

20251021_predicted_nb_updated

2025-10-21

```
#setup dataset and model
nsd_yes_metadata <-
  metadata %>%
  filter(nsd == "Yes") %>%
  filter(., age.in.months != "NA" & da != "NA" & container.title != "NA") %>%
  mutate(da_factor = factor(da),
        container.title = factor(container.title))

nsd_yes_model <-
  glm.nb(is.referenced.by.count ~ da_factor + log(age.in.months) + container.title +
  + container.title*da_factor + log(age.in.months)*da_factor + container.title*log(age.in.months) +
  log(age.in.months)*da_factor*container.title, data = nsd_yes_metadata, link = log)
```

Validation of fit

- How many citations did a paper published in 2021 receive in each journal?

```
nsd_yes_metadata %>%
  filter(year.published == 2021) %>%
  summarize(mean_cites_2021 = mean(is.referenced.by.count),
            median_cites_2021 = median(is.referenced.by.count),
            .by = container.title)
```

```
## # A tibble: 11 x 3
##   container.title               mean_cites_2021 median_cites_2021
##   <fct>                           <dbl>                 <dbl>
## 1 Antimicrobial Agents and Chemotherapy      12.3                  8
## 2 Applied and Environmental Microbiology     13.9                  11
## 3 Infection and Immunity                   9.19                  9
## 4 Journal of Bacteriology                  9.40                  7
## 5 Journal of Clinical Microbiology       16.5                  12
## 6 Journal of Virology                     16.3                  11
## 7 mBio                                21.4                  16
## 8 Microbiology Resource Announcements    2.31                  1
## 9 mSphere                               15.6                  12
## 10 mSystems                             18.1                  14
## 11 Microbiology Spectrum                 13.3                  10
```

Using the existing model

- Remove JMBE, MRA, GA from modeling

- Train on papers <= 10 years old (age.in.months <= 120)
- Re-create figure with and without a common axis
- Previously N = 41,271, now N = 13,911

```
#filter to remove jmbe, mra, ga and for age.in.months <= 120
ten_metadata <-
  nsd_yes_metadata %>%
  filter(journal_abrev != "jmbe" & journal_abrev != "mra" & journal_abrev != "genomea" & age.in.months <= 120)

#sanity check
ten_metadata %>%
  count(journal_abrev)

## # A tibble: 10 x 2
##   journal_abrev     n
##   <chr>           <int>
## 1 aac              1197
## 2 aem              2695
## 3 iai              342
## 4 jb               536
## 5 jcm              699
## 6 jvi              1353
## 7 mbio             1982
## 8 msphere          971
## 9 msystems         1400
## 10 spectrum        2736

ten_metadata %>%
  count(age.in.months) %>%
  tail()

## # A tibble: 6 x 2
##   age.in.months     n
##   <dbl>       <int>
## 1 115            77
## 2 116            68
## 3 117            110
## 4 118            74
## 5 119            83
## 6 120            68

#retrain model
ten_model <-
  glm.nb(is.referenced.by.count ~ da_factor + log(age.in.months) + container.title +
  + container.title*da_factor + log(age.in.months)*da_factor + container.title*log(age.in.months) +
  + log(age.in.months)*da_factor*container.title, data = ten_metadata, link = log)

#get data out of model

age_values <- seq(5, 120, 5)
p_10 <- get_model_data(model = ten_model, type = "pred",
```

```

    terms = c("da_factor", "age.in.months[age_values]", "container.title"),
    colors = "bw") %>%
tibble(da_factor = ifelse(.x == 1, "Data not available", "Data available"), predicted_citations =
age.in.months = .group, container.title = .facet)

#re-create figure with free axes

predicted_plot <-
  ggplot(data = p_10, mapping = aes(x = as.numeric(age.in.months), y = predicted_citations,
                                      color = da_factor)) +
  geom_line(aes(x = age.in.months, y = predicted_citations, group = da_factor)) +
  geom_ribbon(mapping = aes(ymin = conf.low, ymax = conf.high,
                            group = da_factor, fill = da_factor), alpha = 0.2) +
  facet_wrap(~ container.title, nrow = 2,
             labeller = label_wrap_gen(width = 18),
             scale = "free_y") +
  labs(title = "Predicted Number of Citations from GLM.NB",
       subtitle = "Data age.in.months <= 120, removal of JMBE, GA, MRA",
       x = "Age in Months",
       y = "Predicted Number Citations",
       color = "Data availability\nwith 95% CI",
       fill = "Data availability\nwith 95% CI") +
  scale_x_discrete(breaks = seq(12, 120, 12)) +
  theme_classic() +
  theme(legend.position = "bottom" )

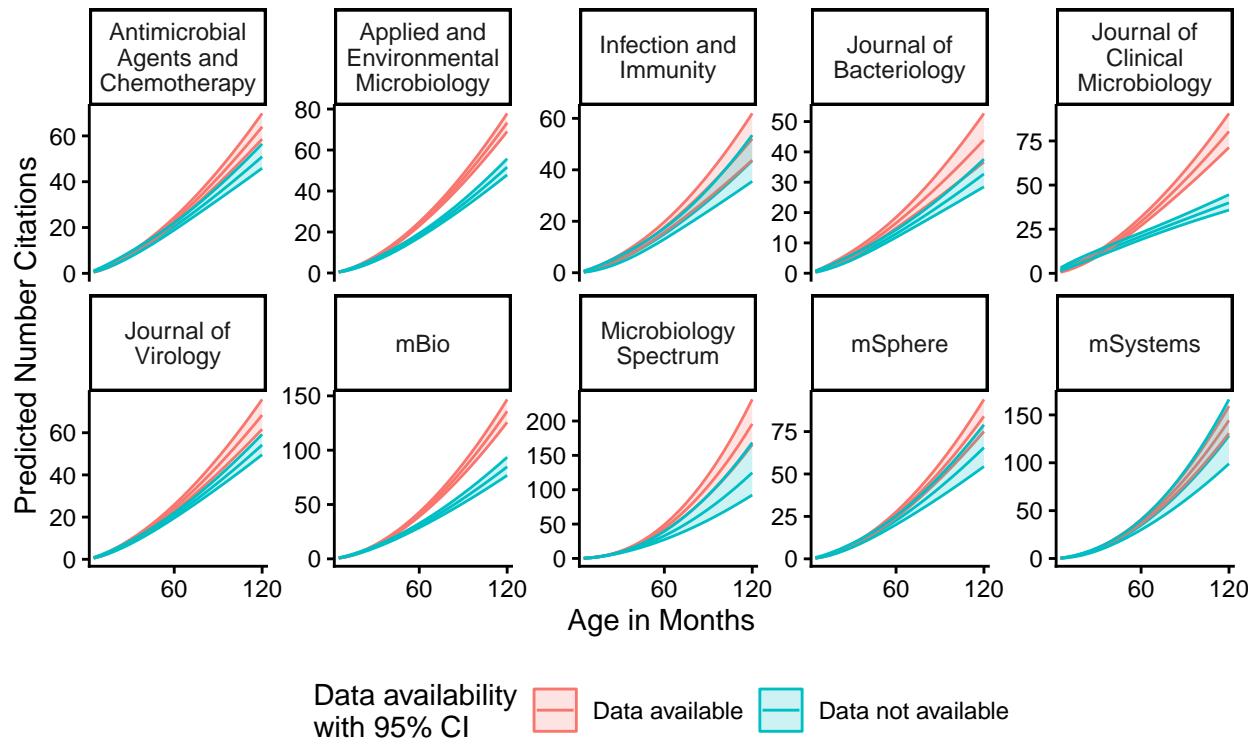
predicted_plot_fixed <-
  ggplot(data = p_10, mapping = aes(x = as.numeric(age.in.months), y = predicted_citations,
                                      color = da_factor)) +
  geom_line(aes(x = age.in.months, y = predicted_citations, group = da_factor)) +
  geom_ribbon(mapping = aes(ymin = conf.low, ymax = conf.high,
                            group = da_factor, fill = da_factor), alpha = 0.2) +
  facet_wrap(~ container.title, nrow = 2,
             labeller = label_wrap_gen(width = 18)) +
  labs(title = "Predicted Number of Citations from GLM.NB",
       subtitle = "Data age.in.months <= 120, removal of JMBE, GA, MRA,
fixed axes",
       x = "Age in Months",
       y = "Predicted Number Citations",
       color = "Data availability\nwith 95% CI",
       fill = "Data availability\nwith 95% CI") +
  scale_x_discrete(breaks = seq(12, 120, 12)) +
  theme_classic() +
  theme(legend.position = "bottom" )

predicted_plot

