Exploratory Analysis — Urology Trials

Carles Raich Bros

2025-09-09

Contents

Script: exploratory_analysis.Rmd

Script to explore and visualize cleaned ClinicalTrials.gov data

Focus: descriptive statistics and plots for urology trials

---

1. Load data

```
clean_file <- params$clean_file
df <- read_csv(clean_file, show_col_types = FALSE)

cat("Loaded", nrow(df), "trials\n")
```

```
## Loaded 7994 trials
```

```
bn <- basename(clean_file)
m <- stringr::str_match(bn, "^clean_trials_(.+)_(\\d{4}-\\d{2}-\\d{2})\\.csv$")
cond_slug <- m[,2]
date_str  <- m[,3]
condition <- stringr::str_replace_all(cond_slug, "_", " ") |> stringr::str_to_title()
```

2. Basic preparation & split trials by status

```
# Phase labels helper
pretty_phase <- function(x) {
  x <- toupper(trimws(x))
  x[x %in% c("N/A", "NA", "", NA)] <- NA
  x <- gsub("PHASE\\s*([0-4])\\s*[,;:_\\-]+\\s*PHASE\\s*([0-4])", "PHASE\\1/PHASE\\2", x, perl = TRUE)
  x <- gsub("EARLY\\s*PHASE\\s*1|EARLY_PHASE_?1", "Early Phase 1", x)
  x <- gsub("^PHASE\\s*([0-4])$", "Phase \\1", x)
  x <- gsub("^PHASE\\s*([0-4])\\s*/\\s*PHASE\\s*([0-4])$", "Phase \\1/\\2", x)
   x <- tools::toTitleCase(tolower(x))
  x
}
df <- df %>%
  dplyr::mutate(Phase = pretty_phase(Phase))

# Intervention helper
extract_intervention_type <- function(x) {
  t <- sub("^\\s*([^;]+);.*$", "\\1", x)
  t[is.na(x) | !nzchar(trimws(x))] <- NA_character_
  t
}

pretty_label <- function(x) {
  ifelse(
    is.na(x),
    NA_character_,
    tools::toTitleCase(gsub("_", " ", tolower(trimws(x))))
  )
}

df <- df %>%
  dplyr::mutate(
    SimpleIntervention = pretty_label(extract_intervention_type(Interventions))
  )

# Status classification
status_finished <- c("COMPLETED","APPROVED_FOR_MARKETING")
status_inprog   <- c("RECRUITING","ACTIVE_NOT_RECRUITING","ENROLLING_BY_INVITATION","NOT_YET_RECRUITING
status_failed   <- c("TERMINATED","WITHDRAWN","SUSPENDED",
                     "TEMPORARILY_NOT_AVAILABLE","NO_LONGER_AVAILABLE")
# Data frames by status
df_finished <- df %>% filter(Status %in% status_finished)
df_inprogress <- df %>% filter(Status %in% status_inprog)
df_failed <- df %>% filter(Status %in% status_failed)

# Numbers
```

```r
n_total  <- nrow(df)
n_finished <- nrow(df_finished)
n_inprogress <- nrow(df_inprogress)
n_failed <- nrow(df_failed)
```

---

3. Discontinued / failed trials

```r
# Opening sentence
pct_failed <- if (n_total > 0) round(100 * n_failed / n_total, 1) else NA_real_
msg_failed <- sprintf(
  "%d %s clinical trial%s were discontinued (terminated/withdrawn/suspended) as of %s (%s%% of downloade
  n_failed,
  condition,
  ifelse(n_failed == 1, "", "s"),
  date_str,
  ifelse(is.na(pct_failed), "NA", pct_failed)
)
cat(msg_failed, "\n")
```

```
## 1089 Prostate Cancer clinical trials were discontinued (terminated/withdrawn/suspended) as of 2025-08
```

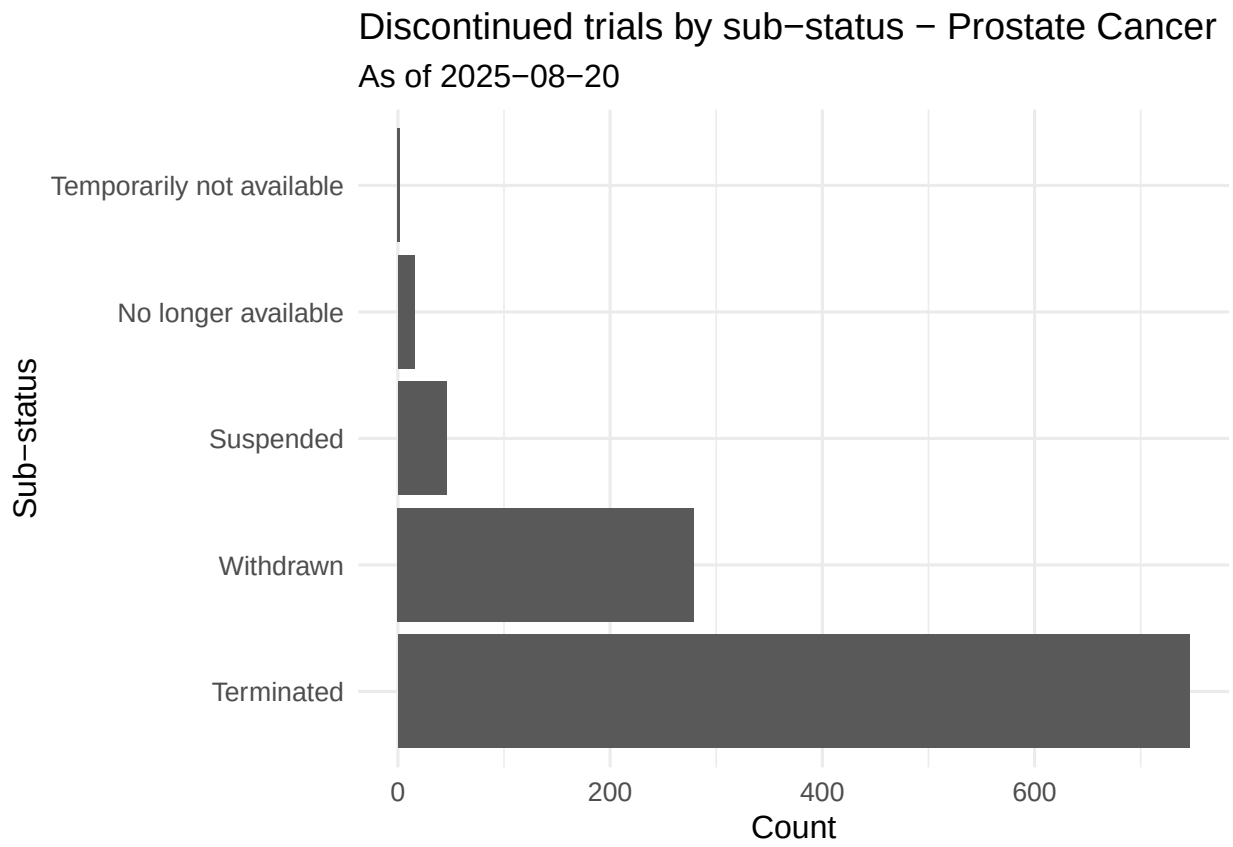
```r
# Breakdown by sub-status
failed_breakdown <- df_failed %>%
  dplyr::filter(!is.na(Status)) %>%
  dplyr::mutate(
    Status = gsub("_", " ", tolower(Status)),
    Status = paste0(toupper(substr(Status, 1, 1)),
                    substr(Status, 2, nchar(Status)))
  ) %>%
  dplyr::count(Status, sort = TRUE)

print(failed_breakdown)
```

```
## # A tibble: 5 x 2
##   Status                      n
##   <chr>                   <int>
## 1 Terminated                746
## 2 Withdrawn                 279
## 3 Suspended                  46
## 4 No longer available        16
## 5 Temporarily not available   2
```

```r
# Bar chart by sub-status
failed_breakdown$Status <- factor(failed_breakdown$Status,
                                  levels = failed_breakdown$Status)

ggplot(failed_breakdown, aes(x = Status, y = n)) +
  geom_col() +
  coord_flip() +
  labs(
    title = paste0("Discontinued trials by sub-status - ", condition),
    subtitle = paste("As of", date_str),
    x = "Sub-status", y = "Count"
  )
```



