RTI Health Solutions

MSR40

A Comparison of the Performance of a Range of Trial-Level Surrogacy Methods

Adrian Vickers
RTI Health Solutions, Manchester, United Kingdom

BACKGROUND

- Surrogate endpoints play an increasingly important role in health technology submissions[1]
- Xie et al.[2] performed a literature review of surrogacy studies and found them to be inadequately conducted
  - Prediction intervals and surrogate threshold effect (STE) were often not presented despite being stated as essential outputs[3]
  - Many studies relied only on weighted linear regression weighted by sample size, which does not contain information on the length of follow-up/censoring
- Hedges[4] showed that the optimal weight is the inverse variance of each effect size, which is used in meta-analyses
- Bujkiewicz et al.[1,5] describe a range of trial-level surrogacy models. However, they do not mention presenting trial-level association plots with prediction intervals or STE

OBJECTIVE

- This research aims to compare the performance of a range of trial-level surrogacy models in 2 different settings

ILLUSTRATIVE EXAMPLES

- 2 data sets were used for this research:
  - Data from 34 clinical trials with a moderate correlation (advanced non-small cell lung cancer [NSCLC] data presented by Vickers et al.[6]). The target endpoint was overall survival (OS), and the surrogate was progression-free survival (PFS)
  - Data from 17 independent trials with a strong correlation (adjuvant gastric cancer data presented by Oba et al.[7]). The target endpoint was OS, and the surrogate was disease-free survival (DFS)

METHODS

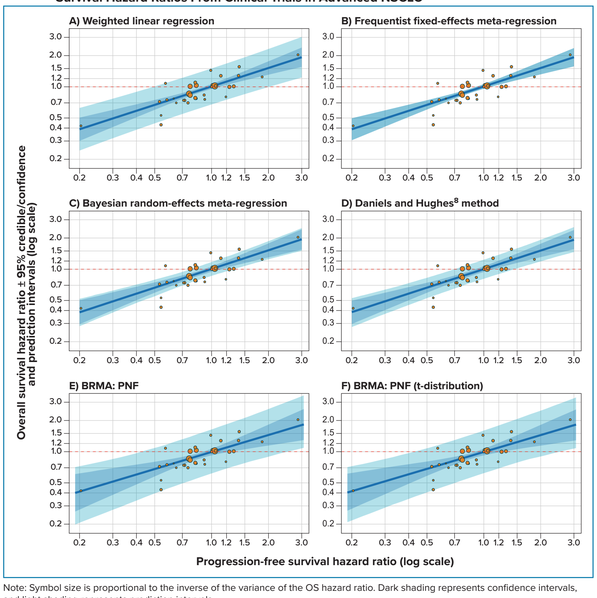
- Trial-level data with log(hazard ratios) and standard errors were used for target and surrogate endpoints
- The trial-level surrogacy models fitted were:
  - Weighted linear regression (weighted by the inverse variance of OS data): assumes no heterogeneity for the surrogate endpoint; prediction intervals estimated from the residual error, so variance in the target endpoint included
  - Fixed-effects meta-regression: assumes no heterogeneity for target or surrogate endpoints
  - Random-effects meta-regression: heterogeneity estimated for the target endpoint but no heterogeneity assumed for the surrogate endpoint
  - Daniels and Hughes[8] method: bivariate normal distribution with standard errors included for target and surrogate endpoints but only heterogeneity estimated for the target endpoint
  - Bivariate random-effects meta-analysis (BRMA) models (product normal formulation [PNF] and PNF [t-distribution]): standard errors included for target and surrogate endpoints, with heterogeneity estimated for both target and surrogate endpoints
- STE estimated: minimum treatment effect on the surrogate necessary to predict a non-zero effect on the true endpoint (the intersection of the 95% prediction interval with the x-axis)[2]
- Leave-one-out cross-validation:
  - Proportion of actual values (simulated from the mean and standard error) that lie within the predicted 95% prediction intervals
  - Probability that the predicted mean was the same as the actual mean based on a t-test for unequal variance
  - Mean absolute difference between predicted and actual values

