WORLD POLITICS
A Quarterly Journal of International Relations

World Politics 53.4, July 2001

Improving Forecasts of State Failure

Gary King Langche Zeng

Abstract

This article offers the first independent scholarly evaluation of the claims, forecasts, and causal inferences of the State Failure Task Force and its efforts to forecast when states will fail. State failure refers to the collapse of the authority of the central government to impose order, as in civil wars, revolutionary wars, genocides, politicides, and adverse or disruptive regime transitions. States that sponsor terrorism or allow it to be organized within their borders are all failed states. This task force, set up at the behest of Vice President Gore in 1994, has been led by a group of distinguished academics working as consultants to the U.S. CIA. State Failure Task Force reports and publications have received attention in the media, in academia, and from public decision makers. The article identifies several methodological errors in the task force work that cause its reported forecast probabilities of conflict to be too large, its causal inferences to be biased in unpredictable directions, and its claims of forecasting performance to be exaggerated. However, the article also finds that the task force has amassed the best and most carefully collected data on state failure to date, and the required corrections provided in this article, although very large in effect, are easy to implement. The article also demonstrates how to improve forecasting performance to levels significantly greater than even corrected versions of its models. Although the matter is still a highly uncertain endeavor, the authors are nevertheless able to offer the first accurate forecasts of state failure, along with procedures and results that may be of practical use in informing foreign policy decision making. The article also describes a number of strong empirical regularities that may help in ascertaining the causes of state failure.

Subject Headings:

- International relations -- Risk assessment.
- International relations -- Forecasting.
- State, The.

Research Note

IMPROVING FORECASTS OF STATE FAILURE

By GARY KING and LANGCHE ZENG*

I. INTRODUCTION

"STATE failure" refers to the complete or partial collapse of state authority, such as occurred in Somalia and Bosnia. Failed states have governments with little political authority or ability to impose the rule of law. They are usually associated with widespread crime, violent conflict, or severe humanitarian crises, and they may threaten the stability of neighboring countries. States that sponsor international terrorism or allow it to be organized from within their borders are all failed states. Since the consequences for the citizens of these states can be very severe and the costs to the international community of rebuilding the states are often substantial, there has long been considerable interest in developing methods of risk assessment and early warning systems in the hope that foreign aid could be directed to prevent states from failing. In 1994, with these goals in mind, the U.S. government, at the behest of Vice President Gore, established and funded the State Failure Task Force, a panel of distinguished academic social scientists, experts in data collection, and consultants in statistical methods. Although the

* The authors have no formal or informal relationship with the State Failure Task Force or the task force's sponsor, the U.S. Central Intelligence Agency. We thank Matt Baum for research assistance; Jim Alt, Aslaug Asgeirsdottir, Bob Bates, Ben Bishin, Lee Epstein, Jim Fearon, Charles Franklin, Jeff Frieden, Kristian Gleditsch, Jack Goldstone, David Laitin, Chris Murray, Kevin Quinn, Ken Scheve, Alan Stain, Ben Valentino, Jonathan Wand, and Mark Woodward for helpful discussions; the State Failure Task Force for collective written comments; Bob Bates for his suggestion that we take on this project; and the National Science Foundation (SBR-9729884, SBR-9753126, and IIS-9874747), the National Institutes of Aging, and the World Health Organization for research support. For providing us copy of the task force data, we thank task force members Bob Bates and Monte Marshall. All data referenced in this article are available at http://gking.harvard.edu, and for making this pos sible we are thankful for the guidance and assistance of Dick Cooper and the efforts of attorneys Kim Budd, Bob Donin, Allan Ryan, and Bob Iulioano in the Harvard University General Council's office.

World Politics 53 (July 2001), 623–58

task force does not use classified information, the data amassed are nonetheless impressive: more than a thousand variables, each carefully collected and documented and many with value added beyond what is available from other sources. (See Appendix 1.) The task force, still in operation, has produced over two hundred pages of widely distributed formal reports and analyses[1] and several published article-length summaries.[2] This work has received attention in the popular news media[3] and "has gained substantial visibility and credibility among those responsible for the analysis of global security and for planning U.S. foreign policy,"[4] an uncommon achievement for quantitative analyses in this field.

