BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Modelling hospital outcome: problems with endogeneity

John L. Moran[1*], John D. Santamaria[2], Graeme J. Duke[3] and The Australian & New Zealand Intensive Care Society (ANZICS) Centre for Outcomes & Resource Evaluation (CORE)

Abstract

Background: Mortality modelling in the critical care paradigm traditionally uses logistic regression, despite the availability of estimators commonly used in alternate disciplines. Little attention has been paid to covariate endogeneity and the status of non-randomized treatment assignment. Using a large registry database, various binary outcome modelling strategies and methods to account for covariate endogeneity were explored.

Methods: Patient mortality data was sourced from the Australian & New Zealand Intensive Society Adult Patient Database for 2016. Hospital mortality was modelled using logistic, probit and linear probability (LPM) models with intensive care (ICU) providers as fixed (FE) and random (RE) effects. Model comparison entailed indices of discrimination and calibration, information criteria (AIC and BIC) and binned residual analysis. Suspect covariate and ventilation treatment assignment endogeneity was identified by correlation between predictor variable and hospital mortality error terms, using the Stata™ "eprobit" estimator. Marginal effects were used to demonstrate effect estimate differences between probit and "eprobit" models.

Results: The cohort comprised 92,693 patients from 124 intensive care units (ICU) in calendar year 2016. Patients mean age was 61.8 (SD 17.5) years, 41.6% were female and APACHE III severity of illness score 54.5(25.6); 43.7% were ventilated. Of the models considered in predicting hospital mortality, logistic regression (with or without ICU FE) and RE logistic regression dominated, more so the latter using information criteria indices. The LPM suffered from many predictions outside the unit [0,1] interval and both poor discrimination and calibration. Error terms of hospital length of stay, an independent risk of death score and ventilation status were correlated with the mortality error term. Marked differences in the ventilation mortality marginal effect was demonstrated between the probit and the "eprobit" models which were scenario dependent. Endogeneity was not demonstrated for the APACHE III score.

Conclusions: Logistic regression accounting for provider effects was the preferred estimator for hospital mortality modelling. Endogeneity of covariates and treatment variables may be identified using appropriate modelling, but failure to do so yields problematic effect estimates.

Keywords: Outcome analysis, Logit, Probit, Linear probability model, Calibration, Endogeneity, Marginal effects

* Correspondence: john.moran@adelaide.edu.au
[1]Department of Intensive Care Medicine, The Queen Elizabeth Hospital, Woodville, Australia
Full list of author information is available at the end of the article

Background

Modelling mortality outcome has been a constant preoccupation within the critical care literature [1] both in terms of predictive models such as the Acute Physiology and Chronic Health Evaluation (APACHE) algorithms [2, 3] and ground up exploratory studies of the impact of covariates of interest [4]. The preferred model has been logistic regression (or logit) [5], rather than probit [6], consistent with the sentiments of Berkson, "Why I prefer logits to probits", expressed 70 years ago [7]. In econometrics, the probit [8] and the linear probability model (LPM) [9] have been extensively used for modelling binary outcomes and such models have occasionally appeared in the biomedical literature [10].

Model validation has also differed between disciplines. Within the biomedical and epidemiological literature extensive discussion has focused around concepts of discrimination and calibration [11–13], whereas in econometrics bias and parameter consistency have been dominant [14–16], to the exclusion of model performance issues such as goodness-of-fit [17], although some issues intersect [18]. Econometrics has paid greater attention to concepts such as endogeneity [19], self-selection [20] and non-randomized treatment assignment [21], although there has been a rapid increase in the biomedical literature devoted to these issues, especially in epidemiology [22]. Previous attention [23] has been drawn to suspected endogeneity in mortality models where length of stay [24] or mortality probability [25] were entered as predictive covariates; such regression of a variable upon its components has been termed a "dubious practice" [26].

The purpose of this study was to explore the performance of regression models, logistic, probit and the LPM in predicting the hospital mortality risk of a large cohort of critically ill intensive care patients whose data was recorded in the ANZICS (Australian and New Zealand Intensive Care Society) Adult Patient Data Base [27]. Machine learning approaches were not considered [28], albeit there is debate as to what constitutes "machine learning" [29, 30]. Performance of both fixed and random effects models of logit and probit was compared with particular attention directed to calibration [13]. The following issues were also canvassed; the potential endogeneity of hospital length of stay (HLOS) and hospital mortality probability (ROD) recorded in the data base and derived from an independent published algorithm [31], and the effect of mechanical ventilation (MV) status (recorded as a binary variable) as an endogenous treatment assignment.

Methods

Ethics statement

Access to the data was granted by the Australian and New Zealand Intensive Care Society (ANZICS) Centre for Outcomes & Resource Evaluation (CORE) Management Committee in accordance with standing protocols; local hospital (The Queen Elizabeth Hospital) Ethics of Research Committee waived the need for patient consent to use their data in this study. The dataset was anonymized before release to the authors by ANZICS CORE custodians of the database. The dataset is the property of the ANZICS CORE and contributing ICUs and is not in the public domain. Access to the data by researchers, submitting ICUs, jurisdictional funding bodies and other interested parties is obtained under specific conditions and upon written request ("ANZICS CORE Data Access and Publication Policy.pdf", http://www.anzics.com.au/Downloads/ANZICS%20CORE%20Data%20Access%20and%20Publication%20Policy%20July%202017.pdf).

Data management

Data was accessed from the ANZICS Adult Patient Database [27]; in this instance for calendar year 2016 and processed as previously described in detail [32].

Statistical analysis

Predictive models

To predict hospital mortality a base parsimonious logistic model (Logit1) was developed with a core set of predictor variables and their interactions, similar to previous papers utilizing data from the ANZICS Adult Patient Database [23, 32]; no automated routine for covariate selection, such as stepwise regression, was used. The covariate set was then supplemented by addition of two covariates: log HLOS (in days) and log risk of death (ROD) derived from a locally validated mortality algorithm (Australian and New Zealand Risk of Death model) [31] and model fit was further ascertained. All continuous variables were centred to improve model convergence. Using the same base covariate set and additions, this process was repeated for the following models:

1. Logistic regression with intensive care unit (ICU) providers as fixed effects (FE), (Logit2)
2. A base probit regression (base: Probit1)
3. Probit regression with intensive care unit (ICU) providers as FE (Probit2)
4. Random effects (RE) logit (Logit3) and probit regression (Probit3) with patients nested within ICU providers considered as RE; that is a random intercept model.
    a. The intra-class correlation (ICC), the correlation between patients in ICU providers [33], was calculated for the null model (unconditional) and the full model (conditional) [34].

5. A base LPM (LPM1), and with ICU providers as FE (LPM2)
   a. For the LPM, predictions were constrained within [0,1] using the linear discriminate function as suggested by Haggstrom [35, 36]: the LPM was estimated by ordinary least squares regression (OLS); the parameters were transformed (multiplied by $K = N/RSS$, where N is sample size, RSS is the residual sum of squares and K is > > 1); predicted probabilities were then generated using logistic regression [37]. The user written Stata command "reg2logit" [38] was utilised. Model indices were provided for this model ("LPM_ldm") and for the vanilla linear regression model with probabilities constrained to the [0,1] range ("LPM [0,1]").
6. Where of interest, predicted mortality probabilities were compared graphically using a limits of agreement (LOA) method, whereby the mean difference and the data were presented as paired differences plotted against pair-wise means. The user written Stata module "concord" was employed [39].

Model performance was assessed thus:

1. The traditional criteria of discrimination (receiver operator characteristic curve area, AUC) and calibration (Hosmer-Lemeshow (H-L) statistic). Although the H-L statistic will invariably be significant ($P < 0.1$ and H-L statistic > 15.99) in the presence of large N and increments to the grouping number (default = 10) of the H-L test have been recommended [40], the default grouping number was used.
   a. Calibration plots (observed binary responses versus predicted probabilities) were undertaken using the user-written Stata module "calibrationbelt" [41]. The relationship of predictions to the true probabilities of the event was formulated with a second logit regression model, based upon a polynomial transformation of the predictions, the degree of the polynomial (beginning with second order) being forwardly selected on the basis of a sequence of likelihood ratio tests. The deviation of the calibration belt from the line of identity is indicated by the reported P value (< 0.05).
2. The potential for overfitting, or shrinkage statistics (determined by in-sample and out-of-sample predictive bias and overfitting, expressed in percentages) was undertaken using the user-written Stata module "overfit" [18, 42]; that is, a focus on predictive calibration. Ten-fold cross-validation with 500 repeated iterations were used.
   a. Under conditions of non-applicability of the algorithm, a more traditional approach was used; development and validation model data sets were generated and various indices were generated on each data set using the user-written Stata module "pmcalplot" [43]: calibration-in-the-large [44], calibration slope, C-statistic for model discrimination and ratio of expected and observed events.
3. Model residual analysis was undertaken using the "binned residual" approach as recommended by Gelman and Hill [45] and implemented in Stata by Kasca [46]: the data was divided into categories (bins) based upon the fitted values and the average residual (observed minus expected value) versus the average fitted value was plotted for each bin; the boundary lines, computed as $2\sqrt{p(1-p)/n}$ where n was the number of points per bin, indicated ±2SE bounds, within which one would expect about 95% of the binned residuals to fall.
4. Model comparison was also undertaken by the Akaike Information Criterion (AIC), with the Bayesian Information Criterion (BIC) for non-nested models; lower values being optimal [47].
5. In view of the burgeoning literature on coefficient comparison between nested and non-nested non-linear probability models [48–50], we undertook full (X-Y) standardisation of logistic, probit and LPM (for both the full sample "LPM (all N)" and LPM [0,1]) coefficients using the "listcoef" Stata user-written command [51, 52]. Graphical display of the standardised coefficients utilised violin plots [53]: box plots incorporating estimated kernel density information via the user-written Stata command "vioplot" [54].
6. The Stata™ command "margins" was used to frame predictions under various scenarios [55, 56]; mortality effect over variables such as MV status was generated with due note of the overlapping 95% CI conundrum that such overlapping does not necessarily indicate lack of statistical difference [57, 58]. Although the analyses were performed using Stata ™ statistical software, similar functionality is provided in R statistical software [59].

Covariate endogeneity and selection bias

Endogeneity arises when there is a correlation between an explanatory variable and the regression error term(s), either in OLS (ordinary least squares) regression [60, 61] or probit and logit [62]; the causes being omitted variables, simultaneity (contemporaneous or past) and

measurement error [60]. The paradigmatic econometric model is the Heckman selection model [63]. The consequences of non-random assignment and (self) selection bias, in terms effect estimate bias, have also been well documented in the biomedical literature [64–66]. Predictor variable endogeneity and the impact of endogenous treatment assignment were formally addressed utilizing the Extended Regression (ERM; "eprobit") module of Stata™ statistical software [67]; in particular, the demonstration of a significant correlation between the error term of the variable in question and the error term of the dependent variable, hospital mortality. The method used by "eprobit" was to apply instrumental variables (IV), [68, 69]) which predict the endogenous variable(s) and have an outcome (mortality) effect via these endogenous variables [70], with robust standard errors ("vce (robust")) as recommended [71]. A third (unverifiable) assumption is that the IV-outcome association is unconfounded [72]. Using a potential outcomes scenario, the ventilation average treatment effect (ATE) and the average treatment effect of the treated (ATET, the mortality of those ventilated as opposed to the counterfactual mortality of these ventilated patients considered to be not ventilated) were estimated. Again, the "margins" command, suitably specified for "eprobit", was used to estimate various scenarios on the probability scale; in particular, comparisons were "fixed" such that for endogenous treatment assignment patients were compared assuming all were ventilated and then all were not-ventilated (a counter-factual scenario). The variance-covariance matrix was specified as "unconditional" [73]; that is, via the linearization method, non-fixed covariates were treated in a way that accounted for their having been sampled, allowing for heteroskedasticity or other violations of distributional assumptions and for correlation among the observations in the same manner as vce (robust).

Stata™ Version 16.1 was used for all analyses and statistical significance was ascribed at $P < 0.05$. For continuous variables, results are presented as mean (SD) unless otherwise indicated.

