phase8b_sensitivity_uncertain.R

abdulbasir

2025-12-09

```r
# phase 8b: sensitivity analysis for uncertain classification
#
# testing if excluding "Uncertain" drugs biased results by reclassifying
# all uncertain drugs as "Other" (conservative assumption) and re-running
# the core ANOVA analysis

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

# loading required libraries
suppressPackageStartupMessages({
  library(dplyr)
  library(readr)
  library(ggplot2)
  library(car)  # for Type III SS (Marginal) on unbalanced designs
})

print_section_header("Phase 8b: Sensitivity Analysis For 'Uncertain' Classification")
```

```
##
## =======================================================================
## Phase 8b: Sensitivity Analysis For 'Uncertain' Classification
## =======================================================================
```

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

# 2. creating sensitivity dataset (Uncertain → Other)
cat("Reclassifying 'Uncertain' drugs as 'Other' (conservative assumption)\n")
```

```
## Reclassifying 'Uncertain' drugs as 'Other' (conservative assumption)
```

```r
sensitivity_data = classified_data %>%
  mutate(
    therapeutic_area_sensitivity = if_else(
      therapeutic_area == "Uncertain",
      "Other",
      therapeutic_area
```

```r
    )
  )

# counting reclassifications
n_reclassified = sum(classified_data$therapeutic_area == "Uncertain")
cat(sprintf("\nReclassified %d drugs from 'Uncertain' to 'Other'\n", n_reclassified))
```

```
##
## Reclassified 297 drugs from 'Uncertain' to 'Other'
```

```r
cat("\nOriginal classification:\n")
```

```
##
## Original classification:
```

```r
print(table(classified_data$therapeutic_area))
```

```
##
##  Oncology     Other Uncertain
##       299       739       297
```

```r
cat("\nSensitivity classification:\n")
```

```
##
## Sensitivity classification:
```

```r
print(table(sensitivity_data$therapeutic_area_sensitivity))
```

```
##
## Oncology     Other
##      299      1036
```

```r
# 3. preparing sensitivity analysis dataset
sensitivity_data = sensitivity_data %>%
  mutate(
    review_type_simplified = if_else(
      grepl("Priority", `Review Designation`, fixed = TRUE),
      "Priority",
      "Standard"
    ),
    therapeutic_area_factor = factor(therapeutic_area_sensitivity, levels = c("Other", "Oncology")),
    review_type_factor = factor(review_type_simplified, levels = c("Standard", "Priority")),
    regulatory_era_factor = factor(
      assign_regulatory_era(`Approval Year`),
      levels = c("Pre-PDUFA", "Early-PDUFA", "Mid-PDUFA", "Post-FDASIA"),
      ordered = TRUE
    ),
    review_time_days_response = review_time_days,
    log_review_time_days_response = log(review_time_days)
  )

cat(sprintf("\nSensitivity analysis sample: n = %d\n", nrow(sensitivity_data)))
```

```
##
## Sensitivity analysis sample: n = 1335

cat(sprintf("Original analysis sample (excluding Uncertain): n = %d\n",
            sum(classified_data$therapeutic_area != "Uncertain")))
```

```
## Original analysis sample (excluding Uncertain): n = 1038
```

```
cat(sprintf("Difference: %d additional drugs included\n", n_reclassified))
```

```
## Difference: 297 additional drugs included
```

```
# 4. running sensitivity ANOVA (matching Phase 6 Model 3)
sensitivity_model = lm(
  log_review_time_days_response ~ therapeutic_area_factor * review_type_factor + regulatory_era_factor,
  data = sensitivity_data
)

sensitivity_anova = car::Anova(sensitivity_model, type = 3)

cat("\nSensitivity ANOVA (Uncertain → Other):\n")
```

```
##
## Sensitivity ANOVA (Uncertain → Other):
```

```
print(sensitivity_anova)
```

```
## Anova Table (Type III tests)
##
## Response: log_review_time_days_response
##                                              Sum Sq   Df    F value   Pr(>F)
## (Intercept)                                  22840.9    1 71993.1125 < 2e-16 ***
## therapeutic_area_factor                          0.6    1     1.8274 0.17667
## review_type_factor                              73.0    1   230.2269 < 2e-16 ***
## regulatory_era_factor                          103.0    3   108.2533 < 2e-16 ***
## therapeutic_area_factor:review_type_factor       0.9    1     2.9344 0.08695 .
## Residuals                                      421.3 1328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# extracting key statistics
f_area_sens = sensitivity_anova["therapeutic_area_factor", "F value"]
p_area_sens = sensitivity_anova["therapeutic_area_factor", "Pr(>F)"]
f_review_sens = sensitivity_anova["review_type_factor", "F value"]
p_review_sens = sensitivity_anova["review_type_factor", "Pr(>F)"]
f_interaction_sens = sensitivity_anova["therapeutic_area_factor:review_type_factor", "F value"]
p_interaction_sens = sensitivity_anova["therapeutic_area_factor:review_type_factor", "Pr(>F)"]
f_era_sens = sensitivity_anova["regulatory_era_factor", "F value"]
p_era_sens = sensitivity_anova["regulatory_era_factor", "Pr(>F)"]

cat("\nKey statistics (sensitivity analysis):\n")
```

```
##
## Key statistics (sensitivity analysis):
```

```
cat(sprintf("  Therapeutic area: F=%.2f, p=%.2e\n", f_area_sens, p_area_sens))
```