```

Predicted Number of Citations from GLM.NB

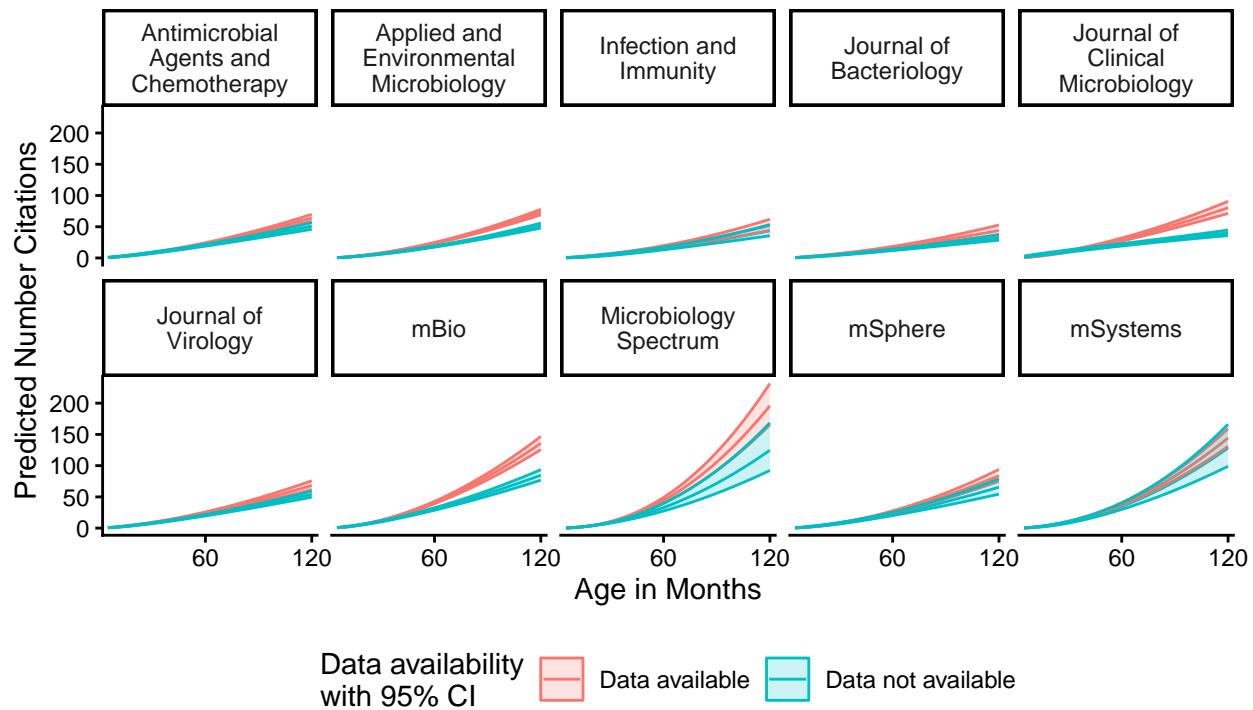
Data age.in.months <= 120, removal of JMBE, GA, MRA