The task force reports were aimed at policymakers, but the research has been of considerable interest to the scholarly community as well. The authors make stunning claims about their success at forecasting these highly heterogeneous and idiosyncratic events, and they draw numerous important inferences about the causes of a critical and understudied political phenomenon. In this article we provide the first independent scholarly evaluation of the methods, analyses, and claims of the State Failure Task Force. We first identify and correct several methodological errors and then show how to use the task force's data to improve forecasts of state failure substantially beyond even appropriately corrected versions of its statistical models. We hope that this article can then help to connect the goals and efforts of the policy and academic communities in understanding and perhaps even addressing this critical global problem. The work analyzed here also touches on an unusually wide range of underutilized methods and relevant methodological issues; we seek to clarify some of these so that scholars can use them more productively.

[1] Daniel C. Esty, Jack Goldstone, Ted Robert Gurr, Pamela T. Surko, and Alan N. Unger, *Working Papers: State Failure Task Force Report* (McLean, Va.: Science Applications International Corporation, 1995); Daniel C. Esty, Jack Goldstone, Ted Robert Gurr, Barbara Harff, Pamela T. Surko, Alan N. Unger, and Robert S. Chen, *The State Failure Task Force Report: Phase II Findings* (McLean, Va.: Science Applications International Corporation, 1998).

[2] Daniel C. Esty, Jack Goldstone, Ted Robert Gurr, Barbara Harff, Pamela T. Surko, Alan N. Unger, and Robert S. Chen, "The State Failure Project: Early Warning Research for U .S. Foreign Policy Planning," in John L. Davies and Ted Robert Gurr, eds., *Preventive Measures: Building Risk Assessment and Crisis Early Warning Systems* (Lanham, Md.: Rowman and Littlefield, 1998); Daniel C. Esty, Jack Goldstone, Ted Robert Gurr, Barbara Harff, Marc Levy, Geoffrey D. Dabelko, Pamela T. Surko, and Alan N. Unger, "The State Failure Report: Phase II Findings," *Environmental Change and Security Project Report* 5 (Summer 1999).

[3] For example, Tim Zimmermann, "CIA Study: Why Do Countries Fall Apart? Al Gore Wanted to Know," *U.S. News and World Report,* March 12, 1996.

[4] Esty et al. (fn. 2, 1998), 27–38; and, e.g., John C. Gannon, *The Global Infectious Disease Threat and Its Implications for the United States* (U.S. National Intelligence Council, http://www.cia.gov/cia/publications/nie/report/nie99-17d.html, 2000).

II. Task Force Data and Models

According to the task force, a "state failure" consists of revolutionary wars ("sustained military conflicts between insurgents and central governments, aimed at displacing the regime"), genocides and politicides ("sustained policies by states or their agents and, in civil wars, by contending authorities that result in the deaths of a substantial portion of members of communal or political groups"), and adverse or disruptive regime transitions ("major, abrupt shifts in patterns of governance, including state collapse, periods of severe regime instability, and shifts toward authoritarian rule)."[5] The authors intentionally included fairly diverse events in their definition of state failure in order to follow the guidelines of policymakers as articulated to the task force. This may be a reasonable starting point, in part, because it increases the number of events in the data set, but also because it assumes that the benefits of having more events outweigh the costs of lower predictive ability and model incoherence resulting from increased heterogeneity in the outcome variable. In this article we use the dependent variable as conceptualized and measured by the task force (in order to isolate the effects of our methodological corrections), but in all likelihood a different causal structure underlies each component of the constructed concept of "state failure." Although our very flexible model will pick up some of these differences, we recommend that future researchers experiment with fitting different models to each component. We return to this issue in the conclusion.

According to the definition used by the task force, 127 state failures had commenced between 1955 and 1998 in some of 195 distinct countries in the data set.[6] Thus, the outcome variable is state failure, which we denote as Y_i, coded 1 for country-years in which a state failure started and 0 for country-years with no failure. Since the goal of the task force was to explain the onset of state failure (incidence rather than prevalence, in epidemiological terms), subsequent years in which a country remained in a state of failure are dropped.

