Incompetents, Accomplices, or Criminals? Panel Fraud in Health Surveys.

Webinar Transcript

November 2024

1 - 46

00:00:02.400 → 00:03:04.680

(webinar setup)

46

00:03:04.680 → 00:03:12.330

Anton: So Hello, everyone, and and welcome to today's webinar on the topic of panel fraud in health surveys.

47

00:03:12.817 → 00:03:19.150

Anton: We're happy to have all of you here. My name is Anton. I'll be one of the the moderators during this webinar.

48

00:03:19.280 → 00:03:26.569

Anton: The duration of the session is 1 h, and we'll be opening up for a Q&A by the last 15 min or so.

49

00:03:26.650 → 00:03:39.000

Anton: So, if you have comments or questions, feel free to pop them into the the Q. And A. Chats, and we'll address it after the presentation. You're also able to leave questions anonymously, if that's more convenient.

50

00:03:39.180 → 00:03:45.269

Anton: But make sure you stay until the end for the Q. And a. And information on how to get the toolkit.

51

00:03:45.730 → 00:03:49.269

Anton: and just a short disclaimer before we start.

52

00:03:49.330 → 00:03:54.210

Anton: Now we operate independently and do not represent any panel.

53

00:03:54.250 → 00:03:57.890

Anton: Not all panels are fraudulent as described.

54

00:03:57.930 → 00:04:02.610

Anton: and sensitive data has been altered to show indicative results. Only.

55

00:04:03.060 → 00:04:10.340

Anton: Now, with that said I will leave it over to Sylvie and Natalia to give a brief introduction here. Thanks.

56

00:04:11.430 → 00:04:39.810

Sylvie: Okay. So we see that we have a global attendance today. So good morning. Good afternoon. Good evening to you. I am Sylvie Magalon, the head of sales and marketing at Survey engine, and let me start by introducing our speakers for today's webinar. Starting with Ben White. Ben is the founder of survey engine with Ludwig and is our Director of technology. He has a background in physics and mathematics.

57

00:04:39.810 → 00:04:44.730

Sylvie: and his career actually began with developing airline scheduling systems.

58

00:04:44.730 → 00:04:59.210

Sylvie: Ben also developed some of the 1st applications in behavioral modeling in state of preference and developed the software for the 1st trials of the best, worst methods.

59

00:04:59.350 → 00:05:05.709

Sylvie: He's the driving force behind the development of the Survey engine research platform.

60

00:05:06.850 → 00:05:35.820

Sylvie: Next, we have Ludwig Butler, co-founder of Survey Engine and the Director of Research as an economist by training he built a solid track record spanning over a decade in decision, modeling and market research. And here at Survey engine he combines his passion for

understanding, human behavior and preferences to really drive evidence-based decision in healthcare.

61

00:05:36.740 → 00:06:04.210

Sylvie: And finally, I'd like to introduce Natalia Bogado, our head of production. Natalia holds a Phd in behavioral psychology and has extensive experience in experimental research, including many global international publications and collaboration. And at survey engine, she leads our team of researchers and data specialists.

62

00:06:06.020 → 00:06:19.389

Sylvie: So this is the agenda for today's webinar. So we're really excited to present to you today, we'll be talking about different types of panels. We'll go over prevention and detections to fraud.

63

00:06:19.390 → 00:06:39.039

Sylvie: and at the end we'll show you a new methods that will actually had quite some good success with it, as well as the toolkit that you can take with you after the webinar, so make sure that you stay tuned until the end, and now I will hand over to my colleague Natalia.

64

00:06:40.170 → 00:06:58.009

Natalia: Thank you, Sylvie. Thank you, Anton, for the introduction, and welcome everyone to this webinar as both the head of production at Survey engine, but also as a researcher myself, that collects and publishes data.

65

00:06:58.310 → 00:07:01.760

Natalia: can tell you that data quality is a constant concern for me

66

00:07:01.840 → 00:07:08.600

Natalia: and and not just for me. But this is a pressing concern across the entire industry, and it is for good reason.

67

00:07:08.720 → 00:07:32.360

Natalia: So a recent study from June, as recent as June, has found that 84% of healthcare researchers reported fraudulent participation in their studies and had to actually exclude data from their studies due to fraud. Another study also found that a decline in usable survey responses in online surveys from 75

68

00:07:32.360 → 00:07:39.190

Natalia: down to just 10% of usable data in recent years. And this is again due to fraud.

69

00:07:39.330 → 00:07:54.669

Natalia: Now these numbers are shocking because they show just how exposed we are to data fraud and how it threatens the work that we do as researchers as well. Now next slide, please, Sylvie.

70

00:07:56.270 → 00:08:22.850

Natalia: So now, so you may ask yourself, why ask why survey engine? Why are we talking to you about this issue. And the short answer is because of our experience. So we've handled thousands of research projects and been cited in hundreds of papers. And we've also been

involved in thousands of healthcare studies where high participant incentives make them a perfect target for fraudsters.

71

00:08:22.980 → 00:08:27.539

Natalia: Now, these studies are also very much scrutinized.

72

00:08:27.800 → 00:08:35.439

Natalia: So any data issues are far more salient in this type of research than in any other area of research.

73

00:08:36.020 → 00:08:38.099

Natalia: And then, finally, and

74

00:08:38.220 → 00:08:51.869

Natalia: most importantly, because our behavioral models like Dce and Dtos, they tend to reveal any behavioral anomalies more than any other methods that are used in behavioral research.

75

00:08:52.393 → 00:09:00.876

Natalia: So why is solving this issue so important? Why does it matter so much? Why are we taking the time to do this webinar

76

00:09:01.720 → 00:09:14.539

Natalia: As healthcare researchers? We have an obligation. We have a responsibility, I would say, to ensure the integrity of the data that we use to ensure the quality of the data that we use in our studies

77

00:09:14.590 → 00:09:30.490

Natalia: because bad data can lead to disastrous outcomes. For instance, rejecting a valid hypothesis because of flawed data, or even worse, accepting an invalid hypothesis because we're using unknowingly using fraudulent data.

78

00:09:30.490 → 00:09:49.100

Natalia: This just risks undermining the very foundations of science and of the production of scientific knowledge more broadly. So the question is, what do we do? Do we abandon online research altogether? And the answer is, no.

