The effect of machine learning tools for evidence synthesis on resource use and time-to-completion

Chris Rose, Norwegian Institute of Public Health (14 Apr 2023)

Generated using git revision: 8cb4a16

# Methods

All statistical analyses were performed as specified in our protocol using Stata 16 (StataCorp LLC, College Station, Texas, USA), except for one secondary analysis (see Protocol Deviations). The study is retrospective, and reviews were not randomized to use recommended ML versus no ML (for example). We therefore modelled ML use as an endogenously assigned treatment predicted by field (healthcare or welfare) and pre-specification (existence of a protocol), as planned. Resource use was analyzed using extended interval regression (Stata's eintreg command) and time-to-completion was analyzed using a likelihood-adjusted-censoring inverse-probability-weighted regression adjustment model (LAC-IPWRA; Stata's stteffects command). Ongoing reviews were right censored at the end of data collection (31 January 2023) and all analyses accounted for this censoring. We had no reason to suspect informative (nonrandom) censoring, so did not model a censoring mechanism. We re-expressed all estimates as ratios (relative resource use and relative time-to-completion) to aid generalization to other institutions. We did this by exponentiating differences in log resource use, and by computing ratios of mean times-to-completion using the delta method. We present two-sided 95% confidence intervals and p-values where appropriate and use a prespecified p < 0.05 significance criterion throughout. We also present the time-to-completion data using Kaplan-Meier estimates of survivor functions.

# Protocol Deviations

We had planned to model ML use as an endogenously assigned treatment in all analyses. However, we chose to deviate from protocol for the secondary analysis of recommended versus non-recommended ML use for the resource use outcome. While there was statistically significant evidence of endogeneity from the corresponding time-to-completion analysis and an exploratory logistic regression, the estimate of relative resource use obtained using the planned model appeared to dramatically overestimate the effect of recommended ML use. We therefore used a model for this analysis that did not account for possible endogeneity.

We updated the preprint version of the protocol during data extraction but before starting the analysis or unblinding the statistician (CJR) to redefine the comparisons in terms of under- and overuse of machine learning. However, only two reviews were judged to have under- or overused machine learning, so it was not possible to perform the revised analyses. We therefore performed and report the analyses as originally planned.

# Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type of ML Use** | **Reviews** | **Sample Mean¹** | **Effect Estimate²** | **p-value** |
| **Resource Use** |  | Person-hours |  |  |
| None | 12 | 140 | 3.7 (0.4 to 37.9) | 0.269 |
| Recommended | 21 | 569 |  |  |
| Non-recommended | 6 | 888 | 0.5 (0.0 to 10.7) | 0.658 |
| Recommended | 21 | 569 |  |  |
| None | 12 | 140 | 0.7 (0.2 to 1.9) | 0.439 |
| Any | 27 | 625 |  |  |
| **Time-to-completion** |  | Weeks |  |  |
| None | 12 | 28.2 | 0.9 (0.5 to 1.6) | 0.753 |
| Recommended | 21 | 27.6 |  |  |
| Non-recommended | 6 | 36.2 | 1.1 (0.7 to 1.9) | 0.658 |
| Recommended | 21 | 27.6 |  |  |
| None | 12 | 28.2 | 0.9 (0.6 to 1.5) | 0.784 |
| Any | 27 | 29.5 |  |  |
| ¹Data are means of samples restricted to completed (uncensored) reviews. ²Estimates are relative resource use and relative time-to-completion, account for right-censored outcomes and, except for the recommended versus non-recommended ML use comparison for the outcome resource use, also account for nonrandom endogenous treatment allocation. All estimates are adjusted for planned meta-analysis. | | | | |

## Kaplan-Meier estimates for Recommended vs No ML Use

## Kaplan-Meier estimates for Recommended vs Non-recommended ML Use

## Kaplan-Meier estimates for Any vs No ML Use

# Appendix — Full Regression Results

## Regression results for rec\_vs\_none with respect to resource

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Coef. | Std. Err. | z | P>|z| | [95% Conf. Interval] | |
| meta\_analysis\_planned |  |  |  |  |  |  |
| Yes | 0.97 | 0.37 | 2.63 | 0.01 | 0.25 | 1.70 |
|  |  |  |  |  |  |  |
| rec\_vs\_none |  |  |  |  |  |  |
| Recommended | 1.31 | 1.19 | 1.11 | 0.27 | -1.01 | 3.64 |
| \_cons | 4.69 | 0.89 | 5.28 | 0.00 | 2.95 | 6.43 |
| rec\_vs\_none |  |  |  |  |  |  |
| field |  |  |  |  |  |  |
| Welfare | 0.90 | 0.48 | 1.89 | 0.06 | -0.03 | 1.83 |
| \_cons | -0.10 | 0.35 | -0.28 | 0.78 | -0.78 | 0.59 |
| var(e.log\_resource1) | 0.58 | 0.15 |  |  | 0.35 | 0.96 |
| corr(e.rec\_vs\_none,e.log\_resource1) | 0.01 | 1.07 | 0.01 | 0.99 | -0.97 | 0.97 |