The task force collected its data via a case-control design, which is especially efficient for rare events data.[7] First, it collects all cases of fail-

[5] Esty et al. (fn. 2, 1998), 27–38.

[6] Fewer than 195 countries appear in the data set in any one year. For example, Germany, East Germany, and West Germany are three separate items in this count, even though for any one year in the data set, either Germany or East and West Germany appear. Countries enter the data set in 1955 or when they first came into existence if later; countries remain in the data set after an episode of failure. In addition, the task force was required by the U.S. government to omit the United States from all analyses. They also omitted countries with fewer than half a million people.

[7] Norman E. Breslow, "Statistics in Epidemiology: The Case-Control Study," *Journal of the American*

ure. Then for each failure it randomly draws three nonfailures from the same year. The advantage of this scheme is clear in comparison with a random sample, which could by chance miss many important events. The 1:3 ratio is not required for case-control studies, although all cases of failure were used and every additional control adds progressively less information after the first, so there is little point of continuing much further if data collection is expensive. The task force then coded hundreds of explanatory variables.[8]

To choose a model, the task force authors lagged all the explanatory variables by two years so that their models would predict two years into the future. They then conducted an extensive search process fitting their entire data set via logistic regression numerous times to different specifications (that is, sets of explanatory variables). They used genetic algorithms, stepwise logistic regression, and other informal procedures to examine other specifications. Listwise deletion was applied to each specification by deleting a country-year if any variable in the model was missing (so that each logit model was run on a different set of country-years). They also report performing a first cut, narrowing the list of variables to thirty-one on the basis of univariate *t* or chi-square tests. Then "combinations of two, three, five, and up to 14 variables contained in the 31-variable set were examined together in an inductive approach to specifying the most accurate analysis or model."[9] Even this second step was constrained by the qualitative knowledge of the authors, as well as by some external rules, since the number of combinations of explanatory variables that could have been tested in this way is over 773 million (if each combination took 10 seconds to run and evaluate, it would take 245 years to complete them all). The summary measure they used to judge the quality of each model was not identified.

This process led the task force to choose a simple logistic regression model with three variables: *democracy* (the standard Polity III democracy and autocracy scores, collapsed into the categories of full democracy, partial democracy, and autocracy and coded as two dummy variables), *trade openness* (the log of imports plus exports as a percentage of GDP), and *infant mortality* (the log of the ratio of the infant mortality rate to the world median). (This entire process was repeated using

Statistical Association 91 (March 1996); Gary King and Langche Zeng, "Explaining Rare Events in International Relations," *International Organization* (forthcoming), preprint at http://gking.harvard.edu.

[8] We focus only on the task force's so-called "global model." Its data set includes 1,231 variables, although many of these are recodes of other variables or markers of problems with individual observations. Although the task force writings indicate that it used only the case-control data, its data set contains at least some information and always Y_t, for every country.

[9] Esty et al. (fn. 2, 1998).

"neural network clustering," the specific version of which the authors do not identify, although they conclude that it does no better than their logit model.)

There is much to criticize in this relatively atheoretical search procedure, but without a theory that rules out hundreds of the variables in the task force data set, uniformly superior statistical procedures do not yet exist (although see West et al.)[10] The task force's ultimate model choice was parsimonious and (aside from the case-control corrections, which had enormous effects) was not easy to surpass.

Since the analyses conducted by the task force and used by policymakers are based on continuously updated data, variables, and methods, we requested and received the data and results from their current model of choice, which still includes the same three variables. In all other ways, too, differences between the data used in their published report and those used in the newer version were minor and always inconsequential for the points discussed herein. Except where noted, we base all our analyses and comparisons on these new data. In the final task force model, 108 failures and 315 nonfailures were included, and 85 other observations were dropped due to the application of listwise deletion.

III. Correcting Methodological Errors

In this section we discuss methodological errors and opportunities for improving on the State Failure Task Force statistical analyses. We consider problems stemming from (1) their case-control design, (2) the way they evaluate forecasting performance, (3) the way they distinguish in-sample fit from out-of-sample forecasts, and (4) the way they treat missing data.