79

00:09:49.750 → 00:09:51.960

Natalia: With the right safeguards in place

80

00:09:52.140 → 00:10:17.549

Natalia: and with the right measures we can. And online research is entirely feasible and remains a powerful tool for advancing scientific knowledge and research. And however, it seems like the responsibility for ensuring data quality has been increasingly sort of brought in house has been increasingly transferred to the researcher.

81

00:10:17.580 → 00:10:28.110

Natalia: And so in this landscape Ben will walk you through some of the ways that we can tackle these challenges in behavioral modeling studies.

82

00:10:28.440 → 00:10:32.760

Natalia: So I will hand this over now to to my colleague, Ben.

83

00:10:36.930 → 00:10:54.860

Ben: The great 1st slide. Thanks, Natalia so, and thanks everybody for joining. So let's start from the beginning. Fraud operates in the dark. Fraudsters don't announce themselves, and they don't share their methods, although strictly, that's not entirely true, as we will show you in a couple of slides from now.

84

00:10:55.364 → 00:10:59.319

Ben: So some of this is going to be guesswork based on our experiences.

85

00:10:59.410 → 00:11:10.019

Ben: So we're going to show you what we've found and how we think fraud works and what we can do about it. And in the end I'll share a technique that's consistently shown to be the best tool that we have

86

00:11:10.670 → 00:11:13.530

Ben: next slide. Thanks, Nata Sylvie.

87

00:11:14.710 → 00:11:30.459

Ben: So when talking about online response, this is often where people start, we have an idealized respondent, which is what we let's say naively, think that that would be the only one. Then we have our professional respondent, the one who does it for the money, but but could well be a valid respondent.

88

00:11:30.540 → 00:11:38.200

Ben: Then we get our unengaged respondents, who may start just bored, or maybe halfway through, because the study is boring, and then, of course, we have our bots.

89

00:11:38.880 → 00:11:42.609

Ben: But of course, this is way too simplistic.

90

00:11:42.950 → 00:11:44.619

Ben: We can go to the next slide.

91

00:11:45.250 → 00:11:49.535

Ben: Virtually all studies are mediated by

92

00:11:50.350 → 00:12:04.829

Ben: well, in this case we're talking about panels, but this could be any intermediary, whether it's pags, whether it could be a campaign on Linkedin or Facebook or your own database. Somebody else. There's somebody in the middle that we are trusting to provide valid respondents.

93

00:12:06.650 → 00:12:12.670

Ben: So in here we've got, we've characterized them as broadly as 4 types of panels

94

00:12:13.060 → 00:12:20.420

Ben: on the far left we have. Well, I've chosen Michael Palin here. This is a man who's never had a bad word spoken about him to represent the ideal

95

00:12:20.480 → 00:12:28.709

Ben: reputable panel. Somebody that has control over the panelists operates in good faith, and if there's a problem we'll jump in and and help you solve it.

96

00:12:29.120 → 00:12:49.270

Ben: Then we have a dear Steve Carell who represents the incompetent panels, people who just basically close their eyes, and then the accomplices and the criminals. So we're not going to hear any more about Michael Palin, because we're here to see see the dodgy characters on the 3 on the left here. So let's meet them. Let's go through them one by one.

97

00:12:50.670 → 00:12:51.899

Ben: Next slide, please.

98

00:12:52.280 → 00:12:55.649

Ben: So let's start with the let's call them the incompetent.

99

00:12:55.770 → 00:13:00.460

Ben: So these are panels who make very large claims about having millions of panelists all over the world.

100

00:13:00.560 → 00:13:06.130

Ben: and you know, when you double, count all their affiliates and partners, this may well be true.

101

00:13:06.450 → 00:13:14.960

Ben: So they're not inherently dishonest. But they've, you know, often got a small list and a long list of partners and affiliates that they then source.

102

00:13:15.560 → 00:13:27.210

Ben: They typically have very weak control or even knowledge of their respondents. They're not particularly interested in policing bad respondents, and this is just simply because they they can't. They don't know them.

103

00:13:27.510 → 00:13:35.480

Ben: And further, there's there's no immediate benefit over trying to police them, rather that they'd rather rely on you to root out the problems.

104

00:13:35.930 → 00:13:38.989

Ben: So I'm going to show you a case study now on the next slide.

105

00:13:39.620 → 00:13:50.100

Ben: It's a good example of this. Now, this was actually a very reputable panel who had just started an affiliate in the Uk. So there's possibly some excuse there. But nevertheless.

106

00:13:50.420 → 00:13:55.489

Ben: so we got the data in, and it was. It was an extremely poor behavioral model.

107

00:13:55.730 → 00:14:03.759

Ben: It was a study on oncology, and we were expecting the respondents to be of retirement age with advanced cancer.

108

00:14:04.040 → 00:14:06.290

Ben: Everything about the data was terrible.

109

00:14:06.300 → 00:14:32.910

Ben: I mean, the behavioral model basically showed no preference for anything but falling over a bit more often. And if the data was to be believed there was a lot of 30 year olds with advanced cancer filling out a survey at 2 Am. Before going to their full-time work, and this was a very easy one to spot even without some of the methods. But I'm going to go through blow by blow. What a typical analysis that Natalia and her team would do if we'd go to the next slide.

110

00:14:35.300 → 00:14:38.439

Ben: So I love this one. This is one of the most powerful methods.

111

00:14:38.660 → 00:15:01.740

Ben: What we've got here is the incidence rate by time it's loosely by time. It's actually by time ordered entry. So in this study, we're expecting an incidence rate of about 2% pretty low, because it was fairly rare, and there was additional screening criteria. But when we look at the incidence rate over time, it starts off very low and then gradually increases, increases

112

00:15:02.340 → 00:15:18.329

Ben: until within the sample. We're getting basically a hundred percent of people who who present with these with the sample frame which is just unbelievable. So what's going on? Either people are getting sick on a daily basis, or what's really happening is

113

00:15:18.470 → 00:15:26.059

Ben: that the respondents? And we think this is a very small group, were learning how to get through the screener, passing that information on.

114

00:15:26.200 → 00:15:29.970

Ben: So the next one of their colleagues could then get through.

