Nonparametric Covariate Adjustment for Receiver Operating Characteristic Curves

Radu Craiu

Department of Statistics
University of Toronto

joint with: Ben Reiser (Haifa) and Fang Yao (Toronto)

IMS - Pacific Rim, Seoul, June 2009

Outline

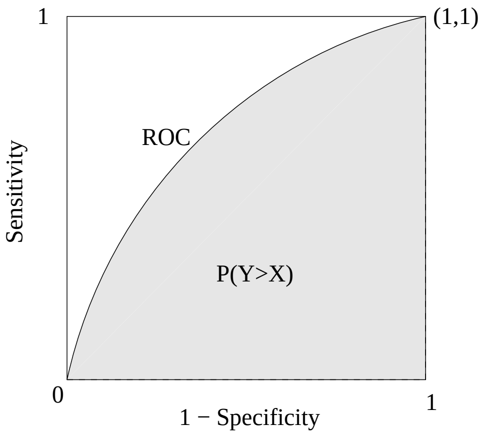
Diagnostic Tests and ROC

- Consider a test designed to differentiate between two classes: diseased and non-diseased.
- Compared to the truth, a.k.a. "the golden rule", one is interested in determining how well the test is performing.
- Given a certain criterion, one can use it to compare different tests and choose the most effective way of separating the two classes.
- All the information available should be used in assessing the test accuracy.

ROC

- Suppose that the test result is r.v. T and depending on whether $T < c$ or $T \geq c$ the test result is considered negative, respectively positive.
- Sensitivity is the true positive rate.
- Specificity is the true negative rate.
- ROC is the plot of Sensitivity against 1-Specificity.
- Different ROC's/tests can be compared using a global univariate summary such as the area under the curve (AUC).
- Bamber (1975) has shown that AUC can be interpreted as the probability that a randomly chosen diseased subject will have a marker (test) value, Y, greater than the value X of a randomly chosen nondiseased subject.

ROC - cont'd



Separating Populations

- More generally, Wolfe and Hogg (1971) have proposed using the $P(Y > X)$ as a measure of the difference between two populations and have argued that this is often more meaningful than looking at mean differences.
- Hauck, Hyslop and Anderson (2000) propose the use of $P(Y > X)$ in assessing treatment effects for clinical trials.

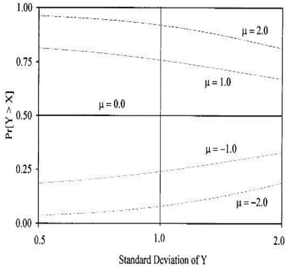


Figure 1. $\Pr[Y > X]$ when X is standard normal and Y is $N(\mu, \sigma^2)$. Each curve corresponds to a value of μ. The ordinate is σ.

- Arises in reliability (Reiser and Guttman, '86).

ROC - cont'd

- Enormous amount of literature dedicated to constructing/comparing ROC's and estimating AUC's under a wide variety of scenarios (Pepe, 2003).
- For this talk of interest is the extra information available for each unit/individual tested.
- For instance, there may be covariate measurements made for each unit tested.
- How to incorporate this information in our assessment?

ROC & Covariates

AI Model the relationship between the ROC/AUC and the covariates directly.

- Loses the connection with the threshold value
- Does not allow prediction of the sensitivity and specificity at a given threshold value conditional on the covariate.
- It does not model covariate effects on the individual marker values.

AII Model the covariate effects on the test values and obtain dependence of AUC on covariates via this. (Faraggi, '03).

A General Regression Model

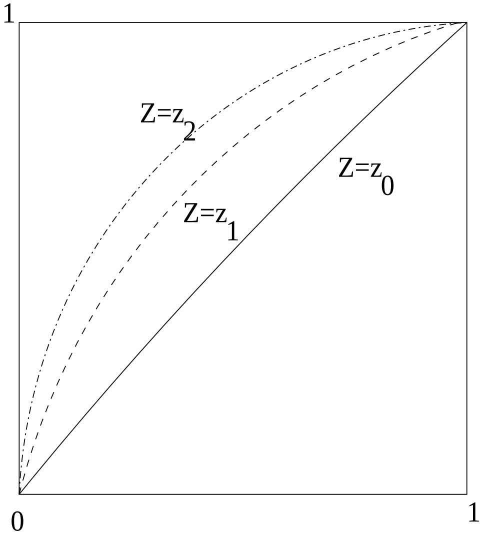
- The test response variable for nondiseased individuals is X and for diseased individuals is Y.

