The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

Check for updates

The Stata Journal (2017)
17, Number 4, pp. 1003–1014

Assessing the calibration of dichotomous outcome models with the calibration belt

Giovanni Nattino
Division of Biostatistics
College of Public Health
The Ohio State University
Columbus, OH
nattino.1@osu.edu

Stanley Lemeshow
Division of Biostatistics
College of Public Health
The Ohio State University
Columbus, OH

Gary Phillips
Center for Biostatistics
The Department of Biomedical Informatics
The Ohio State University
Columbus, OH

Stefano Finazzi
GiViTI Coordinating Center
Laboratory of Clinical Epidemiology
IRCCS Istituto di Ricerche Farmacologiche 'Mario Negri'
Ranica, Italy

Guido Bertolini
GiViTI Coordinating Center
Laboratory of Clinical Epidemiology
IRCCS Istituto di Ricerche Farmacologiche 'Mario Negri'
Ranica, Italy

Abstract. The calibration belt is a graphical approach designed to evaluate the goodness of fit of binary outcome models such as logistic regression models. The calibration belt examines the relationship between estimated probabilities and observed outcome rates. Significant deviations from the perfect calibration can be spotted on the graph. The graphical approach is paired to a statistical test, synthesizing the calibration assessment in a standard hypothesis testing framework. In this article, we present the `calibrationbelt` command, which implements the calibration belt and its associated test in Stata.

Keywords: gr0071, calibrationbelt, logistic regression, calibration, goodness of fit, binary outcome

© 2017 StataCorp LLC

1 Introduction

Statistical models that estimate the probability of binary outcomes are extremely common in many research areas. In particular, logistic regression is probably the most widely used method to generate such models. The reliability of binary outcome models requires two properties to be satisfied (Hosmer, Lemeshow, and Sturdivant 2013). First, a model must be able to distinguish between subjects within the two outcome levels. This property is the "discrimination", which is usually evaluated with the area under the receiver operating characteristic curve. Second, the probabilities estimated by the model must accurately match the true outcome experienced in the data. This second property is the "calibration", and it is the focus of the procedure described here.

The calibration belt is a graphical method that has been recently proposed to evaluate the calibration of binary outcome models. The methodology and its usefulness for calibration assessment are thoroughly described in previous works (Finazzi et al. 2011; Nattino, Finazzi, and Bertolini 2014b, 2016a). In this article, we describe the `calibrationbelt` command, which implements the calibration belt approach.

The calibration belt is a plot depicting the relationship between the model's fit probabilities and the observed proportions of the response. Providing information about the statistical significance of the deviations, the calibration belt outperforms the commonly used graphical approaches such as locally weighted smoothers and plots of observed-expected events across deciles (Nattino, Finazzi, and Bertolini 2014a).

The information conveyed by the graphical approach is synthesized in a formal statistical test. Extensive simulations have shown good performances of the test under several scenarios (Nattino, Finazzi, and Bertolini 2014b, 2016a). Taken together, the calibration belt and the test statistic provide useful information when evaluating the performance of predictive models.

Notably, most of the calibration assessment methods have different assumptions in the internal and external validation settings (that is, whether the model has been fit to the same dataset upon which it is evaluated). The calibration belt approach is not an exception; two different procedures are used for the two contexts, and it is important to apply the correct method to identify model performance.

The rest of the article is organized as follows: In section 2, we explain the distinction between internal and external validation settings in assessing a model's calibration. In section 3, we briefly describe the methodology to generate the calibration belt and the associated statistical test. In section 4, we describe the `calibrationbelt` command and apply it to a dataset of intensive care unit (ICU) patients. Finally, in section 5, we summarize the presented material.

2 Internal and external calibration assessment

The calibration of a model can be evaluated in two different settings. First, the model can be evaluated on the same dataset used to fit the model. This internal assessment is an important step in the process of model development. Second, performance of an existing model can be evaluated in a new dataset unrelated to the original model development. Assessment of model calibration in a new independent sample is known as external validation.

It is important to distinguish these two frameworks because many calibration assessment procedures use different assumptions in the two cases. For example, consider the Hosmer–Lemeshow $\widehat{C}$ test to evaluate the calibration of logistic regression models. If the data are partitioned into g groups, the statistic is distributed as a χ^2 with $g-2$ degrees of freedom when models are evaluated on the developmental sample (Hosmer and Lemeshow 1980). However, when one applies models to independent samples, the degrees of freedom are g (Hosmer, Lemeshow, and Sturdivant 2013).