Case-Control Problems

The task force collects data by selecting on the dependent variable, a procedure known to cause bias.[11] As the task force makes no correction for this problem, all its estimates are therefore biased, in most cases quite severely.

Most obviously, the marginal distribution of Y is biased in case-control sampling. In the present case, the fraction of state failures in the

[10] Mike West, Joseph R. Nevins, Jeffrey R. Marks, Rainer Spang, and Harry Zuzuan, "Bayesian Regression Analysis in the 'Large *p*, Small *n*' Paradigm with Application in DNA Microarray Studies" (Manuscript, Duke University, 2000).

[11] For example, Gary King, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton: Princeton University Press, 1994).

Table 1
State Failure Task Force Logistic Regression Results

	Coefficient	*Standard Error*
Log (infant mortality)	.62	(.20)
Log (trade openness)	−.85	(.22)
Full democracy	−.16	(.40)
Partial democracy	2.00	(.33)
Constant	1.88	(.84)
Prior-corrected constant	−1.12	(.84)

data, $\bar{y}$, is $\bar{y}$ = 0.255, whereas the fraction of failures in the universe of country-years, which we denote τ, is only τ = 0.0168.

Although the costs of taking no action are large, case-control designs (unlike most methods of selecting on Y_i) are exceptionally easy to correct even without additional statistical assumptions or complicated models.[12] Intuitively, we must correct for the way the sample fraction of failures y misrepresents the population fraction, and so the technical fix in any particular statistical model will always involve weighting the sample to move from $\bar{y}$ to τ. Researchers call this procedure *prior correction.* For example, in logistic regression (and any other multiplicative intercept model, such as the neural network committee model we introduce below) prior correction involves merely subtracting $\ln[(1-\tau)/\tau][\bar{y}/(1-\bar{y})]$ from the estimated constant term. Although the slope coefficients require no correction, the nonlinearities of the logit model mean that forecasts and estimates of causal effects (and almost all other quantities of interest) cannot be computed without the constant term.[13] We now demonstrate the bias in the task force's forecasts and causal estimates.

We begin with a replication of the task force's best current model, which appears in Table 1. Substituting into the equation in the previous

[12] King and Zeng (fn. 7).

[13] Let X be a vector of k explanatory variables, including a constant term, and X_0 and X_1 each denote $1 \times k$ vectors of values of the explanatory variables (e.g., with one variable changing and the others remaining constant at their medians between X_0 and X_1). Quantities of interest usually include raw probabilities of failure, relative risks, and first differences. The first difference, $\Pr(Y=1|X_1) - \Pr(Y=1|X_0)$, is the increase in probability, and the relative risk, $\Pr(Y=1|X_1)/\Pr(Y=1|X_0)$, is the factor by which the probability increases, when the explanatory variables change from X_0 to X_1.

In one special case, the relative risk can be approximated indirectly without the constant term via an odds ratio, which in logit is a function of the slopes only. However, this approximation is accurate only as $\tau \rightarrow 0$, which is the assumption that no state is ever at risk of failure, in which case there would not be much point in forecasting state failure in the first place (although the bias can be small if τ is very small). In addition, the assumption implies implausibly that $\Pr(Y=1|X) = 0$ for any X and that all first differences are 0. For details, see Gary King and Langche Zeng, "Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Data," *Statistics in Medicine* (forthcoming), preprint at http://gking.harvard.edu.

TABLE 2
SELECTED PREDICTED PROBABILITIES[a]

State	*Year*	*Task Force*	*Prior Corrected*
Somalia	1988	.45	.04
Comoros	1995	.66	.09
Chile	1973	.56	.06
Ghana	1972	.55	.06
Ecuador	1970	.58	.06
Brazil	1964	.72	.11
Benin	1963	.72	.11

[a]From the task force logistic regression and from the authors' prior-corrected version.

paragraph, we calculate that the correction factor in this case is 3.0. We thus subtract 3.0 from the logit constant to produce the prior-corrected constant in the last line. (This may seem like a small number, but as we shall see, it has a large effect on the quantities of interest.)