$$X|Z = f(Z) + \sqrt{v_1(Z)}\,\epsilon_1, \tag{1}$$

$$Y|Z = g(Z) + \sqrt{v_2(Z)}\,\epsilon_2, \tag{2}$$

- The standardized errors ϵ_1 and ϵ_2 are independent of each other with zero mean and unit variance, and the variance functions $0 < v_1(z) < \infty$ and $0 < v_2(z) < \infty$ for all $z \in \Re$.
- We get a different ROC/AUC for each value of Z!

A Simple Illustration



Normal Noise Assumption

- Errors ϵ_1 and ϵ_2 are normally distributed.

$$A_N(z) = P(Y > X | Z = z) = \Phi\left\{\frac{g(z) - f(z)}{\sqrt{v_1(z) + v_2(z)}}\right\},$$

$$q_N(z) = \Phi\left\{\frac{g(z) - c}{\sqrt{v_2(z)}}\right\}, \qquad 1 - p_N(z) = 1 - \Phi\left\{\frac{c - f(z)}{\sqrt{v_1(z)}}\right\},$$

for a given threshold c.

$$q_N(z) = \Phi\left[\frac{g(z) - f(z) + \sqrt{v_1(z)}\Phi^{-1}\{1 - p_N(z)\}}{\sqrt{v_2(z)}}\right],$$

- The unknown functions f, g, v_1, v_2, are estimated using nonparametric smoothing.

General Noise Assumption

- Motivated by the Mann-Whitney statistic:

$$M_{m,n} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{[0,\infty)}(y_j - x_i)$$

  where $1_{[0,\infty)}(x) = 1$ if $x \geq 0$ and $1_{[0,\infty)}(x) = 0$ otherwise.
- The data for nondiseased and diseased samples is denoted $\{(z_{i,x}, x_i) : i = 1, \ldots, m\}$ and $\{(z_{j,y}, y_j) : j = 1, \ldots, n\}$
- Z values may differ between diseased and non-diseased.
- We want $A(z) = P(Y > X|Z = z)$ for any z in the range of observed values.

General Noise Assumption - cont'd

- We could use the data corresponding to z-values in the neighborhood of z.

$$A_L(z) = \sum_{z_{i,x} \in N(z)} \sum_{z_{j,y} \in N(z)} \frac{1_{[0,\infty)}(y_j - x_i)}{\sum_{i=1}^{m} 1_{N(z)}(z_{i,x}) \sum_{j=1}^{n} 1_{N(z)}(z_{j,y})}$$

- We could also use a fully-nonparametric estimator

$$\hat{A}_{FNP} = \frac{\sum_{j=1}^{n} \sum_{i=1}^{m} 1_{[0,\infty)}(y_j - x_i) K_{h_1}(Z_j - z) K_{h_2}(Z_i - z)}{\sum_{j=1}^{n} \sum_{i=1}^{m} K_{h_1}(Z_j - z) K_{h_2}(Z_i - z)}.$$

- Such local estimators are less efficient and do not take advantage of the model.
- Instead, we propose an estimator that uses the entire data available as well as the models specified.

General Noise Assumption - cont'd

- If we had all the standardized residuals

$$\epsilon_{i,x} = \frac{x_i - f(z_{i,x})}{\sqrt{v_1(z_{i,x})}}, \qquad \epsilon_{j,y} = \frac{y_j - g(z_{j,y})}{\sqrt{v_2(z_{j,y})}},$$

  and if we knew f, g, v_1, v_2 then we could construct working samples $\{x_{i,z}, \ldots, x_{m,z}\}$ and $\{y_{1,z}, \ldots, y_{n,z}\}$ for $Z = z$, as if they were all observed at $Z = z$,

$$x_{i,z} = f(z) + \sqrt{v_1(z)}\epsilon_{i,x}, \qquad y_{j,z} = g(z) + \sqrt{v_2(z)}\epsilon_{j,y}.$$

