WEBVTT

1 00:00:02.040 --> 00:00:03.870 <v Maria>My name is Maria Ciarleglio</v>

2 00:00:03.870 --> 00:00:05.460 and I'm a faculty member

3 00:00:05.460 --> 00:00:07.770 in the Department of Biostatistics

4 00:00:07.770 --> 00:00:10.053 at the Yale School of Public Health.

5 00:00:10.920 --> 00:00:12.270 In this video series,

6 00:00:12.270 --> 00:00:14.910 I will introduce the clinical research process

7 00:00:14.910 --> 00:00:18.903 to prepare you to collaborate with a statistician.

8 00:00:20.370 --> 00:00:21.531 In this third video,

9 00:00:21.531 --> 00:00:24.060 we'll discuss data collection

10 00:00:24.060 --> 00:00:27.330 or how to structure your data in a spreadsheet

11 00:00:27.330 --> 00:00:29.340 so that you can share the data files

12 00:00:29.340 --> 00:00:32.400 with a statistician for analysis.

13 00:00:32.400 --> 00:00:36.721 We'll also discuss variable types and special considerations

14 00:00:36.721 --> 00:00:41.673 for quantitative, categorical, and time-to-event data.

15 00:00:43.140 --> 00:00:45.360 Collecting good data is important

16 00:00:45.360 --> 00:00:46.860 because the data enable you

17 00:00:46.860 --> 00:00:48.870 to answer your research question.

18 00:00:48.870 --> 00:00:50.970 Generally, we're interested in the effect

19 00:00:50.970 --> 00:00:52.697 of one or more exposure

20 00:00:52.697 --> 00:00:55.950 on one or more outcome of interest.

21 00:00:55.950 --> 00:01:00.480 Exposures and outcomes are key variables to collect.

22 00:01:00.480 --> 00:01:02.756 There may be several other descriptive characteristics

23 00:01:02.756 --> 00:01:04.707 that we would like to either report

24 00:01:04.707 --> 00:01:06.900 when we're describing the sample

25 00:01:06.900 --> 00:01:10.620 or control or adjust for in the analysis.

26 00:01:10.620 --> 00:01:12.690 We can also explore these variables

27 00:01:12.690 --> 00:01:14.492 as possible effect modifiers

28 00:01:14.492 --> 00:01:18.780 or as characteristics that define subgroups of interest.

29 00:01:18.780 --> 00:01:20.994 These additional variables are also recorded

30 00:01:20.994 --> 00:01:23.460 during data collection.

31 00:01:23.460 --> 00:01:25.740 What is the structure of this data table?

32 00:01:25.740 --> 00:01:28.500 Specifically, the rows contain information

33 00:01:28.500 --> 00:01:29.880 from different subjects

34 00:01:29.880 --> 00:01:32.760 and the columns contain the different characteristics

35 00:01:32.760 --> 00:01:36.300 or variables collected on the subjects.

36 00:01:36.300 --> 00:01:39.480 This means that the number of rows in your data table

37 00:01:39.480 --> 00:01:41.561 corresponds to the number of participants,

38 00:01:41.561 --> 00:01:43.692 not including the header row

39 00:01:43.692 --> 00:01:47.190 or the row that contains the variable names.

40 00:01:47.190 --> 00:01:49.410 The number of columns in your data table

41 00:01:49.410 --> 00:01:53.550 corresponds to the number of variables in your data table.

42 00:01:53.550 --> 00:01:55.740 There are several features of this table

43 00:01:55.740 --> 00:01:57.570 that I want to discuss further,

44 00:01:57.570 --> 00:02:01.860 but first, let's review a few options available to you

45 00:02:01.860 --> 00:02:04.083 for collecting your study data.

46 00:02:05.280 --> 00:02:07.020 About 90% of the time,

47 00:02:07.020 --> 00:02:10.680 we work with data collected in Excel spreadsheets.

48 00:02:10.680 --> 00:02:13.622 Although Excel was not designed as a data capture tool

49 00:02:13.622 --> 00:02:15.660 for clinical research,

50 00:02:15.660 --> 00:02:19.050 it does provide an easy way to collect,

51 00:02:19.050 --> 00:02:21.183 store and share your data.