---

4. Trials in progress

```r
# Opening sentence
```

```
pct_inprog <- if (n_total > 0) round(100 * n_inprogress / n_total, 1) else NA_real_
msg_inprog <- sprintf(
  "%d %s clinical trial%s are in progress as of %s (%s%% of downloaded trials).",
  n_inprogress,
  condition,
  ifelse(n_inprogress == 1, "", "s"),
  date_str,
  ifelse(is.na(pct_inprog), "NA", pct_inprog)
)
cat(msg_inprog, "\n")
```

```
## 2431 Prostate Cancer clinical trials are in progress as of 2025-08-20 (30.4% of downloaded trials).
```

```
# Phase

n_phase_known_inprog <- sum(!is.na(df_inprogress$Phase))
pct_phase_known_inprog <- if (n_inprogress > 0) round(100 * n_phase_known_inprog / n_inprogress, 1) else

msg_inprog_phase <- sprintf(
  "Phase information available for %d of %d in-progress trials (%s%%).",
  n_phase_known_inprog, n_inprogress,
  ifelse(is.na(pct_phase_known_inprog), "NA", pct_phase_known_inprog)
)
cat(msg_inprog_phase, "\n")
```

```
## Phase information available for 1203 of 2431 in-progress trials (49.5%).
```

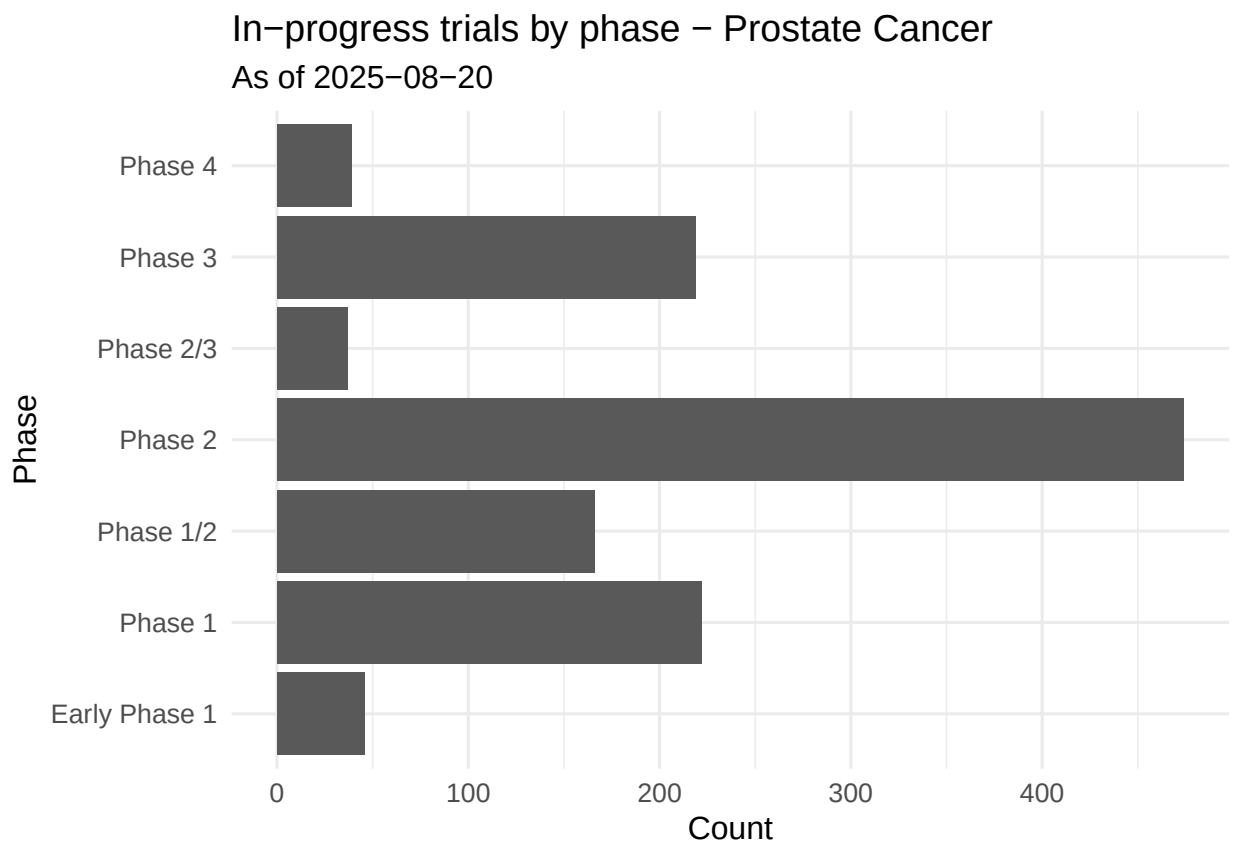
```
inprogress_phase_breakdown <- df_inprogress %>%
  dplyr::filter(!is.na(Phase)) %>%
  dplyr::count(Phase, sort = TRUE)

print(inprogress_phase_breakdown)
```

```
## # A tibble: 7 x 2
##   Phase             n
##   <chr>         <int>
## 1 Phase 2         474
## 2 Phase 1         222
## 3 Phase 3         219
## 4 Phase 1/2       166
## 5 Early Phase 1    46
## 6 Phase 4          39
## 7 Phase 2/3        37
```

```
order_phase <- c("Early Phase 1","Phase 1","Phase 1/2","Phase 2","Phase 2/3","Phase 3","Phase 4")
inprogress_phase_breakdown$Phase <- factor(inprogress_phase_breakdown$Phase, levels = order_phase)

ggplot(inprogress_phase_breakdown, aes(x = Phase, y = n)) +
  geom_col() + coord_flip() +
  labs(title = paste0("In-progress trials by phase - ", condition),
       subtitle = paste("As of", date_str),
       x = "Phase", y = "Count")
```




```
# Intervention

n_interv_known_inprogress <- sum(!is.na(df_inprogress$Interventions) & nzchar(trimws(df_inprogress$Inte
pct_interv_known_inprogress <- if (n_inprogress > 0) round(100 * n_interv_known_inprogress / n_inprogres

msg_inprogress_interv <- sprintf(
  "Intervention information available for %d of %d in-progress trials (%s%%).",
  n_interv_known_inprogress, n_inprogress,
  ifelse(is.na(pct_interv_known_inprogress), "NA", pct_interv_known_inprogress)
)
cat(msg_inprogress_interv, "\n")
```

```
## Intervention information available for 2256 of 2431 in-progress trials (92.8%).
```