Results

Cohort description

The cohort comprised 92,693 patients from 124 intensive care units (ICU) in calendar year 2016; 17% of ICUs were metropolitan (non-tertiary), 32.5% in private, 6.5% rural / regional and 44% were tertiary, as defined in the ANZICS-APD data dictionary [74]. Patient mean age was 61.8 (SD 17.5) years, 41.6%.

were female and APACHE III score 54.5(SD 25.6); 43.7% were ventilated. ICU length of stay was 3.1(SD 4.5) days and HLOS was 11.8(SD 13.0) days. ICU and hospital mortality were 6.45% (95%CI: 6.30, 6.61) and 8.82% (95%CI: 8.64, 9.00) respectively.

Model performance: logit, probit and LPM

For the base logit (Additional file 1), probit and LPM_ldm models, the number of parameters at 110 was large but the shrinkage and overfitting indices did not indicate problematic overfitting (Table 1).

Similarly, the use of ICU providers as FE substantially increased the number of parameters but again there was no evidence of overfitting, although the specification of the FE logit model (Logit2) dominated the FE probit (Probit2). The "overid" module was not applicable to the RE models and a more conventional development / validation data set approach was undertaken; the RE logit model had superior performance, at least by information criteria (BIC). The unconditional ICC in the RE logit model was 0.201 indicating a modest patient correlation within ICUs; not surprisingly, the conditional ICC decreased to 0.018. Patient number for the LPM [0,1] model it was 68,264 as the generated probabilities were < 0 in 24,179 (35.4%) and > 1 in 250 (0.4%). There was little difference in the pattern of the residual graphs between the eight models, except for the vanilla probit model where there was more asymmetry about the null (zero) line as seen in Fig. 1.

Despite having a satisfactory discrimination (AUC: 0.884), the LPM [0,1] model demonstrated poor calibration as displayed by the lack of fit in the residual analysis (Fig. 1). The predicted probabilities of the vanilla logistic and LPM_ldm models were of some interest and were illustrated using a LOA graph (Fig. 2); the differences were exceptionally small, although in opposite directions for vanilla models versus those with ICU provider FE.

Model effect of potentially endogenous covariates

As consideration was also given (see below) to the impact of two potentially endogenous covariates (HLOS and a ROD score), both in log form, a summary of the effect of the addition of these two covariates upon model performance for logit (Logit1, Logit2 and Logit3) is presented in Table 2 and Fig. 3.

Small increments in the AUC and decrements in both AIC and BIC were seen for all logit models compared with the base models (Table 2). There was, however, substantial loss of model calibration and disturbances in residual distribution, more so for the addition to the base model of HLOS than for ROD.

Model coefficients

For the models Logit1, Probit1, LPM (all N) and LPM0, using robust SE [71], the fully standardised β coefficients are seen in Fig. 4. The LPM_ldm model (LPM1) was not used as the β coefficients were transformed.

Table 1 Model performance estimates for the base models

Model title	Logit1	Logit2	Probit1	Probit2	Logit3	Probit3	LPM1	LPM2	LPM0
Regression method	Logisitc	Logit1 + site FE	Probit	Probit1 + site FE	Logit 2 + RE	Probit2+RE	LPM_ldm	LPM_ldm+siteFE	LPM[0,1]
Indices									
Number of patients	92,693	92,693	92,693	92,693	92,693	92,693	92,693	92,716	68,264
Number of parameters	110	234	110	234	107	107	110	234	110
ROC AUC	0.915	0.917	0.915	0.917	0.917	0.917	0.912	0.915	0.884
H-L statistic; *P*-value	0.173	0.044	0.000	0.000	0.073	0.000	0.173	0.103	0.000
Out-of-sample shrinkage %	0.940	1.600	0.940	0.580			0.390	0.360	
In-sample-shrinkage %	0.380	0.360	0.380	−0.510			0.000	0.360	
Overfitting %	0.560	1.250	0.560	1.090			0.390	0.000	
Calibration belt: *P*-value	0.850	0.733	0.000	0.000	0.593	0.000	0.850	0.987	0.000
AIC	33,867.39	33,712.31	33,897.66	33,756.53	33,758.82	33,799.85	33,863.39	33,712.31	
BIC	34,792.25	35,665.78	34,803.62	35,710	34,674.21	34,715.24	34,769.35	35,665.78	
Development set									
CITL					−0.002	0.000			
C-slope					1.005	1.009			
AUC					0.917	0.915			
E:O ratio					1.001	1.000			
Validation set									
CITL					0.002	0.021			
C-slope					1.005	1.006			
AUC					0.916	0.916			
E:O ratio					0.989	0.987			
ICC: unconditional					0.201	0.154			
ICC: conditional					0.018	0.016			
ICC: unconditional					0.201	0.154			
ICC: conditional					0.018	0.016			

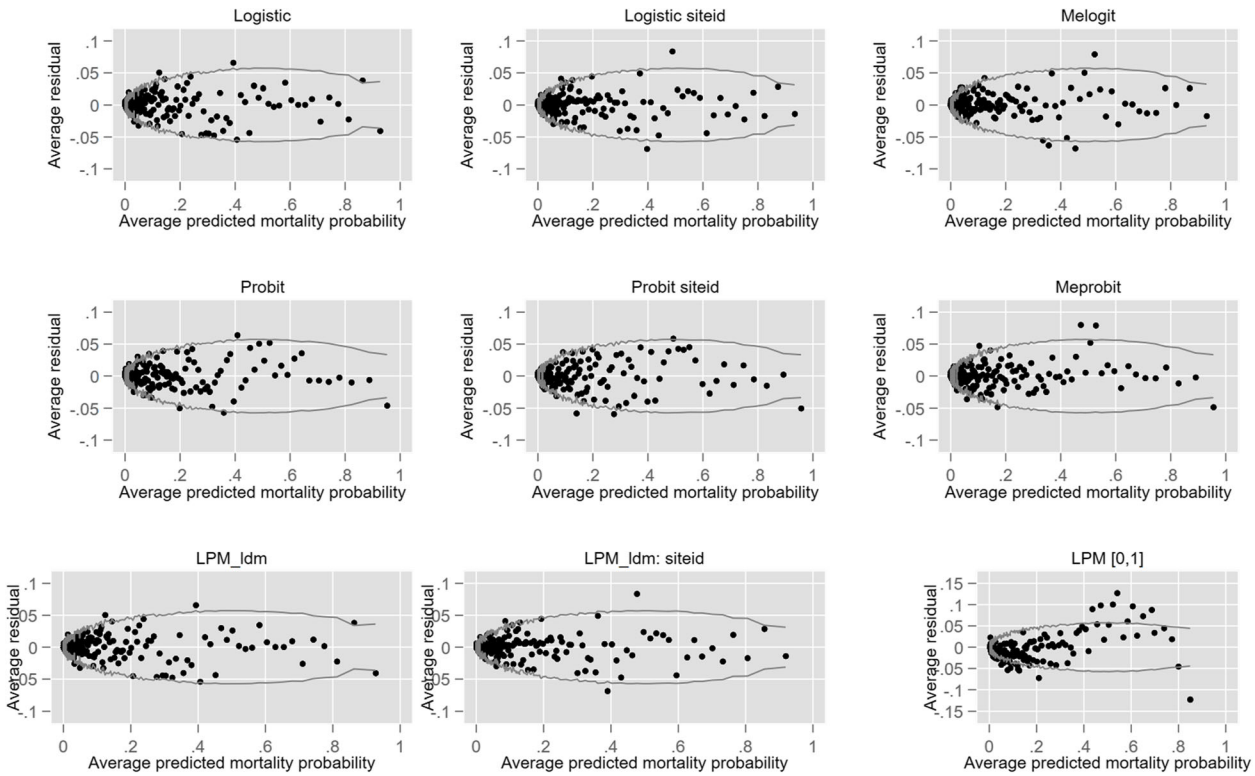

Fig. 1 Binned residual graphs for logistic, probit and LPM models

There was moderate conformity between the density distribution of the four models, but this belied a quantitative comparison using simple regression of the scalar values of the full (*X*-*Y*) standardised β coefficients ($n = 100$) with logistic as the comparator (Table 3).

There was a sizeable overall difference between the average scalar β model coefficients. Of interest, the number of model significant coefficients was 50 in the logit, 55 in the probit, 63 in the LPM (all N) and 69 in the LPM0.

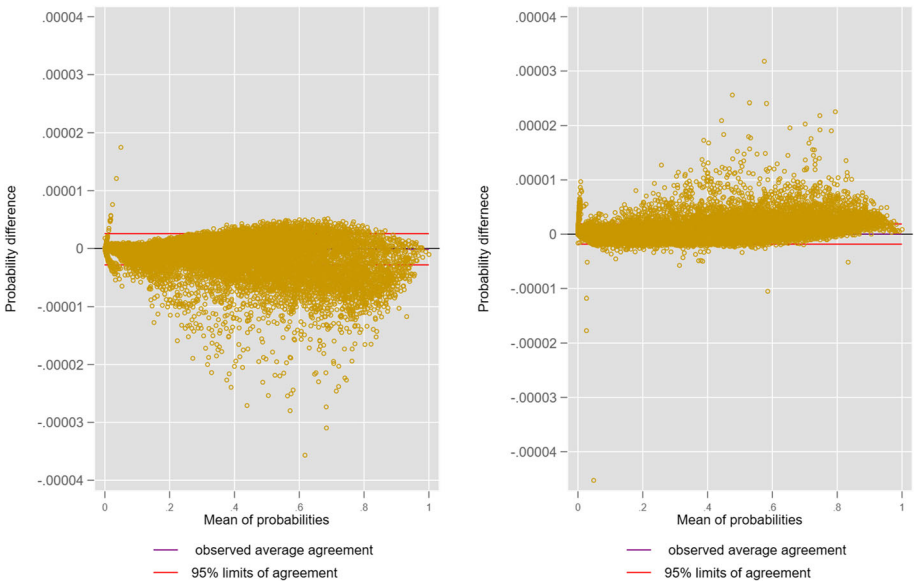

Fig. 2 LOA between logistic and LPM_ldm probabilities:left, base models; right ICU providers as FE ($N = 92{,}693$)

Table 2 Model performance for logit with addition of log HLOS and ROD (NC: not computed)

Model title	Logit1: lnday	Logit1: lnrod	Logit2: lnday	Logit2: lnrod	Logit3:lnday	Logit3:lnrod
Index						
ROC AUC	0.921	0.934	0.923	0.936	0.917	0.935
H-L statistic; P-value	0.000	0.060	0.000	0.013	0.000	0.054
Out-of-sample shrinkage %	0.940	0.020	−1.370	0.750		
In-sample-shrinkage %	0.380	−0.500	−2.350	−0.380		
Overfitting %	0.560	0.510	0.960	1.120		
Calibration belt: P-value	0.000	NC	0.000	0.000	0.000	0.000
AIC	31,306.41	30,574.13	31,077.06	30,492.24	31,148.98	30,523.52
BIC	322,421.1	31,489.52	33,039.96	32,455.96	32,073.81	31,448.36
Development set						
CITL					0.005	0.017
C-slope					1.005	1.004
AUC					0.923	0.935
E:O ratio					0.997	0.998
Validation set						
CITL					0.005	0.017
C-slope					1.005	1.004
AUC					0.923	0.935
E:O ratio					0.997	0.990

