nice layout. 3

Consultants:
David Levy
Chad Pickering

BIOSTATISTICS CONSULTING WORKSHEET

UCLA FIELDING SCHOOL OF PUBLIC HEALTH, DEPARTMENT OF BIOSTATISTICS

Date: 4/12/18

Client Information:

- **Name(s):**
  William Northway, David Edwards, Thomas Colby, and Wayne Dyer
- **Department:**
  Stanford University Medical Center
- **Phone:**
  Contact information unavailable
- **Name of PI:**
  William Northway

Purpose:

- **Statement of research hypotheses:**
  Investigating the factors associated with broncho-pulmonary dysplasia (BPD) in new born infants. The investigation was open-ended, however.
- **Statement of statistical hypotheses:**
  1. Do oxygen or intubation therapy cause BPD?
  2. How are the pathological indices related to BPD?

3. Which background or health variables indicate high risk for contracting BPD?
4. How serious is BPD to an infant's life?

Background & References:

"BPD is a respiratory problem that affects some newborn infants who have had respiratory distress syndrome (RDS) and subsequent oxygen therapy for it. BPD has also been observed in some infants without RDS who have received high levels of oxygen or intubation therapy... BPD is a deterioration of the lung tissue evdienced, for example, by scarring... The severity of the deterioration is measured on a I through IV scale, IV being the most severe... infants reaching stage III or IV [are] classified as having BPD, those reaching only stage I or II as not having BPD," (Miller et al. 1980, *Biostatistics Casebook*).

Descriptions:

- **Study design:**

  A retrospective observational study of a convenience sample from a single academic tertiary care center.

- **Population(s):**

  Infants with respiratory distress syndrome (RDS) who receive ventilatory assistance by intubation.

- **Sample(s):**

  All infants treated at Stanford Medical Center between 1962 and 1973 with clinical and X-ray symptoms of RDS and who received ventilatory assistance by intubation for more than 24 hours, except for those whose records were unavailable.

- **Dependent variable(s):**

  Sex, birth year, one-minute APGAR, estimated gestational age, birthweight, age at onset of respiratory symptoms, age at onset of ventilatory assistance, duration of endotracheal intubation, duration of assisted ventilation, duration of exposure to low concentration O_2, duration of exposure to medium concentration O_2, duration of exposure to high concentration O_2, survival at end of study, survival at follow-up, RDS severity,

highest radiographic BPD stage, bronchiola mucosa, pulmonary interstitium, hyalin membrane, alveolar infiltrate, inflation pattern, BPD stage based on lung tissue, hematoidin

- **Independent variable(s):**

  BPD presence (Q1, Q2 and Q3), survival at end of study and survival at follow-up (Q4)