52 00:02:22.112 --> 00:02:25.505 A nice feature of Excel is the ability to quickly filter

53 00:02:25.505 --> 00:02:28.110 on different variables.

54 00:02:28.110 --> 00:02:30.780 To enable filtering, go to the Data tab.

55 00:02:30.780 --> 00:02:33.570 And then, select the Filter button.

56 00:02:33.570 --> 00:02:35.610 After we turn filtering on,

57 00:02:35.610 --> 00:02:38.010 notice how the cells in the header row

58 00:02:38.010 --> 00:02:39.960 that contain the variable names

59 00:02:39.960 --> 00:02:43.860 now have a dropdown button on the right side of the cell.

60 00:02:43.860 --> 00:02:47.460 If you click this button and select or filter

61 00:02:47.460 --> 00:02:49.560 on the values that you want to view,

62 00:02:49.560 --> 00:02:54.560 for example, suppose sex equals 1 corresponds to males.

63 00:02:55.170 --> 00:02:58.410 If I want to look at the data in our male patients,

64 00:02:58.410 --> 00:03:00.030 filter on the sex variable

65 00:03:00.030 --> 00:03:05.030 and select the value 1 and deselect the value 2.

66 00:03:05.220 --> 00:03:07.220 After filtering, we only see the rows

67 00:03:07.220 --> 00:03:11.190 from the patients with sex equals 1.

68 00:03:11.190 --> 00:03:13.740 Select the Clear filter button

69 00:03:13.740 --> 00:03:16.953 to remove any filters applied in the worksheet.

70 00:03:18.090 --> 00:03:19.920 Another option available to you

71 00:03:19.920 --> 00:03:22.890 for data collection is REDCap.

72 00:03:22.890 --> 00:03:26.272 REDCap is a web-based data collection tool

73 00:03:26.272 --> 00:03:28.800 designed for research data.

74 00:03:28.800 --> 00:03:32.010 The data are securely stored on the cloud.

75 00:03:32.010 --> 00:03:34.320 You will need to request a REDCap account

76 00:03:34.320 --> 00:03:36.630 from the REDCap team at Yale

77 00:03:36.630 --> 00:03:38.220 to create a project.

78 00:03:38.220 --> 00:03:42.120 The REDCap team also provides training and support

79 00:03:42.120 --> 00:03:44.700 and you can find their contact information

80 00:03:44.700 --> 00:03:46.350 on the REDCap website,

81 00:03:46.350 --> 00:03:50.520 portal.redcap.yale.edu.

82 00:03:50.520 --> 00:03:54.960 Let's briefly talk about collecting or recording your data.

83 00:03:54.960 --> 00:03:56.204 In general, variables

84 00:03:56.204 --> 00:03:58.996 are either quantitative or categorical.

85 00:03:58.996 --> 00:04:03.210 We also sometimes collect text fields or notes.

86 00:04:03.210 --> 00:04:06.150 Quantitative variables are numeric variables

87 00:04:06.150 --> 00:04:08.910 such as height, weight, age,

88 00:04:08.910 --> 00:04:11.820 bilirubin, a calculated MELD score.

89 00:04:11.820 --> 00:04:14.640 Technically, dates are also numeric

90 00:04:14.640 --> 00:04:17.490 in the way they're stored in Excel

91 00:04:17.490 --> 00:04:21.300 and most other statistical data analysis software.

92 00:04:21.300 --> 00:04:23.520 In our sample data table,

93 00:04:23.520 --> 00:04:28.083 total bilirubin and INR are quantitative variables.

94 00:04:29.310 --> 00:04:31.170 We also have several dates

95 00:04:31.170 --> 00:04:35.070 including date of birth and date of diagnosis.

96 00:04:35.070 --> 00:04:36.510 In the full data sheet,

97 00:04:36.510 --> 00:04:40.230 we also have intake date into the study.