BIAS IN TASK FORCE FORECASTS

Although the task force repeatedly refers to its forecast numbers as "probabilities," they are not probabilities since they were not prior corrected. For an example of the bias in the numbers produced by the task force, consider the simple case with no explanatory variables (or none with an effect). In this situation a predicted probability of failure, based on the global population of country-years or the corrected case-control sample, would equal 0.0168. However, the uncorrected case-control sample yields an estimate fifteen times larger, an incredible prediction that slightly more than a quarter of the states in the world will fail in any one year. When the task force includes explanatory variables, some of the probabilities extend to 0.89, which is implausibly large for this problem.

Table 2 gives a few examples of the overestimates in the task force's predicted probabilities.[14] The first column of numbers in our table, taken from Esty et al.,[15] is labeled "model score" and is explained as the "the predicted probability according to the model."[16] The second column contains the correct predicted probabilities that we computed.[17]

[14] All country-year predictions were highly biased. For this illustration, we chose a few cases that might be familiar and a few that were less familiar. Readers can easily compute the bias in all other country-years using our methods, a hand calculator, and their tables.

[15] Esty et al. (fn. 1, 1998), Table A-7.

[16] Ibid., 57.

[17] We computed these via prior correction from numbers given in Esty et al. (fn. 1, 1998), by using equation 26 in Gary King and Langche Zeng, "Logistic Regression in Rare Events Data," *Political*

TABLE 3
BIASED AND CORRECTED QUANTITIES OF INTEREST[a]
(WITH 95% CONFIDENCE INTERVALS IN PARENTHESES)

	Relative Risk			*First Difference*		
Autocracy to partial democracy						
task force	3.66	(2.45,	5.61)	.42	(.27,	.55)
corrected	7.04	(3.57,	13.21)	.06	(.03,	.09)
Full to partial democracy						
task force	4.14	(2.37,	7.66)	.43	(.27,	.58)
corrected	8.08	(3.58,	18.59)	.06	(.03,	.10)

[a] The relative risk is the ratio and the first difference is the difference in the probability of state failure when changing the democracy variables, holding other variables constant at their global medians.

By any substantive or statistical measure, the task force estimates are far from accurate and range from 6.54 to 11.25 times too large. The task force authors included several long tables in their reports with numerous estimated probabilities, but unfortunately the lack of prior correction means that every such estimate is incorrect, sometimes by more and sometimes by less than the examples in Table 2. Fortunately, these are easy to correct.

BIAS IN TASK FORCE CAUSAL INFERENCES

We demonstrate the bias in the causal effects estimated from the uncorrected case-control analysis with one key example, highly touted by the task force in all its writings: the effect of democracy on the probability of state failure. Table 3 gives relative risks and first differences (with 95 percent confidence intervals) computed from the original uncorrected model and with appropriate corrections. For example, when a nation moves from autocratic to partial democracy and other variables are held constant at their global medians, the task force's biased estimate is that the probability of state failure more than triples (increases by 3.66). However, the correct estimate (7.04) is nearly twice as large. A similar bias, of about a factor of two (4.14 to 8.08), occurs when moving from full to partial democracy.

Unfortunately, the bias correction does not always increase the size of estimated effects as it happens to with these selected relative risks.

Analysis (forthcoming), preprint at http://gking.harvard.edu. Since the raw data are not needed for this calculation, we stuck to the published version, which was based on a data set that differed slightly from the updated one we used. We also reproduce this with the new data, and there were only very minor differences.

For example, we also computed first differences for these same causal counterfactuals and found the bias to be in the opposite direction. Table 3 shows that a change from autocracy to partial democracy is estimated with the task force's methods to increase the probability of state failure by 0.42, which is an immense effect. The correction brings this down to a modest 0.06. A similar sevenfold change occurs when correcting the first difference for a change from full to partial democracy. (The direction of bias between the relative risk and first difference results may seem contradictory, but (a/b) and ($a - b$) are not constrained mathematically to change in the same direction as the estimates, a and b, change; the directions of the bias may also change in other examples.)