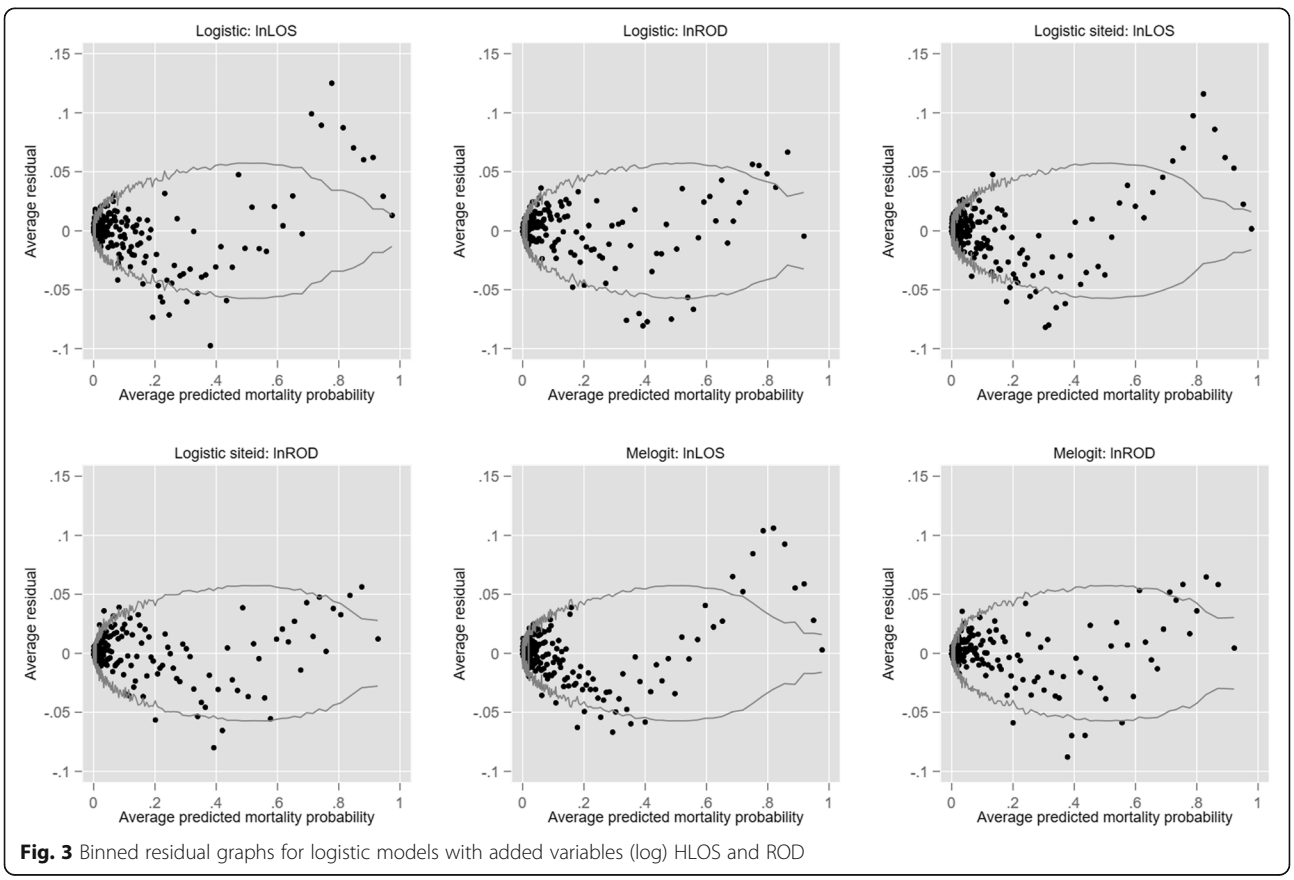

Fig. 3 Binned residual graphs for logistic models with added variables (log) HLOS and ROD

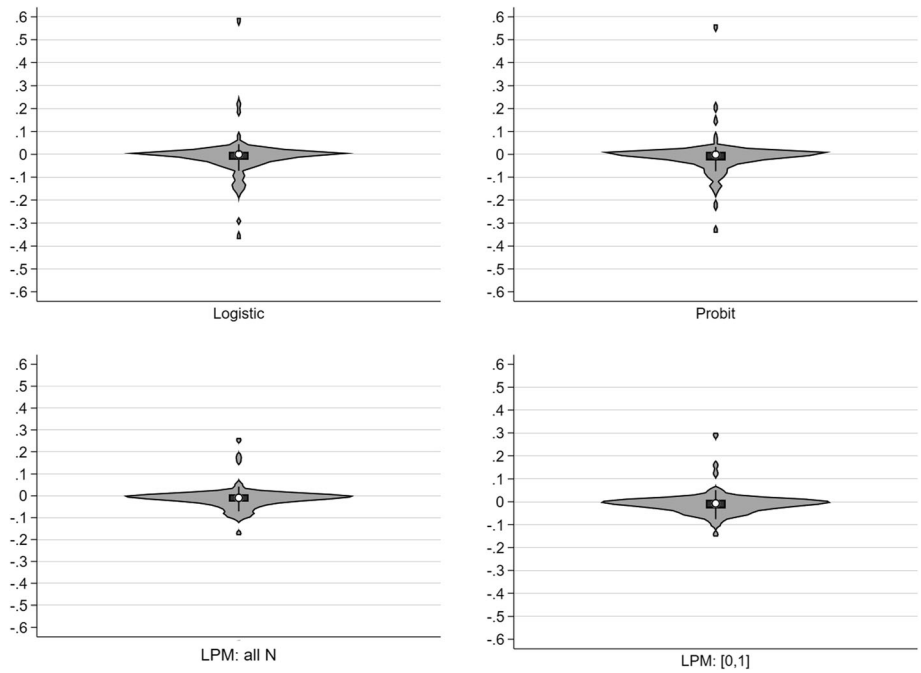

Fig. 4 Violin plots of model coefficients for logistic, probit and LPM models

Endogeneity and non-random assignment

APACHE III severity of illness score As the APACHE III score [2] was a key variable measuring patient severity of illness, the status of this variable with respect to endogeneity was tested using age, hospital level (4 level categorical variable) and APACHE III diagnosis (categorical variable denoting 28 collapsed APACHE III diagnostic codes; see Additional file 1) as IV. There was no evidence for endogeneity; error correlation of APACHE III score v mortality outcome: 0.000(– 1.000, 1.000, $p = 1.0000$).

Endogenous covariates

Models with both risk of death and HLOS as endogenous variables failed to converge and the use of ICU providers as instruments failed to yield marginal estimates after 36 h of computation. The attempt to estimate the MV effect over the span of HLOS and risk of death using margins was also unsuccessful due a nonsymmetric or highly singular matrix. For log HLOS and log ROD as endogenous variables, and MV status as an endogenous treatment assignment, the best model by information criteria used the APACHE III score, hospital level, APACHE III diagnostic categories and annual patient volume (as deciles) as IV, with a substantial reduction (up to 13%) of both AIC and BIC compared with models using a lesser number of IV.

Table 3 Standardised β coefficients, logit versus probit and LPM

Logit	β	*P*-value	95%CI: lower	95%CI: upper
Probit	1.204	0.000	1.186	1.221
LPM: all N	0.028	0.536	−0.061	0.116
LPM: [0,1]	−0.301	0.000	−0.400	−0.202

There was a significant correlation between the error terms of the dependent variable (hospital mortality) and both ventilation status and log HLOS, and between ventilation status and log HLOS, as seen in Table 4. The ATE and ATET were 5.38% (95%CI: 1.33, 9.44) and 4.55% (95%CI: 0.98, 8.13) respectively. For a comparable probit model with log HLOS added as an extra covariate (using the "margins" command), the ventilation mortality effect was 0.48 (95%CI: 0.10, 0.85).

Mechanical ventilation effect

Log HLOS The mortality MV effect over the span of the APACHE III score is shown in Fig. 5 with the log HLOS modelled as an endogenous variable; the comparator is the probit model with log HLOS as an added covariate to the base model. There is an apparent mortality increment across high APACHE III scores for non-ventilation in the probit model, but this is reversed in the "eprobit" model.

The ventilation mortality contrast (absolute difference, MV versus non-ventilated, y-axis: ± about the null difference of 0) for both models is seen in Fig. 6 and exhibits model differences with greater clarity. The mortality increment across high APACHE III

Table 4 Correlation of model error terms ".e" for mortality, ventilation and HLOS

Correlation	Estimate	Robust SE	z-value	*p*-value	95%CI_lower	95%CI_upper
Ventilation.e vs mortality.e	−0.248	0.085	−2.93	0.003	−0.405	−0.076
Log HLOS.e vs mortality.e	−0.315	0.007	−45.50	0.000	−0.328	−0.301
Log HLOS.e vs ventilation.e	0.119	0.005	24.56	0.000	0.109	0.128

scores (range 105–175) for non-ventilation in the probit model reached statistical significance but was quite small. The mortality contrast in the eprobit model was substantial across almost the entire range of APACHE III scores 5–175.

Log ROD For the log risk of death score modelled as an endogenous variable there was significant correlation between the error terms of the dependent model variable (hospital mortality) and both MV status and log risk of death, and between MV status and log risk of death, as seen in Table 5. The ATE and ATET were 3.07% (95%CI: − 0.28, 6.43) and 2.95% (95%CI: − 0.35, 6.24) respectively. For a comparable probit model with log risk of death score added as an extra covariate (using the "margins" command), the MV mortality effect was − 0.58% (95%CI: − 0.93, − 0.23).

The mortality MV effect of the above ERM model is shown in Fig. 7 with the log risk of death modelled as an endogenous variable; the comparator is a probit model with log risk of death as an added covariate to the base model. There was an apparent differential mortality increment for non-ventilation versus MV in the probit model at an APACHE III score of 85, but this reversal was not apparent in the eprobit model. The overall mortality effect of the added variable log risk of death in the probit model was quite modest compared with that of log HLOS.

The MV mortality contrast (MV versus non-ventilated) for both models is seen in Fig. 8, and again demonstrated the difference with more transparency. For the probit model there was a small mortality increment for non-ventilation across APACHE III score 75–195, but this was not reflected in the eprobit model where a marked ventilation mortality increment occurred across APACHE III scores 85–195.

For the "eprobi"t graphic displays, 95%CI span was greater than that of the probit.

Discussion

Of the eight models considered in predicting hospital mortality, logit regression (with or without ICU providers as FE) and RE logit dominated, more so using information criteria indices, in accordance with a recent extensive simulation study [75]. The LPM suffered from

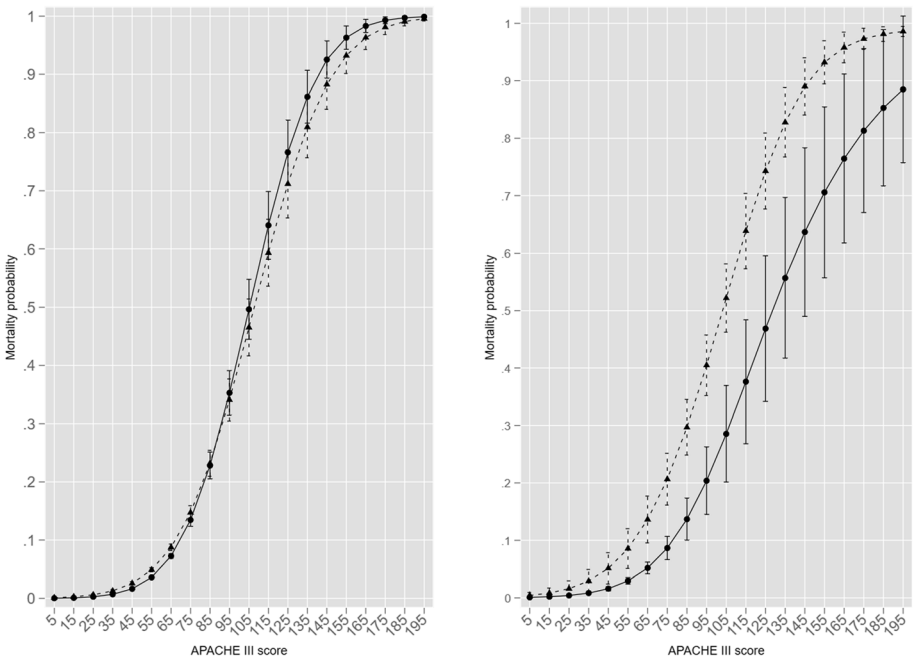

Fig. 5 Mortality MV effect over the span of the APACHE III score with the log HLOS modelled as an endogenous variable for probit model on left, "eprobit" (HLOS endogenous) on right. MV effect as black triangles circles and non-ventilation as solid black circles with 95%CI