RESULTS

Non-Small Cell Lung Cancer

- Figure 1 presents the results of the models fitted to NSCLC data. Table 1 presents correlations, STE, and results from leave-one-out cross-validation
  - Fixed- and random-effects meta-regressions, and Daniels and Hughes[8] method (Figure 1, models B-D)
    - Small prediction intervals that did not cover the range of observed data points
    - Produced high STE values (0.707-0.906)
    - Cross-validation results indicated that these models performed relatively poorly compared with the other models with respect to the measures used (i.e., a relatively low proportion of actual values within the prediction intervals and relatively low probability that the predicted mean was the same as the actual mean)
  - Weighted linear regression and BRMA models (Figure 1, models A, E, and F)
    - Larger prediction intervals that covered the range of observed data points
    - Produced lower STE values (0.398-0.489)
    - Cross-validation results indicated that these models performed relatively well compared with the other models (i.e., a relatively high proportion of actual values within the prediction intervals and relatively high probability that the predicted mean was the same as the actual mean)
    - The BRMA models gave the best overall results

Figure 1. Predicted Trial-Levels Associations From Models Fitted to Overall and Progression-Free Survival Hazard Ratios From Clinical Trials in Advanced NSCLC



Note: Symbol size is proportional to the inverse of the variance of the OS hazard ratio. Dark shading represents confidence intervals, and light shading represents prediction intervals.

Table 1. Predicted Correlations, STEs, and Leave-One-Out Cross-Validation Results From Models Fitted to Overall and Progression-Free Survival Hazard Ratios From Clinical Trials in Advanced NSCLC

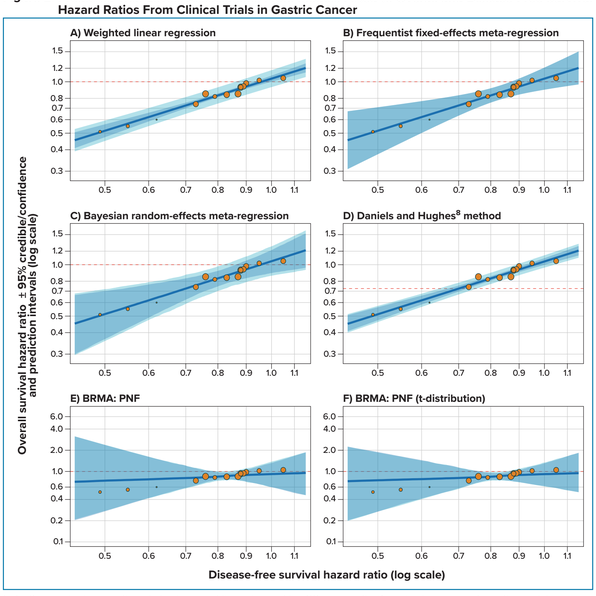
Method	Correlation	Surrogate threshold effect	Proportion of actual values within 95% of the prediction intervals: Median	Probability that the predicted mean equaled the actual mean: Median	Absolute difference between mean predicted value and mean actual value (log scale): Median
Weighted linear regression	0.780	0.489	0.915	0.615	0.143
Frequentist fixed-effects meta-regression	NA	0.906	0.181	0.514	0.143
Bayesian random-effects meta-regression	NA	0.766	0.443	0.527	0.144
Daniels and Hughes[8] method	NA	0.707	0.593	0.558	0.145
BRMA PNF	0.856	0.398	0.963	0.644	0.140
BRMA PNF (t-distribution)	0.848	0.398	0.953	0.659	0.136

NA = not applicable.

Gastric Cancer

- Figure 2 presents the results of the models fitted to gastric cancer data. Table 2 presents correlations and STE, and results from leave-one-out cross-validation
  - The BRMA models did not converge, which was indicated by iteration plots and non-normal parameter distributions, which resulted in more uncertainty in the association between surrogate and target endpoints (Figure 2, models E and F)
  - Weighted linear regression and Daniels and Hughes[8] method (Figure 2, models A and D)
    - Small prediction intervals that covered the range of observed data points but many points were close to the prediction intervals so a new study may not lie within the prediction intervals
    - Produced high STE values (0.859, 0.887)
    - Cross-validation results indicated that these models performed relatively poorly compared with the other models (i.e., a relatively low proportion of actual values within the prediction intervals and relatively low probability that the predicted mean was the same as the actual mean)
  - Fixed- and random-effects meta-regression (Figure 2, models B and C)
    - Large prediction intervals that covered the range of points in the chart, which showed increased uncertainty with less data
    - Random-effects meta-regression produced a lower STE value (0.822) compared with other models
    - Cross-validation results indicated that the random-effects meta-regression appeared to be the best model; it resulted in a relatively high proportion of actual values within the prediction intervals and relatively high probability that the predicted mean was the same as the actual mean

Figure 2. Predicted Trial-Levels Associations From Models Fitted to Overall and Disease-Free Survival Hazard Ratios From Clinical Trials in Gastric Cancer



Note: Symbol size is proportional to the inverse of the variance of the OS hazard ratio. Dark shading represents confidence intervals, and light shading represents prediction intervals.