98 00:04:40.230 --> 00:04:42.510 We recommend that you collect dates

99 00:04:42.510 --> 00:04:44.441 and allow us, the statisticians,

100 00:04:44.441 --> 00:04:47.447 to calculate durations or lengths of time

101 00:04:47.447 --> 00:04:51.410 such as age at diagnosis or age at study baseline

102 00:04:51.410 --> 00:04:54.030 using statistical software.

103 00:04:54.030 --> 00:04:59.030 We often use SAS or R to perform our data analysis.

104 00:04:59.190 --> 00:05:01.954 For example, we can calculate age at diagnosis

105 00:05:01.954 --> 00:05:06.360 using date of birth and date of diagnosis.

106 00:05:06.360 --> 00:05:09.431 The SAS code here adds a new column to the data table

107 00:05:09.431 --> 00:05:12.900 containing age at diagnosis in years.

108 00:05:12.900 --> 00:05:14.910 Same usually goes for variables

109 00:05:14.910 --> 00:05:17.940 such as FIB-4 and MELD score.

110 00:05:17.940 --> 00:05:19.680 We can program the calculation

111 00:05:19.680 --> 00:05:21.435 of these variables in our code

112 00:05:21.435 --> 00:05:24.300 rather than have you perform the calculation

113 00:05:24.300 --> 00:05:26.043 in your Excel spreadsheet.

114 00:05:27.210 --> 00:05:29.370 Categorical variables are ones

115 00:05:29.370 --> 00:05:31.440 where the values the variable can take

116 00:05:31.440 --> 00:05:33.570 are essentially categories.

117 00:05:33.570 --> 00:05:35.664 An example of a categorical variable

118 00:05:35.664 --> 00:05:39.540 is race or sex or gender.

119 00:05:39.540 --> 00:05:42.840 Categorical variables that can take only two levels

120 00:05:42.840 --> 00:05:46.110 are called dichotomous or binary variables.

121 00:05:46.110 --> 00:05:47.760 In our sample data table,

122 00:05:47.760 --> 00:05:51.510 response status, treatment group, race and sex

123 00:05:51.510 --> 00:05:53.280 are categorical variables

124 00:05:53.280 --> 00:05:55.552 but notice that these variables are collected

125 00:05:55.552 --> 00:05:58.020 and coded numerically.

126 00:05:58.020 --> 00:06:01.089 This is a common method for recording categorical data

127 00:06:01.089 --> 00:06:05.910 where each category is given a numerical label.

128 00:06:05.910 --> 00:06:09.570 For example, sex is coded as 1 for males

129 00:06:09.570 --> 00:06:11.520 and 2 for females.

130 00:06:11.520 --> 00:06:15.270 We discourage the use of character variables or text

131 00:06:15.270 --> 00:06:17.490 when collecting categorical data

132 00:06:17.490 --> 00:06:21.300 because our statistical software is case sensitive

133 00:06:21.300 --> 00:06:23.640 when reading character data.

134 00:06:23.640 --> 00:06:26.460 It's important to maintain a data key

135 00:06:26.460 --> 00:06:28.740 that defines the numerical coding.

136 00:06:28.740 --> 00:06:30.540 This is especially important

137 00:06:30.540 --> 00:06:32.700 when sharing the data with others.

138 00:06:32.700 --> 00:06:35.520 Here, our data key would include the definition

139 00:06:35.520 --> 00:06:39.060 of sex equals 1 as corresponding to males

140 00:06:39.060 --> 00:06:43.230 and sex equals 2 as corresponding to females.

141 00:06:43.230 --> 00:06:46.876 We recommend creating a separate tab in the Excel file

142 00:06:46.876 --> 00:06:49.470 that defines the numerical coding

143 00:06:49.470 --> 00:06:52.173 of the categorical variables in the data.

144 00:06:53.070 --> 00:06:57.360 A few notes on naming the variables in your spreadsheet.

145 00:06:57.360 --> 00:06:59.700 Variable names should be descriptive,

146 00:06:59.700 --> 00:07:02.760 making it clear what the variable represents.

147 00:07:02.760 --> 00:07:07.080 SAS variable names may be up to 32 characters in length.

