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

Chris Rose, Norwegian Institute of Public Health (13 Sep 2023)

Generated using git revision: 01409bf

# Methods

Except as noted in Protocol Deviations, the statistical analyses were performed as specified in our protocol using Stata 16 (StataCorp LLC, College Station, Texas, USA). 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 prespecification (existence of a protocol). 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. Normality of residuals from the analyses of resource use were assessed using the Shapiro-Wilk test. To aid generalization to other institutions, we re-expressed analysis results as estimates of ratios (relative resource use and relative time-to-completion). We did this by exponentiating differences in mean 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 (but note that these do not account for nonrandom endogenous treatment assignment and are not adjusted for planned meta-analysis).

# Protocol Deviations

It was not possible to model nonrandom endogenous treatment assignment using both prespecified variables (field and prespecification) in the analyses of resource use because the models did not converge. We therefore used one of the two variables, choosing the variable with the smallest standard error in the model of treatment assignment. Endogenous assignment of any or recommended ML was modelled by field (welfare reviews were generally more likely to use ML) and recommended ML use was modelled by prespecification (reviews with protocols were generally less likely to use recommended ML).

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, too few reviews were judged to have under- or overused ML, so it was not possible to run these analyses. We therefore performed and report the analyses as originally planned.

# Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **None** | **Recommended** | **Non-recommended** | **Recommended** | **None** | **Any** |
| Commissioned reviews | 12 (31%) | 21 (54%) | 6 (15%) | 21 (54%) | 12 (31%) | 27 (69%) |
| Completed reviews | 12 (31%) | 19 (49%) | 5 (13%) | 19 (49%) | 12 (31%) | 24 (62%) |
| **Review type** |  |  |  |  |  |  |
| HTA | 3 ( 8%) | 2 ( 5%) | 2 ( 5%) | 2 ( 5%) | 3 ( 8%) | 4 (10%) |
| Non-HTA | 9 (23%) | 19 (49%) | 4 (10%) | 19 (49%) | 9 (23%) | 23 (59%) |
| **Synthesis type planned** |  |  |  |  |  |  |
| Any (quantitative or qualitative) | 11 (28%) | 19 (49%) | 6 (15%) | 19 (49%) | 11 (28%) | 25 (64%) |
| Pairwise meta-analysis | 4 (10%) | 10 (26%) | 4 (10%) | 10 (26%) | 4 (10%) | 14 (36%) |
| Network meta-analysis | 0 | 0 | 0 | 0 | 0 | 0 |
| **ML used during study identification** |  |  |  |  |  |  |
| Ranking | 0 | 19 (49%) | 6 (15%) | 19 (49%) | 0 | 25 (64%) |
| Classifiers | 0 | 9 (23%) | 3 ( 8%) | 9 (23%) | 0 | 12 (31%) |
| Clustering | 0 | 6 (15%) | 2 ( 5%) | 6 (15%) | 0 | 8 (21%) |
| OpenAlex | 0 | 5 (13%) | 0 | 5 (13%) | 0 | 5 (13%) |
| **ML used during data extraction** |  |  |  |  |  |  |
| Classifiers | 0 | 0 | 0 | 0 | 0 | 0 |
| Clustering | 0 | 1 ( 3%) | 0 | 1 ( 3%) | 0 | 1 ( 3%) |
| Automated data extraction | 0 | 0 | 0 | 0 | 0 | 0 |
| **Other ML functions** | 0 | 0 | 0 | 0 | 0 | 0 |
| Data are number of reviews and percent of all included reviews. | | | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type of ML Use** | **Reviews** | **Sample Mean¹** | **Effect Estimate²** | **p-value** |
| **Resource Use** |  | Person-hours |  |  |
| None | 12 | 140 | 3.71 (0.36 to 37.95) | 0.269 |
| Recommended | 21 | 569 |  |  |
| Non-recommended | 6 | 888 | 0.50 (0.02 to 10.74) | 0.658 |
| Recommended | 21 | 569 |  |  |
| None | 12 | 140 | 0.65 (0.22 to 1.93) | 0.439 |
| Any | 27 | 625 |  |  |
| **Time-to-completion** |  | Weeks |  |  |
| None | 12 | 28.2 | 0.92 (0.53 to 1.58) | 0.753 |
| Recommended | 21 | 27.6 |  |  |
| Non-recommended | 6 | 36.2 | 1.12 (0.67 to 1.89) | 0.658 |
| Recommended | 21 | 27.6 |  |  |
| None | 12 | 28.2 | 0.93 (0.58 to 1.51) | 0.784 |
| Any | 27 | 29.5 |  |  |
| ¹Data are means of samples restricted to completed (uncensored) reviews and do not account for nonrandom endogenous treatment allocation. ²Estimates are relative resource use and relative time-to-completion, and account for right-censored outcomes and nonrandom endogenous treatment allocation. An effect estimate < 1 indicates that recommended or any ML use is associated with less resource use or shorter time-to-completion than to the comparator. All estimates are adjusted for planned meta-analysis. | | | | |

