A Multivariate Logistic Regression Equation to Screen for Diabetes

Development and validation

BAHMAN P. TABAEI, MPH[1]
WILLIAM H. HERMAN, MD, MPH[1,2]

OBJECTIVE — To develop and validate an empirical equation to screen for diabetes.

RESEARCH DESIGN AND METHODS — A predictive equation was developed using multiple logistic regression analysis and data collected from 1,032 Egyptian subjects with no history of diabetes. The equation incorporated age, sex, BMI, postprandial time (self-reported number of hours since last food or drink other than water), and random capillary plasma glucose as independent covariates for prediction of undiagnosed diabetes. These covariates were based on a fasting plasma glucose level ≥126 mg/dl and/or a plasma glucose level 2 h after a 75-g oral glucose load ≥200 mg/dl. The equation was validated using data collected from an independent sample of 1,065 American subjects. Its performance was also compared with that of recommended and proposed static plasma glucose cut points for diabetes screening.

RESULTS — The predictive equation was calculated with the following logistic regression parameters: $P = 1/(1 - e^{-x})$, where $x = -10.0382 + [0.0331$ (age in years) $+ 0.0308$ (random plasma glucose in mg/dl) $+ 0.2500$ (postprandial time assessed as 0 to ≥8 h) $+ 0.5620$ (if female) $+ 0.0346$ (BMI)]. The cut point for the prediction of previously undiagnosed diabetes was defined as a probability value ≥0.20. The equation's sensitivity was 65%, specificity 96%, and positive predictive value (PPV) 67%. When applied to a new sample, the equation's sensitivity was 62%, specificity 96%, and PPV 63%.

CONCLUSIONS — This multivariate logistic equation improves on currently recommended methods of screening for undiagnosed diabetes and can be easily implemented in a inexpensive handheld programmable calculator to predict previously undiagnosed diabetes.

Diabetes Care **25:1999–2003, 2002**

Screening for undiagnosed diabetes is controversial. In 1978, the American Diabetes Association (ADA), the Centers for Disease Control and Prevention, and the National Institutes of Health recommended against screening for diabetes in nonpregnant adults (1). In 1989 and again in 1996, the U.S. Preventive Services Task Force recommended against screening for diabetes in nonpregnant adults (1,2), and in 2001, the ADA recommended against community screening for diabetes (3). Several recent studies have shown that age, sex, BMI, and current metabolic status affect blood glucose levels and have raised concerns about the performance of diabetes screening tests (4–8).

The performance of all screening tests is dependent on the threshold or cut point used to define a positive test. In diabetes screening, choosing a higher glucose cut point reduces sensitivity (probability of a positive screening test given disease) but improves specificity (probability of a negative screening test given absence of disease). Choosing a lower glucose cut point improves sensitivity but reduces specificity. Because the optimal cut point for a positive test may depend on age, sex, BMI, and the time since last food or drink, we propose an alternative approach to interpreting capillary glucose screening tests by developing a multivariate equation using the best combination of readily available data to predict previously undiagnosed diabetes.

RESEARCH DESIGN AND METHODS

— To assess the likelihood of previously undiagnosed diabetes, a predictive equation was developed using data from 1,032 Egyptian subjects without a history of diabetes who participated in the Diabetes in Egypt Project between July 1992 and October 1993 (9). In a household examination, all subjects were assessed for age, sex, height, weight, postprandial time (self-reported number of hours since last food or drink other than water), and random capillary whole blood glucose. On a separate day, fasting plasma glucose (FPG) and plasma glucose 2 h after a 75-g oral glucose load (2-h PG) were measured. Multiple logistic regression analysis was used to develop an equation for prediction of undiagnosed diabetes based on FPG ≥126 mg/dl and/or 2-h PG ≥200 mg/dl. Diabetes risk factors included in the equation were age (years), sex (female), BMI (calculated as weight in kilograms divided by height in meters squared [kg/m^2]), postprandial time (0 to ≥8 h), and random capillary plasma glucose (mg/dl). Age, BMI, and capillary plasma glucose were modeled as continuous variables, postprandial time was modeled as a continuous variable between 0 and 8 h (after which random capillary glucose did not vary as a function of postprandial time), and sex was modeled

From the [1]Department of Internal Medicine, University of Michigan Health System, Ann Arbor, Michigan; and the [2]Department of Epidemiology, University of Michigan Health System, Ann Arbor, Michigan.