- The Covariate-Adjusted Mann-Whitney Estimator (CAMWE) for $A(z)$,

$$A_M(z) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{i=1}^{n} 1_{[0,\infty)}(y_{j,z} - x_{i,z}).$$

General Noise Assumption - cont'd

- The standardized residuals can be estimated using estimates for f, g, v_1 and v_2.
- After obtaining nonparametric estimates of the unknown functions f, g, v_1 and v_2, we do not have to choose other tuning parameters for each covariate value $Z = z$.
- We can calculate the sensitivity and specificity from the working samples for $Z = z$,

$$q_M(z) = \frac{1}{n}\sum_{j=1}^{n} 1_{[0,\infty)}(y_{j,z} \geq c), \quad p_M(z) = \frac{1}{m}\sum_{i=1}^{n} 1_{[0,\infty)}(x_{i,z} \leq c),$$

  for a given threshold c.
- The ROC curves for $Z = z$ can be obtained by plotting $q_M(z)$ versus $1 - p_M(z)$ for all possible values of c.

Nonparametric Smoothing Procedures

- Local polynomial regression for estimating f, g, v_1 and v_2 (Fan and Gijbels, '96).
- The variance functions $v_1(z)$ and $v_2(z)$ for heteroscedastic errors are estimated by fitting local polynomial regression to the squared residuals, $v_{i,x}$ and $v_{j,y}$, $i = 1, \ldots, m, j = 1, \ldots, n$,

$$v_{i,x} = \{x_i - \hat{f}(z_{i,x})\}^2, \quad v_{j,y} = \{y_j - \hat{g}(z_{j,y})\}^2,$$

- All bandwidths are selected using the standard procedure of leave-one-out cross validation.

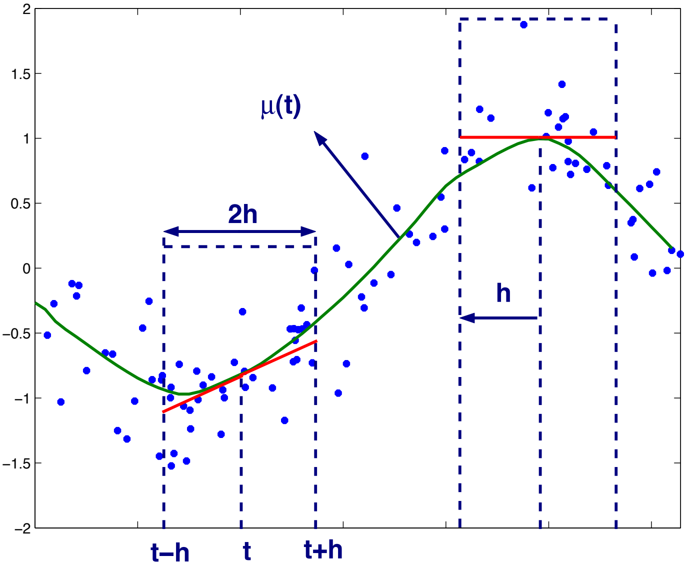
Local Polynomial Regression - short description

- Consider the nondiseased sample $(z_{i,x}, x_i)$, $i = 1, \ldots, m$, which is assumed to consist of i.i.d. realizations from a random vector (Z, X).
- The local polynomial regression estimator of $f(z)$ is obtained by minimizing

$$\sum_{i=1}^{m} \{x_i - \sum_{k=0}^{p} \beta_k (z_{i,x} - z)^k\}^2 K_{h_1}(z_{i,x} - z),$$

where h_1 is a bandwidth controlling the amount of smoothing, and $K_{h_1}(\cdot) = K(\cdot/h_1)/h_1$.

Local Polynomial Regression - short description



Local Polynomial Regression - short description

- In matrix notation let Z_x be the design matrix

$$Z_x = \begin{pmatrix} 1 & (z_{1,x} - z) & \cdots & (z_{1,x} - z)^p \\ \vdots & \vdots & & \vdots \\ 1 & (z_{m,x} - z) & \cdots & (z_{m,x} - z)^p \end{pmatrix},$$

  $W_{x,h_1} = \mathrm{diag}\{K_{h_1}(z_{i,x} - z) : i = 1, \ldots, m\}$ and $\mathbf{x} = (x_1, \ldots, x_m)^T$.