Table 2. Predicted Correlations, STEs, and Leave-One-Out Cross-Validation Results From Models Fitted to Overall and Disease-Free Survival Hazard Ratios From Clinical Trials in Gastric Cancer

Method	Correlation	Surrogate threshold effect	Proportion of actual values within 95% of the prediction intervals: Median	Probability that the predicted mean equaled the actual mean: Median	Absolute difference between mean predicted value and mean actual value (log scale): Median
Weighted linear regression	0.968	0.859	0.455	0.868	0.032
Frequentist fixed-effects meta-regression	NA	0.867	0.500	0.876	0.032
Bayesian random-effects meta-regression	NA	0.822	0.718	0.880	0.030
Daniels and Hughes[8] method	NA	0.887	0.355	0.873	0.030
BRMA PNF	0.301	NA	0.800	0.807	0.070
BRMA PNF (t-distribution)	0.279	NA	0.754	0.717	0.086

DISCUSSION

- Care is needed to ensure data are representative of the type of study requiring surrogacy analyses
  - The data presented here in gastric cancer were from long-term independent trials with median follow-up of 4.9-16.6 years
    - The number of OS and DFS events was similar. The gastric data gave a ratio close to 1:1 between hazard ratios for target and surrogate endpoints. In clinical trial data, we typically observe hazard ratios for DFS and PFS that are better than those for OS such as evidenced in the NSCLC data[6]
    - Ajani et al.[9] conducted a surrogate clinical trial-level study in adjuvant gastric cancer and found less association (e.g., when DFS hazard ratio = 0.6, OS = 0.7), with a correlation of 0.76 and an STE of 0.64

CONCLUSIONS

- Results from different trial-level surrogacy models showed a high degree of variation
- BRMA models make fewer assumptions but may not converge with a small number of studies
- Random-effects meta-regression may be the best choice with a small number of studies and with a high correlation
- Weighted linear regression models gave reasonable predictions and ensure prediction intervals represent 95% of the variance in the data but may not always be the best model choice
- Trial-level association plots, with prediction intervals to show the uncertainty for a new study and STE, are needed so that predictions can be made for future studies

REFERENCES

1. Bujkiewicz, S, Achana, F, Papanikos, T, Riley, RD, Abrams, KR. NICE DSU Technical Support Document 20: multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints. 2019. https://www.sheffield.ac.uk/sites/default/files/2022-02/TSD20-mvmeta-final.pdf. Accessed 5 October 2022.
2. Xie W, Regan M, Buyse M, Halabi S, Kantoff W, Sartar O, et al. Metastasis-free survival Is a strong surrogate of overall survival in localized prostate cancer. J Clin Oncol. 2017;35(27):3097-104.
3. Buyse M, Burzykowski T, Michiels S, Carroll K. Individual- and trial-level surrogacy in colorectal cancer. Stat Methods Med Res. 2008;17(5):467-75.
4. Hedges LV. A random effects model for effect sizes. Psychol Bull. 1983;93:388-95.
5. Bujkiewicz, S, Thompson JR, Spata E, Abrams KR. Uncertainty in the Bayesian meta-analysis of normally distributed surrogate endpoints. State Methods Med Res. 2017;26(5):2287-318.
6. Vickers AD, Winfree KB, Carter GC, Kiiskinen U, Jen M-H, Stull D. Relative efficacy of interventions in the treatment of second-line non-small cell lung cancer: a systematic review and network meta-analysis. BMC Cancer. 2019;19:353.
7. Oba K, Paoletti X, Alberts S, Bang Y-J, Benedetti J, Bleiberg H, et al. Disease-free survival as a surrogate for overall survival in adjuvant trials of gastric cancer: a meta-analysis. J Natl Cancer Inst. 2013; 105: 1600-7.
8. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. Stat Med. 1997; 16:1965-82.
9. Ajani JA, Leung L, Singh P, Kurt M, Kim I, Pourrahmat M-M, et al. Disease-free survival as a surrogate endpoint for overall survival in adults with resectable esophageal or gastroesophageal junction cancer: a correlation meta-analysis. Eur J Cancer. 2022;170:119-30.

CONTACT INFORMATION

Adrian Vickers, PhD, MSc
Statistics Director

RTI Health Solutions
Didsbury, United Kingdom

Email: avickers@rti.org