Address correspondence and reprint requests to William H. Herman, MD, MPH, University of Michigan Health System, Division of Endocrinology and Metabolism, 1500 E. Medical Center Dr., 3920 Taubman Center, Ann Arbor, MI 48109-0354. E-mail: wherman@umich.edu.

Received for publication 21 January 2001 and accepted in revised form 26 June 2002.

Abbreviations: 2-h PG, plasma glucose 2 h after a 75-g oral glucose load; ADA, American Diabetes Association; EPV, events per variable; FPG, fasting plasma glucose; OAPR, odds of being affected given a positive result; PPV, positive predictive value; ROC, receiver-operating characteristic.

A table elsewhere in this issue shows conventional and Système International (SI) units and conversion factors for many substances.

Table 1—*Demographic characteristics of the study populations*

Variable	Egyptian subjects	American subjects
n	1,032	1,065
Age (years)	45 ± 14	46 ± 15
Sex (F)	609 (59)	731 (69)*
BMI (kg/m^2)	29.8 ± 7.6	28.4 ± 6.4
Capillary plasma glucose (mg/dl)	124.7 ± 51.8	100.8 ± 24.1*
Postprandial time (0–8+ h)	3.2 ± 1.7	4.8 ± 2.9*

Data are means ± SD or *n* (%). *Statistically significant at $P < 0.0001$ vs. Egyptian subjects.

as a categorical variable (0 = male and 1 = female). The final mathematical equation provides an estimate of a subject's likelihood of previously undiagnosed diabetes expressed as a probability between 0.0 and 1.0.

The linearity assumption for logistic regression was assessed by categorizing each continuous variable into multiple dichotomous variables of equal units and plotting each variable's coefficient against the midpoint of the variable. We also performed the Mantel-Haenszel χ^2 test for trend. Multicollinearity was assessed using the Pearson correlation coefficient statistic. Accuracy, reliability, and precision of regression coefficients were assessed by calculating the number of events per variable (EPV)—the ratio of the number of outcome events to the number of predictor variables. An EPV number of at least 10 indicates that the estimates of regression coefficients and their CIs are reliable (10,11). The possible interactions among variables were assessed using the Breslow and Day χ^2 test (12).

The −2 log-likelihood ratio test was used to test the overall significance of the predictive equation. The significance of the variables in the model was assessed by the Wald χ^2 test and CIs. The fit of the model was assessed by the Hosmer-Lemeshow goodness of fit χ^2 test (13,14). To assess outliers and detect extreme points in the design space, logistic regression diagnostics were performed by plotting the diagnostic statistic against the observation number using hat matrix diagonal and Pearson and Deviance residuals analyses (13,14).

To select the optimal cut point to define a positive test, a receiver-operating characteristic (ROC) curve was constructed by plotting sensitivity against the false-positive rate (1 − specificity) over a range of cut-point values. Generally, the best cut point is at or near the shoulder of the ROC curve, where substantial gains can be made in sensitivity with only modest reductions in specificity. Sensitivity was defined as the proportion of subjects predicted to have the outcome who really have it (true-positive test) and calculated as [true positives/(true positives + false negatives)] × 100. Specificity was defined as the proportion of subjects predicted not to have the outcome who do not have it (true-negative test) and calculated as [true negatives/(true negatives + false positives)] × 100. Positive predictive value (PPV) was defined as the percentage of individuals with a positive test result who actually have the disease and was calculated as [true positives/(true positives + false positives)] × 100. The odds of being affected given a positive result (OAPR) was defined as the ratio of the number of affected to unaffected individuals among those with positive results and was calculated as true positives/false positives.