BIOSTAT 402A: Assignment 2

Chad Pickering and David Levy

4/16/2018

Below is code that imports the raw text data, cleans the data set for consistency throughout each field, and removes all subjects that are ineligible for the study as described by the clients. As indicated, infants should only be included if they survived longer than 72 hours, resulting in a sample size of 248 for the analysis group.

```
# import data
bpd <- read.csv("bpd.txt", header=FALSE, na.strings="", stringsAsFactors=FALSE)

# cleaning procedures for each field
bpd$patient_id <- substr(bpd$V1, 1, 3)
bpd$sex <- as.factor(substr(bpd$V1, 4, 4))
bpd$birth_year <- as.factor(paste("19", substr(bpd$V1, 5, 6), sep=""))
bpd$apgar <- as.numeric(substr(bpd$V1, 7, 8))
bpd$gest_age <- as.numeric(substr(bpd$V1, 9, 11))/10
bpd$birthwt <- as.numeric(substr(bpd$V1, 12, 15))
bpd$resp_onset_age <- as.numeric(substr(bpd$V1, 16, 17))/10
bpd$vent_onset_age <- as.numeric(substr(bpd$V1, 18, 20))
bpd$intub_dur <- as.numeric(substr(bpd$V1, 21, 24))
bpd$vent_dur <- as.numeric(substr(bpd$V1, 25, 28))
bpd$elev_oxygen_dur_low <- as.numeric(substr(bpd$V1, 29, 33))
bpd$elev_oxygen_dur_med <- as.numeric(substr(bpd$V1, 34, 37))
bpd$elev_oxygen_dur_high <- as.numeric(substr(bpd$V1, 38, 41))
bpd$alive_1975 <- as.factor(substr(bpd$V1, 42, 42))
bpd$surv_dur <- as.numeric(substr(bpd$V1, 43, 47))
bpd$rds_severity <- as.factor(substr(bpd$V1, 48, 48))
bpd$bpd_highstage <- as.numeric(substr(bpd$V1, 49, 50))/10
bpd$bronch_muc <- as.factor(substr(bpd$V1, 51, 51))
bpd$pulm_int <- as.factor(substr(bpd$V1, 52, 52))
bpd$hyalin_mem <- as.factor(substr(bpd$V1, 53, 53))
bpd$alv_infil <- as.factor(substr(bpd$V1, 54, 54))
bpd$infl_pat <- as.factor(substr(bpd$V1, 55, 55))
bpd$bpd_lung <- as.factor(substr(bpd$V1, 56, 56))
bpd$hemat <- as.factor(substr(bpd$V1, 57, 57))

# finalize data set for analysis
bpd <- bpd[!is.na(bpd$patient_id), -1]

# removing infants that did not survive longer than 72 hrs
bpd <- bpd[bpd$surv_dur > 72,]

# generate binary bpd variable
bpd$bpd_prev <- ifelse(bpd$bpd_highstage >= 3, 1, 0)
```

1. Based on the introduction and variable list, complete a Biostatistics Consulting Worksheet.

See attached.

2. Briefly outline statistical procedures that may help answer the four main statistical questions listed on Page 105 of Wolfe's casebook study.

Question 1. Do oxygen or intubation therapy cause BPD?

We essentially have two independent samples here: BPD and no BPD. We should check the sample sizes of both groups:

```
## # A tibble: 2 x 3
##   bpd_prev     n   pct
##      <dbl> <int> <dbl>
## 1     0      174 0.702
## 2     1.00    74 0.298
```

So the groups are unequal, but sample sizes are large enough to assume CLT holds in both cases. Now we should determine whether the variances of our two samples are significantly different:

```
##
##  F test to compare two variances
##
## data:  intub_dur by bpd_prev
## F = 0.078089, num df = 173, denom df = 73, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.05209224 0.11347177
## sample estimates:
## ratio of variances
##          0.07808905
```

```
##
##  F test to compare two variances
##
## data:  elev_oxygen_dur_low by bpd_prev
## F = 0.019786, num df = 173, denom df = 73, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.01319897 0.02875112
## sample estimates:
## ratio of variances
##          0.01978596
```

```
##
##  F test to compare two variances
##
## data:  elev_oxygen_dur_med by bpd_prev
## F = 0.012928, num df = 173, denom df = 73, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.00862405 0.01878564
## sample estimates:
## ratio of variances
##          0.01292791
```

```
##
##  F test to compare two variances
##
## data:  elev_oxygen_dur_high by bpd_prev
```

```
## F = 0.1057, num df = 173, denom df = 73, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.07051436 0.15360039
## sample estimates:
## ratio of variances
##          0.1057048
```

The variances are unequal for all of the intubation duration and elevated oxygen duration variables, so two-sample t-tests with unequal variances can be used with variables intubation duration (hours), elevated oxygen (low, medium, high).

NOTE: We should point out here that the casebook study indicates a sample size of 78 infants with BPD and 170 without BPD. However, these numbers are only attained if a high BPD score of 2.5 is classified as having the disease. Confusingly, the casebook study also indicates that an infant should only be considered to have BPD if a BPD score of 3 or higher is recorded. Thus, we elected to use latter cutoff for the remainder of this analysis, yielding sample sizes of 74 infants with BPD and 174 infants without BPD.

Question 2. How are the pathological indices related to BPD?

We could easily compare means via a 2-sample t-test for the bronch_inuc, pulm_int, hyalin_mem, alv_infil, infl_pat, and hemat variables. Or, because they are discrete values $(1,\ldots,4)$, we could use chi-square tests (2x4 for each) or their non-parametric equivalents if sample size is an issue.