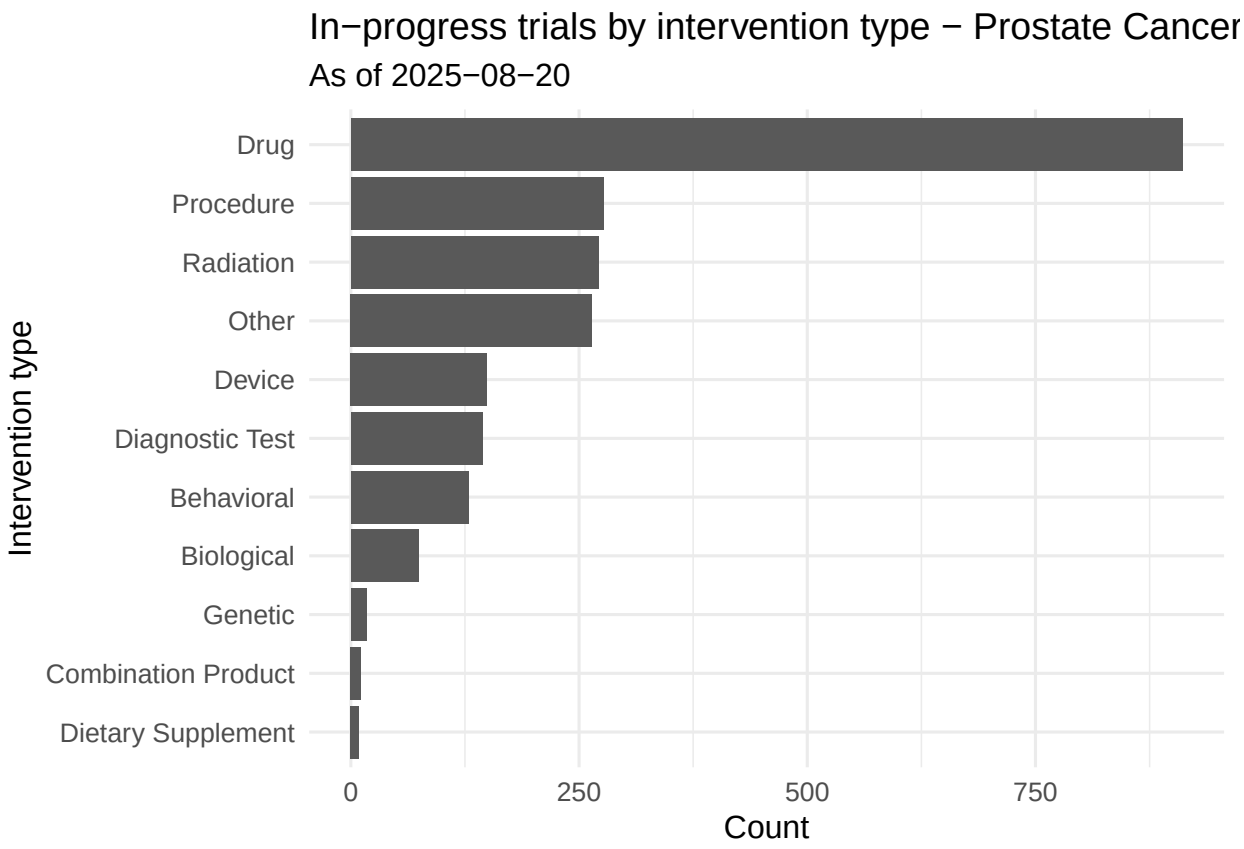
```
inprogress_interv_breakdown <- df_inprogress %>%
  dplyr::filter(!is.na(SimpleIntervention) & nzchar(SimpleIntervention)) %>%
  dplyr::count(SimpleIntervention, sort = TRUE)

print(inprogress_interv_breakdown)
```

```
## # A tibble: 11 x 2
##    SimpleIntervention      n
##    <chr>               <int>
##  1 Drug                  911
##  2 Procedure             277
```

```
##  3 Radiation             271
##  4 Other                 264
##  5 Device                149
##  6 Diagnostic Test       144
##  7 Behavioral            129
##  8 Biological             74
##  9 Genetic                17
## 10 Combination Product    11
## 11 Dietary Supplement      9
```

```r
ggplot(inprogress_interv_breakdown,
       aes(x = forcats::fct_reorder(SimpleIntervention, n), y = n)) +
  geom_col() +
  coord_flip() +
  labs(
    title = paste0("In-progress trials by intervention type - ", condition),
    subtitle = paste("As of", date_str),
    x = "Intervention type", y = "Count"
  )
```




```r
# Sponsors & Collaborators

n_sponsor_known_inprogress <- sum(!is.na(df_inprogress$Sponsor) & nzchar(trimws(df_inprogress$Sponsor)))
pct_sponsor_known_inprogress <- if (n_inprogress > 0) round(100 * n_sponsor_known_inprogress / n_inprogr
```

```
n_collab_known_inprogress <- sum(!is.na(df_inprogress$Collaborators) & nzchar(trimws(df_inprogress$Colla
pct_collab_known_inprogress <- if (n_inprogress > 0) round(100 * n_collab_known_inprogress / n_inprogres

msg_inprogress_sponsor <- sprintf(
  "Sponsor information available for %d of %d in-progress trials (%s%%).",
  n_sponsor_known_inprogress, n_inprogress,
  ifelse(is.na(pct_sponsor_known_inprogress), "NA", pct_sponsor_known_inprogress)
)
cat(msg_inprogress_sponsor, "\n")
```

```
## Sponsor information available for 2431 of 2431 in-progress trials (100%).
```

```
msg_inprogress_collab <- sprintf(
  "Collaborator information available for %d of %d in-progress trials (%s%%).",
  n_collab_known_inprogress, n_inprogress,
  ifelse(is.na(pct_collab_known_inprogress), "NA", pct_collab_known_inprogress)
)
cat(msg_inprogress_collab, "\n")
```

```
## Collaborator information available for 1003 of 2431 in-progress trials (41.3%).
```

```
inprogress_sponsor_top <- df_inprogress %>%
  dplyr::filter(!is.na(Sponsor) & nzchar(trimws(Sponsor))) %>%
  dplyr::transmute(Sponsor = trimws(Sponsor)) %>%
  tidyr::separate_rows(Sponsor, sep = ";") %>%
  dplyr::mutate(Sponsor = trimws(Sponsor)) %>%
  dplyr::filter(nzchar(Sponsor)) %>%
  dplyr::count(Sponsor, sort = TRUE) %>%
  dplyr::mutate(pct = if (n_inprogress > 0) round(100 * n / n_inprogress, 1) else NA_real_,
