in practice.

1052 00:40:40.710 --> 00:40:43.920 So we see here, if we do a mixed scenario where

1053 00:40:43.920 --> 00:40:46.470 we now have calibrated for the global scenarios,

1054 00:40:46.470 --> 00:40:48.210 we've miscalibrated with respect to that.

1055 00:40:48.210 --> 00:40:51.480 We've also not calibrated for interim looks at the data.

1056 00:40:51.480 --> 00:40:53.010 We can actually see that the results for

1057 00:40:53.010 --> 00:40:55.410 the Simon two-stage in the Bayesian design

1058 00:40:55.410 --> 00:40:57.510 without information sharing are very similar

1059 00:40:57.510 --> 00:40:58.650 to what we saw before.

1060 00:40:58.650 --> 00:41:00.180 That's because they don't share information.

1061 00:41:00.180 --> 00:41:02.070 And so in this case with eight null baskets

1062 00:41:02.070 --> 00:41:03.840 into alternative baskets,

1063 00:41:03.840 --> 00:41:06.210 they have very similar responses.

1064 00:41:06.210 --> 00:41:09.060 This contrasts of course with the MEM approach

1065 00:41:09.060 --> 00:41:10.260 or the information sharing approach

1066 00:41:10.260 --> 00:41:11.820 where we actually see now

1067 00:41:11.820 --> 00:41:14.610 many of these results are actually overlapping

1068 00:41:14.610 --> 00:41:17.700 for information sharing and no information sharing.

1069 00:41:17.700 --> 00:41:21.030 What this tells us is that even though we miscalibrated

1070 00:41:21.030 --> 00:41:23.040 up and down the design,

1071 00:41:23.040 --> 00:41:25.680 we are actually able with this more conservative prior

1072 00:41:25.680 --> 00:41:27.450 to down weight borrowing

1073 00:41:27.450 --> 00:41:30.000 and effectuate similar results

1074 00:41:30.000 --> 00:41:34.020 that at lower thresholds for utility monitoring for example

1075 00:41:34.020 --> 00:41:38.190 at 5% can still show potential gains in efficiency relative

1076 00:41:38.190 --> 00:41:40.830 to the Simon design that could likely further be improved

1077 00:41:40.830 --> 00:41:42.333 with actual calibration.

1078 00:41:44.130 --> 00:41:45.030 So just as a reminder,

1079 00:41:45.030 --> 00:41:46.050 we demonstrated today

1080 00:41:46.050 --> 00:41:47.610 and introduced the idea of Simon's two-stage design

1081 00:41:47.610 --> 00:41:51.150 and some alternative methods to compete with them.

1082 00:41:51.150 --> 00:41:53.400 And some just brief discussion and concluding points.

1083 00:41:53.400 --> 00:41:54.510 There is no free lunch

1084 00:41:54.510 --> 00:41:57.030 and this is true regardless of where we are in statistics

1085 00:41:57.030 --> 00:41:59.250 that for example in our designs,

1086 00:41:59.250 --> 00:42:00.930 besides the fact that we miscalibrated

1087 00:42:00.930 --> 00:42:03.840 and made it a bit harder of a comparison for our methods,

1088 00:42:03.840 --> 00:42:05.040 we did try to replicate

1089 00:42:05.040 --> 00:42:06.660 what people might be doing in practice

1090 00:42:06.660 --> 00:42:07.493 or the challenge of

1091 00:42:07.493 --> 00:42:10.350 calibrating these designs into actuality.

1092 00:42:10.350 --> 00:42:13.170 Simon's two-stage design does have a lot of benefits

1093 00:42:13.170 --> 00:42:15.360 from it's ideal characteristics

1094 00:42:15.360 --> 00:42:16.530 that are easy to implement,

1095 00:42:16.530 --> 00:42:19.590 but it is limited in how often it may stop.

1096 00:42:19.590 --> 00:42:20.670 Our Bayesian designs,

1097 00:42:20.670 --> 00:42:22.140 with or without information sharing,

1098 00:42:22.140 --> 00:42:24.480 can lead to reductions in the expected sample size

1099 00:42:24.480 --> 00:42:25.410 in the null basket

1100 00:42:25.410 --> 00:42:27.180 and further could be improved

1101 00:42:27.180 --> 00:42:28.830 if we actually incorporate calibration,

1102 00:42:28.830 --> 00:42:29.970 which we further explored

1103 00:42:29.970 --> 00:42:32.700 in a statistical methods of medical research paper

1104 00:42:32.700 --> 00:42:33.993 published in 2020.

1105 00:42:34.975 --> 00:42:36.240 And so that I have some sources here

1106 00:42:36.240 --> 00:42:37.440 and I thank you for your attention

1107 00:42:37.440 --> 00:42:40.863 and welcome any questions or discussion at this point.