Like the Hosmer–Lemeshow test, the calibration test and belt have different assumptions in the two cases. The most important difference is in the type of models that can be evaluated with the proposed approach. Indeed, the calibration belt and test can be used to evaluate any kind of binary outcome model on external samples. However, only logistic regression models can be evaluated with these methods on the developmental dataset (Nattino, Finazzi, and Bertolini 2016a).

The second difference is in the functional form imposed by the proposed procedure. As will be described in section 3, the calibration belt and test are based on a polynomial regression. The degree of the polynomial link depends on whether the calibration is internally or externally evaluated. Further details are provided in the following section and in Nattino, Finazzi, and Bertolini (2016a).

It is therefore extremely important to recognize the setting where the model is going to be evaluated and to select the options of the `calibrationbelt` command accordingly.

3 The calibration belt and test

In this section, we provide an overview of the methodology behind the calibration belt approach. We consider a sample of size n, where each subject i is characterized by a binary outcome Y_i and by an estimate p_i of $P(Y_i=1)$, the true probability of the positive outcome. We are interested in assessing the calibration of the model, that is, evaluating whether the estimates p_i are compatible with the true probabilities $P(Y_i=1)$.

The way the probabilities p_i are generated depends on the setting. In the internal assessment case, the probabilities are the result of a model developed on the same sample. On the other hand, such probabilities are defined according to an independently developed model in the external validation case.

The approach is based on the estimation of the relationship between the predictions p_i and the true probabilities $P(Y_i = 1)$ with a polynomial logistic regression. In particular, the logit transformation of the predictions p_i is considered, and a logistic regression of the form

$$\text{logit}\left\{P(Y_i = 1)\right\} = \beta_0 + \beta_1 \text{logit}(p_i) + \cdots + \beta_m \{\text{logit}(p_i)\}^m \tag{1}$$

is fit. The logit function is defined as $\text{logit}(p) = \ln\{p/(1-p)\}$.

Notably, the relationship assumed in (1) requires the specification of the degree m of the polynomial. This is an important choice. The assumed relationship could be too simple to describe the real link between predictions and true probabilities if m were fixed at too small a value. Conversely, fixing m too high would lead to the estimation of several useless parameters whenever the relationship is well described by lower-order polynomials.

To overcome the problems associated with a fixed m, we base the procedure on a data-driven forward selection. A low-order polynomial is initially fit, and a sequence of likelihood-ratio tests is used to forwardly identify the degree m. In particular, the degree of the starting model depends on whether the calibration is evaluated internally or on an independent sample. In both cases, the simplest possible polynomial is considered in this stage. This corresponds to the first-order polynomial in the assessment of external calibration but not in the assessment of the developmental dataset. A first-order polynomial is not informative in the latter scenario, because its parameters β_0 and β_1 would always assume the values 0 and 1, regardless of the calibration of the model (Nattino, Finazzi, and Bertolini 2016a). Therefore, the forward selection starts by fitting a second-order polynomial.

Once m is selected, the fit relationship provides information about the calibration of the predictions p_i. Indeed, by definition, a model is perfectly calibrated if $p_i = P(Y_i = 1)$ for each $i = 1, \ldots, n$. Under the link assumed in (1), this identity corresponds to the configuration of the parameters $\beta_0 = \beta_2 = \cdots = \beta_m = 0$ and $\beta_1 = 1$. The idea of the approach is to compare the relationship estimated by fitting the model in (1) with the perfect-calibration link corresponding to the aforementioned choice of the parameters. This comparison can be performed statistically or graphically.

A likelihood-ratio test evaluating the hypothesis H_0: $(\beta_0, \beta_1, \beta_2, \ldots, \beta_m) = (0, 1, 0, \ldots, 0)$ versus the complementary alternative can be used to formally test the calibration of the model. Notably, the distribution of the likelihood-ratio statistic must account for the forward process used to select m. Nattino, Finazzi, and Bertolini (2014b, 2016a) provide the derivation of the distribution of the statistic in the external and internal calibration assessment, respectively.

To graphically assess the calibration, we can represent the fit relationship between the predictions p_i and the true probabilities $P(Y_i = 1)$ with a curve. This curve can be compared with the line associated with the identity of the two quantities, that is, the bisector of the quadrant (45-degree line). A confidence band around the curve, namely, the calibration belt, reflects the statistical uncertainty about the estimate of the curve and allows for the evaluation of the significance of the deviations from the line of perfect calibration.