148 00:07:07.080 --> 00:07:09.270 The first character of the variable name

149 00:07:09.270 --> 00:07:11.910 must begin with an alphabetic character

150 00:07:11.910 --> 00:07:13.380 or an underscore.

151 00:07:13.380 --> 00:07:17.430 Variable names should not begin with a number or symbol.

152 00:07:17.430 --> 00:07:20.220 And finally, the variable name should not contain

153 00:07:20.220 --> 00:07:23.793 any special characters other than the underscore.

154 00:07:24.660 --> 00:07:27.690 Sometimes certain variables are not available

155 00:07:27.690 --> 00:07:29.370 or not collected.

156 00:07:29.370 --> 00:07:31.230 We see in our sample data table

157 00:07:31.230 --> 00:07:33.630 that we have a few missing values.

158 00:07:33.630 --> 00:07:37.080 Missing values occur when we don't have the data available

159 00:07:37.080 --> 00:07:38.520 for that individual.

160 00:07:38.520 --> 00:07:42.630 One or more variables can be missing for an individual.

161 00:07:42.630 --> 00:07:46.483 Here, race is missing for patients 106 and 117.

162 00:07:48.450 --> 00:07:52.530 Date of diagnosis is missing for patient 106.

163 00:07:52.530 --> 00:07:54.900 There are also no ultrasound findings

164 00:07:54.900 --> 00:07:57.690 for the majority of patients.

165 00:07:57.690 --> 00:08:00.000 Notice how missing values are represented

166 00:08:00.000 --> 00:08:02.400 in the data as empty cells.

167 00:08:02.400 --> 00:08:06.180 Do not use an N/A or a period or a dash

168 00:08:06.180 --> 00:08:08.040 to indicate missing values.

169 00:08:08.040 --> 00:08:11.040 Leave the cell blank when the value is missing.

170 00:08:11.040 --> 00:08:13.350 Finally, if any calculated variables

171 00:08:13.350 --> 00:08:15.780 involve variables with a missing value,

172 00:08:15.780 --> 00:08:19.560 then that calculated variable will also be missing.

173 00:08:19.560 --> 00:08:23.040 For example, if total bilirubin is missing for a patient,

174 00:08:23.040 --> 00:08:26.973 then their calculated MELD score will also be missing.

175 00:08:28.140 --> 00:08:31.290 Finally, I want to discuss some special considerations

176 00:08:31.290 --> 00:08:35.940 for collecting endpoint or response data in your study.

177 00:08:35.940 --> 00:08:38.010 Remember that your primary endpoint

178 00:08:38.010 --> 00:08:40.950 answers your primary research question.

179 00:08:40.950 --> 00:08:43.890 You may also be collecting data on secondary,

180 00:08:43.890 --> 00:08:47.370 tertiary, or other exploratory endpoints.

181 00:08:47.370 --> 00:08:50.460 Endpoints are either continuous or quantitative,

182 00:08:50.460 --> 00:08:53.880 categorical or most often dichotomous,

183 00:08:53.880 --> 00:08:57.933 or there's some measure of time to an event of interest.

184 00:08:58.880 --> 00:09:01.320 Looking at quantitative variables,

185 00:09:01.320 --> 00:09:04.515 we can begin by summarizing a quantitative variable

186 00:09:04.515 --> 00:09:06.990 in the full sample.

187 00:09:06.990 --> 00:09:10.115 We can summarize and compare that quantitative variable

188 00:09:10.115 --> 00:09:12.180 in two or more groups

189 00:09:12.180 --> 00:09:15.900 such as the group exposed to a particular intervention

190 00:09:15.900 --> 00:09:18.720 and the group unexposed to that intervention.

191 00:09:18.720 --> 00:09:22.050 And we can also analyze that quantitative variable

192 00:09:22.050 --> 00:09:23.640 in a regression model,

193 00:09:23.640 --> 00:09:27.570 allowing us to control for certain confounders.

194 00:09:27.570 --> 00:09:29.615 In this 2022 hepatology paper,

195 00:09:29.615 --> 00:09:33.330 they explore how TIPS affects PPG

196 00:09:33.330 --> 00:09:35.760 in patients with ascites.