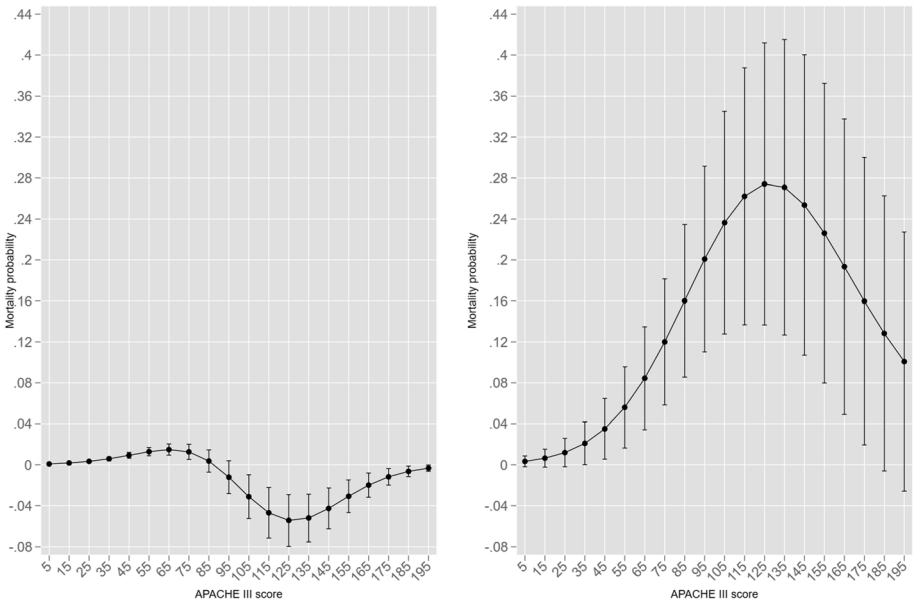

Fig. 6 Ventilation mortality contrast (absolute difference, MV versus non-ventilated, y-axis: ± about the null difference of 0) for probit model on left, "eprobit" (HLOS endogenous) on right. MV contrast effect (ventilated versus not-ventilated) as solid black circles with 95%CI

many predictions outside the unit interval, but the LPM_ldm model demonstrated, perhaps not surprisingly, a performance similar to that of the logit model. HLOS and the ROD score were demonstrated to be endogenous variables and patient ventilation status as an endogenous treatment assignment variable. Marked differences in the MV mortality effect was demonstrated between the vanilla probit and the eprobit models which were scenario dependent. These findings are further discussed.

Logistic regression as the preferred estimator

In biomedicine binary data analysis invariably proceeds using logistic regression in its various forms. Vach notes that "… probit regression and logistic regression give very similar results with respect to the order of the magnitude of the effect estimates" [76]; that is, the familiar scalar multiplier: $\hat{\beta}_{\text{Logit}} \simeq 1.6\hat{\beta}_{\text{Probit}}$ [77]. This belies the demonstrated differences in the fully standardised (X-Y) coefficients of the logistic and probit models in the current study. That the OR is difficult to interpret and is mis-conceived as a RR [78] has become a mantra. However, the interpretation of the probit coefficient is not immediately apparent, being the difference in Z score associated with each one-unit difference in the predictor variable. More generally, it must be noted that the three popular indices of risk, OR, RR and risk difference (RD), are neither related monotonically nor are interchangeable and the "… results based upon one index are generally not translatable into any of the others" [79]. A substantial literature in the social sciences has addressed the problem of coefficient comparison across groups in non-linear probability models, probit and logit, on the basis of unobserved heterogeneity, beginning with the seminal 1999 paper of Allison [80]. We do not pursue this theme [81] further, rather, submit that coefficient non-concordance is a function of the well described non-collapsibility of both odds ratios and probit regression coefficients [56, 82] and may be suitably resolved using marginal effects, including effect derivatives, on the probability scale [16, 83]: "… the output from non-linear models must be converted into marginal effects to be useful. Marginal effects are the (average) changes in the CEF [conditional expectation function: the expectation, or population average, of Y_i

Table 5 Correlation of model error terms ".e" for mortality, ventilation and ROD

Correlation	Estimate	Robust SE	z	*p*-value	95%CI_lower	95%CI_upper
Ventilation.e vs mortality.e	−0.235	0.089	−2.640	0.008	−0.399	−0.055
Log ROD.e vs mortality.e	0.471	0.008	58.29	0.000	0.455	0.486
Log ROD.e vs ventilation.e	−0.052	0.005	−10.35	0.000	−0.062	−0.042

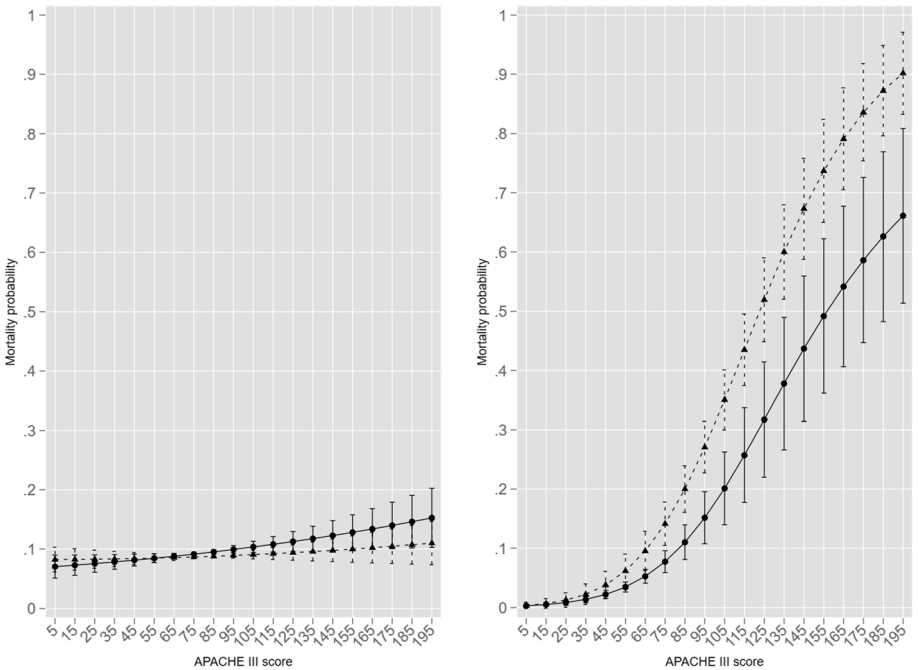

Fig. 7 Mortality MV effect over the span of the APACHE III score with the log ROD modelled as an endogenous variable for probit model on left, "eprobit" (ROD endogenous) on right. MV effect as black triangles and non-ventilation as solid black circles with 95%CI

(dependent variable) with X_i (covariate vector) held fixed] implied by a non-linear model. Without marginal effects, it's hard to talk about the impact on observed dependent variables" [84].

Most models achieved an AUC of ≥0.9 with between model AUC differences being small; the lack of import of such small AUC differences has been canvassed [85]. The primacy of AUC [86] in model assessment, as in machine learning, would appear to be misplaced [87] and calibration indices should be fully incorporated into analysis [88]. Certainly, the graphic residual analysis provided an extra dimension in revealing the effect on

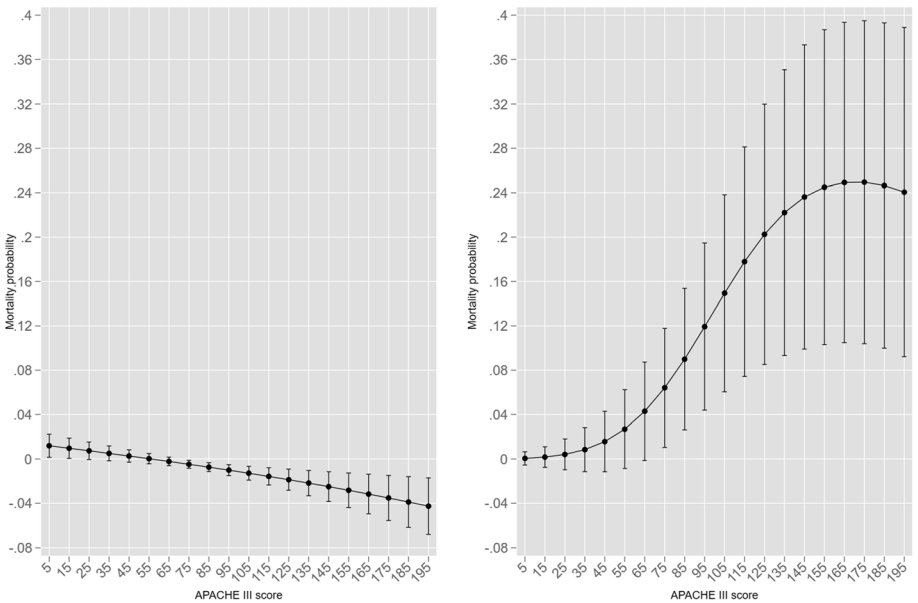

Fig. 8 Ventilation mortality contrast (absolute difference, MV versus non-ventilated, y-axis: ± about the null difference of 0) for probit model on left, "eprobit" (ROD endogenous) on right. MV contrast effect (ventilated versus not-ventilated) as solid black circles with 95%CI

model goodness-of-fit with the addition of the two suspect (see below) endogenous predictor variables, HLOS and ROD score (Table 1 and Fig. 2). The stability of the logit and probit FE estimation, with 123 extra parameters (Table 2 and Fig. 1), was reassuring. There has been considerable debate in both the econometric and statistical literature regarding performance (consistency) of the maximum likelihood estimator in the presence of FE, particularly large group numbers; the "incidental parameters problem" [89]; such concerns may be more apparent than real [90–93].

The choice between vanilla logistic regression (Logit1) and logistic regression with fixed site effects (or dummies [94], Logit 2) and random effects (Logit 3), would appear to depend upon purpose [95]. Transportable models, such as APACHE III [2] and the Australian and New Zealand Risk of Death model [31], eschew site fixed effects for logical reasons. The RE model is "sensible for modelling hierarchical data" [96], perhaps *de rigeur*, and with large data sets the computational demands of implementing, say, adaptive Gauss-Hermite quadrature, can be reasonably overcome by parallelisation, available in Stata™. The RE constraint of independence of provider effect (the random effects) from risk factors is often assumed, but it is plausible that such a correlation may "commonly occur" with consequent estimation bias [97]. Such constraint is not shared by high dimensional logistic FE models, which may have advantage [96], not the least of which is accounting for unobserved heterogeneity and, within the domain of profiling analysis, a smaller error for "exceptional" providers [96, 97]. Such conclusion was endorsed by Roessler et al. [98], who also noted the "sparse literature on fixed effects approaches". Correlated RE models, for instance the Mundlak approach, are estimable for binary outcomes within the generalised linear mixed model framework (GLMM), as in the user-written Stata command "xthybrid" [99] and has been used in hospital outcome analysis [100]. Based upon our findings (Table 2 and Fig. 1), a probit RE model (Probit 3) had no advantage over the logit RE (Logit 3) and the inherent complexities of probit coefficient interpretation would not recommend it, albeit marginal effects on the probability scale are transparent. Moreover, the explained variance (McKelvey & Zavoina [101]) of the two RE models favoured the logit (R^2_{dichot} = 0.62 (logit) versus 0.52 (probit)), where $R^2_{\text{dichot}} = \frac{\sigma^2_F}{\sigma^2_F+\tau^2_0+\sigma^2_R}$; σ^2_F is the linear predictor variance, τ^2_0 is the intercept variance and σ^2_R is the level one residual variance (fixed at $\pi^2/3 = 3.29$ for the logit and 1 for the probit).

With respect to the profiling paradigm, which was not formally addressed, the contemporary choice between so-called "caterpillar plots" of provider effect estimates (plus 95% CI) [100] and funnel plots [102] would appear to have favoured the latter. The confidence intervals of the caterpillar plot "… are appropriate for testing single hypotheses … They are not appropriate for drawing inference about whether a given hospital's performance is different from a set of their peers' performances" [100]. This belies the difference between marginal and simultaneous confidence sets for ranks, whereby simultaneous confidence sets are robust to the latter inferential comparisons [103]. Such confidence sets for ranks have been implemented as "csranks" in both the Stata and R statistical environments.