Concordance and discordance values, derived from the logistic regression analysis, were used to measure the association of predicted probabilities and to check the ability of the model to predict outcome. The higher the value of the concordance and the lower the value of discordance, the greater the ability of the model to predict outcome. To evaluate the overall performance of the equation, we considered several measures of predictive performance, including discrimination and calibration (15–20). Discrimination was defined as the ability of the equation to distinguish high-risk subjects from low-risk subjects and is quantified by the area under the ROC curve (15,19,20). Calibration was defined as whether the predicted probabilities agree with the observed probabilities and is quantified by the calibration slope calculated as [model χ^2 − (df − 1)[/model χ^2 (16,20,21). Well-calibrated models have a slope of ~1, whereas models providing too extreme of predictions have a slope of <1 (17,20).

To validate the equation, we applied it to data that had not been used to generate the equation. Thus, we applied the equation to data collected from 1,065 subjects with no history of diabetes who were studied between September 1995 and July 1998 by health care systems serving communities in Springfield, MA; Robeson County, NC; Providence, Pawtucket, RI; and Central Falls, RI (7). All subjects were assessed for age, sex, height, weight, postprandial time, random capillary plasma glucose, and, on a separate day, FPG and 2-h PG.

To compare the results obtained with the predictive equation and the results obtained with various recommended and proposed random capillary plasma glucose cut points, we applied the equation and those cut points to the combined Egyptian and American datasets. Capillary plasma glucose values were calculated by multiplying capillary whole blood glucose values by 1.14. All statistical analyses were performed using SAS software version 6.12 (SAS Institute, Cary, NC).

RESULTS — Table 1 describes the demographic characteristics of the Egyptian and American subjects. The American participants included Hispanics (58%), non-Hispanic whites (19%), African-Americans (12%), Native Americans (4%), and others (7%). The diabetes predictive equation was calculated with the following logistic regression parameters: $P = 1/(1 - e^{-x})$, where $x = -10.0382 +$ [0.0331 (age in years) + 0.0308 (random plasma glucose in mg/dl) + 0.2500 (postprandial time assessed as 0 to ≥8 h) + 0.5620 (if female) + 0.0346 (BMI)]. Table 2 shows the maximum likelihood estimates for the logistic regression function. The overall significance of the equation by the −2 log-likelihood test was 299.6 ($P = 0.0001$) with 5 df, with 89% concordant pairs and 11% discordant pairs. The Hosmer-Lemeshow goodness of fit test was 5.27 ($P = 0.73$) with 8 df. The EPV number was 134/5 = 26.8. Because no interactions, either alone or in combination, added significantly to the equation, we did not add any of these parameters. No potential outliers were detected, and the equation met the linearity assumption for logistic regression analysis.

Table 2—*Maximum likelihood estimates of logistic regression function*

Variable	Estimated regression coefficient	Estimated SE	Wald$_{\chi^2}$	*P*	Estimated odds ratio	95% CI for odds ratio
Intercept	−10.0382	±0.8123	—	0.0001	—	—
Age (years)	0.0331	±0.009	12.7	0.0004	1.39*	1.16–1.67*
Plasma glucose (mg/dl)	0.0308	±0.003	101.6	0.0001	1.36*	1.28–1.44*
Postprandial time (0–8 h)	0.2500	±0.625	16.0	0.0001	1.28†	1.13–1.45
Sex (F)	0.5620	±0.277	4.1	0.04	1.75	1.02–3.02
BMI (kg/m^2)	0.0346	±0.014	5.8	0.02	1.04†	1.01–1.07

* Estimated odds ratios and 95% CIs for 10-unit increase; †estimated odds ratios and 95% CIs for 1-unit increase

The probability level that provided an optimal cut point was 0.20. Based on the classification table, derived from the logistic regression and ROC curve analysis, sensitivity was 65%, specificity 96%, and PPV 67% (Fig. 1). The area under the ROC curve was 0.88. The calibration slope was (299.6 − 4)/299.6 = 0.99. When applied to a new sample of 1,065 subjects, the equation's sensitivity was 62%, specificity 96%, and PPV 63%. These represented relatively small decrements from the original equation.