197 00:09:35.760 --> 00:09:40.560 They found that mean PPG, portal pressure and IVC pressure

198 00:09:40.560 --> 00:09:43.560 decreased significantly after TIPS.

199 00:09:43.560 --> 00:09:45.720 These are quantitative endpoints

200 00:09:45.720 --> 00:09:49.380 and these measurements are taken at two points in time,

201 00:09:49.380 --> 00:09:51.633 before TIPS and after TIPS.

202 00:09:53.010 --> 00:09:56.190 To analyze the change in these quantitative measures,

203 00:09:56.190 --> 00:09:58.830 we need the measurements recorded in our data

204 00:09:58.830 --> 00:10:01.140 before TIPS and after TIPS.

205 00:10:01.140 --> 00:10:02.940 We can use statistical software

206 00:10:02.940 --> 00:10:06.090 to compute and analyze the responsive interest

207 00:10:06.090 --> 00:10:09.003 which is the change in these measures.

208 00:10:10.260 --> 00:10:12.780 The next type of endpoint we often work with

209 00:10:12.780 --> 00:10:15.960 is a dichotomous or binary endpoint.

210 00:10:15.960 --> 00:10:18.818 For example, the goal could be to determine

211 00:10:18.818 --> 00:10:21.690 if the patient's disease improves

212 00:10:21.690 --> 00:10:23.190 over the course of the study.

213 00:10:23.190 --> 00:10:25.596 This is recorded as a binary variable,

214 00:10:25.596 --> 00:10:28.260 improvement, yes or no.

215 00:10:28.260 --> 00:10:30.210 Similarly, we can look at occurrence

216 00:10:30.210 --> 00:10:34.110 of surgical site infection following liver transplant.

217 00:10:34.110 --> 00:10:35.520 Taking it a step further,

218 00:10:35.520 --> 00:10:38.400 we look for an association between exposure

219 00:10:38.400 --> 00:10:41.880 such as exposure to perioperative antibiotic

220 00:10:41.880 --> 00:10:44.490 compared to intraoperative antibiotic

221 00:10:44.490 --> 00:10:47.220 and development of surgical site infection.

222 00:10:47.220 --> 00:10:50.280 We can also model development of surgical site infection

223 00:10:50.280 --> 00:10:51.990 using predictor variables

224 00:10:51.990 --> 00:10:54.930 such as exposure to a specific treatment

225 00:10:54.930 --> 00:10:58.620 while controlling for potential confounders.

226 00:10:58.620 --> 00:11:02.419 We can use similar methods to analyze categorical responses

227 00:11:02.419 --> 00:11:05.754 with more than two levels and ordinal responses

228 00:11:05.754 --> 00:11:09.333 in which the endpoint categories are ordered.

229 00:11:10.260 --> 00:11:12.990 In this 2019 hepatology paper,

230 00:11:12.990 --> 00:11:15.240 they compare the proportion of patients

231 00:11:15.240 --> 00:11:17.010 with surgical site infection

232 00:11:17.010 --> 00:11:19.362 in those patients receiving 72 hours

233 00:11:19.362 --> 00:11:24.362 of perioperative antibiotics or extended antibiotics,

234 00:11:24.420 --> 00:11:28.110 and those receiving intraoperative antibiotics only

235 00:11:28.110 --> 00:11:30.660 or short antibiotics.

236 00:11:30.660 --> 00:11:32.190 The primary endpoint here

237 00:11:32.190 --> 00:11:35.940 is development of surgical site infection.

238 00:11:35.940 --> 00:11:38.760 They also look at 30-day hospital readmission

239 00:11:38.760 --> 00:11:40.620 and 30-day mortality,

240 00:11:40.620 --> 00:11:43.113 which are also dichotomous responses.

241 00:11:44.400 --> 00:11:46.680 Sometimes we use quantitative data

242 00:11:46.680 --> 00:11:48.984 to define a categorical variable

243 00:11:48.984 --> 00:11:51.510 such as a dichotomous endpoint.