```
##   Therapeutic area: F=1.83, p=1.77e-01
```

```
cat(sprintf("  Review type: F=%.2f, p=%.2e\n", f_review_sens, p_review_sens))
```

```
##   Review type: F=230.23, p=4.47e-48
```

```
cat(sprintf("  Interaction: F=%.2f, p=%.2e\n", f_interaction_sens, p_interaction_sens))
```

```
##   Interaction: F=2.93, p=8.69e-02
```

```
cat(sprintf("  Regulatory era: F=%.2f, p=%.2e\n", f_era_sens, p_era_sens))
```

```
##   Regulatory era: F=108.25, p=1.06e-62
```

```r
# 5. loading original results from Phase 6
model_comparison_file = file.path(RESULTS_DIR, "model_comparison.csv")
model_comparison = load_csv(model_comparison_file)

# extracting original Model 3 results
original_model3 = model_comparison %>% filter(model == "Model 3")

f_area_orig = original_model3 %>% filter(effect == "Therapeutic Area") %>% pull(f_statistic)
p_area_orig = original_model3 %>% filter(effect == "Therapeutic Area") %>% pull(p_value)
f_review_orig = original_model3 %>% filter(effect == "Review Type") %>% pull(f_statistic)
p_review_orig = original_model3 %>% filter(effect == "Review Type") %>% pull(p_value)
f_interaction_orig = original_model3 %>% filter(effect == "Interaction") %>% pull(f_statistic)
p_interaction_orig = original_model3 %>% filter(effect == "Interaction") %>% pull(p_value)
f_era_orig = original_model3 %>% filter(effect == "Regulatory Era") %>% pull(f_statistic)
p_era_orig = original_model3 %>% filter(effect == "Regulatory Era") %>% pull(p_value)

# 6. comparison
cat("\nComparison: Original vs Sensitivity\n")
```

```
##
## Comparison: Original vs Sensitivity
```

```r
comparison_table = data.frame(
  effect = c("Therapeutic Area", "Review Type", "Interaction", "Regulatory Era"),
  f_original = c(f_area_orig, f_review_orig, f_interaction_orig, f_era_orig),
  f_sensitivity = c(f_area_sens, f_review_sens, f_interaction_sens, f_era_sens),
  p_original = c(p_area_orig, p_review_orig, p_interaction_orig, p_era_orig),
  p_sensitivity = c(p_area_sens, p_review_sens, p_interaction_sens, p_era_sens)
) %>%
  mutate(
    f_diff = f_sensitivity - f_original,
```

```r
    f_pct_change = (f_diff / f_original) * 100,
    original_sig = p_original < ALPHA,
    sensitivity_sig = p_sensitivity < ALPHA,
    consistent = original_sig == sensitivity_sig
  )

cat("\nF-statistic comparison:\n")
```

```
##
## F-statistic comparison:
```

```r
print(comparison_table)
```

```
##            effect f_original f_sensitivity   p_original p_sensitivity        f_diff f_pct_change
## 1 Therapeutic Area  18.699579      1.827371 1.678619e-05  1.766685e-01 -16.8722074    -90.22774
## 2      Review Type  77.517028    230.226945 5.465203e-18  4.470575e-48 152.7099167    197.00177
## 3      Interaction   2.433274      2.934354 1.190915e-01  8.694673e-02   0.5010803     20.59284
## 4   Regulatory Era  88.636987    108.253314 4.966764e-51  1.056864e-62  19.6163272     22.13109
##   original_sig sensitivity_sig consistent
## 1         TRUE           FALSE      FALSE
## 2         TRUE            TRUE       TRUE
## 3        FALSE           FALSE       TRUE
## 4         TRUE            TRUE       TRUE
```

```r
# 7. robustness assessment
consistent_count = sum(comparison_table$consistent)
total_effects = nrow(comparison_table)

cat(sprintf("\nConsistency: %d/%d effects (%.1f%%) have consistent significance\n",
            consistent_count, total_effects, 100 * consistent_count / total_effects))
```