The use of the LPM for binary data has generated controversy in the social science and econometric literature, but not in the biomedical; perhaps not surprisingly. However, these issues are addressed here. Firstly, a distinction must be made between the LPM as a preferred model versus its use as an alternative to logistic regression because of OR interpretational differences [104, 105]. We have alluded to this problem above, but it is disconcerting to find in a recent paper that the authors [104], whilst sympathetic to the average marginal effect (AME) as satisfying the criteria of comparability across both models and studies, quote the paper of Mood [56], published in 2010, to the effect that "deriving AME from logistic regression is just a complicated detour". They conclude that "… we explore this procedure no further given its similarity to OLS results and the need for special-purpose routines to no notable advantage" and proceed to offer the LPM and a Poisson working model to compute risk differences and risk ratios, respectively. This ignores the fact that both risk differences and risk RR are collapsible metrics, as opposed to OR and probit coefficients. In Stata™, the "margins" command, introduced in Version 11 (July 2009), is a seamlessly integrated post-estimation tool, albeit it has undergone relevant computational revisions [16].

The question of bias and inconsistency of LPM estimates is somewhat moot: Horace and Oaxaca [14] argued from a theoretical perspective that the LPM was an inconsistent and biased estimator; simulation studies [49, 104, 105] suggest that LPM coefficient estimates were similar in magnitude and significance to that of logistic and probit regression but may be sensitive to continuous covariate distributional shape [106]. In finite examples, as in this study, such similarity was not fully achieved despite using robust standard errors to correct for LPM heteroscedasticity [49]; see Results. Model choice is properly determined by analytic purpose [9]. If outcome probability generation constrained to the unit interval is of importance, for instance the calculation of provider standardised mortality ratios, then the LPM cannot be endorsed, despite recommendations for prediction truncation, which may dramatically reduce study number, 26% in our data set, and converting continuous

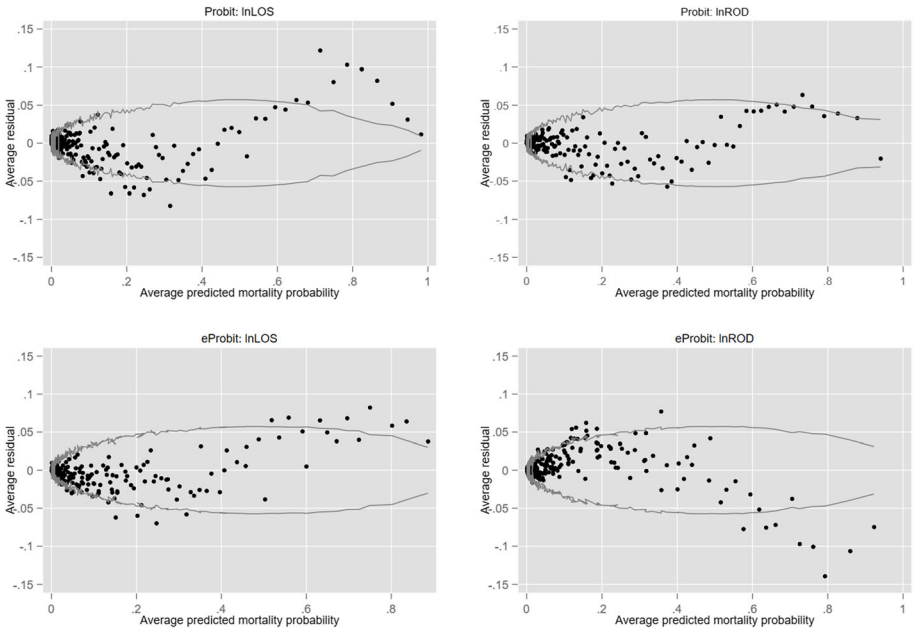

Fig. 9 Model residual analysis: probit (upper panels) and "eprobit" (lower panels) for (log) HLOS and ROD as endogenous variables

variables to factor levels [9, 48, 49, 54, 105]. The utility of a command such as "re2logit" for the purpose of generating probabilities from a LPM consonant with that of logit appears unclear. The method assumes multivariate normality which would not appear to be a fatal weakness [107, 108] and although complete or quasi-complete separation may occur with logistic regression [109, 110] and not with LPM, it was not observed in the current large N study [111]. Separation in logistic regression has been addressed from within the Social and Political Science domains in terms of advocacy for the LPM [94, 112] based upon estimation bias due to data omission. Under conditions of sparse data and separation alternate estimators are currently available, such as Firth's penalised logit and the Mundlak correlated RE formulation which do not incur the prediction penalty of the LPM [94, 113–115]. There may be domain specific preference for the LPM: "… while a non-linear model may fit the CEF for LDVs [limited dependent variables; in this case, binary] more closely than a linear model, when it comes to marginal effects, this probably matters little" [84]. Where probability generation is required, "reg2logit" provides a useful addendum [37].

Endogeneity

Endogeneity, as opposed to exogeneity, is conventionally ascribed to an explanatory variable (x), if the stochastic error (ε) in modelling the dependent variable (y) is not independent of x; that is, if $E(\varepsilon|\, x) \neq 0$, then $E(y|\, x, \varepsilon) \neq E(y|\, x)$ [116]. The causes of endogeneity include omitted variables [63], measurement error, simultaneity (current or past), autocorrelated errors and sample selection [17]; the end result being biased and inconsistent estimates [70]. Endogeneity may occur in the presence [117] or absence of unobserved heterogeneity [118] and is to be distinguished from confounding; endogeneity articulated as

Table 6 Model performance indices, vanilla probit vs "eprobit"

Model	Probit:lnHLOS	Probit: lnROD	eProbit:lnHLOS	eProbit: lnROD
Index				
ROC AUC	0.921	0.934	0.923	0.930
H-L statistic; P-value	0.000	0.000	0.000	0.013
Calibration belt: *P*-value	0.000	0.000	0.001	0.001
CITL	−0.017	− 0.008	−0.050	0.059
C-slope	1.014	1.004	1.096	0.911
AUC	0.920	0.934	0.917	0.913
E:O ratio	1.010	1.004	1.033	0.967

Table 7 Main effects parameter coefficients (*dy/dx*) for probit and eprobit

	Probit				eProbit			
	dy/dx	*P*-value	95% CI (lower)	95% CI (upper)	dy/dx	*P*-value	95% CI (lower)	95% CI (upper)
Age_centered	0.0001	0.5810	−0.0002	0.0004	0.0001	0.3890	−0.0002	0.0004
APIII score_centered	0.0002	0.1570	−0.0001	0.0004	0.0022	0.0000	0.0020	0.0025
Gender								
Female	0.0000				0.0000			
Male	0.0001	0.9700	−0.0029	0.0030	0.0000	0.9810	− 0.0026	0.0026
AP III diagnostic codes								
Cardiovascular_medical	0.0000				0.0000			
Respiratory medical	−0.0056	0.0330	−0.0107	− 0.0005	0.0038	0.4040	−0.0052	0.0128
Liver_GIS_medical	−0.0025	0.5190	−0.0100	0.0050	−0.0166	0.0120	−0.0296	−0.0036
CNS_medical	0.0123	0.0000	0.0063	0.0183	−0.0030	0.6480	−0.0158	0.0098
Sepsis	−0.0103	0.0000	−0.0156	−0.0050	− 0.0128	0.1330	− 0.0296	0.0039
Trauma	0.0050	0.2280	−0.0031	0.0132	−0.0213	0.0180	−0.0388	−0.0037
Metabolic Hormonal	0.0006	0.9110	−0.0107	0.0120	−0.1066	0.0000	−0.1178	−0.0955
Haematologic	−0.0065	0.4590	−0.0237	0.0107	0.0352	0.2230	−0.0214	0.0918
Renal_GUS	−0.0244	0.0000	−0.0349	−0.0140	− 0.0839	0.0000	− 0.0975	−0.0702
Other medical disorders	0.0142	0.3020	−0.0128	0.0413	−0.0472	0.0070	−0.0815	−0.0130
Musculoskeletal / Skin	−0.0074	0.5730	−0.0329	0.0182	−0.0583	0.0000	−0.0879	−0.0287
Cardio-Vascular surgery	−0.0006	0.9070	−0.0103	0.0091	−0.0710	0.0000	−0.0875	−0.0545
Thoracic surgery	0.0264	0.0070	0.0071	0.0457	−0.0536	0.0000	−0.0742	−0.0329
GIS surgery	−0.0015	0.6730	−0.0085	0.0055	−0.0481	0.0000	−0.0615	−0.0346
CNS surgery	0.0141	0.0150	0.0027	0.0254	0.0168	0.1070	−0.0036	0.0373
Traumatic/Orthopaedic surgery	0.0004	0.9470	−0.0107	0.0114	−0.0475	0.0000	−0.0664	−0.0286
Renal_GUS surgery	−0.0176	0.1000	−0.0387	0.0034	−0.0998	0.0000	−0.1148	−0.0848
Gynaecological	0.0834	0.0540	−0.0014	0.1681	−0.1142	0.0000	−0.1380	−0.0904
Musculoskeletal / Skin Surgery	0.0004	0.9530	−0.0138	0.0147	−0.0672	0.0000	−0.0814	−0.0530
Metabolic Surgery	0.0427	0.2800	−0.0348	0.1201	−0.0854	0.0000	−0.1286	−0.0421
Cardiovascular surgery elective	−0.0073	0.1330	−0.0168	0.0022	−0.1179	0.0000	−0.1302	−0.1056
Thoracic surgery elective	−0.0074	0.5570	−0.0320	0.0172	−0.0839	0.0000	−0.1056	−0.0622
GIS surgery elective	−0.0179	0.0040	−0.0300	−0.0059	− 0.0885	0.0000	− 0.0998	−0.0771
CNS surgery elective	0.0227	0.0980	−0.0042	0.0496	−0.0685	0.0000	−0.0916	−0.0453
Traumatic/Orthopaedic surgery el	−0.0345	0.1880	−0.0857	0.0168	−0.0946	0.0000	−0.1302	−0.0591
Renal_GUS surgery elective	0.0220	0.2250	−0.0135	0.0575	−0.0972	0.0000	−0.1185	−0.0760
Gynaecological surgery elective	−0.0491	0.1040	−0.1084	0.0101	−0.1264	0.0000	−0.1414	−0.1114
Musculoskeletal / Skin Surgery el	−0.0058	0.5760	−0.0262	0.0146	−0.0982	0.0000	−0.1151	−0.0813
Annual volume_deciles								
Base	0.0000				0.0000			
2	0.0016	0.6740	−0.0058	0.0090	0.0010	0.7740	−0.0058	0.0077
3	0.0092	0.0180	0.0016	0.0167	0.0064	0.0690	−0.0005	0.0134
4	0.0017	0.7270	−0.0080	0.0114	0.0085	0.1110	−0.0020	0.0190
5	−0.0008	0.8430	−0.0090	0.0074	0.0023	0.5740	−0.0057	0.0104
6	0.0047	0.2750	−0.0038	0.0133	0.0050	0.2680	−0.0038	0.0138
7	0.0032	0.4300	−0.0048	0.0112	0.0129	0.0040	0.0042	0.0217

Table 7 Main effects parameter coefficients (*dy/dx*) for probit and eprobit *(Continued)*

	Probit				eProbit			
	dy/dx	*P*-value	95% CI (lower)	95% CI (upper)	dy/dx	*P*-value	95% CI (lower)	95% CI (upper)
8	0.0020	0.6310	−0.0063	0.0104	0.0066	0.1300	−0.0020	0.0152
9	−0.0030	0.5860	−0.0139	0.0078	−0.0042	0.4180	−0.0145	0.0060
10	−0.0047	0.2280	−0.0125	0.0030	−0.0022	0.6290	−0.0112	0.0068
Hospital classification								
Metropolitan	0.0000				0.0000			
Private	0.0145	0.0000	0.0089	0.0201	0.0124	0.0000	0.0071	0.0178
Rural / Regional	0.0034	0.2420	−0.0023	0.0092	0.0062	0.0260	0.0007	0.0117
Tertiary	0.0164	0.0000	0.0120	0.0209	0.0117	0.0010	0.0049	0.0184
Ventilation status								
Not-ventilated	0.0000				0.0000			
Ventilated	−0.0058	0.0010	−0.0093	−0.0023	0.0309	0.0710	−0.0026	0.0645

"confounding by indication" would appear to be a contradiction [119]. Large sample size ('big data") does not limit the consequences of endogeneity [120, 121]. In the current analysis, where some of the effect of the error term(s) was attributed to the explanatory variable, the optimal course of action would be to "purge" the model of the correlation between the explanatory variable and the error term [19]; to wit, the use of the "eprobit" estimator.