Because the statistical test is based on a likelihood-ratio statistic, the most natural way to define the confidence band is to invert this type of test. Such inversion guarantees the ideal parallelism between the formal statistical test and calibration belt. However, the calculations involved in the inversion of the likelihood-ratio test are computationally intensive for large samples. Fortunately, likelihood-ratio and Wald confidence regions are asymptotically equivalent, and the computations to construct Wald confidence regions are much simpler (Cox and Hinkley 1974). Thus, the construction of the calibration belt plot is implemented using two different algorithms depending on the sample size. If the sample is smaller than 10,000 records, the confidence band is based on the inversion of the likelihood-ratio test, so the same statistical framework is considered for the statistical test and confidence region in small-moderate samples. For samples of 10,000 units or larger, the confidence region is based on the computationally simpler Wald confidence region.

4 The calibrationbelt command

4.1 Syntax

The `calibrationbelt` command is invoked with the following syntax:

```
calibrationbelt [varlist] [if] [, devel(string) cLevel1(#) cLevel2(#)
   nPoints(#) maxDeg(#) thres(#)]
```

The command generates the calibration belt plot and computes the associated statistical test. There are two ways to apply the procedure.

The first way is to pass the variables corresponding to the binary response and to the predicted probabilities of the outcome in the *varlist* (in this order). In this case, the option `devel(`*string*`)` must be specified, reporting whether the predictions have been fit on the dataset under evaluation (`devel("internal")`) or if the assessment consists of an external, independent validation (`devel("external")`). For example, if the variables `Y` and `phat` store the dependent variable of the model and the predicted probabilities fit on the same sample where the calibration assessment is performed, the command can be run with

```
calibrationbelt Y phat, devel("internal")
```

The second way to use `calibrationbelt` is after fitting a logistic regression model (using either the `logit` or `logistic` command). In this case, it is possible to simply run `calibrationbelt` without specifying the argument *varlist* or any other option. The command automatically considers the dependent variable and the predictions of the logistic regression fit as the arguments to be used. Moreover, the procedure assumes that the setting is the assessment of internal calibration.

4.2 Options

`devel(`*string*`)` specifies whether the calibration is evaluated on the same dataset used to fit the model (`devel("internal")`) or is evaluated on external independent samples (`devel("external")`). Depending on whether the *varlist* argument is passed to the command or not, the program may force the user to specify the setting. For further details about internal and external assessment, see section 2.

`cLevel1(#)` sets one of the confidence levels to be considered in the calibration belt plot. A second confidence level can be set with the argument `cLevel2(#)`. The defaults are `cLevel1(95)` and `cLevel2(80)`. A single calibration belt (that is, a plot with a single confidence level) can be generated by specifying only the first argument. For example, setting `cLevel1(0.99)` produces a single calibration belt with confidence level 99%. A double calibration belt with customized pairs of confidence levels can be produced by providing both optional arguments, `cLevel1(#)` and `cLevel2(#)`.

`cLevel2(#)` is described above; see `cLevel1(#)`.

`nPoints(#)` specifies the number of points defining the edges of the belt. The default is `nPoints(100)`. Reducing the number of points can substantially speed up the production of the plot in large datasets. However, this number also affects the estimate of the probabilities where the belt crosses the bisector (that is, the limits of the intervals reported in the table on the plot). Indeed, the greater the value of `nPoints()`, the higher the precision in the estimate of these values. If the production of the belt is too slow, but the analysis requires an iterative construction of many belts, for example, in exploratory analyses, a possible strategy is to decrease the number of points to values much smaller than the default (say, 20 or 50), accounting for the larger uncertainty in the interpretation of the plots. Finally, when the analysis is set up, the number of points can be increased to the default value to achieve more accurate estimates of the potential deviations from the bisector.

`maxDeg(#)` fixes the maximum degree of the polynomial function used to evaluate the model. The default is `maxDeg(4)`. For further information, see section 3.

`thres(#)` sets the threshold parameter involved in the construction of the polynomial function used to evaluate the model. This parameter corresponds to one minus the significance level used when testing the increase of the polynomial order in the forward selection (see section 3 for further details). The default is `thres(0.95)`, specifying a forward selection ruled by a sequence of classic 0.05-level tests. Greater values of `thres()` correspond to more conservative scenarios, where low-order polynomials are preferred to high-order ones. Calibration belts based on lower `thres()` values are more likely to be based on high-order polynomials.