- The local polynomial estimator is given by

$$\hat{f}(z) = \mathbf{e}_1^T (Z_x^T W_{x,h_1} Z_x)^{-1} Z_x W_{x,h_1} \mathbf{x}.$$

- Similarly,

$$\hat{g}(z) = \mathbf{e}_1^T (Z_y^T W_{y,h_2} Z_y)^{-1} Z_y W_{y,h_2} \mathbf{y}.$$

Local Polynomial Regression - short description

- The nonparametric estimators $\hat{v}_1(z)$ and $\hat{v}_2(z)$ are obtained by fitting local polynomial regression to the squared residuals, i.e., the variance observations, $v_{i,x}$ and $v_{j,y}$, $i = 1, \ldots, m, j = 1, \ldots, n$, defined by

$$v_{i,x} = \{x_i - \hat{f}(z_{i,x})\}^2, \qquad v_{j,y} = \{y_j - \hat{g}(z_{j,y})\}^2.$$

- Let b_1 be the bandwidth for $\hat{v}_1(z)$. Let $\mathbf{v}_x = (v_{1,x}, \ldots, v_{m,x})^T$. Then

$$\hat{v}_1(z) = \mathbf{e}_1^T (Z_x^T W_{x,b_1} Z_x)^{-1} Z_x W_{x,b_1} \mathbf{v}_x$$

where $W_{x,b_1} = \text{diag}\{K_{b_1}(z_{i,x} - z) : i = 1, \ldots, m\}$.

- Similar calculations can be done for $\hat{v}_2$.

Bootstrap-based Confidence Bands

- Sample with replacement from the estimated standardized residuals $\{\hat{\epsilon}_{i,x} : i = 1, \ldots, m\}$ and $\{\hat{\epsilon}_{j,y} : j = 1, \ldots, n\}$ to form bootstrap sets $\{\hat{\epsilon}^{(b)}_{i,x}; i = 1, \ldots, m\}$ and $\{\hat{\epsilon}^{(b)}_{j,y} : j = 1, \ldots, n\}$.
- Using the estimated mean and variance functions from the observed data, construct the bootstrapped working samples at covariate value $Z = z$,

$$\hat{x}^{(b)}_{i,z} = \hat{f}(z) + \hat{\epsilon}^{(b)}_{i,x}\sqrt{\hat{v}_1(z)}, \quad \hat{y}^{(b)}_{j,y} = \hat{g}(z) + \hat{\epsilon}^{(b)}_{j,y}\sqrt{\hat{v}_2(z)}, \quad i = 1, \ldots m,$$

- Estimate $A^{(b)}(z)$ using

$$\widehat{A}^{(b)}_M(z) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{[0,\infty)}(\hat{y}^{(b)}_{j,y} - \hat{x}^{(b)}_{i,x}).$$

Then the set $\{\widehat{A}^{(b)}_M(z) : b = 1, \ldots, B\}$ is used to obtain confidence limits for $\widehat{A}(z)$.

Simulations

- For non-diseased individuals:

$$X_i = \alpha_0 + \alpha_1 Z_i + \alpha_2 \sin(Z_i) + \epsilon_i$$

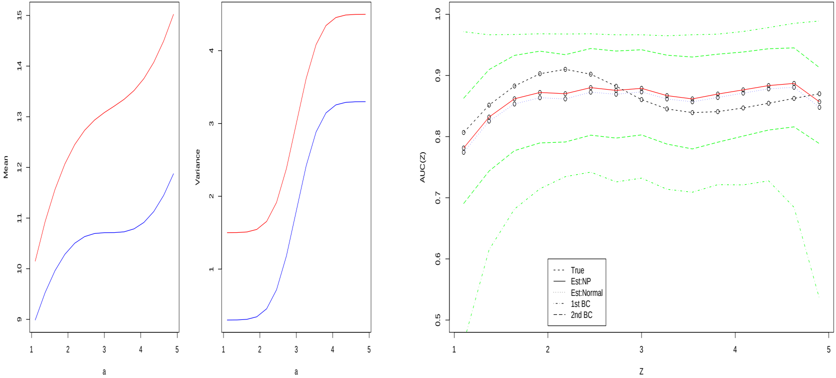
  where the Student(3) deviate ϵ has conditional variance rescaled by $xi_0 + \xi_1 \Phi(\delta_0 + \delta_1 Z_i)$.