                pct = ifelse(is.na(pct), "NA", sprintf("%.1f%%", pct))) %>%
  dplyr::slice_head(n = 10)

cat("Top 10 sponsors (share of in-progress trials)\n")
```

```
## Top 10 sponsors (share of in-progress trials)
```

```
print(inprogress_sponsor_top)
```

```
## # A tibble: 10 x 3
##    Sponsor                                                    n pct
##    <chr>                                                  <int> <chr>
##  1 National Cancer Institute (NCI)                           85 3.5%
##  2 Memorial Sloan Kettering Cancer Center                    51 2.1%
##  3 M.D. Anderson Cancer Center                               49 2.0%
##  4 Mayo Clinic                                               43 1.8%
##  5 University Health Network, Toronto                        33 1.4%
##  6 Jonsson Comprehensive Cancer Center                       31 1.3%
##  7 Novartis Pharmaceuticals                                  24 1.0%
##  8 Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins    23 0.9%
##  9 AstraZeneca                                               21 0.9%
## 10 Chinese University of Hong Kong                           19 0.8%
```

```r
inprogress_collaborators_top <- df_inprogress %>%
  dplyr::filter(!is.na(Collaborators) & nzchar(trimws(Collaborators))) %>%
  dplyr::transmute(Collaborators = trimws(Collaborators)) %>%
  tidyr::separate_rows(Collaborators, sep = ";") %>%
  dplyr::mutate(Collaborator = trimws(Collaborators)) %>%
  dplyr::filter(nzchar(Collaborator)) %>%
  # Exclou tokens TOT en majúscules excepte "NIH"
  dplyr::filter(!(Collaborator != "NIH" & stringr::str_detect(Collaborator, "^[A-ZO-9 &/\\-]+$"))) %>%
  dplyr::count(Collaborator, sort = TRUE) %>%
  dplyr::mutate(pct = if (n_inprogress > 0) sprintf("%.1f%%", 100 * n / n_inprogress) else "NA") %>%
  dplyr::slice_head(n = 10)

cat("Top 10 collaborators (share of in-progress trials)\n")
```

```
## Top 10 collaborators (share of in-progress trials)
```

```r
print(inprogress_collaborators_top)
```

```
## # A tibble: 10 x 3
##    Collaborator                            n pct
##    <chr>                               <int> <chr>
##  1 NIH                                   248 10.2%
##  2 National Cancer Institute (NCI)       234 9.6%
##  3 OTHER_GOV                              61 2.5%
##  4 Bayer                                  36 1.5%
##  5 Merck Sharp & Dohme LLC                27 1.1%
##  6 Prostate Cancer Foundation             24 1.0%
##  7 United States Department of Defense    20 0.8%
##  8 Astellas Pharma Inc                    18 0.7%
##  9 Bristol-Myers Squibb                   18 0.7%
## 10 Pfizer                                 17 0.7%
```