244 00:11:51.510 --> 00:11:54.900 For example, clinically significant portal hypertension

245 00:11:54.900 --> 00:11:58.830 is defined as an HVPG greater than or equal

246 00:11:58.830 --> 00:12:01.710 to 10 millimeters of mercury.

247 00:12:01.710 --> 00:12:04.020 You would collect the quantitative data.

248 00:12:04.020 --> 00:12:06.420 And then, we would create the categorical

249 00:12:06.420 --> 00:12:09.993 or dichotomous variable to use in the analysis.

250 00:12:10.920 --> 00:12:13.890 The final endpoint that we most often see

251 00:12:13.890 --> 00:12:17.730 is a time-to-event or survival endpoint

252 00:12:17.730 --> 00:12:21.024 such as time to death, time to decompensation,

253 00:12:21.024 --> 00:12:24.120 time to recovery or response.

254 00:12:24.120 --> 00:12:26.970 For example, our goal could be to determine

255 00:12:26.970 --> 00:12:29.340 if patients exposed to a new treatment

256 00:12:29.340 --> 00:12:31.290 have longer survival times

257 00:12:31.290 --> 00:12:33.124 or greater likelihood of survival

258 00:12:33.124 --> 00:12:35.220 to a certain time point.

259 00:12:35.220 --> 00:12:37.936 When looking at one group such as the overall sample,

260 00:12:37.936 --> 00:12:40.441 we could report survival probabilities

261 00:12:40.441 --> 00:12:42.582 from a Kaplan-Meier survival curve

262 00:12:42.582 --> 00:12:45.180 or median survival time.

263 00:12:45.180 --> 00:12:46.620 Median survival time

264 00:12:46.620 --> 00:12:49.770 is the time beyond which 50% of the individuals

265 00:12:49.770 --> 00:12:51.870 are expected to survive.

266 00:12:51.870 --> 00:12:55.350 We can naturally extend this to two or more groups

267 00:12:55.350 --> 00:12:57.720 and formally compare the groups,

268 00:12:57.720 --> 00:13:00.180 and we can also build regression models

269 00:13:00.180 --> 00:13:01.830 that estimate the relationship

270 00:13:01.830 --> 00:13:05.490 between rate of the event and exposure variables

271 00:13:05.490 --> 00:13:07.503 such as treatment status.

272 00:13:08.730 --> 00:13:10.669 So far with the other endpoints,

273 00:13:10.669 --> 00:13:13.500 data collection has been pretty intuitive.

274 00:13:13.500 --> 00:13:15.900 However, survival data requires

275 00:13:15.900 --> 00:13:17.550 certain pieces of information

276 00:13:17.550 --> 00:13:20.340 to properly complete the analysis.

277 00:13:20.340 --> 00:13:21.660 In survival data,

278 00:13:21.660 --> 00:13:26.010 the outcome is time to a target event occurring.

279 00:13:26.010 --> 00:13:29.520 Dates are important because we need to compute time

280 00:13:29.520 --> 00:13:32.820 from a specific start point to an event.

281 00:13:32.820 --> 00:13:35.370 To calculate the survival time in your study,

282 00:13:35.370 --> 00:13:37.500 you must precisely define:

283 00:13:37.500 --> 00:13:39.856 the time origin or the starting point,

284 00:13:39.856 --> 00:13:42.300 that is when follow up begins;

285 00:13:42.300 --> 00:13:45.090 the ending event of interest is at death,

286 00:13:45.090 --> 00:13:48.687 decompensation, remission, relapse;

287 00:13:48.687 --> 00:13:51.480 and the measurement scale for the passage of time,

288 00:13:51.480 --> 00:13:54.753 for example, days, months, years.

289 00:13:55.920 --> 00:13:58.163 An important feature of survival data

290 00:13:58.163 --> 00:14:02.430 is that we may have patients who do not experience the event

291 00:14:02.430 --> 00:14:04.230 during the study period.

292 00:14:04.230 --> 00:14:06.120 For example, some patients

293 00:14:06.120 --> 00:14:08.610 may not decompensate during the study.