predicted_plot_fixed

Predicted Number of Citations from GLM.NB

Data age.in.months <= 120, removal of JMBE, GA, MRA,
fixed axes



For each journal separately, overlay citations by paper on model output for DA yes and DA no

```
#lol i just add the data in each geom
journals <- ten_metadata %>%
  count(container.title) %>%
  mutate(container.title = as.character(container.title)) %>%
  dplyr::select(container.title)

j<- 6

for(j in 1:nrow(journals)) {
  #filter metadata for that journal
  j_metadata <- ten_metadata %>%
    filter(container.title == journals$container.title[[j]])

  #filter p_10
  model_data <- p_10 %>%
    filter(container.title == journals$container.title[[j]]) %>%
    mutate(da_factor = ifelse(da_factor == "Data available", "Yes", "No"),
          age.in.months = as.numeric(as.character(age.in.months)))

  #make plot
  plot <-
```

```

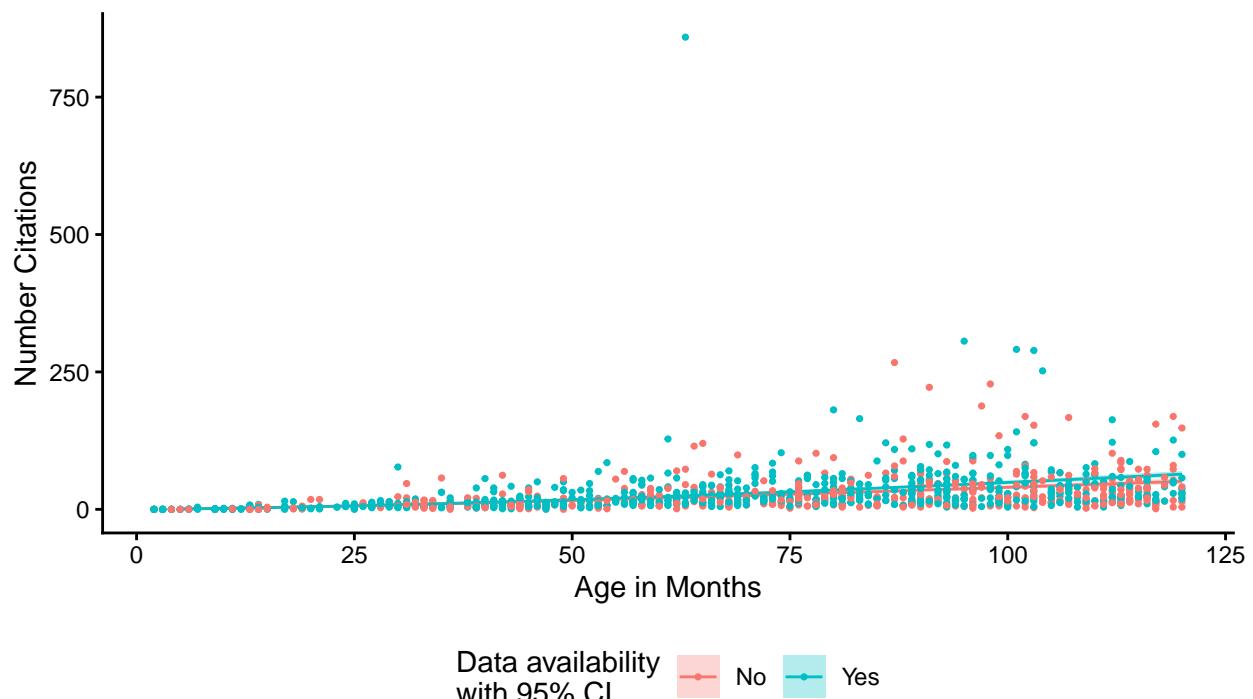
ggplot() +
  # mapping = aes(x = age.in.months, y = predicted_citations,
  #                 color = da_factor)) +
  geom_line(data = model_data, aes(x = age.in.months, y = predicted_citations, group = da_factor, color = da_factor)) +
  geom_ribbon(data = model_data, mapping = aes(x = age.in.months, y = predicted_citations, group = da_factor, fill = da_factor), alpha = 0.3) +
  geom_point(data = j_metadata, aes(x = age.in.months,
                                    y = is.referenced.by.count, color = da_factor),
             size = 0.6) +
  labs(title = paste0("Model vs True Number of Citations from GLM.NB for\n", journals$container.title[1]),
       subtitle = "Data age.in.months <= 120, removal of JMBE, GA, MRA",
       x = "Age in Months",
       y = "Number Citations",
       color = "Data availability\nwith 95% CI",
       fill = "Data availability\nwith 95% CI") +
  # scale_x_discrete(breaks = seq(12, 120, 12)) +
  theme_classic() +
  theme(legend.position = "bottom" )

print(plot)
}