Question 3. Which background or health variables indicate high risk for contracting BPD?

We could use multivariate logistic regression with outcome variable BPD (0 or 1), where variables included are sex, year, APGAR score, gestational age, birthweight, age at onset, and severity of RDS.

Question 4. How serious is BPD to an infant's life?

We could compare duration of survival and survival count between BPD and non-BPD groups if mortality rate information is desired. Mean age of onset for ventiliation assistance can also be compared if a quality of life measure is desired.

survival analysis techniques

3. Perform a descriptive analysis including the following:

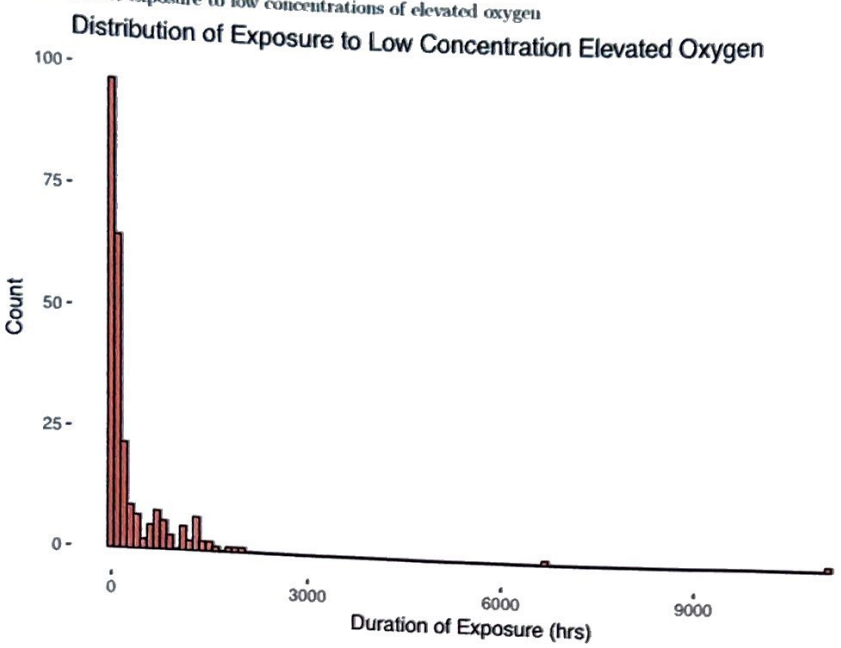
Part a. Demographic table describing the study population.

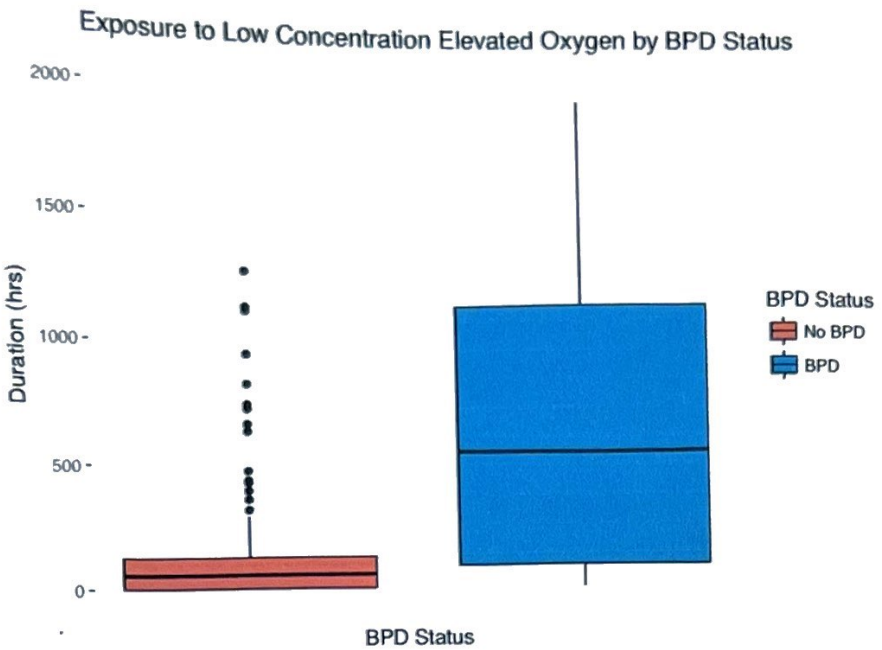
Characteristic	Full cohort (n = 248)	No BPD (n = 174)	BPD (n = 74)
Sex, count (%)			
Female	76 (30.6%)	55 (31.6%)	21 (28.4%)
Male	172 (69.4%)	119 (68.3%)	53 (71.6%)
Birth Year, median (range)	1971 (1962-1973)	1971 (1962-1973)	1969 (1964-1973)
APGAR Score, count (%)			
0	1 (0.4%)	0 (0.0%)	1 (1.4%)
1	15 (6.0%)	13 (7.5%)	2 (2.7%)
2	11 (4.4%)	8 (4.6%)	3 (4.1%)
3	15 (6.0%)	11 (6.3%)	4 (5.4%)
4	14 (5.6%)	10 (5.7%)	4 (5.4%)
5	24 (9.7%)	16 (9.2%)	8 (10.8%)
6	24 (9.7%)	19 (10.9%)	5 (6.8%)
7	31 (12.5%)	22 (12.6%)	9 (12.2%)

