phase3_missing_data.R

abdulbasir

2025-12-09

```r
# phase 3: missing data analysis
#
# detailed investigation into patterns and mechanisms of missing data,
# crucial for assessing bias in the analysis

# sourcing configuration and utilities
source("config.R")
source("utils.R")

# loading required libraries
suppressPackageStartupMessages({
  library(dplyr)
  library(readr)
})

print_section_header("Phase 3: Missing Data Analysis")
```

```
##
## ========================================================================
## Phase 3: Missing Data Analysis
## ========================================================================
```

```r
# 1. loading classified data from phase 2
input_file = file.path(RESULTS_DIR, "fda_analysis_clean_classified.csv")
classified_data = load_csv(input_file)

# 2. analyzing indication missingness
indication_missing = is.na(classified_data$`Abbreviated Indication(s)`)
total_drugs = nrow(classified_data)
missing_count = sum(indication_missing)
missing_rate = (missing_count / total_drugs) * 100

# calculating missingness by regulatory era
classified_data = classified_data %>%
  mutate(regulatory_era = assign_regulatory_era(`Approval Year`))

missingness_by_era = classified_data %>%
  group_by(regulatory_era) %>%
  summarise(
    missing_count = sum(is.na(`Abbreviated Indication(s)`)),
    total_count = n(),
```

```
    .groups = "drop"
  ) %>%
  mutate(missing_rate_pct = (missing_count / total_count) * 100)

cat(paste("Overall indication text missingness:", missing_count, "/", total_drugs,
          sprintf("(%.1f%%)\n", missing_rate)))
```

```
## Overall indication text missingness: 676 / 1335 (50.6%)
```

```
cat("Missingness by regulatory era:\n")
```

```
## Missingness by regulatory era:
```

```
print(missingness_by_era)
```

```
## # A tibble: 4 x 4
##   regulatory_era missing_count total_count missing_rate_pct
##   <chr>                 <int>       <int>            <dbl>
## 1 Early-PDUFA               0         336              0
## 2 Mid-PDUFA               146         262             55.7
## 3 Post-FDASIA             530         530            100
## 4 Pre-PDUFA                 0         207              0
```

```
# 3. identifying post-2000 missingness pattern
pre2000 = classified_data %>% filter(`Approval Year` < 2000)
post2000 = classified_data %>% filter(`Approval Year` >= 2000)

pre2000_missing = sum(is.na(pre2000$`Abbreviated Indication(s)`))
post2000_missing = sum(is.na(post2000$`Abbreviated Indication(s)`))

pre2000_rate = ifelse(nrow(pre2000) > 0, (pre2000_missing / nrow(pre2000)) * 100, 0)
post2000_rate = ifelse(nrow(post2000) > 0, (post2000_missing / nrow(post2000)) * 100, 0)

cat(sprintf("Pre-2000 (<2000): %d/%d missing (%.1f%%)\n",
            pre2000_missing, nrow(pre2000), pre2000_rate))
```

```
## Pre-2000 (<2000): 0/462 missing (0.0%)
```

```
cat(sprintf("Post-2000 ( 2000): %d/%d missing (%.1f%%)\n",
            post2000_missing, nrow(post2000), post2000_rate))
```

```
## Post-2000 ( 2000): 676/873 missing (77.4%)
```

```
cat(sprintf("Key finding: %.1f percentage point increase post-2000\n",
            post2000_rate - pre2000_rate))
```

```
## Key finding: 77.4 percentage point increase post-2000
```

```r
# 4. comparing missingness by review designation
review_missing_crosstab = table(
  classified_data$`Review Designation`,
  indication_missing
)

cat("Missingness by review designation:\n")
```

```
## Missingness by review designation:
```

```r
print(addmargins(review_missing_crosstab))
```

```
##                                         indication_missing
##                                          FALSE TRUE  Sum
##   Priority                                 299  373  672
##   Priority (indication [A] only)             3    2    5
##   Priority (indication [B] only)             1    0    1
##   Priority (used priority review voucher)    0   16   16
##   Standard                                 356  285  641
##   Sum                                      659  676 1335
```

```r
# 5. testing approval year distribution
missing_years = classified_data %>%
  filter(is.na(`Abbreviated Indication(s)`)) %>%
  pull(`Approval Year`)

complete_years = classified_data %>%
  filter(!is.na(`Abbreviated Indication(s)`)) %>%
  pull(`Approval Year`)

cat(sprintf("Missing data: %d-%d (n=%d)\n",
            min(missing_years), max(missing_years), length(missing_years)))
```

```
## Missing data: 2007-2024 (n=676)
```

```r
cat(sprintf("Complete data: %d-%d (n=%d)\n",
            min(complete_years), max(complete_years), length(complete_years)))
```

```
## Complete data: 1985-2007 (n=659)
```

```r
# 6. creating therapeutic area crosstabulation
area_missing_crosstab = table(
  classified_data$therapeutic_area,
  indication_missing
)

cat("Missingness by therapeutic area:\n")
```

```
## Missingness by therapeutic area:
```

```r
print(addmargins(area_missing_crosstab))
```

```
##            indication_missing
##             FALSE TRUE  Sum
##   Oncology     98  201  299
##   Other       561  178  739
##   Uncertain     0  297  297
##   Sum         659  676 1335
```

```r
# 7. running Little's MCAR test
indication_missing_binary = as.integer(is.na(classified_data$`Abbreviated Indication(s)`))