Variables may be conceived by the analyst as endogenous, but it is not evident that in biomedical observational data analysis that particular attention has focused on the modelling consequences [120, 122]. The adverse effect of mechanical ventilation per se has been incorporated seamlessly into mortality prediction models without adjustment for patient selection; that is, a non-random physician treatment decision. The use of mortality probability as an independent variable in mortality prediction would appear to qualify as the regression of a variable upon its components [26]. Prolonged hospital length of stay is conventionally associated with mortality increment but displays a recursive effect or (current) simultaneity. With a large data set, it may not be obvious why the inclusion of one of the two endogenous covariates (HLOS or ROD score) should produce substantial loss of model calibration and disturbances in residual distribution; this may be a signal of an over-parameterised model and / or covariate endogeneity. We previously [123] demonstrated endogeneity of duration of mechanical ventilation in the critically ill ([123], Supplementary Appendix, figure E3) and performance of a tracheostomy as a non-random treatment variable, giving support to the notion that in the critical care domain, the effect of data variables realising complex patient-physician interaction may be endogenous. Similar studies have addressed the endogeneity of ICU admission decisions [124] and therapeutic titration based upon patient severity of illness [120, 121].

Within the limits detailed in Results, substantial differences in both magnitude and direction of the ventilatory effect were demonstrated between the vanilla probit and the "eprobit" models by virtue of accounting for endogeneity. Contrast graphics also possessed merit in that they more clearly demonstrated effect differences obfuscated by seemingly overlapping 95% CI. The difference between the predicted marginal ventilation effects of vanilla probit and eprobit were not accompanied by any substantive improvement in model fit of "eprobit" versus probit and model residual analysis did not substantially favour "eprobit", as seen in Fig. 9 (see also Duke and co-authors [123], Supplementary Appendix, figure E1 AND E2).

Similarly, model performance indices were not substantially different as seen in Table 6.

Not surprisingly, the parameter coefficients for the two models were different, in magnitude and direction, as shown in Table 7 for lnROD considered as an endogenous variable. Estimates, as response derivatives (dy/dx), are displayed for main effects only.

The IV paradigm is not without its limitations [21, 125] and has been subject to recent theoretical re-evaluation from within its archetypal domain, econometrics [126]. Such reviews have been relatively silent on the use of IV with binary outcome models [68, 125], although the LPM has been recommended [127]. The status of IV logistic regression, not implemented in current Stata™, was formally addressed by Foster in 1997 [128] using the Generalized Method of Moments, and more recently by two-stage residual inclusion estimation [129, 130], a preferred method in Mendelian randomisation [131], where identification of causal risk factors is the focus, rather than precise effect estimation [132]. This

being said, two-stage residual inclusion has been shown to be a consistent estimator [129, 130]. IV logistic regression has seen implementation within the R statistical framework in the "naivreg" [133] and "ivtools" [134] packages.

Conclusions

For modelling large scale binary outcome data, logistic regression, particularly the RE model, was the preferred estimator compared with probit and the LPM. The latter estimator cannot be recommended for probability generation. Endogeneity was demonstrated for hospital length of stay, risk of death and for MV treatment status. Accounting for endogeneity produced markedly different effect estimates about patient ventilation status compared with conventional methods. Exploration of and adjustment for endogeneity should be incorporated into modelling strategies, failure to do so may produce results that are "… less likely to be roughly right than they are to be precisely wrong" [120].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-021-01251-8.

Additional file 1.

Acknowledgments

The authors and the ANZICS CORE management committee would like to thank clinicians, data collectors and researchers at the following contributing sites:
Alfred Hospital ICU, Alice Springs Hospital ICU, Armadale Health Service ICU, Austin Hospital ICU, Ballarat Health Services ICU, Bankstown-Lidcombe Hospital ICU, Bendigo Health Care Group ICU, Blacktown Hospital ICU, Box Hill Hospital ICU, Bunbury Regional Hospital ICU, Bundaberg Base Hospital ICU, Caboolture Hospital ICU, Cabrini Hospital ICU, Cairns Hospital ICU, Calvary Adelaide Hospital ICU, Calvary Hospital (Canberra) ICU, Calvary Hospital (Lenah Valley) ICU, Calvary Mater Newcastle ICU, Campbelltown Hospital ICU, Canberra Hospital ICU, Concord Hospital (Sydney) ICU, Dandenong Hospital ICU, Epworth Eastern Private Hospital ICU, Epworth Freemasons Hospital ICU, Epworth Hospital (Richmond) ICU, Fiona Stanley Hospital ICU, Flinders Medical Centre ICU, Flinders Private Hospital ICU, Footscray Hospital ICU, Frankston Hospital ICU, Gold Coast Private Hospital ICU, Gold Coast University Hospital ICU, Gosford Hospital ICU, Gosford Private Hospital ICU, Grafton Base Hospital ICU, Hervey Bay Hospital ICU, Hornsby Ku-ring-gai Hospital ICU, Ipswich Hospital ICU, John Fawkner Hospital ICU, John Flynn Private Hospital ICU, John Hunter Hospital ICU, Joondalup Health Campus ICU, Knox Private Hospital ICU, Latrobe Regional Hospital ICU, Launceston General Hospital ICU, Lismore Base Hospital ICU, Liverpool Hospital ICU, Logan Hospital ICU, Lyell McEwin Hospital ICU, Mackay Base Hospital ICU, Macquarie University Private Hospital ICU, Manly Hospital & Community Health ICU, Maroondah Hospital ICU, Mater Adults Hospital (Brisbane) ICU, Mater Health Services North Queensland ICU, Mater Private Hospital (Brisbane) ICU, Mater Private Hospital (Sydney) ICU, Melbourne Private Hospital ICU, Monash Medical Centre-Clayton Campus ICU, Mount Hospital ICU, Mulgrave Private Hospital ICU, Nambour General Hospital ICU, National Capital Private Hospital ICU, Nepean Hospital ICU, Newcastle Private Hospital ICU, Noosa Hospital ICU, North Shore Private Hospital ICU, Northeast Health Wangaratta ICU, Norwest Private Hospital ICU, Orange Base Hospital ICU, Peninsula Private Hospital ICU, Pindara Private Hospital ICU, Prince of Wales Hospital (Sydney) ICU, Prince of Wales Private Hospital (Sydney) ICU, Princess Alexandra Hospital ICU, Queen Elizabeth II Jubilee Hospital ICU, Redcliffe Hospital ICU, Robina Hospital ICU, Rockhampton Hospital ICU, Rockingham General Hospital ICU, Royal Adelaide Hospital ICU, Royal Brisbane and Women's Hospital ICU, Royal Darwin Hospital ICU, Royal Hobart Hospital ICU, Royal Melbourne Hospital ICU, Royal North Shore Hospital ICU, Royal Perth Hospital ICU, Royal Prince Alfred Hospital ICU, Shoalhaven Hospital ICU, Sir Charles Gairdner Hospital ICU, South West Healthcare (Warrnambool) ICU, St Andrew's Hospital (Adelaide) ICU, St Andrew's Hospital Toowoomba ICU, St Andrew's War Memorial Hospital ICU, St George Hospital (Sydney) CICU, St George Hospital (Sydney) ICU, St George Private Hospital (Sydney) ICU, St John Of God Health Care (Subiaco) ICU, St John Of God Hospital (Geelong) ICU, St John Of God Hospital (Murdoch) ICU, St Vincent's Private Hospital Northside ICU, St Vincent's Hospital (Melbourne) ICU, St Vincent's Hospital (Sydney) ICU, St Vincent's Hospital (Toowoomba) ICU, St Vincent's Private Hospital (Sydney) ICU, St Vincent's Private Hospital Fitzroy ICU, Sunshine Hospital ICU, Sutherland Hospital & Community Health Services ICU, Sydney Adventist Hospital ICU, Tamworth Base Hospital ICU, The Memorial Hospital (Adelaide) ICU, The Northern Hospital ICU, The Prince Charles Hospital ICU, The Queen Elizabeth (Adelaide) ICU, The Wesley Hospital ICU, Toowoomba Hospital ICU, Townsville University Hospital ICU, Tweed Heads District Hospital ICU, University Hospital Geelong ICU, Wagga Wagga Base Hospital & District Health ICU, Warringal Private Hospital ICU, Westmead Hospital ICU, Westmead Private Hospital ICU, Wollongong Hospital ICU.

Authors' contributions

JLM: study design, data collection and analysis, drafting of the manuscript, revising the manuscript, interpretation of results. JDS: review of data analysis, revising the manuscript, interpretation of results. GJD: review of data analysis, revising the manuscript, interpretation of results. All authors had access to the data and to the (Stata) analytic command files and approved the submitted manuscript.

Funding

Local Intensive Care Unit funds only.

Availability of data and materials

The dataset is the property of the ANZICS CORE and contributing ICUs and is not in the public domain. Access to the data by researchers, submitting ICUs, jurisdictional funding bodies and other interested parties is obtained under specific conditions and upon written request ("ANZICS CORE Data Access and Publication Policy.pdf", http://www.anzics.com.au/Downloads/ANZICS%20CORE%20Data%20Access%20and%20Publication%20Policy%20July%202017.pdf).

Declarations

Ethics approval and consent to participate

Access to the data was granted by the Australian and New Zealand Intensive Care Society (ANZICS)) Centre for Outcomes & Resource Evaluation (CORE) Management Committee in accordance with standing protocols; local hospital (The Queen Elizabeth Hospital) Ethics of Research Committee waived the need for patient consent to use their data in this study. The data set was anonymized before release to the authors by ANZICS CORE custodians of the database. The dataset is the property of the ANZICS CORE and contributing ICUs and is not in the public domain.

Competing interests

The authors declare no competing interests

Author details

[1]Department of Intensive Care Medicine, The Queen Elizabeth Hospital, Woodville, Australia. [2]Department of Critical Care Medicine, St Vincent's Hospital (Melbourne), Fitzroy, Australia. [3]Intensive Services, Eastern Health, Box Hill, Australia.