Characteristic	Full cohort (n = 248)	No BPD (n = 174)	BPD (n = 74)
8	37 (14.9%)	27 (15.5%)	10 (13.5%)
9	25 (10.1%)	19 (10.9%)	6 (8.1%)
10	3 (1.2%)	3 (1.7%)	0 (0.0%)
Gest. age, median (IQR)	33 (31 - 36)	33 (31 - 36)	32.5 (31 - 35)
Birthweight, median (IQR)	1786 (1340-2329)	1786 (1335-2396)	1778 (1360-2233)
Resp. symptoms (age at onset)	0.0 (0.0 - 0.5)	0.0 (0.0 - 0.5)	0.0 (0.0 - 0.3)
Vent. assist. (age at onset)	26 (12 - 40)	25 (11.25 - 38.75)	26.5 (13.25 - 43)

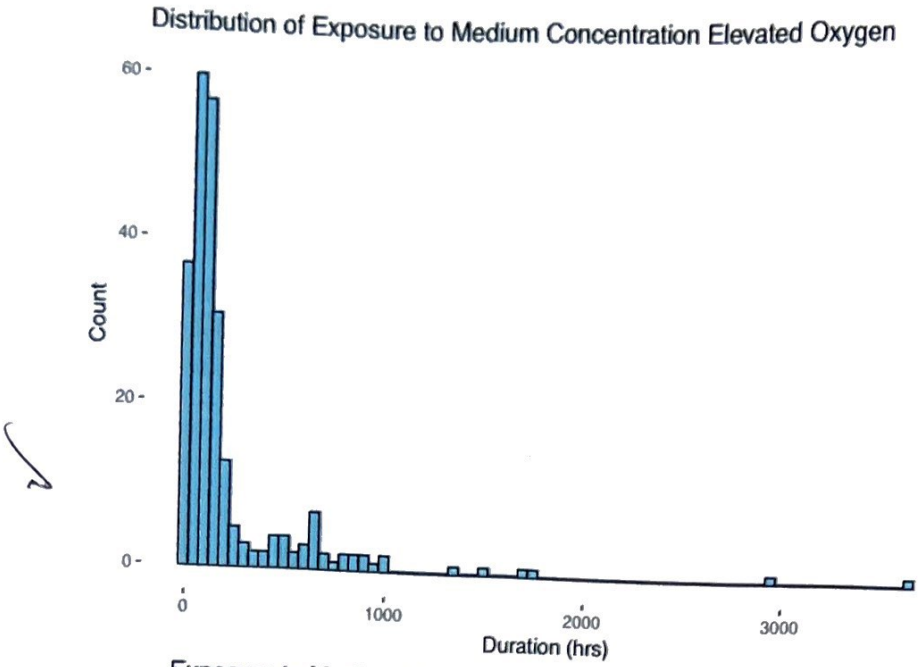
Part b. Plots (histograms or box plots) and summary statistics.