115

00:15:30.200 → 00:15:32.290

Ben: We're we're almost certain of this.

116

00:15:32.790 → 00:15:37.500

Ben: So this is just one example. On the next slide. I'll show you a different analysis of the same data.

117

00:15:39.240 → 00:15:52.950

Ben: So what we're looking at here is the expected reading speed by page number in the survey. So the blue line is the expected reading speed. It's fairly simple. It's basically based on the amount of textual content in the survey.

118

00:15:53.410 → 00:16:05.499

Ben: And the bottom green line is the actual time spent. As you see, there's absolutely no correlation. Whether there was 3 pages long, 50 questions or just one question. It was the same amount of time. So this is an advanced version, I suppose, of the speeders.

119

00:16:05.880 → 00:16:15.880

Ben: and there's a scatterplot on the top right? You can see there's absolutely no correlation. So this is, this has been pretty good, and I'll show you this later in a in a false negative example as well.

120

00:16:17.344 → 00:16:25.949

Ben: And look. As I said this, this really was an easy one. In fact, with the wash up. We believe it was precisely 2 individuals

121

00:16:26.280 → 00:16:38.299

Ben: in the same time zone, actually, as the Indian subcontinent, which was not Australia, setting up fake accounts unpoliced, purely for the incentive, which would have been several $1,000 each

122

00:16:38.370 → 00:16:40.190

Ben: if they'd got through it.

123

00:16:41.220 → 00:16:48.480

Ben: Next, we're going to go on to the next one of our rogues gallery. And this is the accomplice. And we're really getting to the crooks now.

124

00:16:49.030 → 00:16:54.990

Ben: and actually on the very next screen. Not yet, Sylvie, but on the very next screen. I've got something quite surprising to show everybody.

125

00:16:55.438 → 00:16:59.340

Ben: No, no, we gotta. We stay on that one back to the the accomplice.

126

00:16:59.480 → 00:17:00.093

Ben: Thank you.

127

00:17:01.210 → 00:17:04.519

Ben: So no. Did no reveal back.

128

00:17:05.240 → 00:17:11.650

Ben: Okay, so these are regular panels, although they present as regular panels.

129

00:17:12.380 → 00:17:18.490

Ben: And they've got similar claims to the incompetent. But these guys are clever and they're sneaky.

130

00:17:18.920 → 00:17:29.950

Ben: So as we're setting up the survey with these these chaps, we have to tell them how to pass the screeners so that their project managers can actually test the survey. And at this point it's business as usual.

131

00:17:29.980 → 00:17:33.849

Ben: But then what they do is next slide big reveal.

132

00:17:35.130 → 00:17:37.290

Ben: They tell their respondents.

133

00:17:37.450 → 00:17:40.899

Ben: And this is a real screen grab of one of these characters.

134

00:17:41.660 → 00:17:49.939

Ben: and it's a screen grab of their admin system. They accidentally shared with us. And let's just take a moment to have a look. What's really going on. This should terrify you.

135

00:17:50.610 → 00:18:06.509

Ben: This is a this is a Prescreen that they send out to the people they're going to send into our survey, and they have a number of just standard questions. But the last question they actually give them direct, explicit information on how not only to get into this study, but what to do once they're in it.

136

00:18:06.880 → 00:18:08.420

Ben: This is their text.

137

00:18:08.440 → 00:18:15.570

Ben: This is a quality check, and you will need to select none of the above to proceed. Please do the same once you are in the main survey.

138

00:18:16.830 → 00:18:24.570

Ben: So this. This stuff is very, very hard to pick, because these are real people going through there. And real humans are much harder to find than bots.

139

00:18:25.450 → 00:18:33.030

Ben: So this is a particularly difficult one to spot. We were. We were lucky in this case. There was other things that were wrong with the data, but this is why

140

00:18:33.090 → 00:18:36.519

Ben: you can get bad data from from real people

141

00:18:37.340 → 00:18:47.287

Ben: on the next screen. I'm gonna we're gonna go into the criminals. Oh, sorry. Stop! Stop so just quickly. This is what that data might look like from from

142

00:18:47.850 → 00:18:59.529

Ben: these accomplices that you would get a consistently high incidence rate compared to other panels. Now, this could also be that they're actually very good, and they've got good targeting. So it's not definitive

143

00:19:00.580 → 00:19:03.130

Ben: if we go to the next slide. Thanks. So we?

144

00:19:05.170 → 00:19:13.560

Ben: And now we're getting into the proper bad guys, and we are to some extent speculating here. But but I'm these are. These are real criminals here.

145

00:19:14.230 → 00:19:24.528

Ben: So the way that they work is that we think that they are co-opting those weak or incompetent panels rather than setting up

146

00:19:25.300 → 00:19:27.420

Ben: companies which would leave a footprint.

147

00:19:27.980 → 00:19:32.039

Ben: They use bots, and we think probably more often click farms.

148

00:19:32.180 → 00:19:35.699

Ben: which are actually easier and cheaper to program than bots are.

149

00:19:36.140 → 00:19:42.730

Ben: They've probably got a bunch of side hustles and they're reusing their technology stack. They're probably doing fake rating scams as well.

150

00:19:43.490 → 00:19:50.630

Ben: But they are actually vulnerable. And let me just walk you through a speculative path. We'll stay on the screen for a second.

151

00:19:51.270 → 00:19:56.899

Ben: So let's imagine there's a high value health study that pops up an invitation on a more or less credible panel.

152

00:19:57.040 → 00:19:58.499

Ben: Let's say it's 50 bucks.

153

00:19:58.760 → 00:20:02.680

Ben: Now, on this panel is one of these criminals, and they see this.

154

00:20:03.000 → 00:20:06.219

Ben: They verify that this is a this is a good target.

155

00:20:06.380 → 00:20:11.710

Ben: and then they proceed to set up hundreds of fake accounts, to dominate the survey and share the proceeds.

156

00:20:12.470 → 00:20:18.579

Ben: Now, what might such a speculated path look like, Salvi if you could go to the next slide?

157

00:20:20.140 → 00:20:27.209

Ben: So this is our incidence rate chart again. And this is, I love this so much, and this is split by 2 different panel sources.

158

00:20:27.660 → 00:20:33.959

Ben: Now the green one is someone with a stable incidence rate, which is what you'd expect. It should be independent of the last.