- For diseased individuals we consider the model

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 \sin(Z_i) + \beta_3 \sqrt{Z_i - 1} + \eta_i,$$

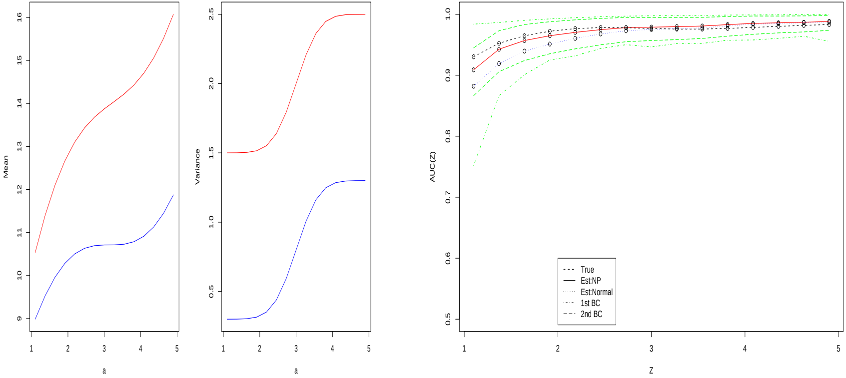
  with η Student(3) with conditional variance $\text{var}(\eta_i|Z_i) = \text{var}(\epsilon_i|Z_i) + \gamma$.

Simulations-cont'd



Scenario 1: $n = 40,\ \beta_0 = \alpha_0 = 0,\ \alpha_1 = \alpha_2 = \beta_2 = \beta_1 = 3,\ \beta_3 = 1$
$\xi_0 = 0.3,\ \xi = 3,\ \delta_1 = 2,\ \delta_0 = -6, \gamma = 1.2$

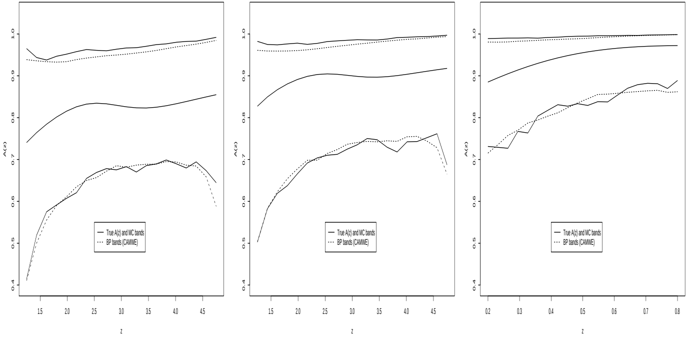
Simulations-cont'd



Scenario 2: $n = 100$, $\beta_0 = \alpha_0 = 0$, $\alpha_1 = \alpha_2 = \beta_2 = \beta_1 = 1.5$, $\beta_3 = 2.5$
$\xi_0 = 0.3$, $\xi = 1$, $\delta_1 = 2$, $\delta_0 = -6$,$\gamma = 1.2$

Simulations-cont'd

Confidence Bands for errors distributed: normal (L), t_3 (C) and lognormal (R)



Example: White Onions Data

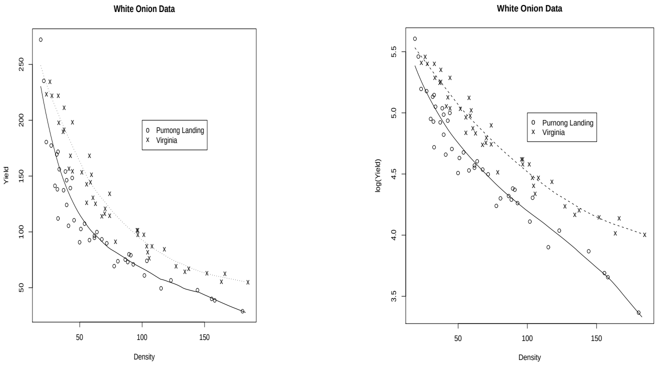
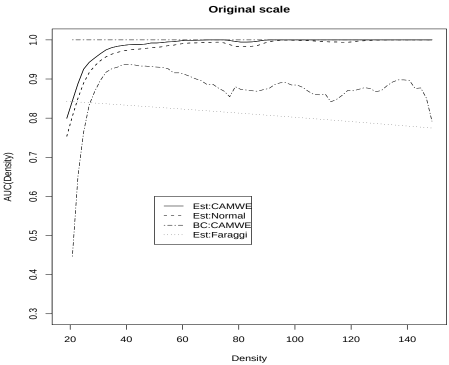