# test 1: chi-square test - is missingness independent of regulatory era?
era_missing_contingency = table(
  classified_data$regulatory_era,
  indication_missing_binary
)
chi2_era_result = chisq.test(era_missing_contingency)
chi2_era = chi2_era_result$statistic
p_era = chi2_era_result$p.value

# test 2: chi-square test - is missingness independent of review designation?
review_missing_contingency = table(
  classified_data$`Review Designation`,
  indication_missing_binary
)
chi2_review_result = chisq.test(review_missing_contingency)
```

```
## Warning in chisq.test(review_missing_contingency): Chi-squared approximation may be incorrect
```

```r
chi2_review = chi2_review_result$statistic
p_review = chi2_review_result$p.value

# test 3: point-biserial correlation between missingness and approval year
correlation = cor(indication_missing_binary, classified_data$`Approval Year`)
n = length(indication_missing_binary)
t_stat = correlation * sqrt((n - 2) / (1 - correlation^2))
p_corr = 2 * pt(-abs(t_stat), df = n - 2)

cat("Testing missing completely at random (MCAR) hypothesis:\n")
```

```
## Testing missing completely at random (MCAR) hypothesis:
```

```r
cat("Test 1: missingness vs regulatory era\n")
```

```
## Test 1: missingness vs regulatory era
```

```r
cat(sprintf("  Chi-square = %.2f, p-value = %.2e\n", chi2_era, p_era))
```

```
##   Chi-square = 1076.39, p-value = 4.81e-233
```

```
cat(sprintf("  Result: missingness %s associated with era\n",
            ifelse(p_era < 0.001, "is", "is not")))
```

```
##   Result: missingness is associated with era
```

```
cat("\nTest 2: missingness vs review designation\n")
```

```
##
## Test 2: missingness vs review designation
```

```
cat(sprintf("  Chi-square = %.2f, p-value = %.2e\n", chi2_review, p_review))
```

```
##   Chi-square = 33.00, p-value = 1.19e-06
```

```
cat(sprintf("  Result: missingness %s associated with review type\n",
            ifelse(p_review < 0.001, "is", "is not")))
```

```
##   Result: missingness is associated with review type
```

```
cat("\nTest 3: point-biserial correlation (missingness ~ approval year)\n")
```

```
##
## Test 3: point-biserial correlation (missingness ~ approval year)
```

```
cat(sprintf("  Correlation = %.3f, p-value = %.2e\n", correlation, p_corr))
```

```
##   Correlation = 0.889, p-value = 0.00e+00
```

```
cat(sprintf("  Result: %s correlation\n",
            ifelse(abs(correlation) > 0.5, "strong", "weak")))
```

```
##   Result: strong correlation
```

```
cat(sprintf("\nConclusion: data is %s missing completely at random\n",
            ifelse(p_era < 0.001, "NOT", "")))
```

```
##
## Conclusion: data is NOT missing completely at random
```

```
cat("Missingness is systematic and related to time period\n")
```

```
## Missingness is systematic and related to time period
```

```r
# 8. saving results
results_data = data.frame(
  metric = c(
    "total_drugs", "missing_count", "missing_rate_pct",
    "pre2000_missing_rate", "post2000_missing_rate",
    "chi2_era", "p_era", "chi2_review", "p_review",
    "correlation", "p_correlation", "is_mcar"
  ),
  value = c(
    total_drugs, missing_count, missing_rate,
    pre2000_rate, post2000_rate,
    chi2_era, p_era, chi2_review, p_review,
    correlation, p_corr, ifelse(p_era >= 0.001, 1, 0)
  )
)

output_file = file.path(RESULTS_DIR, "missing_data_patterns.csv")
save_csv(results_data, output_file)
```

```
## saving results to: /Users/abdulbasir/Downloads/Experimental AI/fda-oncology-approval-analysis/results
```

```r
cat("\nPhase 3 complete\n")
```

```
##
## Phase 3 complete
```