## Regression results for rec\_vs\_none with respect to time

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Robust |  |  |  |  |
| \_t | Coef. | Std. Err. | z | P>|z| | [95% Conf. Interval] | |
| ATE |  |  |  |  |  |  |
| rec\_vs\_none |  |  |  |  |  |  |
| (Recommended vs None) | -3.02 | 9.65 | -0.31 | 0.75 | -21.93 | 15.90 |
| POmean |  |  |  |  |  |  |
| rec\_vs\_none |  |  |  |  |  |  |
| None | 36.34 | 7.68 | 4.73 | 0.00 | 21.28 | 51.40 |

## Regression results for rec\_vs\_nonrec with respect to resource

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Coef. | Std. Err. | z | P>|z| | [95% Conf. Interval] | |
| meta\_analysis\_planned |  |  |  |  |  |  |
| Yes | 0.32 | 0.28 | 1.11 | 0.27 | -0.24 | 0.87 |
|  |  |  |  |  |  |  |
| rec\_vs\_nonrec |  |  |  |  |  |  |
| Recommended | -0.69 | 1.56 | -0.44 | 0.66 | -3.76 | 2.37 |
| \_cons | 6.94 | 1.28 | 5.42 | 0.00 | 4.43 | 9.45 |
| rec\_vs\_nonrec |  |  |  |  |  |  |
| prespecified |  |  |  |  |  |  |
| Yes | -0.26 | 0.56 | -0.47 | 0.64 | -1.37 | 0.84 |
| \_cons | 0.92 | 0.44 | 2.11 | 0.03 | 0.07 | 1.78 |
| var(e.log\_resource1) | 0.49 | 0.21 |  |  | 0.21 | 1.15 |
| corr(e.rec\_vs\_nonrec,e.log\_resource1) | 0.24 | 1.24 | 0.20 | 0.84 | -0.98 | 0.99 |

## Regression results for rec\_vs\_nonrec with respect to time

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Robust |  |  |  |  |
| \_t | Coef. | Std. Err. | z | P>|z| | [95% Conf. Interval] | |
| ATE |  |  |  |  |  |  |
| rec\_vs\_nonrec |  |  |  |  |  |  |
| (Recommended vs Non-recommended) | 3.83 | 8.79 | 0.44 | 0.66 | -13.39 | 21.06 |
| POmean |  |  |  |  |  |  |
| rec\_vs\_nonrec |  |  |  |  |  |  |
| Non-recommended | 30.87 | 5.66 | 5.45 | 0.00 | 19.77 | 41.97 |

## Regression results for any\_vs\_none with respect to resource

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Coef. | Std. Err. | z | P>|z| | [95% Conf. Interval] | |
| meta\_analysis\_planned |  |  |  |  |  |  |
| Yes | 0.99 | 0.31 | 3.17 | 0.00 | 0.38 | 1.61 |
|  |  |  |  |  |  |  |
| any\_vs\_none |  |  |  |  |  |  |
| Any | -0.43 | 0.55 | -0.77 | 0.44 | -1.52 | 0.66 |
| \_cons | 5.97 | 0.41 | 14.42 | 0.00 | 5.16 | 6.79 |
| any\_vs\_none |  |  |  |  |  |  |
| field |  |  |  |  |  |  |
| Welfare | 0.59 | 0.31 | 1.89 | 0.06 | -0.02 | 1.20 |
| \_cons | 0.28 | 0.25 | 1.13 | 0.26 | -0.21 | 0.76 |
| var(e.log\_resource1) | 1.34 | 0.55 |  |  | 0.61 | 2.98 |
| corr(e.any\_vs\_none,e.log\_resource1) | 0.95 | 0.08 | 12.47 | 0.00 | 0.25 | 1.00 |

## Regression results for any\_vs\_none with respect to time

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Robust |  |  |  |  |
| \_t | Coef. | Std. Err. | z | P>|z| | [95% Conf. Interval] | |
| ATE |  |  |  |  |  |  |
| any\_vs\_none |  |  |  |  |  |  |
| (Any vs None) | -2.48 | 9.15 | -0.27 | 0.79 | -20.41 | 15.45 |
| POmean |  |  |  |  |  |  |
| any\_vs\_none |  |  |  |  |  |  |
| None | 37.97 | 7.98 | 4.76 | 0.00 | 22.34 | 53.60 |