This same problem also occurs in the simpler context of comparing the raw numbers the task report reports as probabilities, since the correct probabilities are not preserved in the levels or even in the ratios of or differences between their numbers. For example, when the explanatory variables change from the profile of Somalia in 1988 to that of Brazil in 1964 (both taken from the task force and Table A-7 reproduced in our Table 2), the relative risk increases by a factor of 0.72/0.45 = 1.6 according to their numbers, but a much larger 0.11/0.04 = 2.75 when appropriately corrected. Similarly, the first difference indicates an increase in the probability of 0.72 − 0.45 = 0.27 according to the task force but of only 0.11 − 0.04 = 0.07 when appropriately corrected.[18]

Task force estimates without prior correction do preserve the ranking of the (in-sample) probabilities, but this ranking by itself is not useful for any policy purpose: it indicates only which country has a higher risk of failure, not whether any country is at high enough risk to warrant spending money or risking troops. And, as we will see in the next section, the rankings are not preserved in real out-of-sample forecast probabilities.

These results show that the causal estimates in the task force reports are unreliable and the biases are in otherwise unexpected directions and magnitudes. The biases in other comparisons of relative risks and first differences that we calculated (not shown) vary widely. Fortunately, the

[18] In our discussions with the task force, we learned that they sometimes estimated relative risks in Table 3 indirectly and approximately via an odds ratio (where prior correction is unnecessary; see fn. 13), rather than directly and without prior correction, as assumed here. The indirect approach is also biased except when the expected population of failures becomes 0. The indirect approximation (and even the phrase "odds ratio") is never mentioned in the task force reports or other publications, but if the task force had used it for its written work, then its relative risk estimates computed from the logistic regression in Table 1 are more accurate than indicated in our Table 3. However, the task force estimates of relative risks, such as those computed from the probabilities in Table 2 and described in the text above, would be as biased, and their estimates of probabilities and first differences would be considerably less accurate than we indicate.

corrections we provide can easily be used in future work to generate accurate figures.

All quantities labeled "causal effects" in this section are calculated based on the assumptions that the counterfactuals necessary for making causal inferences are correct and that the task force's model is appropriate. For the first, the task force assumes that infant mortality and trade openness are causally prior to the level of democracy. This means that if we could exogenously change the level of democracy in a country, then infant mortality and trade openness would not change as a result, as assumed in Table 3. For the second, the task force assumes that if any causes of state failure exist that are causally prior to and uncorrelated with democracy, then they are included in the equation; if an explanatory variable exists that meets these conditions other than democracy, trade openness, and infant mortality, the task force model has additional biases. The first assumption is implausible and unfortunately very hard to correct. The second assumption is by definition unverifiable but is considerably more plausible given the task force's extensive search for other predictive variables. We continue this discussion in Section VI.

Evaluating Forecasting Success

When appropriately corrected, the logistic regression models used by the task force give estimates of the probability of state failure conditional on chosen values of the explanatory variables $\hat{\pi}$ $\Pr(Y = 1|X)$. A separate step, governed by decision theory, is required to decide on the basis of $\hat{\pi}$ whether the state in question will fail or not.

Let C denote the *c*ost of mispredicting a state failure as a nonfailure relative to the cost of mispredicting a nonfailure as a failure. Decision theory tells us that whatever C is, the optimal prediction (in the sense of minimizing total expected cost) is $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and $Y = 0$ otherwise.[19] Hence, if the two possible mispredictions are equally costly, then $C = 1$ and we would predict that a state will fail when $\hat{\pi} > 0.5$. However, if the cost of mispredicting a state failure is, say, twice as costly as mispredicting a nonfailure, then $C = 2$, and an optimal decision process would predict state failure whenever $\hat{\pi} > 1/3$.

Only by applying decision theory in this way can we compare model outputs to the data and judge our success in prediction. The key, however, is that the value of C must be decided *independently* of the data and statistical results. The task force violated this rule and instead "di-

[19] For example, B. D. Ripley, *Pattern Recognition and Neural Networks* (New York: Cambridge University Press, 1996).

vided errors evenly between 'false positives' and 'false negatives.'" Of course, the only way to "divide errors evenly" is to inspect the actual values of *Y* and change *C* on that basis, which is inappropriate. This is easy to see when forecasting out of sample for genuine policy purposes, since we would not know the future values of *Y* when *C* was being chosen.