159

00:20:34.370 → 00:20:40.989

Ben: But the red line is the the scammers learning how to get through, and eventually eventually

160

00:20:41.270 → 00:20:49.029

Ben: getting into the study very quickly. This gives them a huge numerical advantage, and Ludwig is going to share a particularly terrifying

161

00:20:49.380 → 00:20:56.139

Ben: consequence of this when he starts. But this is what it might look like.

162

00:20:56.450 → 00:20:59.390

Ben: Let's go to the next slide. Same study.

163

00:21:00.210 → 00:21:13.070

Ben: And this is one of the ways of trapping them. I suppose this is activity chart by day. So that 1st little hump is the pilot. Study all good. The researchers check their data, then they go for the main study, which is the second group.

164

00:21:13.300 → 00:21:18.699

Ben: And then this huge spike here is the is the criminals having worked out

165

00:21:18.900 → 00:21:30.659

Ben: how to get through it and setting up the 100 accounts. There's a certain lag time there. And so what we're tending to find, or we suspect, anyway, is that the fraud happens late in the piece, not in the early. It gives them time to

166

00:21:30.850 → 00:21:33.449

Ben: get their act together and set up hundreds of accounts.

167

00:21:34.380 → 00:21:40.510

Ben: So let's go to the next slide. Let's just put all this together, and how we think the landscape looks in fraud.

168

00:21:41.550 → 00:21:43.209

Ben: So here they all are again.

169

00:21:45.050 → 00:21:54.850

Ben: We've got our reputable people. There's Michael Palin recruiting people, applying strict controls like requiring bank accounts and verified addresses.

170

00:21:54.990 → 00:22:05.060

Ben: There's the accomplice there, supercharging ordinary respondents through to to supercharge them and get get more of his panelists through.

171

00:22:05.340 → 00:22:15.820

Ben: We've got the hapless Steve Carell there, just just smiling and and just letting everybody through without a care. And then there's the criminal, and the criminal is feeding this

172

00:22:16.310 → 00:22:25.979

Ben: through the incompetent panels, and either getting the money or learning how to get through quicker and quicker. And we think this is probably how it's working.

173

00:22:27.200 → 00:22:33.549

Ben: and of course, you guys are going to get these slides later if you want to look at them in some sort of detail. And this is at least a working hypothesis.

174

00:22:34.700 → 00:22:38.250

Ben: So look, what can we do about this? If we go to the next slide?

175

00:22:40.540 → 00:22:42.430

Ben: Well, we've got

176

00:22:42.740 → 00:22:50.949

Ben: 2 areas we can look at 1st is just preventing it outright, and many of these will be familiar to you. There's in the screening. We can use things like captures.

177

00:22:51.490 → 00:23:02.820

Ben: test questions, trap questions, cookies and Vpns. They're all vulnerable to certain, you know, to bypass, obviously catches won't get real humans. Which is why we think that click farms are involved.

178

00:23:03.760 → 00:23:10.330

Ben: but they're not to be forgotten. On the right hand side is the contractual side, and it can easily be forgotten. Obviously vendor vetting.

179

00:23:10.390 → 00:23:17.890

Ben: limiting partners or no partners and having certain amounts of transparency on the data can help.

180

00:23:18.820 → 00:23:20.340

Ben: We go to the next slide.

181

00:23:21.410 → 00:23:26.770

Ben: and these will be very familiar. These are sort of standard analyses that are done routinely

182

00:23:26.890 → 00:23:35.120

Ben: speeders and pattern responses, literacy, literacy consistency tests like that. And these happen after you've collected the data.

183

00:23:35.240 → 00:23:40.190

Ben: Now, these particular ones suffer from the fact that they are all individual based.

184

00:23:40.810 → 00:23:52.699

Ben: And they're effectively rule based. And one of the problems is you can start to introduce biases here. I mean, maybe speeders are valid. They're just getting through quicker, and I'll show you an example of valid speed is getting through.

185

00:23:52.740 → 00:23:57.639

Ben: Maybe, you know, thanks very much. Great survey is actually not a nonsense textual response.

186

00:23:57.780 → 00:24:02.139

Ben: And so they suffer from from from being individually based.

187

00:24:02.220 → 00:24:03.760

Ben: If we can go to the next slide.

188

00:24:05.830 → 00:24:16.060

Ben: so we can by relaxing that a little bit, and and to try to identify groups rather than individuals, allows us a lot more latitude in statistical methods.

189

00:24:16.100 → 00:24:32.770

Ben: And the other thing is that fraud typically happens by some kind of group. We say panels, but it can also be by wave, it could be by mode of access or other things. So by broadening a little bit, we open up the floor to better techniques.

190

00:24:33.270 → 00:24:35.830

Ben: If we can go to the next slide. Thanks, Sulvi.

191

00:24:38.320 → 00:24:46.010

Ben: So this is the one we've already been through the incidence rate check. I really recommend this. This is really good to to detect

192

00:24:46.180 → 00:24:51.430

Ben: deliberate passing of the screeners. I mean, frauds is a

193

00:24:51.610 → 00:24:55.039

Ben: their time is valuable, too. They're not going to sit there and pretend to be

194

00:24:55.070 → 00:25:02.400

Ben: 19 invalid respondents for the one that goes through. They they get to try and get through and and get the incentive, which is where it is

195

00:25:03.370 → 00:25:04.939

Ben: on the next one

196

00:25:04.960 → 00:25:15.399

Ben: is really more a supporting method. This is diurnal activity. It can't really be used as a method on its own. But this is an example of activity on the survey

197

00:25:16.390 → 00:25:19.399

Ben: for a study in Australia, and

198

00:25:19.430 → 00:25:26.760

Ben: what we're expecting is if they're real respondents to for people to be doing this generally during the waking hours. What we're seeing here is a lot of people

199

00:25:26.790 → 00:25:41.019

Ben: doing the survey in the wee hours, and this is particularly useful if the fraudsters are in a different time zone to where the study is being conducted, it was easy, because it was Australia, and this was shifted by about 6 h.

200

00:25:43.052 → 00:25:47.380

Ben: On the next one is the next slide.