4.3 An example: ICU data from the GiViTI network

To provide an example of applying the `calibrationbelt` command, we use a dataset of 1,000 ICU patients admitted to Italian ICUs. This dataset is a subsample of the cohort of patients enrolled in the Margherita-PROSAFE project, an international observational study established to monitor the quality of care in ICU. The ongoing project is based on the continuous data collection of clinical information in about 250 units. This collaborative effort was promoted by the GiViTI network (*Gruppo Italiano per la valutazione degli interventi in Terapia Intensiva*, Italian Group for the Evaluation of the Interventions in Intensive Care Units).

The variables of the dataset include hospital mortality and the 15 covariates that compose the simplified acute physiology score, a widely used prognostic model for hospital mortality in ICU patients (Le Gall, Lemeshow, and Saulnier 1993). The 15 variables include patient demographic information, comorbidities, and clinical information. The actual values of the variables have been modified to protect subject confidentiality.

In the example described in the following sections, we use the available sample to fit a logistic regression model using these predictors. The dataset is randomly split in 750 records for model development and 250 patients for external validation. First, we fit the model on the developmental sample, and we evaluate the internal calibration with the calibration belt approach (section 4.4). Then, that model is applied to the validation sample, and the external calibration is assessed (section 4.5).

4.4 calibrationbelt for internal validation

The dataset was randomly split into developmental and validation subsets of 750 and 250 patients, respectively. The logistic regression model with the available predictors is fit to the developmental sample.

```
. use icudata
(ICU patients from the international GiViTI network)
. set seed 101
. generate random = runiform()
. sort random, stable
. generate extValidation = (_n>750)
. quietly logit outcome ib3.adm ib1.chronic ib1.age ib5.gcs
>                 ib3.BP ib3.HR ib1.temp ib3.urine ib1.urea
>                 ib2.WBC ib2.potassium ib2.sodium ib3.HCO3
>                 ib1.bili ib1.paFiIfVent if extValidation == 0
```

Considering the demonstrative purposes of this example, we evaluate only the calibration of the fit model, without considering the other fundamental model-building steps. The calibration belt and the related test for the internal assessment can be produced using the second method described in section 4.1 by simply typing `calibrationbelt` after fitting the model.

```
. calibrationbelt
-----------------------------------------------------------
               GiViTI Calibration Belt
Calibration belt and test for internal validation:
the calibration is evaluated on the training sample.
Sample size: 750
Polynomial degree: 2
Test statistic:  1.54
p-value: 0.2142
-----------------------------------------------------------
```

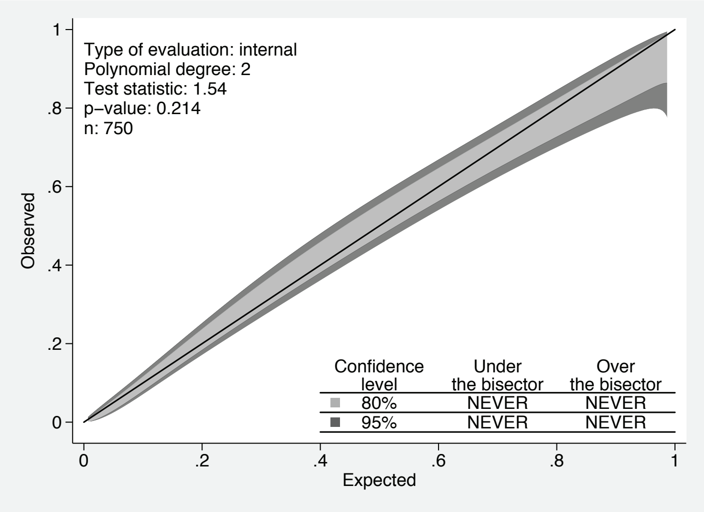

Figure 1. Calibration belt plot on the developmental sample

The output of the program reports the value of the statistic (1.54) and the p-value (0.21) of the test. These results suggest that the hypothesis of good calibration is not rejected (at the classically adopted 0.05 level). Similar conclusions can be drawn from the interpretation of the produced plot, reported in figure 1. We note that both the 80% and 95% calibration belts encompass the bisector over the whole range of the predicted probabilities. This suggests that the predictions of the model do not significantly deviate from the observed rate in the developmental sample (that is, that the model's internal calibration is acceptable).