The diabetes predictive equation performed better than the various proposed static random capillary plasma glucose cut points for a positive test when applied to the combined population with 10% prevalence of undiagnosed diabetes (the prevalence observed in the combined Egyptian and American data sets) (Table 3). In general, the equation yielded higher sensitivity, identified more new cases (true positives), and missed fewer new cases (false negatives) than the static capillary plasma glucose cut points ≥140, ≥150, ≥160, ≥170, and ≥180 mg/dl. The equation yielded higher specificity and identified fewer false-positive cases than the static capillary plasma glucose cut points ≥110, ≥120, ≥130, ≥140, and ≥150 mg/dl. The equation yielded higher PPV and OAPR than the static capillary plasma glucose cut points ≥110, ≥120, ≥130, ≥140, ≥150, ≥160, and ≥170 mg/dl.

CONCLUSIONS — The performance of all screening tests depends on the cut points used to define a positive test. The choice of a higher cut point leaves more cases undetected, and the choice of a lower cut point classifies more healthy individuals as abnormal (5). Currently, there are no widely accepted or rigorously validated cut points to define positive screening tests for diabetes in nonpregnant adults (6). The ADA has recommended a random capillary whole blood glucose cut point of ≥140 mg/dl (capillary plasma glucose ≥160 mg/dl), and Rolka et al. (7) have recommended a random capillary plasma glucose cut point of ≥120 mg/dl.

Optimal cut points for random capillary glucose tests depend on age, sex, BMI, and postprandial time (6,7). Multivariate equations incorporate multiple pieces of diagnostic information and can provide a flexible alternative to static cut points for the definition of a positive test (21). We have developed a multivariate predictive equation based on age, sex, BMI, postprandial time, and capillary plasma glucose levels to assess the likelihood of previously undiagnosed diabetes. The equation was 65% sensitive and 96% specific. In validation testing, the equation was 62% sensitive and 96% specific. Predictive equations rarely perform as well with new data as with the data with which they were developed because during development, the equation maximizes the probability of predicting the values in the original dataset. When testing an equation, the important factor is the size of the decrement in performance. The relatively small decrement in sensitivity and unchanged specificity suggest that the equation has both external validity and generalizability (21).

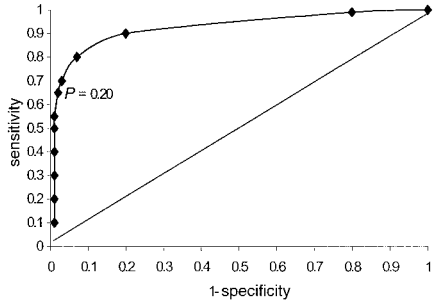

Figure 1—*ROC curve. Points on the ROC curve represent the probability levels generated from the logistic regression analysis that was used to select the optimal cut point. A probability value of 0.20 provided a sensitivity of 65% and a specificity of 96%. Sensitivity and specificity of risk factors for the prediction of previously undiagnosed diabetes based on FPG ≥126 mg/dl and/or 2-h PG ≥200 mg/dl were estimated using the multiple regression model described in the text, in which FPG and/or 2-h PG were modeled as a function of age, random plasma glucose, postprandial time, sex, and BMI. Screening tests that discriminate well between diabetic and nondiabetic individuals aggregate toward the upper left corner of the ROC curve. The area under the curve quantifies how well the screening test correctly distinguishes a diabetic from a nondiabetic individual; the greater the area under the curve, the better the performance of the screening test. A diagonal reference line (area under the curve = 0.50) defines points where a test is no better than chance in identifying individuals with diabetes.*

A decision regarding acceptable levels of sensitivity and specificity involves weighting the consequences of leaving cases undetected (false negatives) and classifying healthy individuals as abnormal (false positives) (22,23). Like the ADA-recommended plasma glucose cut point of 160 mg/dl, the logistic equation provided high specificity (96%) (Table 3). Compared with the ADA-recommended cut point of 160 mg/dl, the logistic equation improved sensitivity (44 and 63%, respectively) (Table 3). Compared with the plasma glucose cut point of 120 mg/dl, the logistic equation improved specificity (77 and 96%, respectively) but was less sensitive (76 and 63%, respectively) (Table 3).