---

5. Trials completed

```r
# Opening sentence

pct_finished <- if (n_total > 0) round(100 * n_finished / n_total, 1) else NA_real_
msg_finished <- sprintf(
  "%d %s clinical trial%s have been completed as of %s (%s%% of downloaded trials).",
  n_finished,
  condition,
  ifelse(n_finished == 1, "", "s"),
  date_str,
  ifelse(is.na(pct_finished), "NA", pct_finished)
```

```
)
cat(msg_finished, "\n")
```

```
## 3596 Prostate Cancer clinical trials have been completed as of 2025-08-20 (45% of downloaded trials)
```

```
# Phase

n_phase_known_finished <- sum(!is.na(df_finished$Phase))
pct_phase_known_finished <- if (n_finished > 0) round(100 * n_phase_known_finished / n_finished, 1) else

msg_finished_phase <- sprintf(
  "Phase information available for %d of %d completed trials (%s%%).",
  n_phase_known_finished, n_finished,
  ifelse(is.na(pct_phase_known_finished), "NA", pct_phase_known_finished)
)
cat(msg_finished_phase, "\n")
```

```
## Phase information available for 2050 of 3596 completed trials (57%).
```

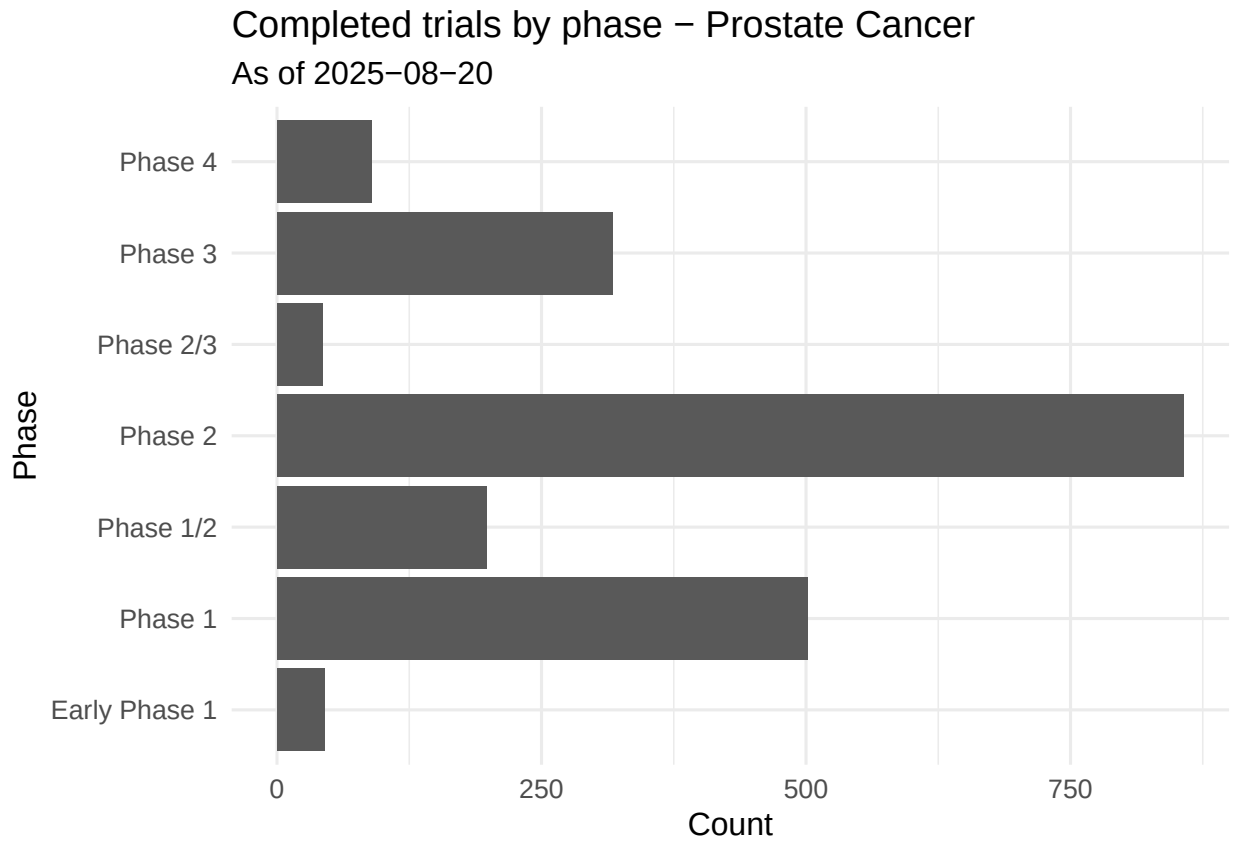
```
finished_phase_breakdown <- df_finished %>%
  dplyr::filter(!is.na(Phase)) %>%
  dplyr::count(Phase, sort = TRUE)

print(finished_phase_breakdown)
```

```
## # A tibble: 7 x 2
##   Phase             n
##   <chr>         <int>
## 1 Phase 2         857
## 2 Phase 1         501
## 3 Phase 3         317
## 4 Phase 1/2       198
## 5 Phase 4          89
## 6 Early Phase 1    45
## 7 Phase 2/3        43
```

```
order_phase <- c("Early Phase 1","Phase 1","Phase 1/2","Phase 2","Phase 2/3","Phase 3","Phase 4")
finished_phase_breakdown$Phase <- factor(finished_phase_breakdown$Phase, levels = order_phase)

ggplot(finished_phase_breakdown, aes(x = Phase, y = n)) +
  geom_col() + coord_flip() +
  labs(title = paste0("Completed trials by phase - ", condition),
       subtitle = paste("As of", date_str),
       x = "Phase", y = "Count")
```




```r
# Duration

finished_dates <- df_finished %>%
  transmute(
    Start = lubridate::ymd(StartDate, quiet = TRUE),
    PComp = lubridate::ymd(PrimaryCompletionDate, quiet = TRUE),
    Comp  = lubridate::ymd(CompletionDate, quiet = TRUE)
  ) %>%
  mutate(
    End = coalesce(Comp, PComp),
    DurationMonths = if_else(
      !is.na(Start) & !is.na(End) & End >= Start,
      time_length(interval(Start, End), "months"),
      as.numeric(NA)
    )
  )

n_start_known_finished <- sum(!is.na(finished_dates$Start))
pct_start_known_finished <- if (n_finished > 0) round(100 * n_start_known_finished / n_finished, 1) else

n_comp_known_finished <- sum(!is.na(finished_dates$Comp))
pct_comp_known_finished <- if (n_finished > 0) round(100 * n_comp_known_finished / n_finished, 1) else

n_pcomp_known_finished <- sum(!is.na(finished_dates$PComp))
pct_pcomp_known_finished <- if (n_finished > 0) round(100 * n_pcomp_known_finished / n_finished, 1) else
```

```r
cat(sprintf("Start date available for %d of %d completed trials (%s%%).\n",
            n_start_known_finished, n_finished,
            ifelse(is.na(pct_start_known_finished), "NA", pct_start_known_finished)))
```

```
## Start date available for 1609 of 3596 completed trials (44.7%).
```

```r
cat(sprintf("Completion date available for %d of %d completed trials (%s%%).\n",
            n_comp_known_finished, n_finished,
            ifelse(is.na(pct_comp_known_finished), "NA", pct_comp_known_finished)))
```

```
## Completion date available for 1835 of 3596 completed trials (51%).
```

```r
cat(sprintf("Primary completion date available for %d of %d completed trials (%s%%).\n",
            n_pcomp_known_finished, n_finished,
            ifelse(is.na(pct_pcomp_known_finished), "NA", pct_pcomp_known_finished)))
```

```
## Primary completion date available for 1759 of 3596 completed trials (48.9%).
```

```r
n_duration_finished <- sum(!is.na(finished_dates$DurationMonths))
pct_duration_finished <- if (n_finished > 0) round(100 * n_duration_finished / n_finished, 1) else NA_r
mean_duration_m   <- if (n_duration_finished > 0) round(mean(finished_dates$DurationMonths, na.rm = TRU
median_duration_m <- if (n_duration_finished > 0) round(stats::median(finished_dates$DurationMonths, na

cat(sprintf("Study duration computable for %d of %d completed trials (%s%%). Mean: %s months (median: %
            n_duration_finished, n_finished,
            ifelse(is.na(pct_duration_finished), "NA", pct_duration_finished),
            ifelse(is.na(mean_duration_m), "NA", mean_duration_m),
            ifelse(is.na(median_duration_m), "NA", median_duration_m)))
```

```
## Study duration computable for 1588 of 3596 completed trials (44.2%). Mean: 49.3 months (median: 38.9
```