On the basis of its post hoc adjustment of *C*, the task force concludes: "The models classify correctly about 70% of historical cases. A model with 70% accuracy two years in advance would correctly identify about two out of three failures and two out of three stable countries." (Using the task force's methods with its new data results in nearly the same number.) Since *C* cannot be adjusted post hoc in real forecasting, these figures are overstated.

Although the task force computes a threshold after the fact without knowingly choosing *C*, we can still back out its implicit choice. Our calculation indicates that its procedures assume that the costs of mispredicting a state failure is $C = 60.2$ times more costly than mispredicting a nonfailure.[20] This value for *C* would cause a bilateral or international aid agency to "waste" funds on sixty countries not at risk for failure for every one that really is at risk. The task force's given perspective was to use its data and methods to help the U.S. foreign policy establishment narrow its focus and direct foreign aid at a small number of high-risk countries in hopes of making a difference. However, only about three states fail in any one year. As such, $C = 60$ means that the focus of foreign policy would almost not be reduced at all from the list of all countries in the world (191) to $3 \times 60 = 180$. With aid dollars as restricted as they are, this seems like an implausible summary of the political or economic situation and is probably not a useful decision rule.[21]

What is a reasonable summary of the task force's forecasting performance? If $C = 1$, the most commonly used value in other contexts but probably too small here, the task force would correctly classify 0 percent of failures accurately. If *C* were large enough, they could correctly classify as much as 100 percent of failures, at the cost of mispredicting many nonfailures. Similarly, predicting that states will never fail would have accurately predicted 98.3 percent of all state-years (that is, 100

[20] Esty et al. (fn. 1, 1998) report using a 0.26 cutting point, and they use 0.25 in their new data (which may conceivably indicate that they intended, although failed, to assign $C = 3$) (p. 57). This, by applying equation 26 from King and Zeng (fn. 17), translates to 0.01634 and thus implies that $C = 1/0.01634 - 1 = 60.2$.

[21] Of course, from the perspective of the people in countries at high risk, $C = 60$ might even be too small. A very useful future project would be to survey policymakers to measure their values for *C*. In all probability, *C* varies to some extent over people, countries, and time, but there surely are some patterns that would be helpful in evaluating future forecasting efforts.

percent of nonfailures and 0 percent of failures). In Section IV we use a method of evaluating the performance of forecasting models that works when we are unsure of an appropriate choice for C (or when different people might choose different values).

Forecasting versus Causal Structure

The primary goal of the task force is to make accurate forecasts, but it also draws causal inferences from the same models. Although many researchers seem to think that both goals cannot be accomplished with the same statistical models, this is inaccurate. Indeed, the only way that forecasts can remain accurate far into the future is if the causal structure giving rise to the data remains stable. That means that any claim to accurate forecasts is also implicitly a claim about causal structure. It is true that forecasts are often made using proxy variables (such as infant mortality) and possibly even theoretically uninteresting measures, but, almost by definition, prediction efforts not based in some way on causal structure will fail in the long run when the causal structure inevitably deviates from the convenient measures with which they were once correlated. The classic methodological warning about association not being causation applies equally well to forecasting efforts.

Similarly, almost all causal models that have been specified in international relations implicitly claim that the causal structure being estimated is stable and will remain so for at least some time into the future. Since a finding about a causal structure that changes unpredictably over time is of dubious value, most causal claims imply that accurate forecasts are possible; and, indeed, accurate forecasts are often the most powerful observable implications of the same causal models and can be used as validation for them.

Although there are models that can discern causal structure and in theory are unable to forecast, they are quite unusual. The theory of efficient markets in financial economics is the leading example. In the field of international conflict, Gartzke[22] and Bernstein et al.[23] develop theories which imply that forecasting should be impossible. However, without an explicit theory like this or some knowledge of the kinds of structural breaks that may occur in the future, we must regard models that make causal inferences as also capable of forecasting. If they are not in practice, then their value as causal models must also be ques-

[22] Erik Gartzke, "War Is in the Error Term," *International Organization* 53 (Summer 1999).