4.5 calibrationbelt for external validation

The calibration of the fitted model is then evaluated on the records set aside for the external validation of the model. We generate the predicted probabilities resulting from the model fit in section 4.4, and we run the `calibrationbelt` command on the external sample using the `if` qualifier. Here we use the first method described in section 4.1 to invoke the command, that is, passing the variables containing the binary response (`outcome`) and the predicted probabilities (`phat`) to the program. The option `devel("external")` specifies the setting of external calibration assessment.

```
. predict phat, pr
. calibrationbelt outcome phat if extValidation==1, devel("external")
-------------------------------------------------------------
               GiViTI Calibration Belt
Calibration belt and test for external validation:
the calibration is evaluated on an external, independent sample.
Selection: extValidation==1
Sample size: 250
Polynomial degree: 1
Test statistic: 26.30
p-value: 0.0000
-------------------------------------------------------------
```

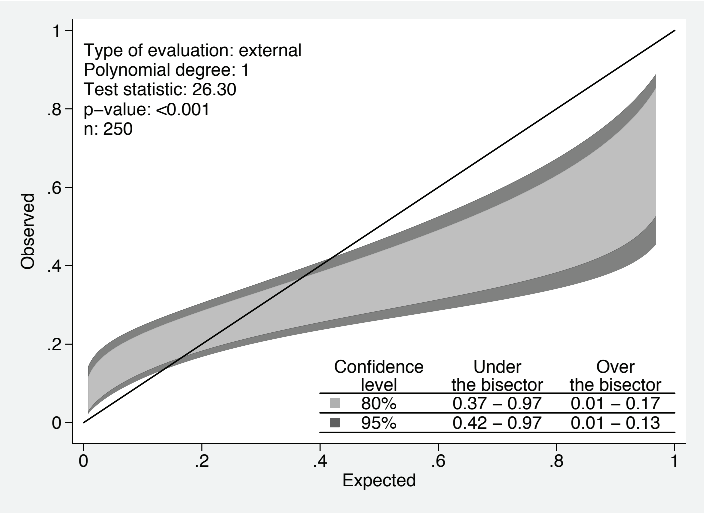

Figure 2. Calibration belt plot on the external sample

The output suggests that the fitted model is not well calibrated in the validation sample. The *p*-value is extremely small (less than 0.0001), which rejects the hypothesis of satisfactory fit even with very conservative significance levels. The produced calibration belts (reported in figure 2) provide interesting information. Because they lie above and do not include the bisector for small predictions, the predictions of the model significantly underestimate the actual risk in the low range of probabilities. With 95% and 80% confidence, the mortality rates are underestimated for estimated probabilities smaller than 0.13 and 0.17, respectively (see the table in the bottom-right corner of the plot). However, the model also overestimates the mortality rates for high predicted probabilities. Indeed, the calibration belts are below the bisector for probabilities higher than 0.42 and 0.37 with 95% and 80% confidence, respectively.

These results suggest that this model performed poorly in a sample different from the one on which the model was developed. A careful application of the recommended model-building steps would be necessary to achieve a better-fitting model.

5 Discussion

We presented the `calibrationbelt` command, which implements the calibration belt approach to provide important information about the calibration of binary outcome models. The calibration belt plot spots any deviation from the perfect fit, pointing out the direction of possible miscalibrations. Conveying information about the statistical significance of the deviations, the method outperforms the existing approaches to graphically evaluate binary outcome models (Nattino, Finazzi, and Bertolini 2014a).

The graphical method is paired to a statistical test, resulting in a p-value reflecting the model under consideration. Tests and belts often return concordant outputs: nonsignificant tests are often associated with the belt encompassing the 45-degree lines and significant tests with the belt deviating from the bisector. However, there are cases where the output of belt and test can disagree. Whenever the $(1-\alpha)$ 100%-calibration belt deviates from the bisector, the statistical test is always significant at the α-level (that is, the p-value is smaller than α). On the other hand, a significant test at the α-level might correspond to a belt that does not deviate significantly from the 45-degree line at any point. The reason for the potential disagreement is the difference between the two approaches in terms of power. The calibration belt is a two-dimensional projection of the multidimensional polynomial relationship used to test the fit of the model. Being assessed in the multidimensional parameter space, the statistical test takes advantage of the full information available. On the other hand, the calibration belt, as a graphical projection, is associated with a loss of information and lower power. A simulation study investigating the agreement between test and belt is described by Nattino, Finazzi, and Bertolini (2014b). The results confirmed this theoretical explanation, showing that the calibration belt behaves slightly more conservatively than the test. In particular, the chances of having a discordant output between belt and test increase with the increase of the polynomial order.