201

00:25:47.540 → 00:25:56.689

Ben: So this is the reading speed. But here's a counter example. Let me just explain this. This is the good example actually, for why you might want to keep speed, as in.

202

00:25:57.740 → 00:26:04.099

Ben: So this is the same study split by the 2 major groups, one in the Uk and one in Colombia.

203

00:26:04.200 → 00:26:13.229

Ben: And what we're seeing here is that they're following the same pattern. The expected pattern. Those big spikes in the middle are the very long boring, informed consent pages.

204

00:26:13.530 → 00:26:22.060

Ben: and they are tracking together, but they are consistently taking about twice as long as the Uk sample, and

205

00:26:22.270 → 00:26:31.449

Ben: given that this study was actually on a monoculture study on coffee plantations. The simple explanation is that the Colombians were actually considering the realities more closely

206

00:26:31.480 → 00:26:42.789

Ben: rather than, let's say automatically, virtue signalling greens, credentials with no knowledge of what a coffee plantation meant. And the correlation is very high. So this is another technique. That's pretty good.

207

00:26:44.253 → 00:26:45.040

Ben: Next month.

208

00:26:50.510 → 00:26:53.610

Ben: This is our best method at the moment.

209

00:26:54.280 → 00:26:58.199

Ben: So what we're looking at here is a behavioral model of a

210

00:26:58.230 → 00:27:02.959

Ben: well, it actually doesn't matter. I'll explain that in a second. But this is a behavior model split along some.

211

00:27:03.440 → 00:27:11.230

Ben: some variable what? We, an irrelevant variable in this case it's whether they're using incognito mode.

212

00:27:11.400 → 00:27:16.119

Ben: Now, we would not expect those 2 models to be different. We'd expect them to be the to be the same.

213

00:27:16.730 → 00:27:25.580

Ben: Now, in addition, because this is actually a health valuation study, we know the the directions, and they're well published. The benchmarks are there.

214

00:27:25.620 → 00:27:42.619

Ben: and even just common sense will just tell you, and I'll just read them off for you. What they are. The top one is life expectancy, which everybody would like more of, and the rest are all increasingly negative attributes of health, from mobility, pain, depression, and so on, and we would expect them to be negative.

215

00:27:43.800 → 00:28:01.500

Ben: So on the right hand side when we removed all the incognito people. We are getting the right Valence, in the right direction. It's a good model on the left is, if you take them out, you've got a completely nonsense model. And in this particular case we're able to say, Look, that's a that's a basis for removing them.

216

00:28:02.720 → 00:28:05.159

Ben: But we can go a little bit further than this.

217

00:28:05.250 → 00:28:10.610

Ben: and this will be my last slide here. If we could go to the last one or my last one.

218

00:28:13.460 → 00:28:20.440

Ben: I think these are out of order. If we just go one more, I think the I'll go forward and backwards. Sorry about this guys, I'm just gonna just do.

219

00:28:23.090 → 00:28:25.990

Ben: So we we just go to the next one and I'll come back to the talking

220

00:28:26.330 → 00:28:28.639

Ben: no, next next to next.

221

00:28:30.610 → 00:28:31.590

Ben: That's it.

222

00:28:33.050 → 00:28:41.449

Ben: So I've done this backwards a bit. But that's all right. So we use discrete choice experiments, and they're extremely sensitive to choice behavior.

223

00:28:42.312 → 00:28:45.480

Ben: If we segment on an irrelevant, variable

224

00:28:46.141 → 00:28:51.219

Ben: we would expect it to be to to line up.

225

00:28:51.370 → 00:28:57.129

Ben: But when we get different models on some irrelevant variable, we know that there's something wrong with the data.

226

00:28:57.160 → 00:29:15.180

Ben: Now, the good thing about that is, we can know the main priority, and we can step through them whether it's screen resolution whether it's the time that they started the survey which panel they're from, and this gives us a really good way to at least determine if there's something generally wrong, and then when we start to dive in deeper, we can, we can then work that out.

227

00:29:15.400 → 00:29:18.620

Ben: If we could just go backwards and then forwards. That'd be great. So we

228

00:29:21.260 → 00:29:41.410

Ben: so look, there's a there's a toolkit. I'm not going to list them all. They'll be on the website. And in the email that you you get just that will can help you to start thinking about how you might tools that you might might use and I've been through all those already, and it's really just a summary of those. So finally, we'll go to the next slide for the 3rd time. If we could.

229

00:29:43.300 → 00:29:44.840

Ben: 1 1 before.

230

00:29:46.280 → 00:30:04.310

Ben: Thank you. Very good. So this is the method that we're working on, and what you see there is a lovely picture of Ludwig Butler, who's going to take over in a second with a poster on this method we presented at Isport in Barcelona, and it's something we're actively pursuing to use on every study. So

231

00:30:04.600 → 00:30:09.759

Ben: we finally got there. And now I'm gonna hand over to Ludwig to to wrap up.

232

00:30:11.250 → 00:30:34.759

Ludwig: Thanks a lot, Ben, and I will start on the next slide just to give you a very kind of quick reminder on, on what it need, what it, what it is in the background that may happen to the studies while we're leaving them out there unprotected. So

233

00:30:35.980 → 00:30:56.880

Ludwig: what do we know about the fraudsters in the panel, we, as Ben laid out, we don't have a lot of visibility. What is happening there. But what we do know is that basically no source or no panel comes without at least some degree of bad actors, and we also found that the fraudsters

234

00:30:56.980 → 00:31:08.760

Ludwig: by nature and by learning and by coached, and finding the leading questions in your screeners, may have a much higher likelihood to begin with, to passing a screener.

235

00:31:08.990 → 00:31:37.019

Ludwig: And what this does is that it gives the fraudsters a headwind to enter any study, and while the legitimate respondents enter with the expected rate, the fraudsters come in at a much higher pace. Now on the left of the graph, if you have a high incidence study, you may see that the level of fraud in this study is still at the expected level of what you would assume from a panel on the outset. But

236

00:31:37.840 → 00:31:45.220

Ludwig: if we are looking even at the best panel with a very low rate of fraudsters in there.

237

00:31:45.260 → 00:32:12.309

Ludwig: in a low incidence setting, like in the most of the healthcare studies that we do. What happens is, as we move along to lower incidences, this increased ir of the fraudsters really is kicking in and bringing even a very good panel to actually the same level of fraud you may have expected from one of the incompetent or accomplice panels that we have there.