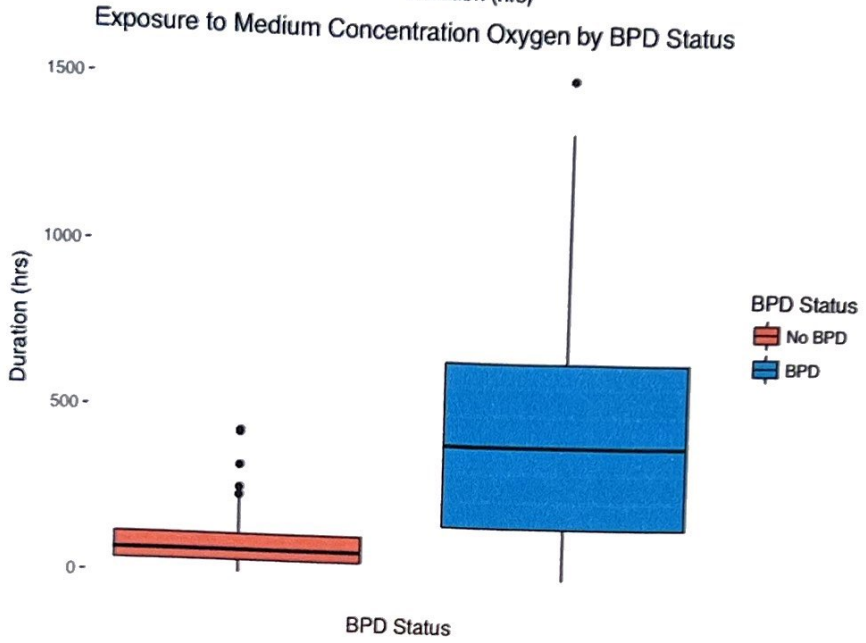
i. hours of exposure to low concentrations of elevated oxygen