```
##
## Consistency: 3/4 effects (75.0%) have consistent significance
```

```r
# 8. interpretation
cat("\nRobustness interpretation:\n")
```

```
##
## Robustness interpretation:
```

```r
therapeutic_area_f_change = comparison_table$f_pct_change[comparison_table$effect == "Therapeutic Area"]

if (f_area_sens > 30) {
  cat(sprintf("Therapeutic area F-statistic: %.2f → %.2f (%.1f%% change)\n",
              f_area_orig, f_area_sens, therapeutic_area_f_change))
  cat("→ Classification is robust to 'Uncertain' exclusion\n")
  cat("  Including all 'Uncertain' drugs as 'Other' still yields a strong therapeutic area effect (F > 3
  cat("  This indicates that keyword-based oncology classification was not overly exclusive\n")
} else if (f_area_sens < 10) {
  cat(sprintf("Therapeutic area F-statistic: %.2f → %.2f (%.1f%% change)\n",
```

```r
                 f_area_orig, f_area_sens, therapeutic_area_f_change))
  cat("→ Classification may have been too exclusive\n")
  cat("  Including 'Uncertain' drugs as 'Other' dramatically reduces the therapeutic area effect (F < 1
  cat("  Many drugs classified as 'Uncertain' may actually be oncology drugs that keyword list failed t
  cat("  Recommendation: Manually review 'Uncertain' drugs and expand keyword list\n")
} else {
  cat(sprintf("Therapeutic area F-statistic: %.2f → %.2f (%.1f%% change)\n",
                 f_area_orig, f_area_sens, therapeutic_area_f_change))
  cat("→ Partial sensitivity to classification\n")
  cat("  The therapeutic area effect is somewhat reduced but still present\n")
  cat("  'Uncertain' drugs are likely a mix of both oncology and non-oncology\n")
}
```

```
## Therapeutic area F-statistic: 18.70 → 1.83 (-90.2% change)
## → Classification may have been too exclusive
##    Including 'Uncertain' drugs as 'Other' dramatically reduces the therapeutic area effect (F < 10)
##    Many drugs classified as 'Uncertain' may actually be oncology drugs that keyword list failed to ca
##    Recommendation: Manually review 'Uncertain' drugs and expand keyword list
```

```r
# overall conclusion
if (consistent_count == total_effects) {
  cat("\nOverall conclusion: All effects robust - results not sensitive to 'Uncertain' classification d
} else {
  cat(sprintf("\nOverall conclusion: %d/%d effects changed significance\n",
                 total_effects - consistent_count, total_effects))
  cat("Results show some sensitivity to 'Uncertain' classification\n")
  cat("Consider manual review of 'Uncertain' drugs for improved accuracy\n")
}
```

```
##
## Overall conclusion: 1/4 effects changed significance
## Results show some sensitivity to 'Uncertain' classification
## Consider manual review of 'Uncertain' drugs for improved accuracy
```

```r
# 9. visualization: sensitivity_uncertain_comparison.png
# preparing data for horizontal comparison plot
comparison_long = data.frame(
  Effect = rep(comparison_table$effect, 2),
  Analysis = factor(
    rep(c("Main (Uncertain excluded)", "Sensitivity (Uncertain→Other)"), each = nrow(comparison_table))
    levels = c("Main (Uncertain excluded)", "Sensitivity (Uncertain→Other)")
  ),
  F_statistic = c(comparison_table$f_original, comparison_table$f_sensitivity),
  p_value = c(comparison_table$p_original, comparison_table$p_sensitivity),
  significant = c(comparison_table$original_sig, comparison_table$sensitivity_sig)
)

# reordering effects by original F-statistic (descending)
effect_order = comparison_table$effect[order(-comparison_table$f_original)]
comparison_long$Effect = factor(comparison_long$Effect, levels = effect_order)