The strength of evidence for endogeneity varied by comparison. For example, the correlation between use of any versus no ML and resource use was 0.95 (95% CI 0.25 to 1.00; p<0.0001) while that for recommended versus no ML use was 0.01 (95% CI -0.97 to 0.97; p=0.99). However, as far as possible, we chose to account for possible endogeneity according to our protocol.

Due to the smaller than anticipated sample size, none of the effect estimates were sufficiently precise to be able to conclude that use of recommended or any ML is associated with more or less resource use, or longer or shorter time-to-completion, compared to no or non-recommended ML use (i.e., all confidence intervals include the null). For resource use, point estimates favor recommended and any ML use over non-recommended and no ML use, while no ML use is favored over recommended ML use. For time-to-completion, point estimates favor recommended and any ML use over no ML use, while non-recommended ML use is favored over recommended ML use. The estimates are generally but not always consistent with the sample means and Kaplan-Meier plots. However, the sample means may be quite misleading due to possible confounding (nonrandom endogenous treatment assignment), do not account for censoring, and are not adjusted for the effect of planned meta-analysis, which is associated with more resource use and longer time-to-completion. The Kaplan-Meier plots show censored reviews, but do not account for endogeneity and are not adjusted for planned meta-analysis. Recall that we estimated ratios of means, while Kaplan-Meier plots are generally interpreted in terms of quantiles (e.g., median time-to-completion).

## 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

# Discussion

Results for the resource use outcome are a little challenging to interpret. The effect estimates suggest that recommended or any ML use is associated with less resource use compared to non-recommended or no ML use, but the sample means for the any versus no ML use comparison suggest the opposite (but could be quite misleading). The effect estimate suggests that recommended ML use is associated with substantially more resource use compared to no ML use, and this is also reflected by the sample means. We find this result quite surprising, but it could be explained by confounding that we have not been able to account for. For example, perhaps reviews that did not use ML did not do so because they were judged to be "easy", and "easy" reviews are not resource intensive.

Results for the time-to-completion outcome are somewhat easier to interpret. The effect estimates suggest that recommended or any ML use is associated with shorter time-to-completion compared to no ML use. These results are consistent with our experience and with one of the associated sets of sample means. The point estimate for the recommended versus non-recommended comparison slightly favors non-recommended ML use, but the sample means strongly suggest otherwise. Because the estimate is imprecise, and the point estimate is close to the null, it would not surprise us if the direction of effect is wrong.

We think the results for the time-to-completion outcome are somewhat easier to interpret than those for resource use because review commission and completion dates (which are used to define time-to-completion) are well-defined and easy to measure, while resource use is essentially a self-reported outcome. Researchers at our institute use a web- or mobile app-based system to allocate hours worked to specific projects. There is likely to be inter-researcher differences in reporting, which could lead to substantial variation in outcome measurements. In addition, confounding may occur if researchers who under- or over-allocate hours are also more likely to use or not use ML. While confounding may also occur for time-to-completion outcomes, it may be easier to account for in analysis. For these reasons, we suggest that future studies specify well-defined time-to-completion outcomes as being of primary interest.

# Appendix — Full Regression Results

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Coefficient | 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 | Coefficient | 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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Coefficient | 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 | Coefficient | 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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Coefficient | 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 | Coefficient | 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 |