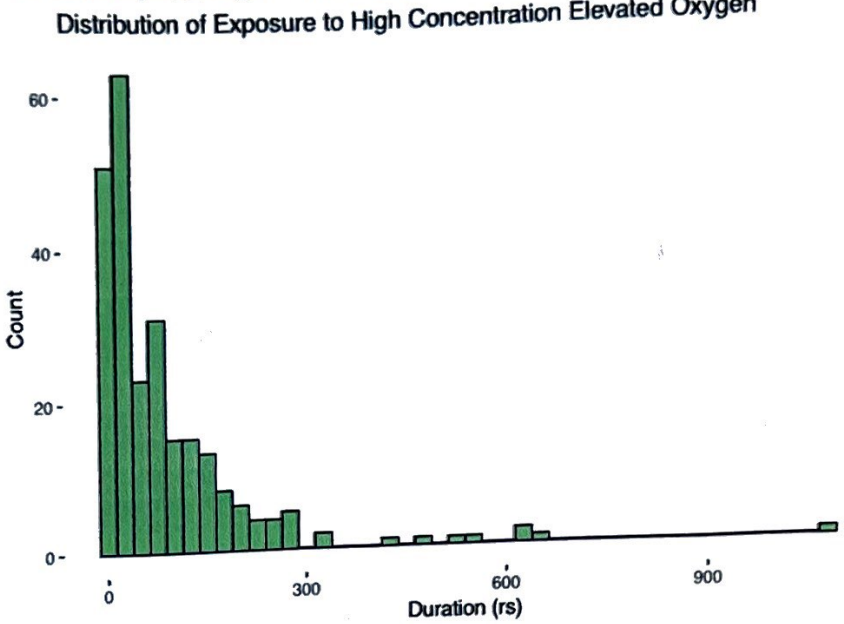


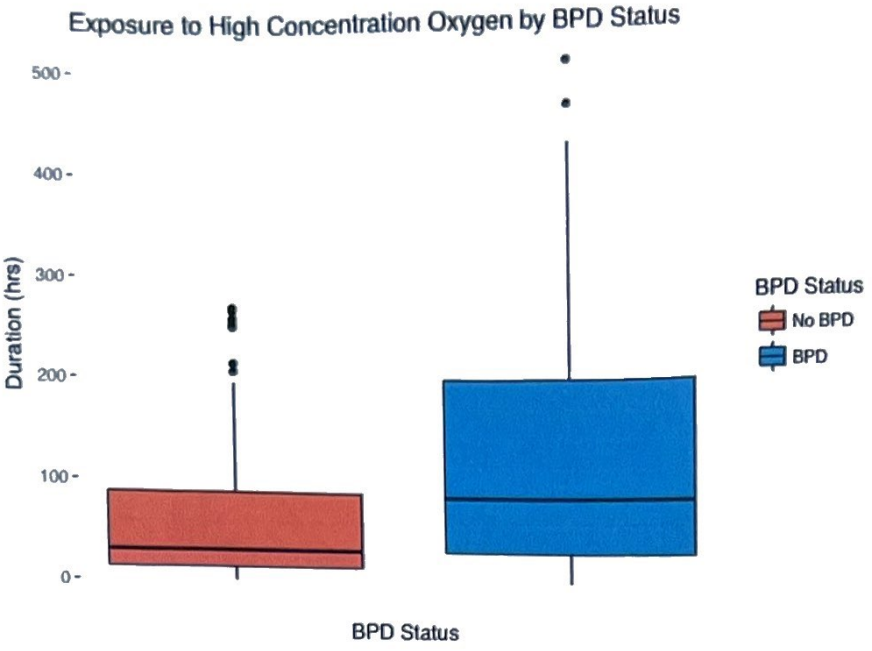
ii. hours of exposure to medium concentrations of elevated oxygen


Distribution of Exposure to Medium Concentration Elevated Oxygen
Count
60
40
20
0
0
1000
2000
3000
Duration (hrs)

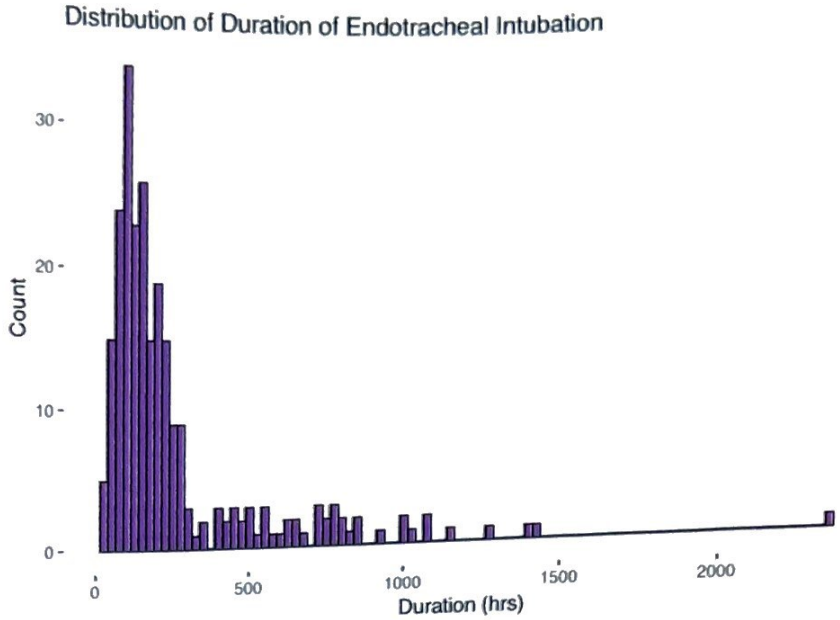

Exposure to Medium Concentration Oxygen by BPD Status
Duration (hrs)
1500
1000
500
0
BPD Status
No BPD
BPD
BPD Status