```

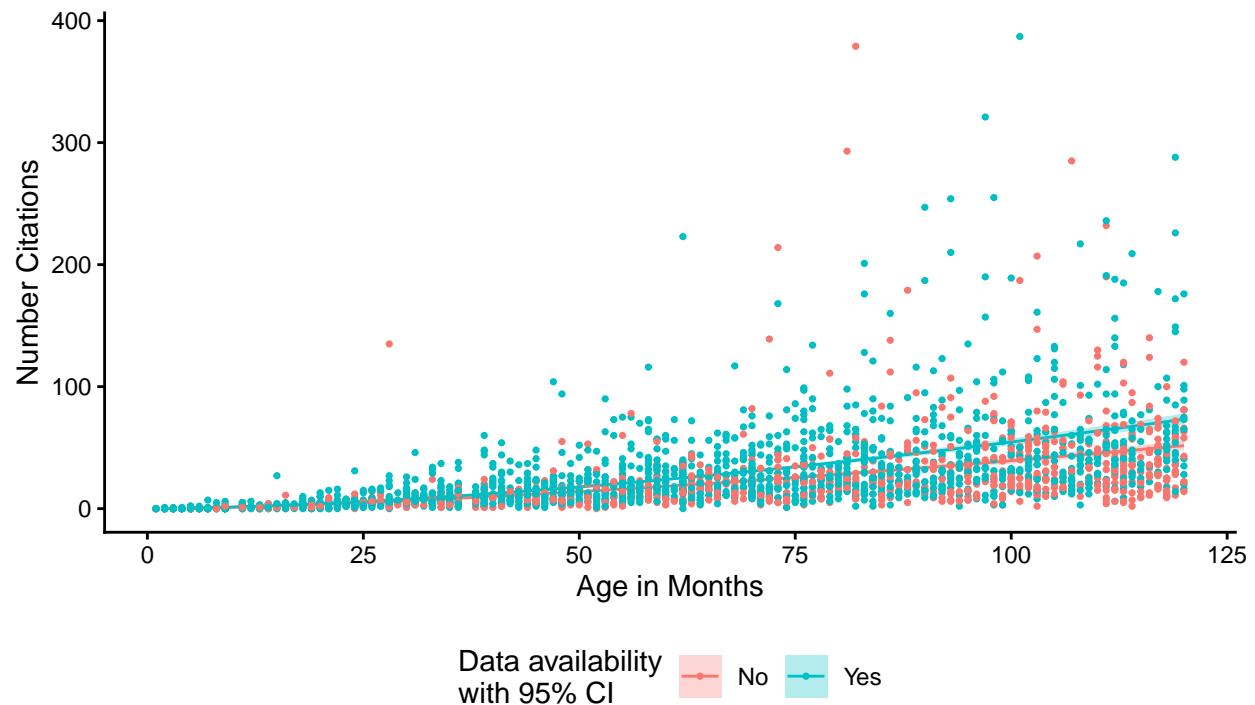
Model vs True Number of Citations from GLM.NB for Antimicrobial Agents and Chemotherapy

Data age.in.months <= 120, removal of JMBE, GA, MRA



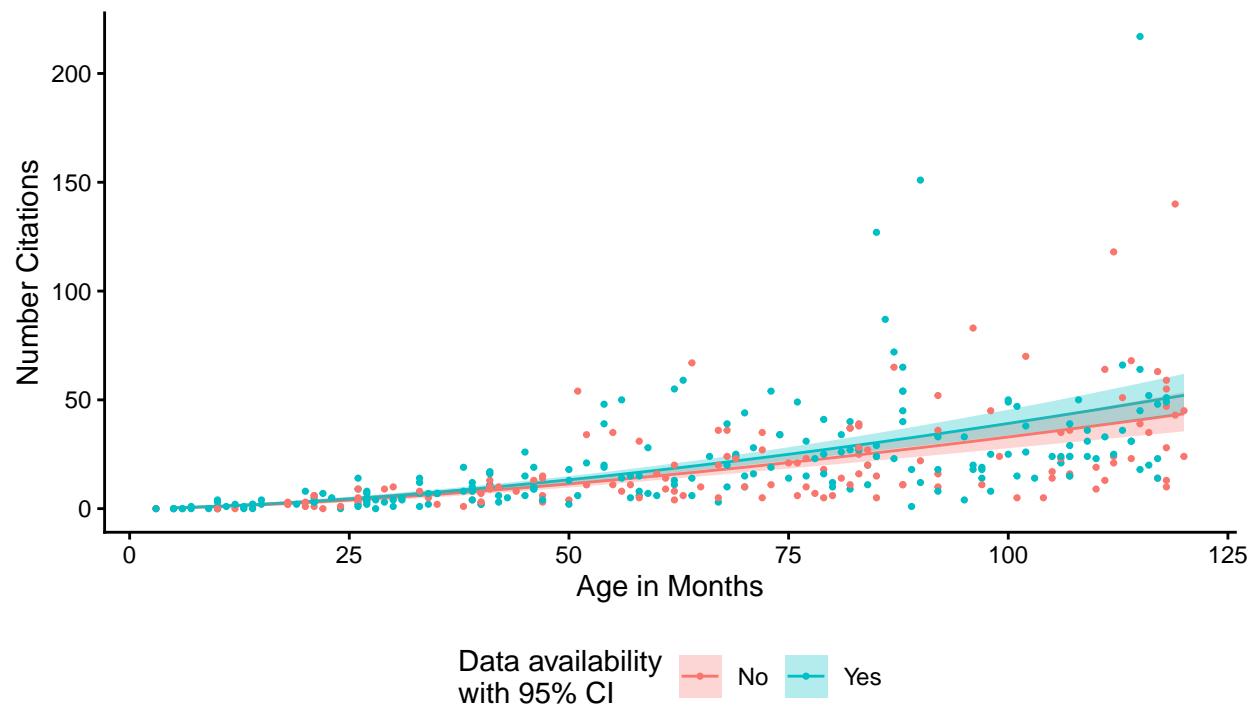
Model vs True Number of Citations from GLM.NB for Applied and Environmental Microbiology

