```
  shrinkage         = 0.01,
  bag.fraction      = 0.8,
  cv.folds          = 5,
  stop.method       = "es.mean",
  n.cores           = parallel::detectCores()
)

formula_abg      <- reformulate(covars_gbm, response = "has_abg")
formula_vbg      <- reformulate(covars_gbm, response = "has_vbg")
```

Chunk propensity-config runtime: 0.01 s

Creating subset_data

```
set.seed(123)
rows_to_keep <- round(nrow(stata_data) * 1) #1 for real run
subset_data <- stata_data[sample(nrow(stata_data), rows_to_keep), ]

subset_data <- subset_data %>%
  filter(encounter_type != 1)

table(subset_data$encounter_type)
```

```

     2      3 
171727 343559 
```

```
dim(subset_data)
```

```
[1] 515286    546
```

Chunk sample-subset-data runtime: 6.18 s

Generating Codebook for the Full Dataset

```r
message("Generating codebook for the dataset...")
```

```
Generating codebook for the dataset...
```

```r
study_codebook <- codebookr::codebook(
  stata_data,
  title = "Full TrinetX",
  subtitle = "Dataset Documentation",
  description = "This dataset contains patient-level records from the TrinetX database.
                 It has been processed and converted from the original Stata file."
)
codebook_file <- file.path(data_dir_name, "codebookr.docx")
print(study_codebook, codebook_file)
message("Codebook saved as 'codebookr.docx' in the data directory.")
```

```
Codebook saved as 'codebookr.docx' in the data directory.
```

Chunk codebook-export-full runtime: 97.13 s

New Variable - Death at 60 days

```r
subset_data <- subset_data %>%
  mutate(
    ## 1. Did the patient die?
    died = if_else(!is.na(death_date), 1L, 0L),

    ## 2. Absolute death date (if death_date is an offset)
    death_abs = if_else(!is.na(death_date),
                        encounter_date + death_date,
                        as.Date(NA)),

    ## 3. Year month (YM) for encounter and death
    enc_ym   = floor_date(encounter_date, unit = "month"),
    death_ym = floor_date(death_abs      , unit = "month"),
    ## 4. Reference censoring date: 1 Jun 2024
    ref_ym = ymd("2024-06-01"),

    ## 5. Months from encounter to death or censoring
    months_death_or_cens = case_when(
      !is.na(death_ym) ~ interval(enc_ym, death_ym) %/% months(1),
      TRUE             ~ interval(enc_ym, ref_ym)   %/% months(1)
    ),

    ## 6. Remove impossible values
    months_death_or_cens = if_else(
      months_death_or_cens < 0 | months_death_or_cens > 16,
      NA_integer_, months_death_or_cens
    ),

    ## 7. Death within one or two months
    died_1mo = if_else(died == 1 & months_death_or_cens <  1, 1L, 0L),
    died_2mo = if_else(died == 1 & months_death_or_cens <= 1, 1L, 0L),

    ## 8. Month of death (missing if censored)
    death_time = if_else(died == 1, months_death_or_cens, NA_integer_),

    ## 9. Death within 60 days (new variable)
    death_60d = if_else(died == 1 & death_abs <= (encounter_date + days(60)), 1L, 0L)
  ) %>%
  select(-enc_ym, -death_ym)

subset_data <- subset_data %>%
  mutate(
    death_60d = if_else(died == 1 & death_abs <= (encounter_date + days(60)), 1L, 0L)
  )
```

Chunk derive-death-60d runtime: 1.65 s

```r
table(subset_data$death_60d, useNA = "ifany")
```

```

     0      1 
461485  53801 
```

```r
prop.table(table(subset_data$death_60d, useNA = "ifany"))
```

```

      0       1 
0.89559 0.10441 
```

```r
summary(subset_data$death_60d)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.1044  0.0000  1.0000 
```