Highly specific screening tests minimize the number of false-positive results but increase the number of false-negative results. They are preferable if the failure to

Table 3—*Comparison of the performance of the predictive equation and static capillary plasma glucose cut points*

Screening test	Sensitivity (%)	Specificity (%)	PPV (%)	OAPR	True positive	False positive	False negative
Equation	63	96	64	1.75	63	36	37
Capillary plasma glucose							
≥110 mg/dl	84	65	21	0.27	84	315	16
≥120 mg/dl	76	77	27	0.37	76	207	24
≥130 mg/dl	63	87	35	0.54	63	117	37
≥140 mg/dl	55	92	43	0.76	55	72	45
≥150 mg/dl	50	95	53	1.11	50	45	50
≥160 mg/dl	44	96	55	1.22	44	36	56
≥170 mg/dl	42	97	60	1.56	42	27	58
≥180 mg/dl	39	98	68	2.17	39	18	61

True positive = new cases = prevalence × sensitivity × n; false positive = 1 − prevalence × 1 − specificity × n; false negative = missed cases = prevalence × 1-sensitivity × n, where prevalence of undiagnosed diabetes is 10% and n = 1,000.

make an early diagnosis and initiate treatment does not have dire health consequences, if a disease is uncommon in the population, and if false-positive results can harm the subject physically, emotionally, or financially. Type 2 diabetes is often slowly progressive and is not associated with complications in the short term. Individuals with initial false-negative screening tests will be identified as abnormal on rescreening, particularly if they have progressive glucose intolerance. In addition, undiagnosed diabetes is uncommon: in a representative sample of the U.S. population 40–74 years of age, undiagnosed diabetes, defined by FPG ≥140 mg/dl or 2-h PG ≥200, was present in only 6.7% (24). False-positive screening tests require further diagnostic tests that are inconvenient, expensive, and time-consuming. For these reasons, we believe that the predictive equation, which is highly specific, is preferable to a static glucose cut point of 120 mg/dl, which is much less specific. We also believe that the predictive equation is preferable to a static glucose cut point of 160 mg/dl because, given comparable high specificity, it is much more sensitive.

PPV and OAPR are measures of the performance of a diagnostic test that depend on the prevalence of the disease in the screened population and on the sensitivity and specificity of the test (22,25,26). However, unlike sensitivity and specificity, they are not properties of the screening test itself, but of its application. The multivariate predictive equation provided a PPV of 64% and an OAPR of 1.75. These results were better than those obtained with all static plasma glucose cut points <180 mg/dl and indicate that among those with a positive test, 64% actually have diabetes (true positives), and the odds of having a true-positive test result are 1.75 times greater than the odds of having a false-positive result (Table 3). Tests with an OAPR <1 identify fewer true positives than false positives.

In summary, by incorporating relevant risk factor data, the predictive equation performs better in the general population than any single glucose cut point. The multivariate equation can be implemented with a number of inexpensive, programmable, handheld calculators. We programmed the formula and coefficients presented in RESEARCH DESIGN AND METHODS into a TI-83 graphic and scientific calculator (Texas Instruments, Dallas, TX). To obtain a probability value, the user enters the values for age (years), capillary plasma glucose (mg/dl), postprandial time (0 to ≥8 h), BMI (kg/m^2), and sex (0 = male and 1 = female). The calculator prompts the user by displaying the coefficient for the variable that should be entered next. The result displayed is the calculated probability that a subject has previously undiagnosed diabetes (a number between 0.0 and 1.0). The programming is available on request. Using this device and a glucose meter, a health care professional can perform a quick point-of-care assessment of the probability of undiagnosed diabetes in either a public health or clinical setting.

Acknowledgments— This work was supported by the U.S. Agency for International Development and the Egyptian Ministry of Health under PASA (Participating Agency Service Agreement) 236-0102-P-HI-1013-00, the Michigan Diabetes Research and Training Center under grant DK-20572, and the Centers for Disease Control and Prevention.