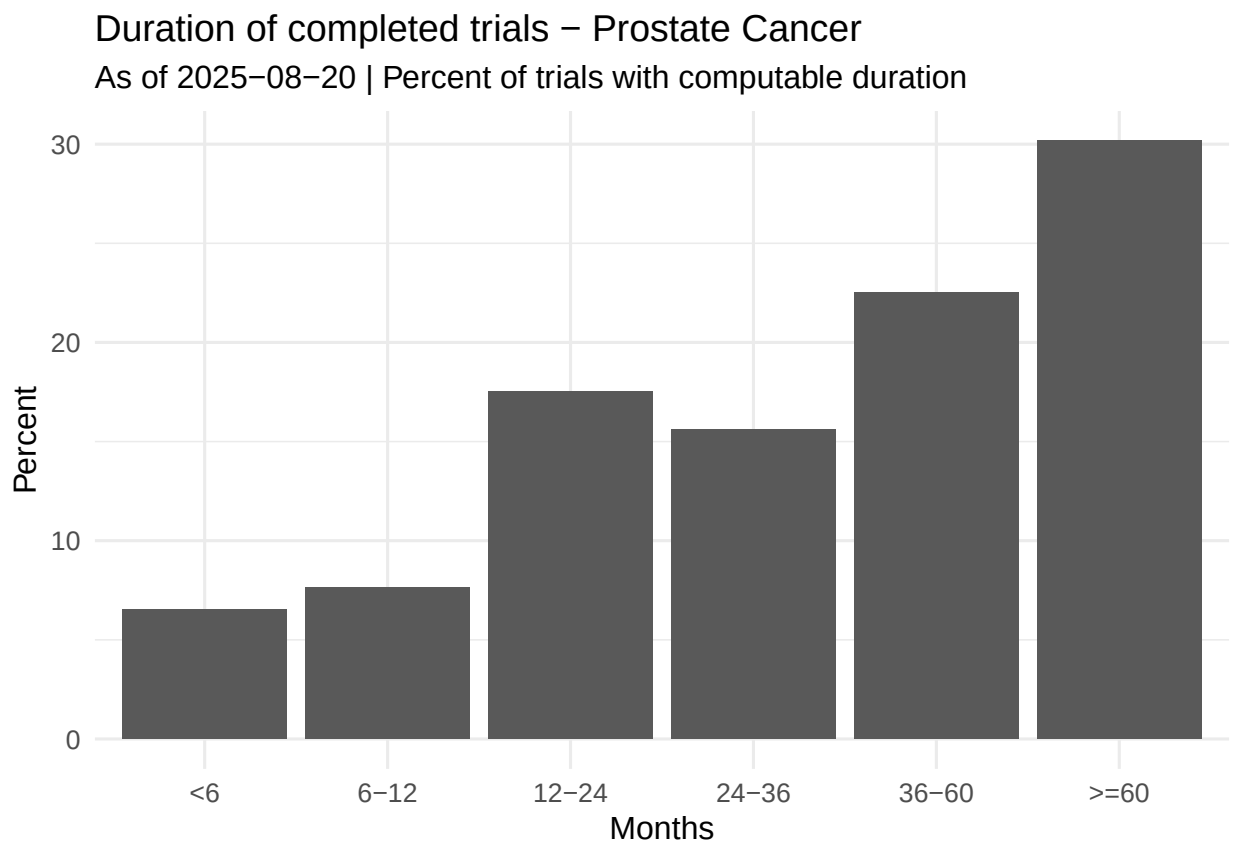
```r
duration_breakdown <- finished_dates %>%
  dplyr::filter(!is.na(DurationMonths), DurationMonths >= 0) %>%
  dplyr::mutate(
    Bucket = cut(DurationMonths,
                 breaks = c(-Inf, 6, 12, 24, 36, 60, Inf),
                 labels = c("<6", "6-12", "12-24", "24-36", "36-60", " 60"),
                 right = FALSE)
  ) %>%
  dplyr::count(Bucket, .drop = FALSE) %>%
  dplyr::mutate(
    pct_num = if (n_duration_finished > 0) 100 * n / n_duration_finished else NA_real_,
    pct = ifelse(is.na(pct_num), "NA", sprintf("%.1f%%", pct_num))
  )

print(duration_breakdown)
```

```
## # A tibble: 6 x 4
##   Bucket     n pct_num pct
##   <fct>  <int>   <dbl> <chr>
```

```
## 1 <6        104    6.55 6.5%
## 2 6-12      121    7.62 7.6%
## 3 12-24     278   17.5  17.5%
## 4 24-36     248   15.6  15.6%
## 5 36-60     358   22.5  22.5%
## 6  60       479   30.2  30.2%
```

```r
ggplot(duration_breakdown, aes(x = Bucket, y = pct_num)) +
  geom_col() +
  labs(
    title = paste0("Duration of completed trials - ", condition),
    subtitle = paste("As of", date_str, "| Percent of trials with computable duration"),
    x = "Months", y = "Percent"
  )
```




```r
# Enrollment

n_enroll_known_finished <- sum(!is.na(df_finished$Enrollment))
cat(sprintf("Enrollment available for %d of %d completed trials.\n",
            n_enroll_known_finished, n_finished))
```

```
## Enrollment available for 3503 of 3596 completed trials.
```

```r
if (n_enroll_known_finished > 0) {
  enroll_vals <- df_finished$Enrollment[!is.na(df_finished$Enrollment)]
  cat(sprintf("Enrollment range: %d-%d; mean: %.1f participants.\n",
              min(enroll_vals), max(enroll_vals), mean(enroll_vals)))
}
```

```
## Enrollment range: 1-5940299; mean: 3660.2 participants.
```

```r
# Sex (if proceeds)
skip_sex <- grepl("(prostate|prostatic|testicular|testis|testicle)", tolower(condition))
if (!skip_sex) {
  sex_tbl <- df_finished %>%
    dplyr::transmute(SexRaw = toupper(trimws(Sex))) %>%
    dplyr::mutate(
      SexCat = dplyr::case_when(
        is.na(SexRaw) | SexRaw %in% c("", "N/A", "NA") ~ NA_character_,
        grepl("ALL", SexRaw) ~ "All",
        grepl("FEMALE", SexRaw) & !grepl("MALE", SexRaw) ~ "Female only",
        grepl("MALE", SexRaw)   & !grepl("FEMALE", SexRaw) ~ "Male only",
        TRUE ~ "Other/Complex"
      )
    )

  n_known_sex <- sum(!is.na(sex_tbl$SexCat))
  pct_known_sex <- if (n_finished > 0) round(100 * n_known_sex / n_finished, 1) else NA_real_

  cat(sprintf("Sex eligibility available for %d of %d completed trials (%s%%).\n",
              n_known_sex, n_finished,
              ifelse(is.na(pct_known_sex), "NA", pct_known_sex)))

  sex_levels <- c("All","Male only","Female only","Other/Complex")

  sex_breakdown <- sex_tbl %>%
    dplyr::filter(!is.na(SexCat)) %>%
    dplyr::mutate(SexCat = factor(SexCat, levels = sex_levels)) %>%
    dplyr::count(SexCat, .drop = FALSE) %>%
    dplyr::mutate(pct = if (n_known_sex > 0) round(100 * n / n_known_sex, 1) else NA_real_)

  print(sex_breakdown)

  ggplot(sex_breakdown, aes(x = SexCat, y = n)) +
    geom_col() +
    labs(
      title = paste0("Completed trials by sex eligibility - ", condition),
      subtitle = paste("As of", date_str),
      x = "Sex eligibility", y = "Count"
    )
} else {
  cat("Sex eligibility breakdown skipped (condition is prostatic/testicular).\n")
}
```

```
## Sex eligibility breakdown skipped (condition is prostatic/testicular).
```

```r
# Intervention

n_interv_known_finished <- sum(!is.na(df_finished$Interventions) & nzchar(trimws(df_finished$Interventi
pct_interv_known_finished <- if (n_finished > 0) round(100 * n_interv_known_finished / n_finished, 1) el

msg_finished_interv <- sprintf(
  "Intervention information available for %d of %d completed trials (%s%%).",
  n_interv_known_finished, n_finished,
  ifelse(is.na(pct_interv_known_finished), "NA", pct_interv_known_finished)
)
cat(msg_finished_interv, "\n")
```