# creating horizontal comparison plot
```

```r
robustness_plot = ggplot(
  comparison_long,
  aes(y = Effect, x = F_statistic, fill = Analysis)
) +
  geom_col(position = position_dodge(width = 0.75), width = 0.65, alpha = 0.92, color = "white") +
  geom_text(
    aes(label = sprintf("F=%.2f | p=%.4f%s",
                        F_statistic,
                        p_value,
                        ifelse(significant, " *", ""))),
    position = position_dodge(width = 0.75),
    hjust = -0.05,
    size = 3.3,
    fontface = "bold"
  ) +
  scale_fill_manual(
    values = c(
      "Main (Uncertain excluded)" = "#2c2c2c",
      "Sensitivity (Uncertain→Other)" = "#d62728"
    )
  ) +
  labs(
    title = "Robustness Check: Effect of 'Uncertain' Classification",
    subtitle = sprintf("Main (n=%d, Uncertain excluded) vs Sensitivity (n=%d, Uncertain→Other)",
                        original_n, n_reclassified + original_n),
    x = "F-Statistic",
    y = "Effect",
    fill = "Analysis"
  ) +
  coord_cartesian(xlim = c(0, max(comparison_long$F_statistic) * 1.25)) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", size = 15, hjust = 0),
    plot.subtitle = element_text(size = 11, hjust = 0, color = "gray30"),
    plot.background = element_rect(fill = "white", color = NA),
    panel.background = element_rect(fill = "white", color = NA),
    panel.grid.major.x = element_line(color = "gray90", linewidth = 0.5),
    panel.grid.major.y = element_blank(),
    panel.grid.minor = element_blank(),
    legend.position = "top",
    legend.title = element_text(face = "bold", size = 11),
    legend.text = element_text(size = 10),
    axis.text.y = element_text(size = 11, face = "bold"),
    axis.text.x = element_text(size = 10),
    axis.title = element_text(face = "bold", size = 12)
  ) +
  annotate(
    "text",
    x = max(comparison_long$F_statistic) * 1.22,
    y = 0.5,
    label = "* = p < 0.05",
    hjust = 1,
    vjust = 0,
```
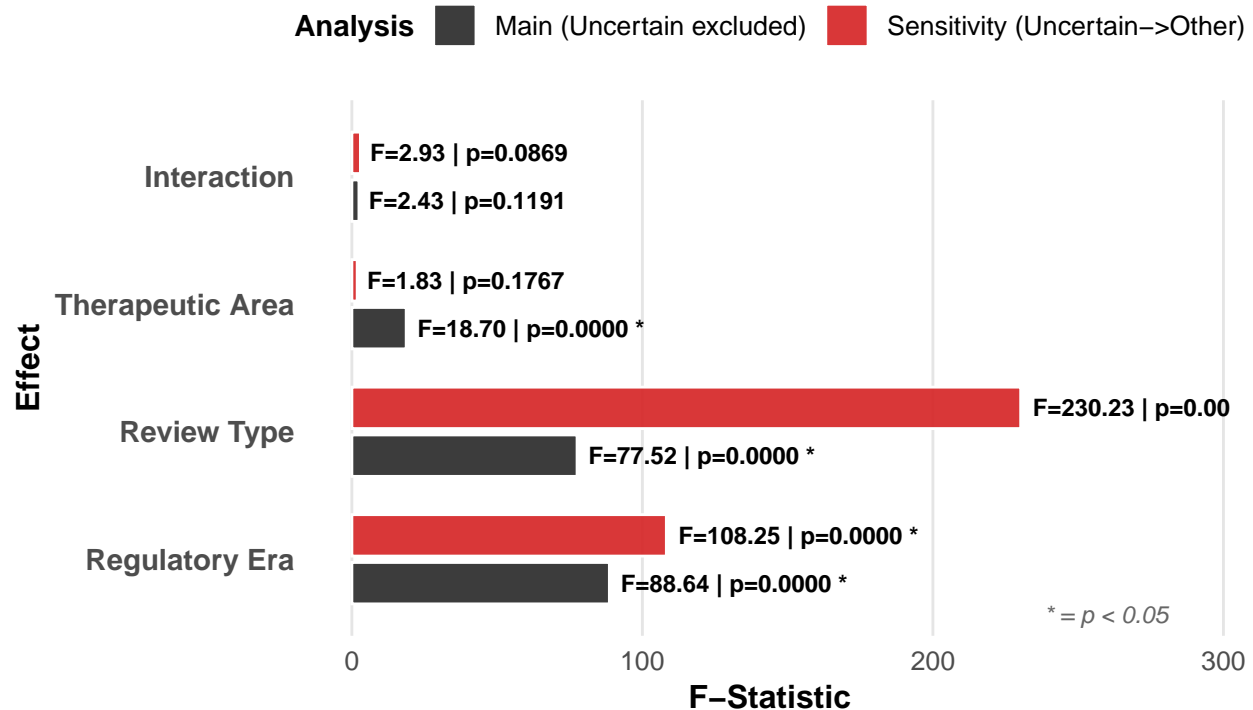
```
    size = 3,
    fontface = "italic",
    color = "gray40"
  )

print(robustness_plot)
```

Robustness Check: Effect of 'Uncertain' Classifica

Main (n=1038, Uncertain excluded) vs Sensitivity (n=1335, Uncertain–>C

Analysis ■ Main (Uncertain excluded) ■ Sensitivity (Uncertain–>Other)


```
ggsave(
  file.path(FIGURES_DIR, "sensitivity_uncertain_comparison.png"),
  plot = robustness_plot,
  width = 10,
  height = 6,
  dpi = DPI
)

cat("Saved: sensitivity_uncertain_comparison.png\n")
```

Saved: sensitivity_uncertain_comparison.png

```
# 10. saving results
output_file = file.path(RESULTS_DIR, "sensitivity_uncertain_comparison.csv")
save_csv(comparison_table, output_file)
```

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

```r
cat("\nPhase 8b complete\n")
```

```
## 
## Phase 8b complete
```