238

00:32:12.880 → 00:32:30.520

Ludwig: so that underlines kind of importance of not leaving these studies out there simply by hoping to have, you know, found a good panel that you can work with, but really actively go on and manage the the quality and the prevention of the studies.

239

00:32:30.912 → 00:32:33.540

Ludwig: So can we go to the next slide? Silly, please?

240

00:32:35.400 → 00:32:41.870

Ludwig: And so some of the methods that we've that we've shown in the beginning were

241

00:32:41.990 → 00:32:59.579

Ludwig: honestly also a few of them like lucky finds like this screenshot that Ben presented. This is something that can be great if you have it, but you cannot rely on an ad hoc method to actually be happy that you found that one little bit where the fraudster tripled over and left a trace.

242

00:33:00.218 → 00:33:04.400

Ludwig: So instead. And this is the good news on it.

243

00:33:04.480 → 00:33:34.259

Ludwig: we can actually do better than that. And of course, we also need to. So the solution that we're using internally and continue to develop is based on the methods that Ben just described. The irrelevant segment modeling and combine this with the other tool that we've introduced before. And the key is really to have these implemented at a standardized and then automated way before the respondents actually enter the study.

244

00:33:34.800 → 00:33:49.350

Ludwig: And what we're doing by having the ability to run these checks in real time and against published benchmarks of these models is that we can potentially detect these pockets of bad data and fraud

245

00:33:49.430 → 00:34:13.059

Ludwig: to investigate them before they come online. Our long term goal here is to elevate these methods, so we can bring them onto a citable level that allows the researchers to employ them and use them properly. But this is kind of where we are in our machine room, if you will, and developing these methods forward.

246

00:34:13.130 → 00:34:16.760

Ludwig: So, Sylvia, you can move to the next slide.

247

00:34:18.190 → 00:34:22.210

Ludwig: We've summarized some some key takeaways here.

248

00:34:22.916 → 00:34:27.080

Ludwig: And I'll just briefly go through them. So if you

249

00:34:27.130 → 00:34:37.960

Ludwig: basically see that fraud is often concentrated by source or panel, this is one of the underlying topics of today. So given that working with multiple ones

250

00:34:38.040 → 00:34:45.380

Ludwig: can actually be better than just one. So you can utilize these segment models, and compare actually between sources.

251

00:34:45.870 → 00:35:02.000

Ludwig: And unfortunately, as we've also seen in quite a few studies, even very highly reputable panels are at risk of being infiltrated by fraudsters, and having multiple partners on their end again, may increase the fraud rate as

252

00:35:02.070 → 00:35:13.030

Ludwig: at scale. This is highly profitable for the fraudsters to enter in a way that they can find. So these chains of partners in the background are an obvious invitation for them.

253

00:35:13.620 → 00:35:29.920

Ludwig: What we've also seen is that simple things like the incidence rates should not increase. So as researchers, you need to absolutely be sure and informed about the mechanisms of screening, and how people actually enter your studies in order to assess for fraud or quality.

254

00:35:30.050 → 00:35:47.809

Ludwig: and if you do that ideally, you have some way of doing that in real time, and look at the quality reports, because finding broad post data collection, and in the analysis is basically too late, and nothing can be done anymore other than maybe the contractual measures and policing.

255

00:35:48.010 → 00:35:48.990

Ludwig: So

256

00:35:49.110 → 00:36:12.799

Ludwig: if you rely simply on the stated vague quality claims that a lot of vendors put out there, you might be better informed to do some independent verification of the quality of these sources, and only then, if anything else fails, you may want to rely on your friends in the legal department and have them build contracts that include the risks

257

00:36:13.232 → 00:36:31.829

Ludwig: in in them. So you can protect yourself, and on the other side, use that knowledge also in your vendor, search and building the agreements upfront to have the most what you've possible at protection level in both prevention and detection of fraud.

258

00:36:32.230 → 00:36:44.729

Ludwig: and with this, and neatly ahead of time, I will hand over back to Anton to maybe open the floor for some questions.

259

00:36:45.370 → 00:36:55.119

Anton: Super. Thank you. Ben and Ludwig. So over to the Q. And A, we have a actually a number of questions that that came up here during the presentation.

260

00:36:56.700 → 00:37:00.689

Anton: Start with this one question for for Natalia.

261

00:37:01.210 → 00:37:06.669

Anton: What is one of the 1st things that I will spot if my data is fraudulent.

262

00:37:07.900 → 00:37:29.199

Natalia: Well, this is a great question. Well, 1st of all, thank you to the presenters. That was very informative to see it all laid out like that. That was actually really great. I hope the audience as well enjoy this. Now, to answer that question. There are common technical markers

263

00:37:29.200 → 00:37:42.150

Natalia: like duplicate Ips and Odd Start times, which is all things that were mentioned during the presentation. High. VPN. Use, or, I'm thinking, repetitive answers.

264

00:37:42.150 → 00:37:59.340

Natalia: illogical responses, and mainly one of the most salient things is preferences that don't match your target group. So this is a little bit what was being said before. If you have a patient group. And it's giving you answers that don't like. They prefer

265

00:37:59.340 → 00:38:15.489

Natalia: cancer, or they prefer pain rather than preferring feeling. Well, this is already a flag. So this would be the 1st things to pop in a data set that has a high fraud rate. For sure.

266

00:38:16.290 → 00:38:18.100

Anton: That's great. Thank you, Natalia.

267

00:38:18.700 → 00:38:27.550

Anton: and question for for living. Then how do you make sure that you don't mix good answers with fraudulent answers?

268

00:38:28.180 → 00:38:29.050

Ludwig: You don't.

269

00:38:29.870 → 00:38:49.340

Ludwig: So I mean, that's part of, I think, what we mentioned in the survey there. So the key is to collect background variables and metadata. So you can actually see where these respondents come from.

270

00:38:49.340 → 00:39:18.569

Ludwig: So if you have them all based on just one source, and you have no idea on where they came from, and how you can somehow distinguish that. Then you also have no way to identify the sources. But if you do, and you collect that metadata with the survey upfront. You then have the ability to do those split models and other ways to trace back where they came from, and then exactly say, wait a second. There might be a bad source mixed in.