Data age.in.months <= 120, removal of JMBE, GA, MRA



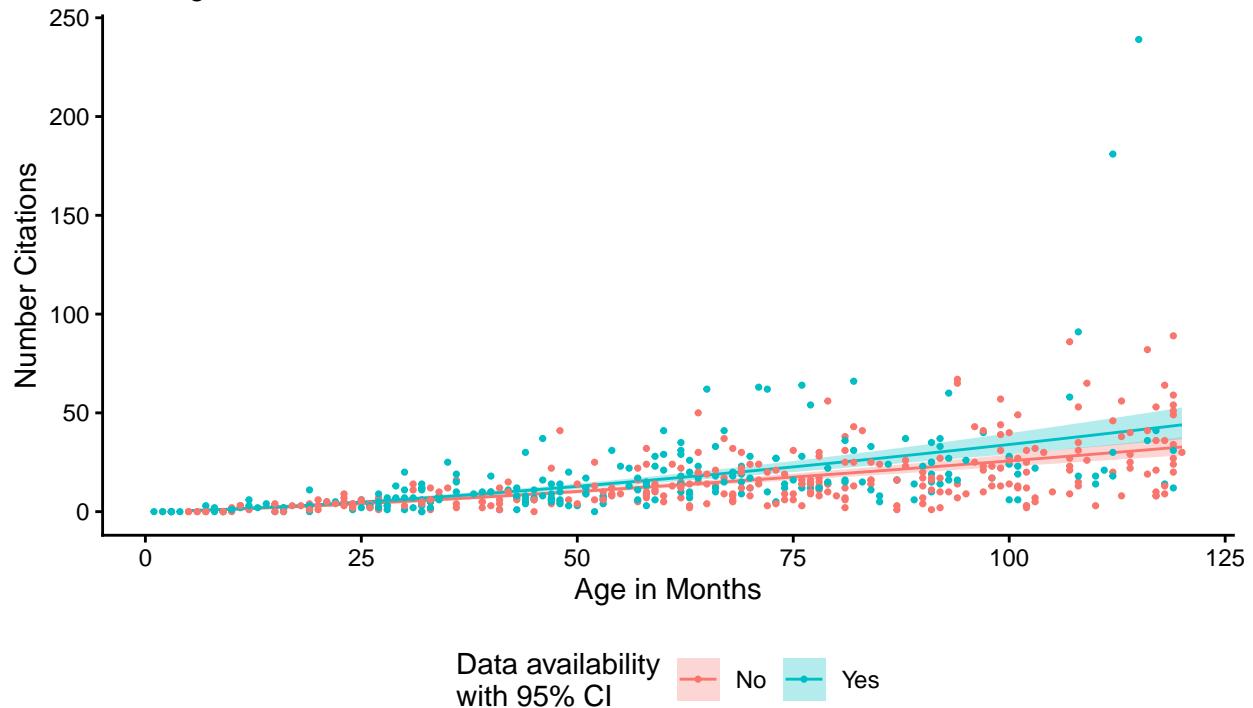
Model vs True Number of Citations from GLM.NB for Infection and Immunity

Data age.in.months <= 120, removal of JMBE, GA, MRA



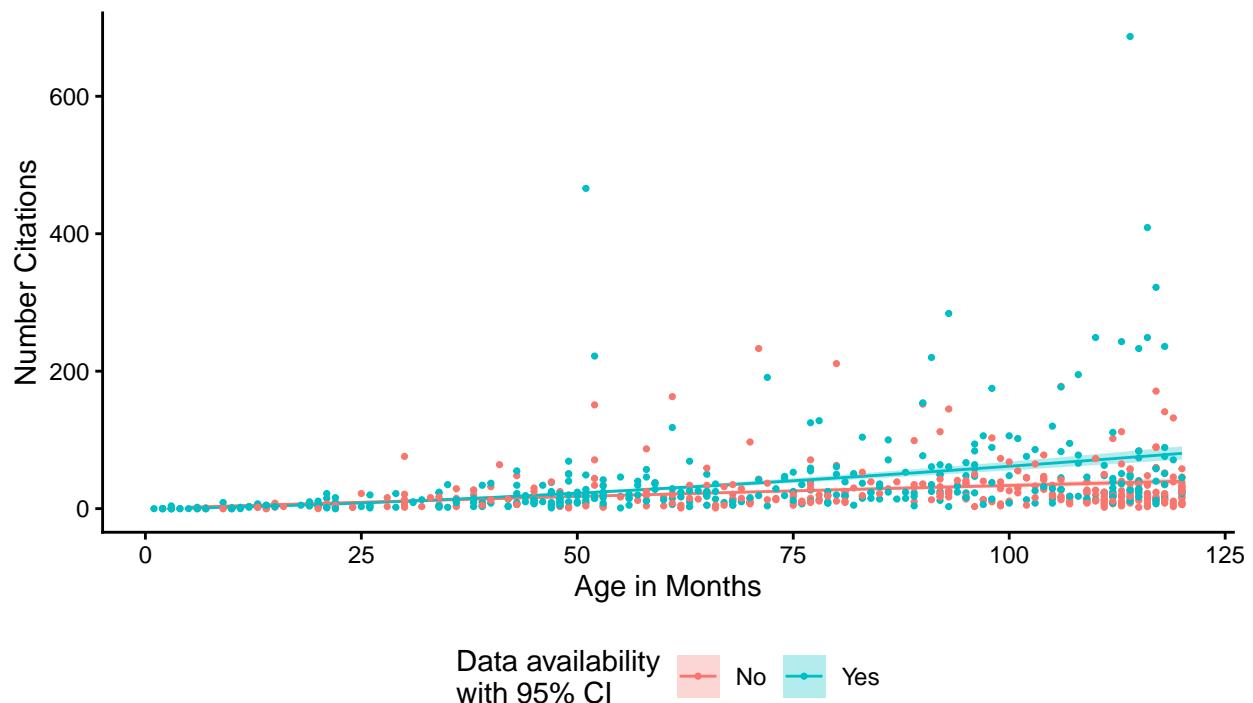
Model vs True Number of Citations from GLM.NB for Journal of Bacteriology