294 00:14:08.610 --> 00:14:10.860 Those patients are censored.

295 00:14:10.860 --> 00:14:13.260 Here, patient one enters the study

296 00:14:13.260 --> 00:14:16.500 and experiences the event at three months.

297 00:14:16.500 --> 00:14:18.480 Patient two enters the study

298 00:14:18.480 --> 00:14:20.640 and is either lost a follow up

299 00:14:20.640 --> 00:14:23.638 or withdraws from the study before we observe the event.

300 00:14:23.638 --> 00:14:26.842 Patient two is censored at his withdrawal

301 00:14:26.842 --> 00:14:31.620 or at the last time we knew he was event free at two months.

302 00:14:31.620 --> 00:14:33.570 Patient three enters the study

303 00:14:33.570 --> 00:14:35.400 and does not experience the event

304 00:14:35.400 --> 00:14:37.530 before the end of the study.

305 00:14:37.530 --> 00:14:39.180 Patient three is censored

306 00:14:39.180 --> 00:14:43.500 at the administrative end of the study at four months.

307 00:14:43.500 --> 00:14:47.310 It's straightforward to compute survival time in patient one

308 00:14:47.310 --> 00:14:49.530 because he experienced the event,

309 00:14:49.530 --> 00:14:52.770 so his survival time is his event date

310 00:14:52.770 --> 00:14:54.753 minus his entry date.

311 00:14:55.590 --> 00:14:56.940 For those censored,

312 00:14:56.940 --> 00:14:58.980 there's no time-to-event

313 00:14:58.980 --> 00:15:00.980 but we want to account for time at risk

314 00:15:00.980 --> 00:15:03.600 because they're at risk for the event.

315 00:15:03.600 --> 00:15:06.323 But the fact that they don't experience the event

316 00:15:06.323 --> 00:15:09.330 before their loss or withdrawal

317 00:15:09.330 --> 00:15:12.120 or end of follow up is meaningful.

318 00:15:12.120 --> 00:15:14.160 We compute their survival time

319 00:15:14.160 --> 00:15:18.750 as their censoring date minus their entry date.

320 00:15:18.750 --> 00:15:22.470 Survival endpoints must contain two variables:

321 00:15:22.470 --> 00:15:24.090 an event indicator,

322 00:15:24.090 --> 00:15:27.840 which is usually referred to as a censoring indicator

323 00:15:27.840 --> 00:15:29.970 and the patient's survival time

324 00:15:29.970 --> 00:15:31.620 which is time-to-event

325 00:15:31.620 --> 00:15:33.900 for those experiencing the event

326 00:15:33.900 --> 00:15:37.170 or time to censoring for those censored.

327 00:15:37.170 --> 00:15:39.630 Patient one experiences the event

328 00:15:39.630 --> 00:15:42.251 so his event indicator is equal to 1.

329 00:15:42.251 --> 00:15:44.768 Patient's 2 and 3 are censored

330 00:15:44.768 --> 00:15:48.630 so their event indicator is equal to 0.

331 00:15:48.630 --> 00:15:51.060 And survival time is time from entry

332 00:15:51.060 --> 00:15:55.023 to the event or censoring, whichever occurs first.

333 00:15:56.220 --> 00:15:59.340 We recommend that you collect the relevant dates

334 00:15:59.340 --> 00:16:02.250 and allow us to compute survival time.

335 00:16:02.250 --> 00:16:04.500 Here we have the sample toy data set

336 00:16:04.500 --> 00:16:06.240 that we looked at earlier.

337 00:16:06.240 --> 00:16:08.639 The goal is to analyze time to response

338 00:16:08.639 --> 00:16:10.710 and describe the probability

339 00:16:10.710 --> 00:16:13.320 of treatment and response over time.

340 00:16:13.320 --> 00:16:17.610 The first variable needed to define the survival endpoint

341 00:16:17.610 --> 00:16:20.760 is the event or censoring indicator.

342 00:16:20.760 --> 00:16:23.205 Here we identify which subjects experienced

343 00:16:23.205 --> 00:16:28.080 the event of interest using the variable called response.