271

00:39:18.590 → 00:39:29.430

Ludwig: Of course there might be some respondents that look legitimate, but if you can identify the whole source is wrong, then that is the way to the way to go in these kind of mixed data sets.

272

00:39:30.740 → 00:39:31.600

Anton: Okay?

273

00:39:32.160 → 00:39:39.320

Anton: And yeah, here, here's a good one for Natalia. Are there any safe recruitment sources.

274

00:39:41.600 → 00:40:03.020

Natalia: Well, unfortunately, no. Unfortunately, fraud exists, and we've observed in our experience at survey engine in all the routes that we've used, paid panels, vendors, even Phds, quite shockingly. So the bottom line is like, if there's an incentive, there is potential for fraud.

275

00:40:04.000 → 00:40:06.270

Anton: Right and.

276

00:40:06.910 → 00:40:31.490

Ludwig: Yeah, sorry. I might add that there is no, although you could then say, well, then, we want to do all the research without incentives, but that's also not possible. Sometimes even the the Irbs or the fair market value would dictate that you have to pay somebody. So there is always that element that we cannot simply go for one silver bullet.

277

00:40:33.760 → 00:41:02.819

Natalia: Yeah, but other ideas. And I'm putting this out there as ways to mitigate, for instance, in patient recruitment, offering non-monetary incentives, like donations to charitable organizations or to Phgs can be a good way, because that holds no incentive for a fraudster, but it does hold

some attraction and some motivation for an actual legitimate patient. So those are avenues that can also be explored.

278

00:41:03.440 → 00:41:04.230

Anton: Right?

279

00:41:04.860 → 00:41:18.029

Anton: And yeah, we're getting some more questions here. We'll start with this one. So, Ludwig, are you able to tell us a bit more about the segmentation of irrelevant respondents. How is it done in principle.

280

00:41:18.710 → 00:41:34.019

Ludwig: So maybe I can. I can say I can do a few words about the in principle, but then I might hand over to Ben for the actual. So in principle, we're, we're looking at the segmentation of

281

00:41:34.020 → 00:41:58.640

Ludwig: irrelevant variables that should not matter in the modeling. So it's not about irrelevant respondents per se. It's about a variable that should not have anything to do or have any influence in a group behavior in the modeling. So when we have those variables, and I just give you an example. It could be globally the time of day that 2 sources at the same time send a sample in.

282

00:41:58.730 → 00:42:13.760

Ludwig: or it could be a browser setting that we have like a device type or something like that, that should not matter to the actual health behavior in the background. But, Ben, maybe you want to share some more technical explanation of that.

283

00:42:13.760 → 00:42:21.960

Ben: Yeah, look, it's it's the sort of thing I can imagine a lot of objections to what what this approach does is. It changes the question to

284

00:42:22.030 → 00:42:47.539

Ben: easy question to answer, which is is that variable, really irrelevant? And a colleague recently said, Oh, my my 80 year old mother uses windows nt, and uses this particular browser, I mean, that's fair enough, but what what we can do is at least find there's some anomalous behavior, and then we can answer that question. Is that really a reasonable thing to say? To have someone have wildly different preferences because they happen to use Internet explorer.

285

00:42:47.610 → 00:43:16.509

Ben: That's the 1st thing. So it's going to be your one's choice of how reasonable those irrelevant variables are. But typically, there's a set of easily accessible ones. So time is a really good one. But time period, a very simple thing that we've applied even just last week, was just simply split the sample into first, st half second half around the median, and we get a weird model at the tail end, which matches what we spoke about earlier. That that's

286

00:43:16.610 → 00:43:19.600

Ben: it's loaded with more fraudsters.

287

00:43:20.150 → 00:43:32.920

Ben: The idea is that you choose them in advance, and then you can literally cycle through if you've got the right tools of just systematically building models, breaking them down by these different types until you find a match that something is not right.

288

00:43:33.060 → 00:43:49.699

Ben: And the nice thing about this, compared to, let's say latent class methods is, you get an actionable, variable, like panel X browser mode, y, you know, wave a and the more metadata the more opportunity you have to do those splits.

289

00:43:50.990 → 00:43:51.870

Anton: That's great.

290

00:43:52.230 → 00:44:11.480

Anton: and we have another good question here. Perhaps this one is is for for you as well, Ben, so you use Dce results to estimate potential fraudsters. Do you use the client's results, or do you compute a utility? Model yourselves? In the latter case? Do you compare your model.

291

00:44:11.480 → 00:44:40.539

Ben: Okay? Well, for within the scope of a contractual contract and a partnership with our clients that's often requested of us, they say, look, you guys have got the right tools to build these immediately. We don't. We are. Our analysts are on holidays. We have the right tools to do this, we can do a a fairly good m and L model in real time. So we would do that with their consent within knowledge. Obviously, we can't present any of that information.

292

00:44:41.428 → 00:44:53.150

Ben: Upfront. But the idea is to bring a standard Dce, which are fantastic behavioral modeling instruments earlier on, and use a standardized

293

00:44:54.420 → 00:45:22.129

Ben: measure like the fact. 8 d. Or the Eq. 5 d. 5 l. Or something like that which has got a wealth of data on to then have at least the basis for comparing health preference, and that could be done outside of the project, the client, and because it's before the study happens. So we have a lot more freedom to do that. But that's the idea. But all every one of these would be done within, with the the knowledge and the consent and the cooperation of the of the client.

294

00:45:23.270 → 00:45:24.200

Anton: Excellent.

295

00:45:24.560 → 00:45:28.720

Anton: And yeah, another question here, Ben, I think this one goes to you as well.

296

00:45:28.810 → 00:45:33.639

Anton: Can I use AI or machine learning to detect fraudulent data.

297

00:45:34.130 → 00:46:03.039

Ben: Yes. But like all AI, it's got AI likes a lot of data that's marked, that's trained. And then AI is basically a summarization tool that can do that. So if you do have to be lucky enough to have data that you know categorically, these are fraudsters, and these aren't. You could put them

into these into a machine learning model and get something meaningful. And this is something we are actually currently exploring. The problems, of course, are

298

00:46:03.330 → 00:46:15.650

Ben: 3 problems. There isn't a vast amount of data to start off with. It's very sparse in that. One study may ask for gender, another one may not, and so on. So you don't have the same.