1108 00:42:55.860 --> 00:42:58.610 <v Man>Thank you so much. Any questions from the room?</v>

1109 00:43:11.520 --> 00:43:14.160 <v Student>Okay, so yeah, I have questions.</v>

1110 00:43:14.160 --> 00:43:18.030 So in the example you just showed,

1111 00:43:18.030 --> 00:43:22.350 all the like the task becomes so, can be achievable, right?

1112 00:43:22.350 --> 00:43:24.840 So if the baskets,

1113 00:43:24.840 --> 00:43:27.840 they are expected to have different benefits (indistinct),

1114 00:43:27.840 --> 00:43:32.840 and say the 10 basket (indistinct)

1115 00:43:32.961 --> 00:43:37.350 some other basket MEMs would allow a bigger benefit,

1116 00:43:37.350 --> 00:43:41.010 how will the (indistinct)

1117 00:43:44.700 --> 00:43:45.753 scenarios?

1118 00:43:48.360 --> 00:43:49.260 <v Alex>Yeah, well, I think,</v>

1119 00:43:49.260 --> 00:43:50.550 if I understood your question correctly

1120 00:43:50.550 --> 00:43:53.970 and I misheard through the phone, let me know,

1121 00:43:53.970 --> 00:43:56.280 but if we have different sample sizes for baskets,

1122 00:43:56.280 --> 00:43:58.470 which actually really corresponds

1123 00:43:58.470 --> 00:44:00.690 to what we've seen in practice for real basket trials

1124 00:44:00.690 --> 00:44:02.310 where they have fairly

1125 00:44:02.310 --> 00:44:04.983 wide range of sample sizes in each basket.

1126 00:44:05.880 --> 00:44:06.870 I think what we would see,

1127 00:44:06.870 --> 00:44:08.880 and let me see if I can pop back quickly to the

1128 00:44:08.880 --> 00:44:12.720 mixed scenario results here just to illustrate some ideas.

1129 00:44:12.720 --> 00:44:13.920 One of the concepts here that,

1130 00:44:13.920 --> 00:44:15.547 so we did explicitly look at that to say like,

1131 00:44:15.547 --> 00:44:18.030 "Well, what if one basket never gets beyond seven

1132 00:44:18.030 --> 00:44:20.010 of the 25," let's say.

1133 00:44:20.010 --> 00:44:21.240 But what we can infer is that

1134 00:44:21.240 --> 00:44:23.460 if a basket stopped early for futility,

1135 00:44:23.460 --> 00:44:26.220 it essentially has a smaller sample size to contribute

1136 00:44:26.220 --> 00:44:28.920 to any analysis whether or not it was a

1137 00:44:28.920 --> 00:44:31.530 falsely stopped basket that had a 30% effect

1138 00:44:31.530 --> 00:44:33.630 or it was truly a null basket.

1139 00:44:33.630 --> 00:44:36.390 And so we do see in this case that the method

1140 00:44:36.390 --> 00:44:39.330 averaging over those ideas of differential sample sizes

1141 00:44:39.330 --> 00:44:41.160 based off of soft baskets

1142 00:44:41.160 --> 00:44:42.810 does seem to be borrowing,

1143 00:44:42.810 --> 00:44:44.790 appropriately depending on the context.

1144 00:44:44.790 --> 00:44:46.980 So like the mixed scenario results here suggests

1145 00:44:46.980 --> 00:44:49.830 limited borrowing in the presence of that uncertainty

1146 00:44:49.830 --> 00:44:50.910 from the global scenario

1147 00:44:50.910 --> 00:44:52.530 because we didn't calibrate for anything else

1148 00:44:52.530 --> 00:44:56.010 it does show more of a benefit of the stopping rate

1149 00:44:56.010 --> 00:44:58.290 and other properties incorporating that data

1150 00:44:58.290 --> 00:45:00.240 even in small sample sizes.

1151 00:45:00.240 --> 00:45:01.650 And there's also been some other work

1152 00:45:01.650 --> 00:45:03.870 and illustrations done by Dr. Emily Zebra

1153 00:45:03.870 --> 00:45:06.030 at the Cleveland Clinic with who I work

1154 00:45:06.030 --> 00:45:08.700 about some of the re-analysis of oncology trials

1155 00:45:08.700 --> 00:45:11.280 that do show even in small basket sizes,

1156 00:45:11.280 --> 00:45:14.100 we can move that significance evaluation

1157 00:45:14.100 --> 00:45:16.283 into a more clinically meaningful realm.

1158 00:45:26.312 --> 00:45:30.312 <v Wayne>Thanks, so do we have other questions?</v>

1159 00:45:57.093 --> 00:46:01.469 Okay, so (indistinct) that's (indistinct).

1160 00:46:01.469 --> 00:46:06.469 Okay, so since there are no questions let's stop here.

1161 00:46:06.699 --> 00:46:09.116 (indistinct)

1162 00:46:16.028 --> 00:46:18.445 <v Alex>Yeah. Thank you all.</v>