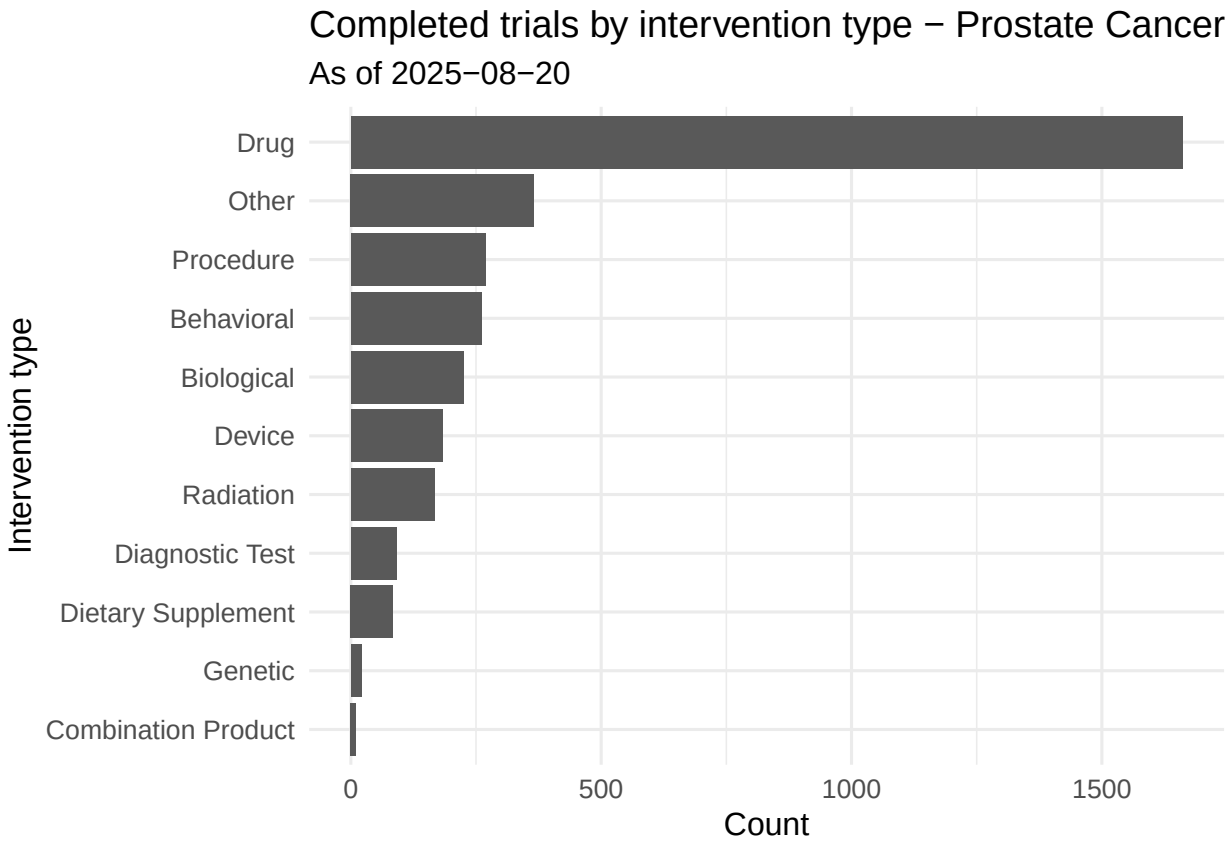
```
## Intervention information available for 3338 of 3596 completed trials (92.8%).
```

```r
finished_interv_breakdown <- df_finished %>%
  dplyr::filter(!is.na(SimpleIntervention) & nzchar(SimpleIntervention)) %>%
  dplyr::count(SimpleIntervention, sort = TRUE)

print(finished_interv_breakdown)
```

```
## # A tibble: 11 x 2
##    SimpleIntervention      n
##    <chr>               <int>
##  1 Drug                 1661
##  2 Other                 366
##  3 Procedure             269
##  4 Behavioral            261
##  5 Biological            225
##  6 Device                183
##  7 Radiation             167
##  8 Diagnostic Test        91
##  9 Dietary Supplement     84
## 10 Genetic                21
## 11 Combination Product    10
```

```r
ggplot(finished_interv_breakdown,
       aes(x = forcats::fct_reorder(SimpleIntervention, n), y = n)) +
  geom_col() +
  coord_flip() +
  labs(
    title = paste0("Completed trials by intervention type - ", condition),
    subtitle = paste("As of", date_str),
    x = "Intervention type", y = "Count"
  )
```




```
# Sponsors & Collaborators

n_sponsor_known_finished <- sum(!is.na(df_finished$Sponsor) & nzchar(trimws(df_finished$Sponsor)))
pct_sponsor_known_finished <- if (n_finished > 0) round(100 * n_sponsor_known_finished / n_finished, 1)

n_collab_known_finished <- sum(!is.na(df_finished$Collaborators) & nzchar(trimws(df_finished$Collaborat
pct_collab_known_finished <- if (n_finished > 0) round(100 * n_collab_known_finished / n_finished, 1) el

msg_finished_sponsor <- sprintf(
  "Sponsor information available for %d of %d completed trials (%s%%).",
  n_sponsor_known_finished, n_finished,
  ifelse(is.na(pct_sponsor_known_finished), "NA", pct_sponsor_known_finished)
)
cat(msg_finished_sponsor, "\n")
```

```
## Sponsor information available for 3596 of 3596 completed trials (100%).
```

```
msg_finished_collab <- sprintf(
  "Collaborator information available for %d of %d completed trials (%s%%).",
  n_collab_known_finished, n_finished,
  ifelse(is.na(pct_collab_known_finished), "NA", pct_collab_known_finished)
)
cat(msg_finished_collab, "\n")
```

```
## Collaborator information available for 1441 of 3596 completed trials (40.1%).
```

```r
finished_sponsor_top <- df_finished %>%
  dplyr::filter(!is.na(Sponsor) & nzchar(trimws(Sponsor))) %>%
  dplyr::transmute(Sponsor = trimws(Sponsor)) %>%
  tidyr::separate_rows(Sponsor, sep = ";") %>%
  dplyr::mutate(Sponsor = trimws(Sponsor)) %>%
  dplyr::filter(nzchar(Sponsor)) %>%
  dplyr::count(Sponsor, sort = TRUE) %>%