Chunk death-60d-summary runtime: 0.04 s

2.2 2) Baseline tables

2.2.1 2.1 Table 1A and 1B:

```r
# Robust derivation of analysis variables + helper for Table 1 production
# ---------------------------------------------------------------------------

# helper: label binary 0/1 → "No"/"Yes"
bin_lab <- function(x) factor(x, levels = c(0, 1), labels = c("No", "Yes"))

subset_data <- subset_data %>%
  mutate(
    ## ensure 0/1 numerics (avoids factor-level coercion)
    across(c(has_abg, has_vbg, hypercap_on_abg, hypercap_on_vbg),
           ~ as.numeric(as.character(.))),

    ## derive ABG / VBG hypercapnia groups
    abg_group
    = case_when(
      has_abg == 0                         ~ "No ABG",
      has_abg == 1 & hypercap_on_abg == 0  ~ "ABG_NoHypercapnia",
      has_abg == 1 & hypercap_on_abg == 1  ~ "ABG_Hypercapnia",
      TRUE                                 ~ "Missing"
    ),
    vbg_group = case_when(
      has_vbg == 0                         ~ "No VBG",
      has_vbg == 1 & hypercap_on_vbg == 0  ~ "VBG_NoHypercapnia",
      has_vbg == 1 & hypercap_on_vbg == 1  ~ "VBG_Hypercapnia",
      TRUE                                 ~ "Missing"
    ),

    ## factorise groups with explicit NA/Missing level
    abg_group = factor(
      abg_group,
      levels = c("No ABG", "ABG_NoHypercapnia", "ABG_Hypercapnia", "Missing")
    ),
    vbg_group = factor(
      vbg_group,
      levels = c("No VBG", "VBG_NoHypercapnia", "VBG_Hypercapnia", "Missing")
    ),

    ## labelled covariates
    sex_label      = factor(sex, levels = c(0, 1), labels = c("Female", "Male")),
    race_ethnicity_label     = factor(
      race_ethnicity,
      levels = c(0, 1, 2, 3, 4, 5, 6),
      labels = c("White", "Black or African American", "Hispanic",
                 "Asian", "American Indian", "Pacific Islander", "Unknown")
    ), location_label     = factor(
      location,
      levels = c(0, 1, 2, 3),
      labels = c("South", "Northeast" ,"Midwest", "West")
    ), encounter_type_label = factor(
      encounter_type,
      levels = c(2, 3),
      labels = c("Emergency", "Inpatient")
    ),
    osa_label      = bin_lab(osa),
    asthma_label   = bin_lab(asthma),
    copd_label     = bin_lab(copd),
    chf_label      = bin_lab(chf),
    nmd_label      = bin_lab(nmd),
    phtn_label     = bin_lab(phtn),
    ckd_label      = bin_lab(ckd),
    diabetes_label = bin_lab(dm)
  )

# variables to summarise
vars <- c(
  "age_at_encounter", "curr_bmi", "sex_label", "race_ethnicity_label", "location_label",
  "osa_label", "asthma_label", "copd_label", "chf_label", "nmd_label",
  "phtn_label", "ckd_label", "diabetes_label", "encounter_type_label", "vbg_co2", "paco2"
)

# Table 1 constructor
make_table1 <- function(data, group_var, caption = "") {
  group_sym <- rlang::sym(group_var)

  data %>%
    filter(!is.na(!!group_sym),                 # drop explicit NA
           !!group_sym != "Missing") %>%        # drop "Missing" cohort
    droplevels() %>%                            # trim empty factor levels
    select(all_of(c(group_var, vars))) %>%
    gtsummary::tbl_summary(
      by   = !!group_sym,
      type = list(sex_label ~ "categorical"),
      statistic = list(
        gtsummary::all_continuous()  ~ "{mean} ± {sd}; {N_miss}/{N_obs} missing ({p_miss}%)",
        gtsummary::all_categorical() ~ "{n} ({p}%)"
      ),
      digits   = list(gtsummary::all_continuous() ~ 1),
      missing  = "no"                              # no gtsummary missing column/row
    ) %>%
    gtsummary::modify_header(label = "**Variable**") %>%
    gtsummary::modify_caption(caption)
}

# build tables
table1A <- make_table1(subset_data, "abg_group", caption = "Table 1A: ABG cohorts")
table1B <- make_table1(subset_data, "vbg_group", caption = "Table 1B: VBG cohorts")

table1A
```

```
table1B
```