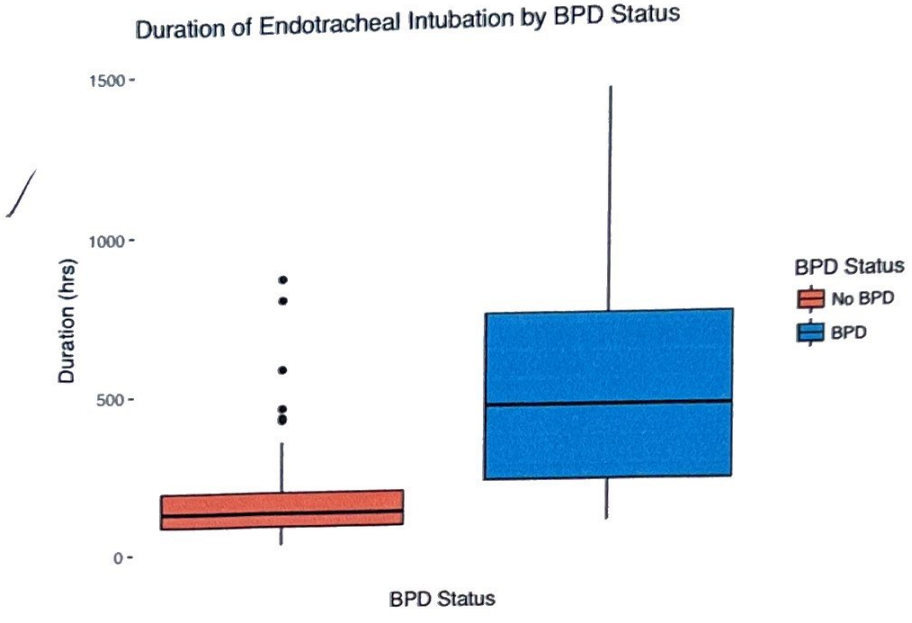
iii. hours of exposure to high concentrations of elevated oxygen