Despite the usefulness of the statistical method, the procedure has been implemented only in an R package so far (Nattino et al. 2016b). The `calibrationbelt` command provides a user-friendly way to generate the calibration belt in Stata. The informativeness of the plot generated with the easy-to-use command can be an invaluable tool in developing better models.

6 References

Cox, D. R., and D. V. Hinkley. 1974. *Theoretical Statistics*. London: Chapman & Hall.

Finazzi, S., D. Poole, D. Luciani, P. E. Cogo, and G. Bertolini. 2011. Calibration belt for quality-of-care assessment based on dichotomous outcomes. *PLOS ONE* 6: e16110.

Hosmer, D. W., Jr., and S. Lemeshow. 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics—Theory and Methods* 9: 1043–1069.

Hosmer, D. W., Jr., S. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.

Le Gall, J.-R., S. Lemeshow, and F. Saulnier. 1993. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA* 270: 2957–2963.

Nattino, G., S. Finazzi, and G. Bertolini. 2014a. Comments on 'Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers' by Peter C. Austin and Ewout W. Steyerberg. *Statistics in Medicine* 33: 2696–2698.

———. 2014b. A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. *Statistics in Medicine* 33: 2390–2407.

———. 2016a. A new test and graphical tool to assess the goodness of fit of logistic regression models. *Statistics in Medicine* 35: 709–720.

Nattino, G., S. Finazzi, G. Bertolini, C. Rossi, and G. Carrara. 2016b. *givitiR: The GiViTI Calibration Test and Belt.* R package version 1.3. http://CRAN.R-project.org/package=givitiR.

About the authors

Giovanni Nattino holds a master's of science degree in applied mathematics and is currently pursuing a PhD in biostatistics from the Ohio State University. He received a postgraduate certificate in Biomedical Research from the Mario Negri Institute for Pharmacological Research, where he has worked as a statistician for four years. He has experience in statistical applications in the areas of critical care, infant mortality, and emergency department admission. His research interests include modeling of clinical data and causal inference in observational studies.

Stanley Lemeshow, PhD, has been with the Ohio State University since 1999 as a biostatistics professor in the School of Public Health and the Department of Statistics, director of the biostatistics core of the Comprehensive Cancer Center, and director of the University's Center for Biostatistics. He served as founding Dean of the Ohio State University College of Public Health from 2003–2013. His biostatistics research includes statistical modeling of medical data, sampling, health disparities, and cancer prevention.

Gary Phillips is a consulting biostatistician who retired from the Center for Biostatistics in the Department of Biomedical Informatics at The Ohio State University College of Medicine. He has 14 years of consulting experience and has worked on both large and small databases in the areas of critical care medicine, oncology, surgery, pharmacy, veterinary medicine, and social work. He specializes in statistical techniques involving logistic regression, linear regression, time to event analysis, and longitudinal analysis.

Stefano Finazzi is a theoretical physicist with a PhD degree in astrophysics. He has worked in the areas of quantum field theory in curved spacetime and quantum optics. He is currently a researcher at the Laboratory of Clinical Epidemiology, Mario Negri Institute for Pharmacological Research (Bergamo, Italy), and he is pursuing a PhD in life sciences from the Open University, UK. His current research interests include mathematical modeling of physiological systems and statistical applications in the area of critical care.

Guido Bertolini, MD, is the head of the Laboratory of Clinical Epidemiology at the Mario Negri Institute for Pharmacological Research (Bergamo, Italy) and of the GiViTI (Italian Group for the Evaluation of Interventions in Intensive Care Medicine) Coordinating Center. He has served as expert for different institutions, such as the EMA (European Medicines Agency), the National Bioethics Committee, and the AIFA (Italian Medicines Agency). He has been a contract professor in "Research methods and statistical analysis" at the postgraduate school in Anesthesia and Intensive Care at the Universities of Brescia and Milan and an adjunct professor in "Methodology for research and training in health services" at the University of Bergamo. He reviews grant applications for several organizations, including the European Commission. He has served as principal investigator of five research projects funded by the Italian Ministry of Health and the European Union. His major fields of expertise are quality-of-care assessment, comparative effectiveness research, infection control, and traumatic brain injury. He has spoken at more than 300 national and international conferences and has authored over 90 peer-reviewed articles.