Data age.in.months <= 120, removal of JMBE, GA, MRA



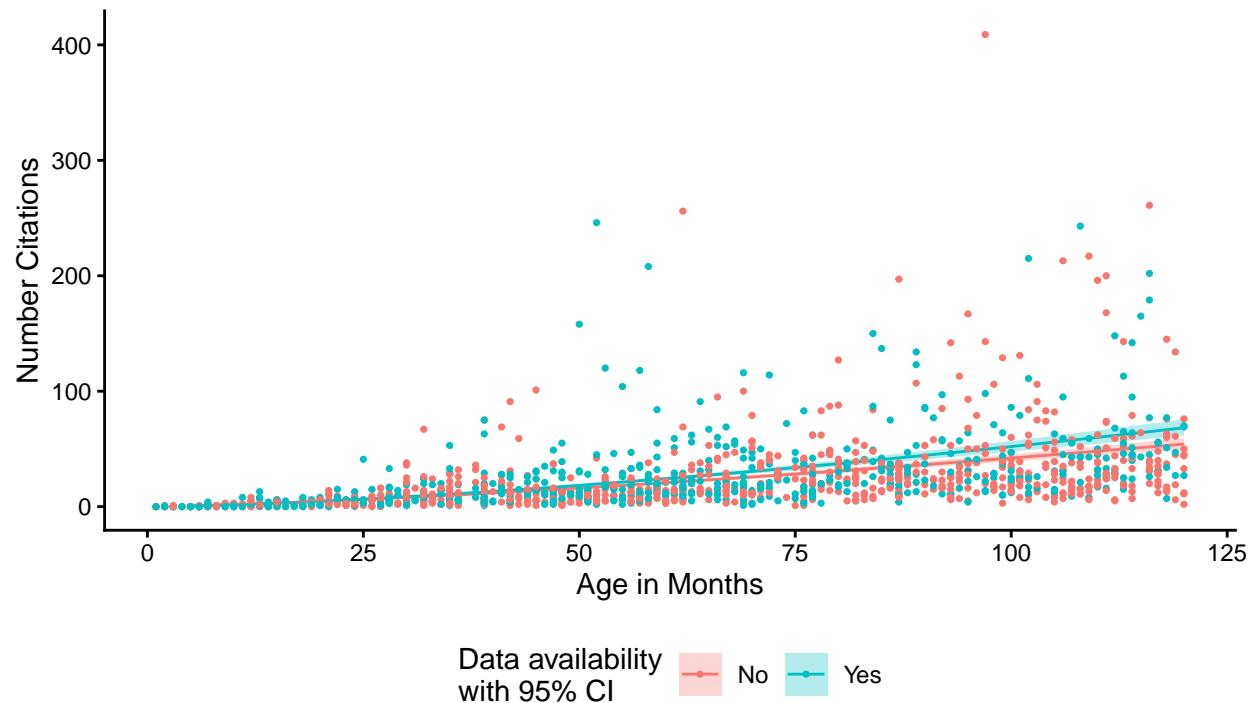
Model vs True Number of Citations from GLM.NB for
Journal of Clinical Microbiology

Data age.in.months <= 120, removal of JMME, GA, MRA



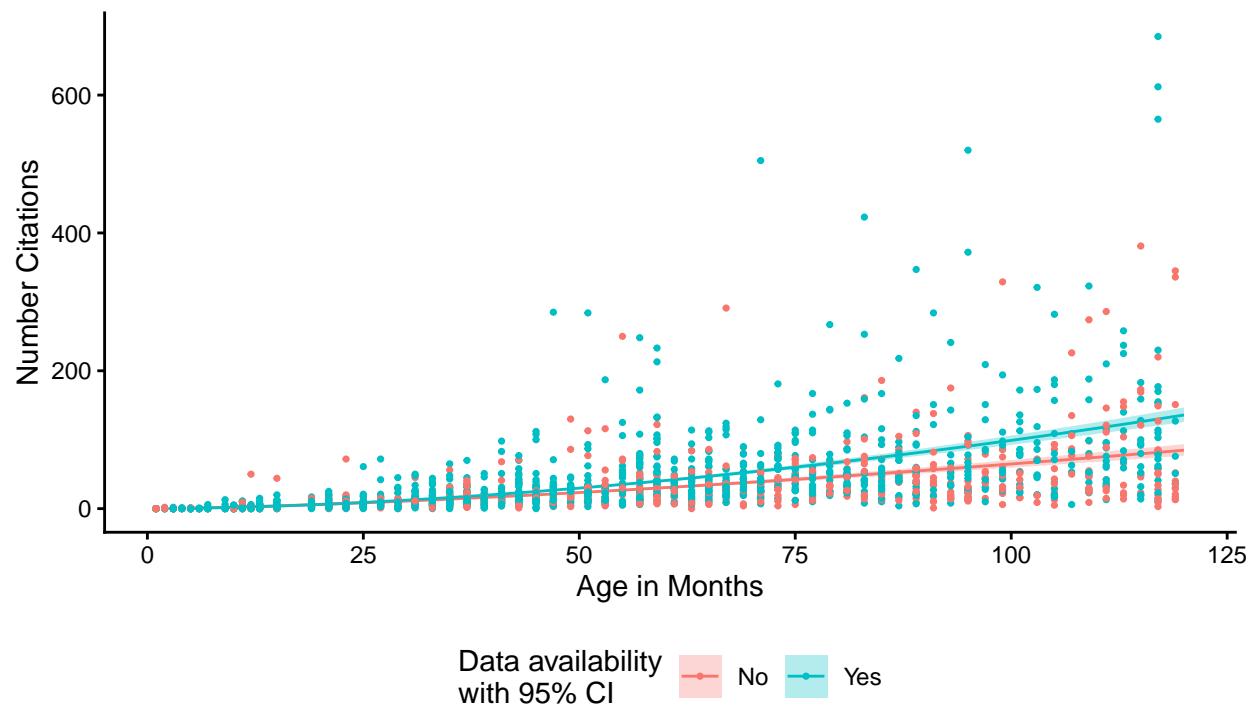
Model vs True Number of Citations from GLM.NB for Journal of Virology