344 00:16:28.080 --> 00:16:32.280 Response equal to 1 if the subject experienced the event

345 00:16:32.280 --> 00:16:36.570 and 0 if the subject does not experience the event.

346 00:16:36.570 --> 00:16:41.370 Next, collect intake date or the start of follow up.

347 00:16:41.370 --> 00:16:43.860 In the subjects who experienced the event,

348 00:16:43.860 --> 00:16:45.870 the variable response date

349 00:16:45.870 --> 00:16:48.480 records the date the event occurred.

350 00:16:48.480 --> 00:16:50.820 Notice that this variable is missing

351 00:16:50.820 --> 00:16:53.523 for those who do not experience the event.

352 00:16:54.450 --> 00:16:55.950 For censored patients,

353 00:16:55.950 --> 00:16:57.810 we compute their survival time

354 00:16:57.810 --> 00:17:01.590 as the time between intake and censoring.

355 00:17:01.590 --> 00:17:03.720 Censoring time is the last time

356 00:17:03.720 --> 00:17:06.360 we knew the subject was event free.

357 00:17:06.360 --> 00:17:10.590 In this case, we can use the last visit date for a patient.

358 00:17:10.590 --> 00:17:14.190 You can record last visit date for all patients

359 00:17:14.190 --> 00:17:17.310 and we would use this information as the censoring date

360 00:17:17.310 --> 00:17:20.340 for those who do not experience the event.

361 00:17:20.340 --> 00:17:22.350 Notice that the last visit dates

362 00:17:22.350 --> 00:17:27.300 for subjects 106 and 107 are after their response dates.

363 00:17:27.300 --> 00:17:30.420 But when we compute the time variable for these subjects,

364 00:17:30.420 --> 00:17:34.920 we would use time from intake to the response date.

365 00:17:34.920 --> 00:17:38.010 You could also leave the last visit date blank

366 00:17:38.010 --> 00:17:40.860 for those subjects who experienced the event.

367 00:17:40.860 --> 00:17:44.130 We will use the response column and the time column

368 00:17:44.130 --> 00:17:47.550 as the response variable in the survival analysis.

369 00:17:47.550 --> 00:17:50.970 The response variable indicates if the time variable

370 00:17:50.970 --> 00:17:55.083 represents time to event or time to censoring.

371 00:17:56.430 --> 00:17:58.950 In the 2022 hepatology paper

372 00:17:58.950 --> 00:18:02.070 looking at the effect of TIPS on PPG

373 00:18:02.070 --> 00:18:03.690 in patients with ascites,

374 00:18:03.690 --> 00:18:06.330 we know that they looked at PPG reduction

375 00:18:06.330 --> 00:18:08.070 as a continuous endpoint

376 00:18:08.070 --> 00:18:11.250 but they also looked at survival or time to death

377 00:18:11.250 --> 00:18:14.250 in patients with ascites resolution after TIPS.

378 00:18:14.250 --> 00:18:15.450 That's the purple curve.

379 00:18:15.450 --> 00:18:19.579 Compared to those with a persistent need for paracentesis,

380 00:18:19.579 --> 00:18:22.710 six weeks after TIPS, the red curve,

381 00:18:22.710 --> 00:18:25.860 they found significantly improved survival

382 00:18:25.860 --> 00:18:28.263 in those with ascites resolution.

383 00:18:29.790 --> 00:18:32.215 In this video, we discussed data collection

384 00:18:32.215 --> 00:18:35.490 and variable types and specifically looked

385 00:18:35.490 --> 00:18:38.610 at quantitative endpoints, dichotomous endpoints,

386 00:18:38.610 --> 00:18:41.130 and time-to-event endpoints.

387 00:18:41.130 --> 00:18:43.290 Now that we've seen different examples

388 00:18:43.290 --> 00:18:46.218 of common endpoint types in clinical research,

389 00:18:46.218 --> 00:18:50.370 in our next video, the fourth video in this series,

390 00:18:50.370 --> 00:18:54.510 we'll discuss an important step in the study design process

391 00:18:54.510 --> 00:18:57.063 and that's sample size determination.