Received: 9 January 2021 Accepted: 9 March 2021
Published online: 21 June 2021

References

1. Power G, Harrison DA. Why try to predict ICU outcomes? Curr Opin Crit Care. 2014;20(5):544–9. https://doi.org/10.1097/MCC.0000000000000136.
2. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system. Risk prediction of hospital mortality

for critically ill hospitalized adults. Chest. 1991;100(6):1619–36. https://doi.org/10.1378/chest.100.6.1619.
3. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006;34(5):1297–310. https://doi.org/10.1097/01.CCM.0000215112.84523.F0.
4. Solomon P, Kasza J, Moran J. ANZICS CfOaRE: identifying unusual performance in Australian and New Zealand intensive care units from 2000 to 2010. BMC Med Res Methodol. 2014;14(1):53. https://doi.org/10.1186/1471-2288-14-53.
5. Hilbe J. Logistic regression models. Boca Raton: Taylor & Francis Group; 2009. https://doi.org/10.1201/9781420075779.
6. Bliss CI. The method of probits. Science. 1934;79(2037):38–9. https://doi.org/10.1126/science.79.2037.38.
7. Berkson J. Why I prefer logits to probits. Biometrics. 1951;7(4):327–9. https://doi.org/10.2307/3001655.
8. Cameron AC, Trivedi PK. Binary outcome models. In: Microeconometrics Using Stata: Revised Edition. College Station: Stata Press; 2010. p. 459–89.
9. Hellevik O. Linear versus logistic regression when the dependent variable is a dichotomy. Qual Quant. 2009;43(1):59–74. https://doi.org/10.1007/s11135-007-9077-3.
10. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection Bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA. 2007;297(3):278–85. https://doi.org/10.1001/jama.297.3.278.
11. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1–W73. https://doi.org/10.7326/M14-0698.
12. Harrell FE Jr. Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis. 2nd ed. New York: Springer International Publishing; 2015. https://doi.org/10.1007/978-3-319-19425-7.
13. Harrison DAP, Brady ARM, Parry GJP, Carpenter JRD, Rowan KD. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Crit Care Med. 2006;34(5):1378–88. https://doi.org/10.1097/01.CCM.0000216702.94014.75.
14. Horrace WC, Oaxaca RL. Results on the bias and inconsistency of ordinary least squares for the linear probability model. Econ Lett. 2006;90(3):321–7. https://doi.org/10.1016/j.econlet.2005.08.024.
15. Chen G, Tsurumi H. Probit and Logit model selection. Commun Stat. 2010;40(1):159–75. https://doi.org/10.1080/03610920903377799.
16. Bland JR, Cook AC. Random effects probit and logit: understanding predictions and marginal effects. Appl Econ Lett. 2019;26(2):116–23. https://doi.org/10.1080/13504851.2018.1441498.
17. Qin D. Let's take the bias out of econometrics. J Econ Methodol. 2019;26(2):81–98. https://doi.org/10.1080/1350178X.2018.1547415.
18. Bilger M, Manning WG. Measuring overfitting in nonlinear models: a new method and an application to health expenditures. Health Econ. 2015;24(1):75–85. https://doi.org/10.1002/hec.3003.
19. Briscoe J, Akin J, Guilkey D. People are not passive acceptors of threats to health: endogeneity and its consequences. Int J Epidemiol. 1990;19(1):147–53. https://doi.org/10.1093/ije/19.1.147.
20. Hazlett C. Estimating causal effects of new treatments despite self-selection: the case of experimental medical treatments. J Causal Inference. 2019;7:1.
21. Hernan MA, Robins JM. Instruments for causal inference: an Epidemiologist's dream? Epidemiology. 2006;17(4):360–72. https://doi.org/10.1097/01.ede.0000222409.00878.37.
22. Hernan MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health. 2006;60(7):578–86. https://doi.org/10.1136/jech.2004.029496.
23. Moran J, Solomon P. A review of statistical estimators for risk-adjusted length of stay: analysis of the Australian and new Zealand intensive care adult patient data-base, 2008-2009. BMC Med Res Methodol. 2012;12(1):68. https://doi.org/10.1186/1471-2288-12-68.
24. Knaus WA, Wagner DP, Zimmerman JE, Draper EA. Variations in mortality and length of stay in intensive care units. Ann Intern Med. 1993;118(10):753–61. https://doi.org/10.7326/0003-4819-118-10-199305150-00001.
25. Render ML, Kim HM, Deddens J, Sivaganesin S, Welsh DE, Bickel K, et al. Variation in outcomes in veterans affairs intensive care units with a computerized severity measure. Crit Care Med. 2005;33(5):930–9. https://doi.org/10.1097/01.CCM.0000162497.86229.E9.
26. Basu AP, Manning WGP. Issues for the Next Generation of Health Care Cost Analyses. Med Care. 2009;47(7_Supplement_1):S109–14.
27. Stow PJ, Hart GK, Higlett T, George C, Herkes R, McWilliam D, et al. Development and implementation of a high-quality clinical database: the Australian and new Zealand Intensive Care Society adult patient database. J Crit Care. 2006;21(2):133–41. https://doi.org/10.1016/j.jcrc.2005.11.010.
28. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.
29. Bian J, Buchan I, Guo Y, Prosperi M. Statistical thinking, machine learning. J Clin Epidemiol. 2019;116:136–7. https://doi.org/10.1016/j.jclinepi.2019.08.003.
30. Van Calster B, Verbakel JY, Christodoulou E, Steyerberg EW, Collins GS. Statistics versus machine learning: definitions are interesting (but understanding, methodology, and reporting are more important). J Clin Epidemiol. 2019;116:137–8. https://doi.org/10.1016/j.jclinepi.2019.08.002.
31. Paul E, Bailey M, Pilcher D. Risk prediction of hospital mortality for adult patients admitted to Australian and New Zealand intensive care units: development and validation of the Australian and New Zealand risk of death model. J Crit Care. 2013;28(6):935–41. https://doi.org/10.1016/j.jcrc.2013.07.058.
32. Moran JL, Solomon PJ. (ANZICS) ftACfOaRECotAaNZICS: fixed effects modelling for provider mortality outcomes: analysis of the Australia and New Zealand intensive care society (ANZICS) adult patient database. PLoS One. 2014;9:e102297. https://doi.org/10.1371/journal.pone.0102297.
33. Gelman A, Hill J. Data analysis using regression and Multilelvel/ hierarchal models. New York: Cambridge University Press; 2007.
34. Rabe-Hesketh S, Skrondal A. Random intercept models with covariates. In: Multilevel and longitudinal modeling using Stata volume 1: continuous responses. 3rd ed. College Station, TX: Stata Press; 2012. p. 123–71.
35. Allison PD, Williams RA, Hippel V: Better Predicted Probabilities from Linear Probability Models. Available @ https://wwwstatacom/meeting/us20/slides/us20_Allisonpdf; downloaded 15th September 2020 2020.
36. Haggstrom GW. Logistic regression and discriminant analysis by ordinary least squares. J Bus Econ Stat. 1983;1(3):229–38.
37. Allison PD: Better Predicted Probabilities from Linear Probability Models. Available @https://statisticalhorizonscom/better-predicted-probabilities; Downloaded 7th November 2020 2020.
38. von Hippel P, Williams R, Allison P: reg2logit -- Approximate logistic regression parameters using OLS linear regression. Avaiable @ https://econpapersrepecorg/software/bocbocode/S458865htm; Downloaded 7th November 2020 2020.
39. Cox NJ, Steichen T: CONCORD: Stata module for concordance correlation. Statistical Software Components S404501, Boston College Department of Economics; Version 310, revised 10 Nov 2010.
40. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. Stat Med. 2013;32(1):67–80. https://doi.org/10.1002/sim.5525.
41. Nattino G, Lemeshow S, Phillips G, Finazzi S, Bertolini G. Assessing the calibration of dichotomous outcome models with the calibration belt. Stata J. 2017;17(4):1003–14. https://doi.org/10.1177/1536867X1801700414.
42. Bilger M: overfit: module to calculate shrinkage statistics to measure overfitting as well as out- and in-sample predictive bias. @ http://econpapersrepecorg/scripts/searchpf?ft=overfit; Downloaded 1st March 2016.
43. Esnor J, Snell KI, Martins EC: ovefit: Stata module to produce calibration plot of prediction model performance. Statistical Software Components S458486, Boston College Department of Economics; revised 04 January 2020. 2020.
44. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925–31. https://doi.org/10.1093/eurheartj/ehu207.
45. Gelman A, Hill J. Logistic Regression. In: Data analysis using Regression and Multilelvel/ Hierarchal Models. New York: Cambridge University Press; 2007. p. 79–108.
46. Kasza J. Stata tip 125: binned residual plots for assessing the fit of regression models for binary outcomes. Stata J. 2015;15(2):599–604. https://doi.org/10.1177/1536867X1501500219.

47. Kuha J. AIC and BIC: comparisons of assumptions and performance. Sociol Methods Res. 2004;33(2):188–229. https://doi.org/10.1177/0049124103262065.
48. Breen R, Karlson KB, Holm A. Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models. In: Cook KS, Massey DS, editors. Annual Review of Sociology, vol. 44; 2018. p. 39–54.
49. Chatla SB, Shmueli G. An Extensive Examination of Regression Models with a Binary Outcome Variable. J Assoc Inf Syst. 2017;18(4):1.
50. Horowitz JL, Savin NE. Binary response models: Logits, Probits and Semiparametrics. J Econ Perspect. 2001;15(4):43–56. https://doi.org/10.1257/jep.15.4.43.
51. Long JS, Freese J. Methods of interpretation. In: Regression Models for Categorical Dependent Variables using Stata. College Station: Stata Press; 2014. p. 133–84.
52. Karlson KB. Another look at the method of Y-standardization in Logit and Probit models. J Math Sociol. 2015;39(1):29–38. https://doi.org/10.1080/0022250X.2014.897950.
53. Hintze JL, Nelson RD. Violin plots: a box plot-density trace synergism. Am Stat. 1998;52(2):181–4.
54. Winter N, Nichols A: VIOPLOT: Stata module to produce violin plots. @ http://econpapersrepec.org/scripts/search/searchasp?ft=vioplot2010, Accessed June 2010.
55. Williams R. Using the margins command to estimate and interpret adjusted predictions and marginal effects. Stata J. 2012;12(2):308–31. https://doi.org/10.1177/1536867X1201200209.
56. Mood C. Logistic regression: why we cannot do what we think we can do, and what we can do about it. Eur Sociol Rev. 2010;26(1):67–82. https://doi.org/10.1093/esr/jcp006.
57. Wolfe R, Hanley J. If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. Can Med Assoc J. 2002;166(1):65–6.
58. Long JS, Freese J. Models for binary outcomes: Interpretation. In: Regression Models for Categorical Dependent Variables using Stata. College Station: Stata Press; 2014. p. 227–308.
59. Leeper TJ, Arnold J, Arel-Bundock V: margins: Marginal Effects for Model Objects: version 0.3.23. Available @ https://cranr-projectorg/web/packages/margins/indexhtml 2018.
60. Roberts MR, Whited TM: Endogeneity in Empirical Corporate Finance. Soimon School Working Paper No FR11–29; Available at SSRN: https://ssrncom/abstract=1748604 2012.
61. Abdallah W, Goergen M, O'Sullivan N. Endogeneity: how failure to correct for it can cause wrong inferences and some remedies. Br J Manag. 2015;26(4):791–804. https://doi.org/10.1111/1467-8551.12113.
62. Cameron AC, Trivedi PK. Endoegenous regressors. In: Microeconometircs in Stata: Revise Edition. edn. Clloege Station: Stata Press; 2010. p. 479–86.
63. Koné S, Bonfoh B, Dao D, Koné I, Fink G. Heckman-type selection models to obtain unbiased estimates with missing measures outcome: theoretical considerations and an application to missing birth weight data. BMC Med Res Methodol. 2019;19(1):231. https://doi.org/10.1186/s12874-019-0840-7.
64. Odgaard-Jensen J, Vist GE, Timmer A, Kunz R, Akl EA, Schünemann H, et al. Randomisation to protect against selection bias in healthcare trials. Cochrane Database Syst Rev. 2011;4:MR000012.
65. Shadish WR, Clark MH, Steiner PM. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. J Am Stat Assoc. 2008;103(484):1334–43. https://doi.org/10.1198/016214508000000733.
66. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615–25. https://doi.org/10.1097/01.ede.0000135174.63482.43.
67. StataCorp CST: Stata extended regression models reference manual release 16. Available @ https://wwwstatacom/manuals/ermpdf; Accessed 19th September 2020.
68. Lousdal ML. An introduction to instrumental variable assumptions, validation and estimation. Emerg Themes Epidemiol. 2018;15(1):1. https://doi.org/10.1186/s12982-018-0069-7.
69. Martens EP, Pestman WR, Klungel OH. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study (p n/a) by Peter C. Austin, Paul Grootendorst, Sharon-Lise T. Normand, Geoffrey M. Anderson, Statistics in Medicine, Published Online: 16 June 2006. Stat Med. 2007;26(16):3208–10. https://doi.org/10.1002/sim.2618.
70. Mennemeyer ST. Can econometrics rescue epidemiology? Ann Epidemiol. 1997;7(4):249–50. https://doi.org/10.1016/S1047-2797(97)00021-5.
71. Freedman DA. On the so-called "Huber Sandwich estimator" and "robust standard errors". Am Stat. 2006;60(4):299–302. https://doi.org/10.1198/000313006X152207.
72. Dahlqwist E, Kutalik Z, Sjölander A. Using instrumental variables to estimate the attributable fraction. Stat Methods Med Res. 2019;29(8):2063–73. https://doi.org/10.1177/0962280219879175.
73. StataCorp: margins. Marginal means, predictive margins, and marginal effects. Available @ https://wwwstatacom/manuals13/rmarginspdf 2019.
74. ANZICS CORE - Adult patient database: APD data dictionary: version 5.10, March 2020. Available @ https://wwwanzicscomau/adult-patient-database-apd/; downloaded 7th September 2020.
75. Wynants L, Vergouwe Y, Van Huffel S, Timmerman D, Van Calster B. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. Stat Methods Med Res. 2018;27(6):1723–36. https://doi.org/10.1177/0962280216668555.
76. Vach W. Specific Regression Models. In: Regression models as a Tool in Medical research. edn. Boca Raton: CRC Press; 2013. p. 407–8.
77. Cameron AC, Trivedi PK. Comparioson of binary models and parameter estimates. In: Microeconometircs in Stata: Revise Edition. edn. Clloege Station: Stata Press; 2010. p. 465–6.
78. Davies HT, Crombie IK, Tavakoli M. When can odds ratios mislead? BMJ. 1998;316(7136):989–91. https://doi.org/10.1136/bmj.316.7136.989.
79. Feng C, Wang B, Wang H. The relations among three popular indices of risks. Stat Med. 2019;38(23):4772–87. https://doi.org/10.1002/sim.8330.
80. Allison PD. Comparing logit and probit coefficients across groups. Sociol Methods Res. 1999;28(2):186–208. https://doi.org/10.1177/0049124199028002003.
81. Kuha J, Mills C. On group comparisons with logistic regression models. Sociol Methods Res. 2020;49(2):498–525. https://doi.org/10.1177/0049124117747306.
82. Burgess S. Estimating and contextualizing the attenuation of odds ratios due to non collapsibility. Commun Stat Theory Methods. 2017;46(2):786–804. https://doi.org/10.1080/03610926.2015.1006778.
83. Cameron AC, Trivedi PK. Nonlinear regression methods. In: Microeconometircs in Stata: Revise Edition. edn. Clloege Station: Stata Press; 2010. p. 341–54.
84. Angrist JD, Pischke JS. Making regression make sense. In: Mostly harmless econometrics: An empiricist's companion. edn. Princeton: Princeton University Press; 2008. p. 27–110.
85. Moran JL, Santamaria J. Reconsidering lactate as a sepsis risk biomarker. PLoS One. 2017;12(10):e0185320. https://doi.org/10.1371/journal.pone.0185320.
86. van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Initiative S: Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:1.
87. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. JAMA. 2018;320(1):27–8. https://doi.org/10.1001/jama.2018.5602.
88. Cortese G. How to use statistical models and methods for clinical prediction. Ann Transl Med. 2020;8:4.
89. Neyman J, Scott EL. Consistent estimates based on partially consistent observations. Econometrica. 1948;16(1):1–32. https://doi.org/10.2307/1914288.
90. Greene WH: Estimating Econometric Models With Fixed Effects. 2001 @ http://www.sternnyu.edu/eco/wkpapers/workingpapers01/01-10Greene.doc., Accessed 13 Oct 2010.
91. Greene W. The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. Econ J. 2004;7(1):98–119. https://doi.org/10.1111/j.1368-423X.2004.00123.x.
92. Moran JL, Solomon PJ. (ANZICS) ftACfOaRECotAaNZICS: fixed effects Modelling for provider mortality outcomes: analysis of the Australia and new Zealand Intensive Care Society (ANZICS) adult patient Data-Base. PLoS One. 2014;9(7):e102297. https://doi.org/10.1371/journal.pone.0102297.
93. Mroz TA, Zayats YV. Arbitrarily normalized coefficients, information sets, and false reports of biases in binary outcome models. Rev Econ Stat. 2008;90(3):406–13. https://doi.org/10.1162/rest.90.3.406.
94. Timoneda JC. Estimating group fixed effects in panel data with a binary dependent variable: how the LPM outperforms logistic regression in rare