Data age.in.months <= 120, removal of JMME, GA, MRA



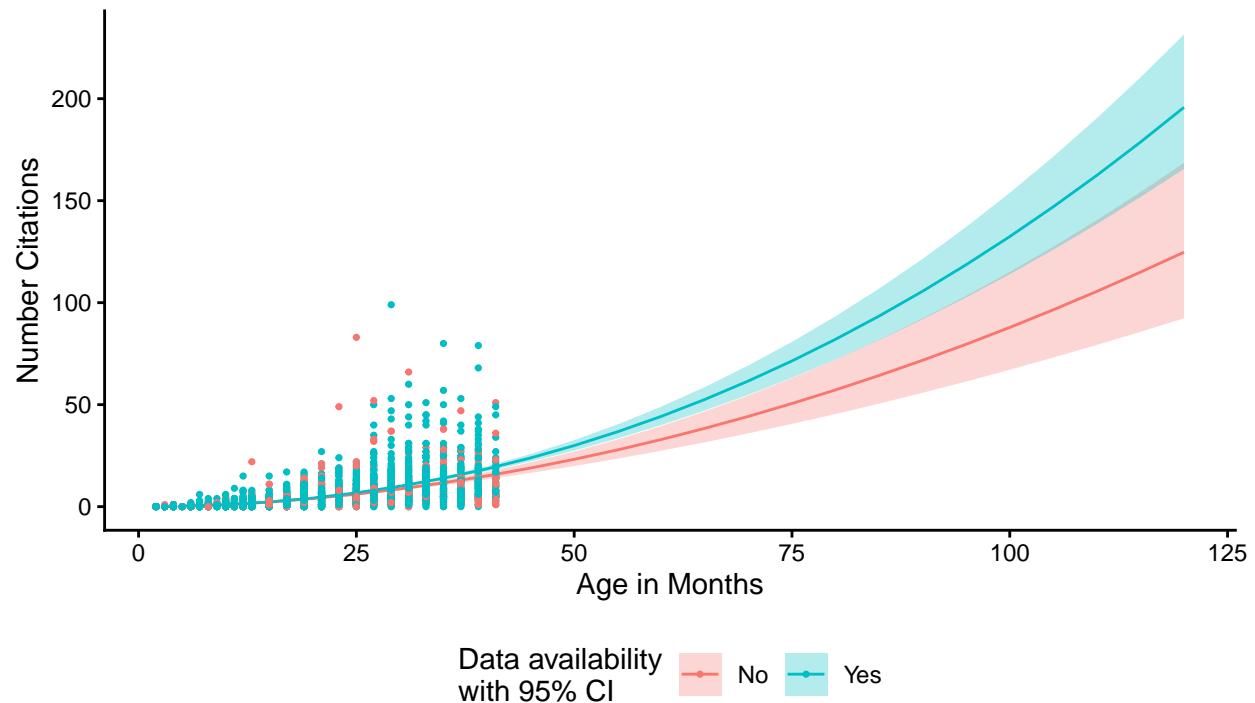
Model vs True Number of Citations from GLM.NB for mBio

Data age.in.months <= 120, removal of JMBE, GA, MRA



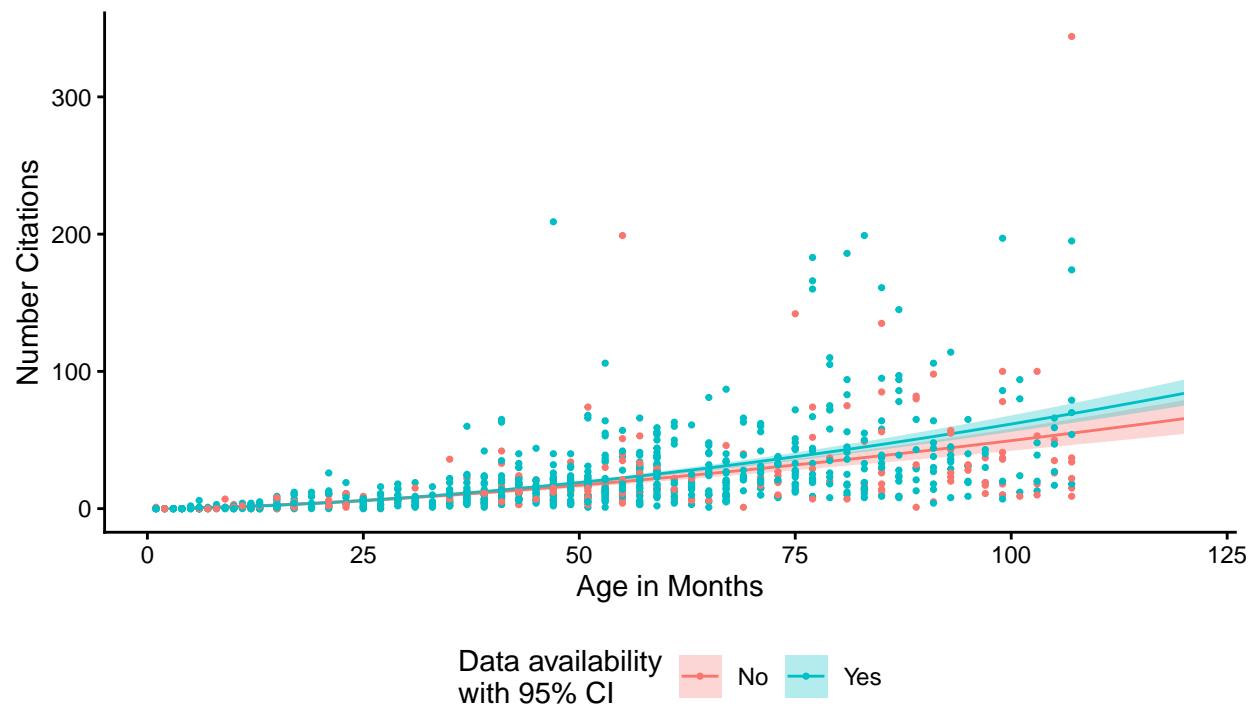
Model vs True Number of Citations from GLM.NB for Microbiology Spectrum