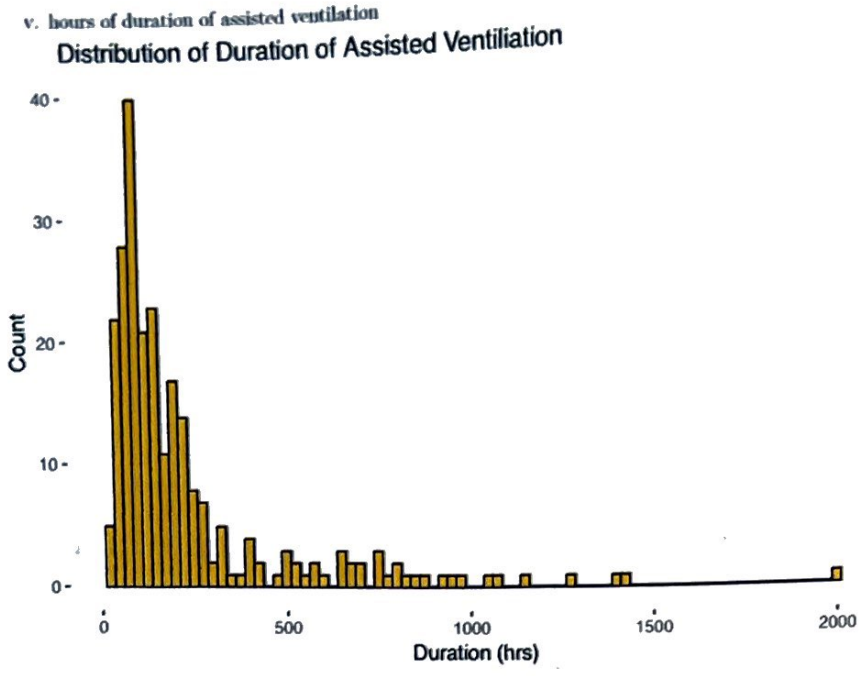


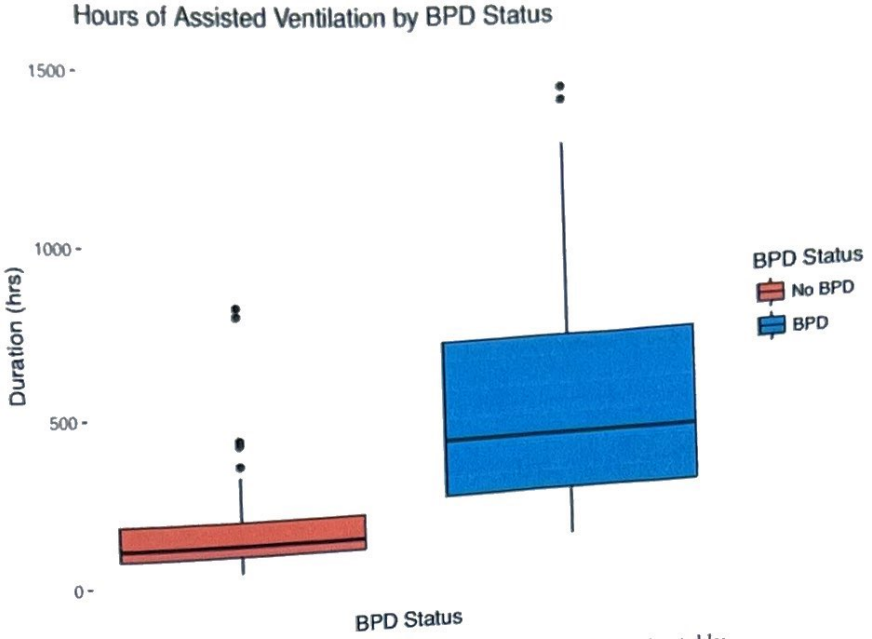
iv. hours of duration of endotracheal intubation


Distribution of Duration of Endotracheal Intubation
Count
30
20
10
0
0
500
1000
1500
2000
Duration (hrs)


Duration of Endotracheal Intubation by BPD Status
Duration (hrs)
1500
1000
500
0
BPD Status
No BPD
BPD
BPD Status

v. hours of duration of assisted ventilation





Summary statistics for the variables plotted above are presented in the following table:

Characteristic	Full cohort (n = 248)	No BPD (n = 174)	BPD (n = 74)
Low O_2, hrs, median (IQR)	84.5 (3-278.75)	56 (2-126.5)	546 (91-1134)
Med O_2, hrs, median (IQR)	100 (50-179.5)	70.5 (39.25-119)	410.5 (160.25-665)
High O_2, hrs, median (IQR)	47.5 (16-117.5)	31 (14-88.5)	86 (30.5-204.25)
Intubation, hrs, median (IQR)	159 (99-263.25)	128.5 (86.5-193.75)	456 (223.75-741.5)
Assisted vent, hrs, median (IQR)	145 (91.5-243.5)	109 (78.25-182)	390 (222-688)

Part c. Counts of RDS and radiographic BPD scores.

The number of infants classified as having various stages of RDS is summarized in the table below:

RDS severity, count (%)	n = 248
0 (no disease seen)	2 (0.8%)
1	48 (19.4%)
2	40 (16.1%)
3	120 (48.4%)
4	15 (6.0%)
5 (very severe disease)	22 (8.9%)

The number of infants falling into each stage (or between stages if status was indeterminate) of BPD is summarized below. Note that infants were classified as having BPD if they were radiographically determined

to have stage III or stage IV disease. These criteria produced counts of 74 infants with BPD and 174 infants without BPD.

Radiographic BPD score, count (%)	$n = 248$
1.0 (stage I)	125 (50.4%)
1.5 (stage I-II)	11 (4.4%)
2.0 (stage II)	34 (13.7%)
2.5 (stage II-III)	4 (1.6%)
3.0 (stage III)	34 (13.7%)
4.0 (stage IV)	40 (16.1%)

Part d. Missing data and outliers.

Variables with large numbers of missing data values (>20% missing) were identified and summarized in the tabled below:

Variable	Missing value, count (%)
APGAR	48 (19.4%)
Bronchiola mucosa	164 (66.1%)
Pulmonary interstitium	152 (61.3%)
Hyalin membranes	194 (78.2%)
Alveolar infiltrate	162 (65.3%)
Inflation pattern	178 (71.8%)
BPD stage by tissue	152 (61.3%)
Hematoidin	205 (82.7%)

watch denominators

Using the standard rule of thumb that an observation is an extreme outlier if it falls more than 1.5 interquartile ranges below the the first quartile or more than 1.5 interquartile ranges above the third quartile, we determined that the following variables have extreme outlying observations:

Birthweight, age at onset of respiratory symptoms, age at onset of ventilatory assistance, duration of exposure to oxygen at all three levels and survival duration.

Values?