Q for Bill:

* Multiple testing correction
* Phylogeny outgroup

Data in general

* The original Raw dataset contains only 98 samples, although 100 patients are reported. Why ?
* Carl’s dataset seem to
  + include 35\_S35, which should be treated as “contamination” and excluded (%GBS= 13)
  + Excludes 26\_S26 and 93\_S93. Why ?
  + 86\_S86 has days\_at\_dx = 4651. It should be 4590, since date of Diagnosis is 2020-02-25 and DoB is 2007-08-02
  + 59\_S59 is isolated from “blood” instead of “both blood and meningitis”, and thus as “No Meningitis”, instead of “Meningitis”
  + 21\_S2, 34\_S34, and 51\_S51 are classified as not having Neutropenia although they have ANC <1.50 *\*10^3 cells/mm3* (1.32, 1.49 and 1.29, respectively)
  + Just a detail: Leukocytosis and Leucopenia were coded such that 0= Leukocytosis (or Leucopenia) and 1= No Leukocytosis (or No Leucopenia). This may appear counter-intuitive
* Any differences in Genomic characteristics are due to my updating of Bioinformatics methods to meet gold standards.

Figure1:

* I propose a histogram showing serotype and CC frequency over time, but not in ICU vs Non-ICU, because this is relevant only in later parts. Notably, whether different serotypes present among ICU-admitted vs non-ICU admitted patients falls more under correlates of severe disease. Thoughts ?

Figure 2:

* For the global phyogeny, were are the accessions used ? Are they the same as these: <https://docs.google.com/spreadsheets/d/15x8nOByTBejwWHnZJ4Dg2_G-THJx4EBVFFQ78J3JSqg/edit?usp=sharing> ?

Table 1:

* The presumed causal direction of the associations investigated is from “Infant” and “Meningitis” to the various laboratory metrics, and from the various laboratory metrics to “ICU admission”. For the binary exposures of “Infant” and “Meningitis”, we used Wilcoxon test to investigate their association with continuous laboratory metrics and logistic regression for the binary metrics.For the binary outcome of “ICU admission”, we used logistic regression to investigate the effect of both continuous and binary laboratory metrics.
* Ideally, due to the different direction of the causal relationships presented in this table, it should be separated into 2 tables, and contain risk and odds rations + 95%

CIs.

Table 2:

* Do we want to add an adjusted OR ? i.e. to control for potential confounders ?
* Do we want to check for association with
  + Meningitis?
  + age of disease onset ?

**Supplementary Table 4:**

* Replaced with intended. Is this what you had in mind ?
* Currently, the isolates with “Known STs with mismatches” are types as CC=NA. Should I match them with their corresponding CC?

**Supplementary Table 5:**

* Should I keep the SIP1a and SIP3a rows ?
* Would you additionally like a column of the total among EOD+LOD , or infants alone ?

# Same patient phylogeny

A diagram of a tree

Description automatically generated with medium confidence

# Outgroup Test

## GAS

A graph with colorful dots and numbers

Description automatically generatedA chart with colorful dots and numbers

Description automatically generated

## S.Pneumo

A graph with colored dots

Description automatically generated with medium confidenceA chart with colorful dots

Description automatically generated