299

00:46:15.790 → 00:46:36.030

Ben: The same data sets, and the 3rd is the hardest one is. You just don't know whether they're they're fraudsters. But look, here's another little kind of insight. Is that machine learning at the limit is the same as a dce. They're the same models. The difference is Dce are directed, whereas the machine learning isn't so much directed looking for patterns.

300

00:46:36.030 → 00:46:47.780

Ben: In fact, Dce, are from a modeling perspective superior when you know what you're looking for. But I think it's a fantastic idea. The trick is to get a large data set that's categorized as a training set

301

00:46:47.800 → 00:46:53.389

Ben: which we we have partially. We do have some sets. But you need a lot of data to do that. Yeah.

302

00:46:54.700 → 00:46:55.720

Anton: Understood.

303

00:46:56.559 → 00:47:06.860

Anton: And here's a question. I believe this one goes to Ludwig. We are often tied to regulatory directives. How does survey engine ensure compliance.

304

00:47:08.670 → 00:47:09.470

Ludwig: Okay.

305

00:47:10.014 → 00:47:21.989

Ludwig: I think that that's a little a little broader and and more looking from the outset rather than you know. At at one of these very specific tools that we presented today.

306

00:47:22.700 → 00:47:26.830

Ludwig: The idea behind all of these

307

00:47:28.350 → 00:47:50.889

Ludwig: fraud prevention techniques is that you don't want to be accused of cherry picking and just cleaning a data set behind closed doors to then ship something that just looks too perfect, and everybody knows that there is always some noise, something going on in the data sets. And it's really about the transparency from our side

308

00:47:50.890 → 00:48:02.000

Ludwig: to say, exactly, okay, this is happening in your study. This is what we're going to do, and if we clarify that upfront, so, for example, bring that into a protocol level.

309

00:48:02.160 → 00:48:11.459

Ludwig: then we have the a clear pathway of a predefined answer to these effects that you can see in the data set.

310

00:48:11.740 → 00:48:19.100

Ludwig: So then we stay on the side of not doing it behind the you know the black box, but have it transparently out there.

311

00:48:19.501 → 00:48:27.939

Ludwig: Yeah, that would be my my response on how to yeah. Work with that on the regulatory level data or on the compliance side.

312

00:48:28.500 → 00:48:30.049

Anton: That's great. Thank you.

313

00:48:30.936 → 00:48:39.430

Anton: So we still have time for for any any final questions. Do we have any any final questions from from the audience?

314

00:48:43.330 → 00:48:44.790

Ludwig: If not, I would have one.

315

00:48:45.480 → 00:48:46.190

Anton: Please go ahead.

316

00:48:46.783 → 00:49:09.319

Ludwig: This might be more for Natalia. You you mentioned before that also. Basically, other sources are deceptible. Of fraud. I mean, we talked a lot about panels here today and that. But maybe you can give us a a brief insight into what what that means from your experience.

317

00:49:11.769 → 00:49:20.559

Natalia: So I imagine what you're asking is, where? What are the sources? Where are we seeing fraud happening? Is this your question?

318

00:49:20.560 → 00:49:21.250

Ludwig: Yeah, yeah.

319

00:49:21.250 → 00:49:50.410

Natalia: Yeah, so well as it's been described and and discussed today, what's the situation with panels and panels that are using partners? What was surprising to me when I started working. Here was the situation with Phs, because I was hoping that this was going to be a a hundred percent safe route. So it was. It was shocking to see how they are also being infiltrated by this criminal, and and what what Ben was describing as the criminal actors.

320

00:49:51.180 → 00:49:55.770

Natalia: And so it seems like the only safe routes

321

00:49:55.880 → 00:50:10.630

Natalia: that are left are real world and basically going straight to the source, right and finding, especially when it comes to patient recruitment, the only is just going to the source and finding the patient where the patients are, which

322

00:50:10.650 → 00:50:23.719

Natalia: could well be phs, but also clinical centers sites. So basically trying other ways of recruitment that put you in more direct contact with the patients.

323

00:50:23.890 → 00:50:36.360

Natalia: But that's that's a that's a situation that we're seeing at the moment where there's clearly no, you cannot take any route for for granted. That's

324

00:50:36.460 → 00:50:39.619

Natalia: that's the situation. I don't know if this is where your question was.

325

00:50:39.780 → 00:51:04.819

Ludwig: Yeah, I think so. And it and it probably underlines that some of the studies will need still to depend on, you know, a way to recruit larger amounts of respondents. If you're if you're not only looking for a handful, and that's where these methods, and having something in place is a super important topic. Yeah.

326

00:51:05.010 → 00:51:05.630

Natalia: Yes.

327

00:51:06.590 → 00:51:09.830

Anton: Yeah. Any other questions here?

328

00:51:09.830 → 00:51:10.400

Natalia: Yeah.

329

00:51:10.410 → 00:51:13.429

Anton: Before we start to to wrap up today's session.

330

00:51:15.080 → 00:51:15.690

Anton: No.

331

00:51:16.520 → 00:51:27.694

Anton: in that case a big thank you. To Ben Ludwig and Natalia, and everyone who who took took some time to join us today?

332

00:51:28.730 → 00:51:36.809

Anton: basically, we. We hope that you guys found value in this session. You'll be receiving a toolkit shortly in a post Webinar email.

333

00:51:37.000 → 00:51:40.209

Anton: Please follow our Linkedin page for more updates on

334

00:51:40.230 → 00:51:44.159

Anton: coming webinars and other events that might be be relevant to you.

335

00:51:44.630 → 00:51:47.189

Anton: And that's it. Thank you for joining.

336

00:51:48.620 → 00:51:53.979

Ludwig: And thanks a lot, Anton, for guiding us through the early afternoon in Germany.

337

00:51:55.420 → 00:51:56.470

Anton: Thank you. Everybody.

338

00:51:56.470 → 00:51:58.040

Natalia: Everybody. Thank you.

339

00:51:58.040 → 00:51:59.040

Natalia: Thanks a lot.

340

00:51:59.680 → 00:52:00.490

Ludwig: Bye.