events data. Soc Sci Res. 2021;93:102486. https://doi.org/10.1016/j.ssresearch.2020.102486.
95. Shmueli G. To explain or to predict? Stat Sci. 2010;25(3):289–310.
96. Chen Y, Senturk D, Estes JP, Campos LF, Rhee CM, Dalrymple LS, et al. Performance characteristics of profiling methods and the impact of inadequate case-mix adjustment. Commun Stat Simul Comput. 2019;2019:1.
97. Kalbfleisch J, Wolfe R. On monitoring outcomes of medical providers. Stat Biosci. 2013;5(2):286–302. https://doi.org/10.1007/s12561-013-9093-x.
98. Roessler M, Schmitt J, Schoffer O. Ranking hospitals when performance and risk factors are correlated: A simulation-based comparison of risk adjustment approaches for binary outcomes. PLoS One. 2019;14:12.
99. Schunck R, Perales F. Within- and between-cluster effects in generalized linear mixed models: a discussion of approaches and the xthybrid command. Stata J. 2017;17(1):89–115. https://doi.org/10.1177/1536867X1701700106.
100. Danks L, Duckett SJ: All complications should count: Using our data to make hospitals safer (Methodological supplement). Available @ https://grattaneduau/wp-content/uploads/2018/02/897-All-complications-should-count-methodological-supplementpdf; Downloaded 19th February 2021 2018.
101. Snijders TAB, Bosker RJ. Discrete Dependent Variables. In: Multilevel Ahalysis: an introduction to basic and advanced multilevel modeling. 2nd ed. London: Sage Publications Inc; 2012. p. 289–320.
102. Neuburger J, Cromwell DA, Hutchings A, Black N, van der Meulen JH. Funnel plots for comparing provider performance based on patient-reported outcome measures. BMJ Qual Saf. 2011;20(12):1020–6. https://doi.org/10.1136/bmjqs-2011-000197.
103. Mogstad M, Romano JP, Shaikh AM, Wilhelm D: Inference on Ranks with Applications to Mobility Across Neighborhoods and Academic Achievement Across Countries. Available @ https://bfiuchicagoedu/wp-content/uploads/BFI_WP_202016pdf; Downloaded 16th Feb 2021 2020.
104. Uanhoro JO, Wang Y, Oconnell AA. Problems With Using Odds Ratios as Effect Sizes in Binary Logistic Regression and Alternative Approaches. J Exp Educ. 2019;1:1.
105. Huang FL. Alternatives to logistic regression models in experimental studies. J Exp Educ. 2019:1–16. https://doi.org/10.1080/00220973.2019.1699769.
106. Holm A, Ejrnæs M, Karlson K. Comparing linear probability model coefficients across groups. Qual Quant. 2015;49(5):1823–34. https://doi.org/10.1007/s11135-014-0057-0.
107. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis. 1967;20(7):511–24. https://doi.org/10.1016/0021-9681(67)90082-3.
108. Press SJ, Wilson S. Choosing between logistic regression and discriminant analysis. J Am Stat Assoc. 1978;73(364):699–705. https://doi.org/10.1080/01621459.1978.10480080.
109. Allison PD: Convergence Failures in Logistic Regression. Available @http://wwwpeoplevcuedu/~dbandyop/BIOS625/Convergence_Logisticpdf; downloaded 7 Nov 2020 2008.
110. Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression: causes, consequences, and control. Am J Epidemiol. 2018;187(4):864–70. https://doi.org/10.1093/aje/kwx299.
111. Šinkovec H, Geroldinger A, Heinze G. Bring more data!-a good advice? Removing separation in logistic regression by increasing sample size. Int J Environ Res Public Health. 2019;16(23):4658. https://doi.org/10.3390/ijerph16234658.
112. Beck N. Estimating grouped data models with a binary-dependent variable and fixed effects via a Logit versus a linear probability model: the impact of dropped units. Polit Anal. 2020;28(1):139–45. https://doi.org/10.1017/pan.2019.20.
113. Crisman-Cox C. Estimating substantive effects in binary outcome panel models: a comparison. J Polit. Vol. 0 Issue 0 Pages 000-000. https://doi.org/10.1086/709839.
114. Cook SJ, Hays JC, Franzese RJ. Fixed effects in rare events data: a penalized maximum likelihood solution. Polit Sci Res Methods. 2020;8(1):92–105. https://doi.org/10.1017/psrm.2018.40.
115. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. BMJ. 2016;352:i1981.
116. Liu X, Chen W, Chen T, Zhang H, Zhang B. Marginal effects and incremental effects in two-part models for endogenous healthcare utilization in health services research. Health Serv Outcome Res Methodol. 2020;20(2–3):111–39. https://doi.org/10.1007/s10742-020-00211-x.
117. Zohoori N, Savitz DA. Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. Ann Epidemiol. 1997;7(4):251–7. https://doi.org/10.1016/S1047-2797(97)00023-9.
118. Berg GD, Mansley EC. Endogeneity bias in the absence of unobserved heterogeneity. Ann Epidemiol. 2004;14(8):561–5. https://doi.org/10.1016/j.annepidem.2003.09.020.
119. Leisman DE. Ten pearls and pitfalls of propensity scores in critical care research: a guide for clinicians and researchers. Crit Care Med. 2019;47(2):176–85. https://doi.org/10.1097/CCM.0000000000003567.
120. Leisman DE. The goldilocks effect in the ICU-when the data speak, but not the truth. Crit Care Med. 2020;48(12):1887–9. https://doi.org/10.1097/CCM.0000000000004669.
121. de Grooth H-J, Girbes ARJ, van der Ven F, Oudemans-van Straaten HM, Tuinman PR, de Man AME. Observational research for therapies titrated to effect and associated with severity of illness: misleading results from commonly used statistical methods*. Crit Care Med. 2020;48(12):1720–8. https://doi.org/10.1097/CCM.0000000000004612.
122. Zohoori N. Does endogeneity matter? A comparison of empirical analyses with and without control for endogeneity. Ann Epidemiol. 1997;7(4):258–66. https://doi.org/10.1016/S1047-2797(97)00022-7.
123. Duke GJ, Moran JL, Santamaria JD, Roodenburg O. Safety of the endotracheal tube for prolonged mechanical ventilation. J Crit Care. 2021;61:144–51. https://doi.org/10.1016/j.jcrc.2020.10.018.
124. Kim S-H, Chan CW, Olivares M, Escobar G. ICU admission control: an empirical study of capacity allocation and its implication for patient outcomes. Manag Sci. 2015;61(1):19–38. https://doi.org/10.1287/mnsc.2014.2057.
125. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables application and limitations. Epidemiology. 2006;17(3):260–7. https://doi.org/10.1097/01.ede.0000215160.88317.cb.
126. Qin D. Resurgence of the Endogeneity-backed instrumental variable methods. Econ Open Access Open Assess E-J. 2015;9:1.
127. Angrist JD, Pischke JS. Mostly harmless econometrics: an empiricist's companion. Princeton: Princeton University Press; 2008. https://doi.org/10.2307/j.ctvcm4j72.
128. Foster EM. Instrumental variables for logistic regression: an illustration. Soc Sci Res. 1997;26(4):487–504. https://doi.org/10.1006/ssre.1997.0606.
129. Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. J Health Econ. 2008;27(3):531–43. https://doi.org/10.1016/j.jhealeco.2007.09.009.
130. Koladjo BF, Escolano S, Tubert-Bitter P. Instrumental variable analysis in the context of dichotomous outcome and exposure with a numerical experiment in pharmacoepidemiology. BMC Med Res Methodol. 2018;18(1):61. https://doi.org/10.1186/s12874-018-0513-y.
131. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. Stat Methods Med Res. 2017;26(5):2333–55. https://doi.org/10.1177/0962280215597579.
132. Burgess S, Collaboration CCG. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. Stat Med. 2013;32(27):4726–47. https://doi.org/10.1002/sim.5871.
133. Fan Q, Zhong W. Nonparametric additive instrumental variable estimator: a group shrinkage estimation perspective. J Bus Econ Stat. 2018;36(3):388–99. https://doi.org/10.1080/07350015.2016.1180991.
134. Sjolander A, Martinussen T. Instrumental variable estimation with the R package ivtools. Epidemiol Methods. 2019;8(1):20180024. https://doi.org/10.1515/em-2018-0024.

Publisher's Note