Figure: *Spanish Onion Data with response on: the orginal scale (left) the logarithmic scale (right).*

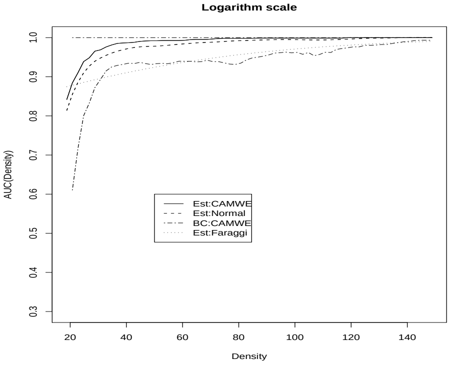
Example

Figure: *Comparison of estimated dependency between AUC and density obtained using the nonparametric approach with and without normal noise with the parametric estimation of the same dependency assuming a normal linear regression model.*



Example

Figure: *Response is on the logarithmic scale.*



Asymptotic Results - Normal Error

Convergence in the Normal Error Case

If $n/m \to \infty$,

$$\sqrt{mh_1}(\hat{A}_N(z) - A_N(z)) \to N(B_1(z), V_1(z)).$$

If $n/m \to 0$,

$$\sqrt{nh_2}(\hat{A}_N(z) - A_N(z)) \to N(B_2(z), V_2(z)).$$

If $n/m \to c \in (0, \infty)$,

$$\sqrt{mh_1}(\hat{A}_N(z) - A_N(z)) \to N(B_3(z), V_3(z)).$$

Under stronger assumptions the convergence of $\hat{A}_N(z) - A_N(z)$ to 0 holds almost surely.

Asymptotic Results - General Error

Step I - Convergence of the hypothetical estimator

Take

$$A_M(z) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{i=1}^{n} 1_{[0,\infty)}(y_{j,z} - x_{i,z})$$

where

$$x_{i,z} = f(z) + \sqrt{v_1(z)}\epsilon_{i,x}, \quad y_{j,z} = g(z) + \sqrt{v_2(z)}\epsilon_{j,y}.$$

Then if $n/m \to \lambda$ for some $0 < \lambda < \infty$, $\xi(z) > 0$

$$\sqrt{m+n}\{A_M(z) - A(z)\} \xrightarrow{D} N(0, \xi(z))$$

where $\lambda^* = 1/(1+\lambda)$.

Asymptotic Results - General Error

Step II - L^2 Consistency

For a given z

$$E[\{\widehat{A}_M(z) - A_M(z)\}^2] \longrightarrow 0.$$

Step I + Step II

$$E[\{\widehat{A}_M(z) - A(z)\}^2] \longrightarrow 0.$$

References

- Bamber, D. C. (1975) , "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *J. Math. Physiol.*, 12, 387–415.
- Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall.
- Faraggi, D. (2003), "Adjusting receiver operating curves and related indices for covariates," *The Statistician*, 52, 179–192.
- Hauck, W., Hyslop, T., and Anderson, S. (2000), "Generalized treatment effects for clinical trials," *Statist. Medicine*, 19, 887–899.
- Pepe, M. S. (2003) , *The Statistical Evaluation of Medical Tests for Classification and Prediction* Oxford Statistical Sciences Series.
- Reiser, B. and Guttman, I. (1986), "Stastistical-Inference for Pr(Y-less-than-X) - The normal case," *Technometrics*, 28, 253–257.
- Wolfe, D., and Hogg, R. (1971), "Constructing Statistics and reporting data," *Amer. Statistician*, 25, 27–30.