Data age.in.months <= 120, removal of JMBE, GA, MRA



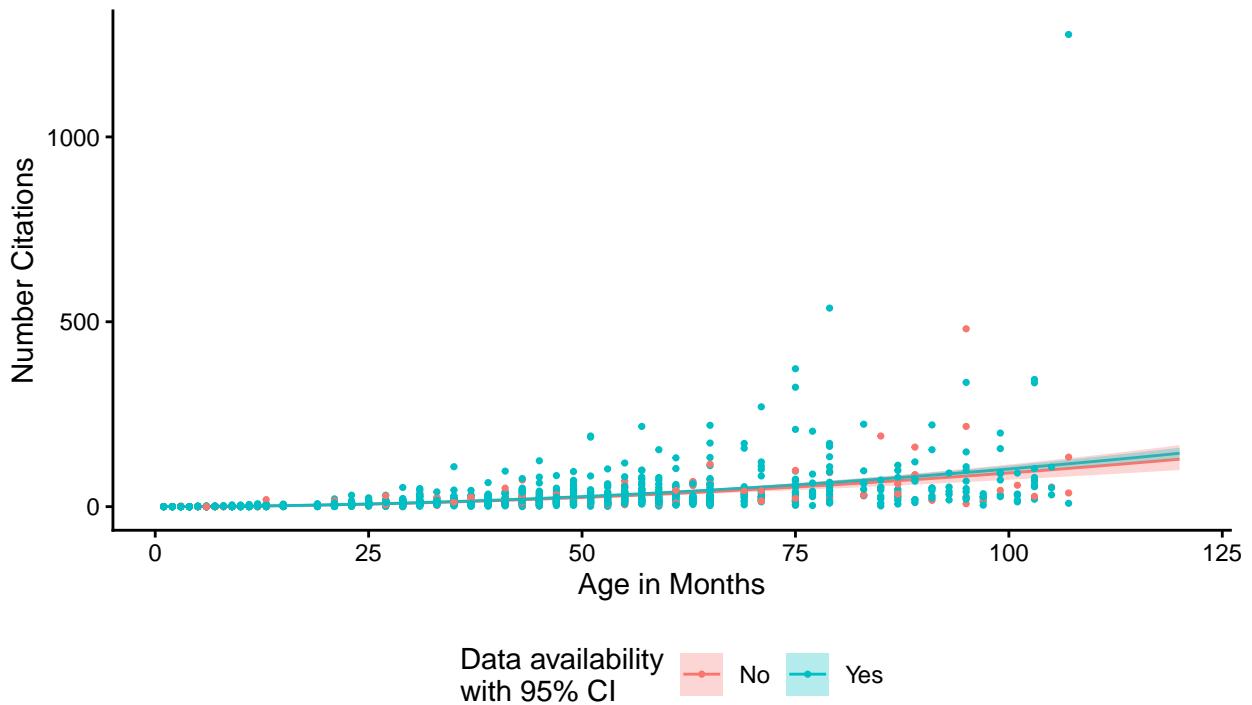
Model vs True Number of Citations from GLM.NB for mSphere

Data age.in.months <= 120, removal of JMBE, GA, MRA



Model vs True Number of Citations from GLM.NB for mSystems

Data age.in.months <= 120, removal of JMME, GA, MRA



Create a new model for sequence data vs no new sequence data

```
#setup dataset and model

#filter dataset for no nas, filter out jmbe, mra, ga, age in months <= 120
nsd_model_metadata <-
  metadata %>%
  filter(., age.in.months != "NA" & nsd != "NA" & container.title != "NA") %>%
  filter(journal_abrev != "jmbe" & journal_abrev != "mra" & journal_abrev != "genomea" & age.in.months <= 120)
  mutate(nsd_factor = factor(nsd),
        container.title = factor(container.title))

nsd_model <-
  glm.nb(is.referenced.by.count ~ nsd_factor + log(age.in.months) + container.title +
  + container.title*nsd_factor + log(age.in.months)*nsd_factor + container.title*log(age.in.months) +
  log(age.in.months)*nsd_factor*container.title, data = nsd_model_metadata, link = log)

# make plots for each journal

p_nsd <- get_model_data(model = nsd_model, type = "pred",
                         terms = c("nsd_factor", "age.in.months[age_values]", "container.title"),
                         colors = "bw") %>%
```

```

tibble(nsd_factor = ifelse(.\$x == 1, "Contains New Seq Data", "No New Seq Data"), predicted_citation
age.in.months = .\$group, container.title = .\$facet)

predicted_plot_nsd <-
  ggplot(data = p_nsd, mapping = aes(x = as.numeric(age.in.months), y = predicted_citations,
                                      color = nsd_factor)) +
  geom_line(aes(x = age.in.months, y = predicted_citations, group = nsd_factor)) +
  geom_ribbon(mapping = aes(ymin = conf.low, ymax = conf.high,
                            group = nsd_factor, fill = nsd_factor), alpha = 0.2) +
  facet_wrap(~ container.title, nrow = 2,
             labeller = label_wrap_gen(width = 18),
             scale = "free_y") +
  labs(title = "Predicted Number of Citations from GLM.NB",
       subtitle = "NSD Model",
       x = "Age in Months",
       y = "Predicted Number Citations",
       color = "New Seq Data\nwith 95% CI",
       fill = "New Seq Data\nwith 95% CI") +
  scale_x_discrete(breaks = seq(12, 120, 12)) +
  theme_classic() +
  theme(legend.position = "bottom" )

predicted_plot_nsd <-
  ggplot(data = p_nsd, mapping = aes(x = as.numeric(age.in.months), y = predicted_citations,
                                      color = nsd_factor)) +
  geom_line(aes(x = age.in.months, y = predicted_citations, group = nsd_factor)) +
  geom_ribbon(mapping = aes(ymin = conf.low, ymax = conf.high,
                            group = nsd_factor, fill = nsd_factor), alpha = 0.2) +
  facet_wrap(~ container.title, nrow = 2,
             labeller = label_wrap_gen(width = 18),
             ) +
  labs(title = "Predicted Number of Citations from GLM.NB",
       subtitle = "NSD Model",
       x = "Age in Months",
       y = "Predicted Number Citations",
       color = "New Seq Data\nwith 95% CI",
       fill = "New Seq Data\nwith 95% CI") +
  scale_x_discrete(breaks = seq(12, 120, 12)) +
  theme_classic() +
  theme(legend.position = "bottom" )
predicted_plot_nsd

```

Predicted Number of Citations from GLM.NB

NSD Model