Chunk derive-table1-cohorts runtime: 7.07 s

Generating Word Doc for Table 1A & 1B

```
ft_table1A <- as_flex_table(table1A)
ft_table1B <- as_flex_table(table1B)

doc <- read_docx() %>%
  body_add_par("Table 1A. Baseline Characteristics by ABG Group", style = "heading 1") %>%
  body_add_flextable(ft_table1A) %>%
  body_add_par("Table 1B. Baseline Characteristics by VBG Group", style = "heading 1") %>%
  body_add_flextable(ft_table1B)

print(doc, target = "Table1_ABG_VBG.docx")
```

Chunk export-table1a-table1b-word runtime: 0.60 s

Variable	**No ABG** N = 328,044[1]	**ABG_NoHypercapnia** N = 129,429[1]	**ABG_Hypercapnia** N
Age (years)	58.1 ± 18.1; 0.0/328,044.0 missing (0.0%)	60.8 ± 17.1; 0.0/129,429.0 missing (0.0%)	62.1 ± 16.4; 0.0/57,813.0 mi
Current BMI kg/m2	32.3 ± 8.7; 184,223.0/328,044.0 missing (56.2%)	28.6 ± 6.9; 75,826.0/129,429.0 missing (58.6%)	29.8 ± 7.9; 33,496.0/57,813.0 n
sex_label			
Female	169,023 (52%)	57,767 (45%)	27,116 (47%)
Male	159,021 (48%)	71,662 (55%)	30,697 (53%)
race_ethnicity_label			
White	200,033 (61%)	81,357 (63%)	39,784 (69%)
Black or African American	62,418 (19%)	19,197 (15%)	8,082 (14%)
Hispanic	23,548 (7.2%)	7,464 (5.8%)	2,757 (4.8%)
Asian	4,880 (1.5%)	2,739 (2.1%)	789 (1.4%)
American Indian	1,971 (0.6%)	1,768 (1.4%)	316 (0.5%)
Pacific Islander	460 (0.1%)	162 (0.1%)	56 (<0.1%)
Unknown	34,734 (11%)	16,742 (13%)	6,029 (10%)
location_label			
South	138,843 (42%)	70,729 (55%)	32,694 (57%)
Northeast	93,209 (28%)	23,262 (18%)	12,975 (22%)
Midwest	22,924 (7.0%)	10,703 (8.3%)	4,844 (8.4%)
West	73,068 (22%)	24,735 (19%)	7,300 (13%)
osa_label	60,653 (18%)	17,709 (14%)	11,965 (21%)
asthma_label	48,456 (15%)	13,049 (10%)	8,268 (14%)
copd_label	60,214 (18%)	21,195 (16%)	18,846 (33%)
chf_label	59,770 (18%)	25,469 (20%)	16,219 (28%)
nmd_label	11,891 (3.6%)	5,861 (4.5%)	2,487 (4.3%)
phtn_label	23,854 (7.3%)	10,513 (8.1%)	7,347 (13%)
ckd_label	54,528 (17%)	24,849 (19%)	11,769 (20%)
diabetes_label	93,007 (28%)	37,426 (29%)	18,521 (32%)
encounter_type_label			
Emergency	142,713 (44%)	19,196 (15%)	9,818 (17%)
Inpatient	185,331 (56%)	110,233 (85%)	47,995 (83%)
VBG PCO2	45.5 ± 10.5; 233,430.0/328,044.0 missing (71.2%)	42.0 ± 11.2; 91,782.0/129,429.0 missing (70.9%)	57.4 ± 18.4; 40,411.0/57,813.0 r
Arterial PCO2	NA ± NA; 328,044.0/328,044.0 missing (100.0%)	35.5 ± 6.1; 0.0/129,429.0 missing (0.0%)	58.5 ± 20.4; 0.0/57,813.0 mi

[1] Mean ± SD; N Missing/No. obs. missing (% Missing); n (%)