References

1. Herron CA: Screening in diabetes mellitus: report of the Atlanta workshop. *Diabetes Care* 2:357–362, 1979
2. U.S. Preventive Services Task Force: *Guide to Clinical Preventive Services: Screening for Diabetes Mellitus*. 2nd ed. Baltimore, MD, Williams and Wilkins, 1996, p. 193–208
3. American Diabetes Association: Screening for diabetes (Position Statement). *Diabetes Care* 24 (Suppl. 1):S21–S25, 2001
4. Engelgau MM, Aubert RE, Thompson TJ, Herman WH: Screening for NIDDM in nonpregnant adults: a review of principles, screening tests, and recommendations. *Diabetes Care* 18:1601–1618, 1995
5. Engelgau MM, Thompson TJ, Herman WH, Boyle JP, Aubert RE, Kenny SJ, Badran A, Sous ES, Ali MA: Comparison of fasting and 2-hour glucose and HbA_{1c} levels for diagnosing diabetes: diagnostic criteria and performance revisited. *Diabetes Care* 20:785–791, 1997
6. Engelgau MM, Nayaran KMV, Herman WH: Screening for type 2 diabetes. *Diabetes Care* 23:1563–1580, 2000
7. Rolka DB, Nayaran KMV, Thompson TJ, Goldman D, Lindenmayer J, Alich K, Bacall D, Benjamin EM, Lamb B, Stuart DO, Engelgau MM: Performance of recommended screening tests for undiagnosed diabetes and dysglycemia. *Diabetes Care* 24:1899–1903, 2001
8. Engelgau MM, Thompson TJ, Smith PJ, Herman WH, Aubert RE, Gunter EW, Wetterhall SF, Sous ES, Ali MA: Screening for diabetes mellitus in adults. *Diabetes*

Care 18:463–466, 1995
9. Herman WH, Ali MA, Aubert RE, Engelgau MM, Kenny SJ, Gunter EW, Malarcher AM, Brechner RJ, Wetterhall SF, DeStefano F, Thompson TJ, Smith PJ, Badran A, Sous ES, Habib M, Hegazy M, abd el Shakour S, Ibrahim AS, el Moneim el Behairy A: Diabetes mellitus in Egypt: risk factors and prevalence. *Diabet Med* 12:1126–1131, 1995
10. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 12:1373–1379, 1996
11. Bagley SC, White H, Golomb BA: Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *J Clin Epidemiol* 54:979–985, 2001
12. SAS Institute Inc.: SAS/STAT User's Guide, Version 6. Vol. 2., 4th ed., Cary, NC, SAS Institute Inc., 1990
13. SAS Institute Inc.: SAS/STAT software: changes and enhancements through release 6.12. Cary, NC, SAS Institute Inc., 1997
14. Hosmer DW, Lemeshow S: *Applied Logistic Regression.* 2nd ed. New York, Wiley, 2000
15. Steyerberg EW, Harrell FE Jr, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF: Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 54:774–781, 2001
16. Altman DG, Royston P: What do we mean by validating a prognostic model? *Stat Med* 19:453–473, 2000
17. Steyerberg EW, Eijkemans MJC, Harrell FE Jr, Habbema JDF: Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* 19: 1059–1079, 2000
18. Harrell FE Jr, Lee KL, Mark DB: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 15:361–387, 1996
19. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. *Ann Intern Med* 130:515–524, 1999
20. Steyerberg EW, Eijkemans MJC, Houwelingen JC, Lee KL, Habbema JD: Prognosyic models based on literature and individual patient data in logistic regression analysis. *Stat Med* 19:141–160, 2000
21. Katz MH: *Multivariable Analysis: A Practical Guide for Clinicians*. Cambridge, U.K., Cambridge University Press, 1999
22. Fletcher RH, Fletcher SW, Wagner EH: *Clinical Epidemiology: The Essentials*. 3rd ed. Baltimore, MD, Williams and Wilkins, 1996
23. Hennekens CH, Buring JE: *Epidemiology in Medicine*. Boston, MA, Little Brown, 1987
24. Harris MI, Flegal KM, Cowie CC, Eberhardt MS, Goldstein DE, Little RR, Wiedmeyer H-M, Byrd-Holt DD: Prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in U.S. adults. *Diabetes Care* 21:518–524, 1998
25. Greenberg RS, Daniels SR, Flanders WD, Eley JW, Boring JR III: *Medical Epidemiology*. New York, McGraw Hill, 2001
26. Cuckle HS, Wald N: Tests using single markers. In *Antenatal and Neonatal Screening*. 2nd ed. Wald N, Leck I, Eds. Oxford, U.K., Oxford University Press, 2000, p. 3–22