  dplyr::mutate(pct = if (n_finished > 0) sprintf("%.1f%%", 100 * n / n_finished) else "NA") %>%
  dplyr::slice_head(n = 10)

cat("Top 10 sponsors (share of completed trials)\n")
```

```
## Top 10 sponsors (share of completed trials)
```

```r
print(finished_sponsor_top)
```

```
## # A tibble: 10 x 3
##    Sponsor                                                    n pct
##    <chr>                                                  <int> <chr>
##  1 National Cancer Institute (NCI)                          147 4.1%
##  2 Memorial Sloan Kettering Cancer Center                    88 2.4%
##  3 University Health Network, Toronto                        63 1.8%
##  4 Bayer                                                     59 1.6%
##  5 AstraZeneca                                               56 1.6%
##  6 M.D. Anderson Cancer Center                               49 1.4%
##  7 Dana-Farber Cancer Institute                              42 1.2%
##  8 Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins  42 1.2%
##  9 Ferring Pharmaceuticals                                   38 1.1%
## 10 Duke University                                           34 0.9%
```

```r
finished_collaborators_top <- df_finished %>%
  dplyr::filter(!is.na(Collaborators) & nzchar(trimws(Collaborators))) %>%
  dplyr::transmute(Collaborators = trimws(Collaborators)) %>%
  tidyr::separate_rows(Collaborators, sep = ";") %>%
  dplyr::mutate(Collaborator = trimws(Collaborators)) %>%
  dplyr::filter(nzchar(Collaborator)) %>%
  dplyr::filter(!(Collaborator != "NIH" & stringr::str_detect(Collaborator, "^[A-Z0-9 &/\\-]+$"))) %>%
  dplyr::count(Collaborator, sort = TRUE) %>%
  dplyr::mutate(pct = if (n_finished > 0) sprintf("%.1f%%", 100 * n / n_finished) else "NA") %>%
  dplyr::slice_head(n = 10)

cat("Top 10 collaborators (share of completed trials)\n")
```

```
## Top 10 collaborators (share of completed trials)
```

```r
print(finished_collaborators_top)
```

```
## # A tibble: 10 x 3
##    Collaborator                          n pct
##    <chr>                             <int> <chr>
##  1 NIH                                 472 13.1%
```

```
##  2 National Cancer Institute (NCI)       434 12.1%
##  3 OTHER_GOV                              56 1.6%
##  4 United States Department of Defense    33 0.9%
##  5 Sanofi                                 29 0.8%
##  6 National Institutes of Health (NIH)    24 0.7%
##  7 Astellas Pharma Inc                    22 0.6%
##  8 AstraZeneca                            22 0.6%
##  9 Princess Margaret Hospital, Canada     21 0.6%
## 10 Bristol-Myers Squibb                   19 0.5%
```