Variable	**No VBG** N = 365,623[1]	**VBG_NoHypercapnia** N = 105,646[1]	**VBG_Hypercapnia** N
Age (years)	59.4 ± 17.8; 0.0/365,623.0 missing (0.0%)	58.1 ± 17.8; 0.0/105,646.0 missing (0.0%)	61.0 ± 16.7; 0.0/44,017.0 mi
Current BMI kg/m2	31.8 ± 8.5; 192,892.0/365,623.0 missing (52.8%)	28.7 ± 7.2; 69,615.0/105,646.0 missing (65.9%)	29.3 ± 7.9; 31,038.0/44,017.0 m
sex_label			
Female	184,619 (50%)	48,931 (46%)	20,356 (46%)
Male	181,004 (50%)	56,715 (54%)	23,661 (54%)
race_ethnicity_label			
White	241,114 (66%)	55,100 (52%)	24,960 (57%)
Black or African American	61,814 (17%)	19,199 (18%)	8,684 (20%)
Hispanic	22,951 (6.3%)	8,354 (7.9%)	2,464 (5.6%)
Asian	5,439 (1.5%)	2,293 (2.2%)	676 (1.5%)
American Indian	2,128 (0.6%)	1,683 (1.6%)	244 (0.6%)
Pacific Islander	543 (0.1%)	110 (0.1%)	25 (<0.1%)
Unknown	31,634 (8.7%)	18,907 (18%)	6,964 (16%)
location_label			
South	196,774 (54%)	30,426 (29%)	15,066 (34%)
Northeast	65,537 (18%)	44,405 (42%)	19,504 (44%)
Midwest	24,891 (6.8%)	9,178 (8.7%)	4,402 (10%)
West	78,421 (21%)	21,637 (20%)	5,045 (11%)
osa_label	65,748 (18%)	15,634 (15%)	8,945 (20%)
asthma_label	49,810 (14%)	13,419 (13%)	6,544 (15%)
copd_label	70,950 (19%)	16,459 (16%)	12,846 (29%)
chf_label	68,964 (19%)	20,573 (19%)	11,921 (27%)
nmd_label	14,796 (4.0%)	3,754 (3.6%)	1,689 (3.8%)
phtn_label	27,731 (7.6%)	8,534 (8.1%)	5,449 (12%)
ckd_label	61,091 (17%)	21,290 (20%)	8,765 (20%)
diabetes_label	101,173 (28%)	33,665 (32%)	14,116 (32%)
encounter_type_label			
Emergency	124,405 (34%)	34,711 (33%)	12,611 (29%)
Inpatient	241,218 (66%)	70,935 (67%)	31,406 (71%)
VBG PCO2	NA ± NA; 365,623.0/365,623.0 missing (100.0%)	40.1 ± 6.6; 0.0/105,646.0 missing (0.0%)	60.2 ± 12.6; 0.0/44,017.0 mi
Arterial PCO2	42.4 ± 15.5; 233,430.0/365,623.0 missing (63.8%)	38.6 ± 15.4; 68,334.0/105,646.0 missing (64.7%)	52.7 ± 19.6; 26,280.0/44,017.0 r

[1] Mean ± SD; N Missing/No. obs. missing (% Missing); n (%)

2.2.2 2.2 Table 1 (Overall ABG/VBG status)

```
# Status factors (column labels are taken from factor levels)
subset_data <- subset_data %>%
  mutate(
    abg_status = factor(has_abg, levels = c(0, 1),
                        labels = c("Did not get ABG", "Did get ABG")),
    vbg_status = factor(has_vbg, levels = c(0, 1),
                        labels = c("Did not get VBG", "Did get VBG"))
  )

# ABG table with "Everyone" column first
tbl1_abg <- subset_data %>%
  select(all_of(vars), abg_status) %>%
  gtsummary::tbl_summary(
    by = abg_status,
    type = list(sex_label ~ "categorical"),
    statistic = list(
      gtsummary::all_continuous()  ~ "{mean} ± {sd}; {N_miss}/{N_obs} missing ({p_miss}%)",
      gtsummary::all_categorical() ~ "{n} ({p}%)"
    ),
    digits   = list(gtsummary::all_continuous() ~ 1),
    missing  = "no"
  ) %>%
  gtsummary::add_overall(last = FALSE, col_label = "Everyone") %>%
  gtsummary::modify_header(label = "**Variable**")

# VBG table (no "Everyone" here)
tbl1_vbg <- subset_data %>%
  select(all_of(vars), vbg_status) %>%
  gtsummary::tbl_summary(
    by = vbg_status,
    type = list(sex_label ~ "categorical"),
    statistic = list(
      gtsummary::all_continuous()  ~ "{mean} ± {sd}; {N_miss}/{N_obs} missing ({p_miss}%)",
```