[23] Steven Bernstein, Richard Ned Lebow, Janice Gross Stein, and Steven Weber, "God Gave Physics the Easy Problems: Adapting Social Science to an Unpredictable World," *European Journal of International Relations* 6 (March 2000).

tioned. As such, scholars would do well to judge all models in terms of their forecasting success, regardless of the purpose for which they were originally developed.

In practice, of course, social science research is difficult. Optimally, the world produces a data set to compute forecasts, and new, truly out-of-sample data sets arrive daily that we can use to check the model continually and thereby sequentially improve the forecasts. A large number of out-of-sample tests enable us to rule out random chance and overfitting in accounting for any forecasting success.

Unfortunately, many applications offer only one data set, and in such situations it is difficult to know whether our statistical model is detecting the causal structure or the idiosyncratic features of the particular sample drawn. In the present case the task force had only one data set and tested countless specifications on it. At least according to the task force reports, no out-of-sample tests were conducted. As such, the odds are high that the task force overfit the idiosyncrasies in its data rather than the underlying structure, although it was quite disciplined in keeping to a parsimonious model. Indeed, the task force was careful to explain that its "models are based on historical analysis. It remains to he demonstrated that they will be equally accurate in identifying prospective cases of state failure."[24] Despite these cautionary words, however, academics and policymakers have read the models as making accurate forecasts.

The only way to be reasonably certain about whether its models can forecast would be to wait for more data to come in, but this will take many years, since state failures are rare events. By the time enough data arrive, the international community may miss many opportunities to prevent these disastrous events. In addition, we might also reasonably expect the actual underlying causal structure to have changed to some degree if we wait, and so waiting is not an effective option.

Instead, in our work below we follow standard procedure in forecasting studies and divide the observed sample into two parts, 1955–90 for fitting (or "training") a statistical model and 1991–98 for out-of-sample testing. We use the case-control data for our training set and the entire world for our test set. This means that our out-of-sample test set is a different time period as well as a different set of countries, making it especially difficult. Reserving multiple test sets would have been even better, but the rareness of events (127 in the entire period and only 27 since 1991) makes this infeasible. Overfitting and optimistic assess-

[24] Esty et al. (fn. 2, 1998), 27–38.

ments are still possible with only one test set, but it is considerably less likely than by evaluating out-of-sample forecasts from in-sample data only. Our choice of the point at which to split the sample between training and test sets is arbitrary, but given the massive changes in the world at about that time, from a cold war to post–cold war international regime, it may be the hardest test available to us. We also report the results of a variety of other stringent tests at the end of Section IV.

Missing Data

Listwise deletion, which the task force uses, is well known to be an inefficient procedure for dealing with missing data (since so many data were discarded). It also biases forecasts and causal inferences unless some implausible assumptions hold.[25] In the task force's final model, one of five observations was discarded. Bias also seems quite likely, since the state-years deleted were not representative of those included. For example, in the global data, 1.68 percent of state-years witnessed failures, but after deleting observations with at least one missing value on their three explanatory variables, this figure rose more than 50 percent to 2.58 percent (even though their dependent variable was fully observed). This would also seem to indicate that valuable information exists in a missingness indicator variable that could be recovered with a better procedure.

The problem of missing data in this application and the effects of listwise deletion on the task force results appear to be more severe than the consequences of dropping a nonrandom 20 percent of state-years from the final model. Listwise deletion also constrained the choice of model. As the authors write: "In many cases, we found that the gaps in the range of particular variables were so great that any possible gains in prediction were offset by statistical uncertainties or missing data problems associated with measuring those additional variables."[26] Because the task force has produced the best collections of data on state failure in existence, this problem results solely from its choice of a statistical procedure for dealing with missing data.

Since valuable information remains in the discarded cases and variables and since we wish to avoid the other problems with listwise deletion, we use multiple imputation[27] to impute the missing values (along

[25] Gary King, James Honaker, Anne Joseph, and Kenneth Scheve, "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review* 95 (March 2001).

[26] Esty et al. (fn. 1, 1998), 29.

[27] Donald Rubin, *Multiple Imputations for Nonresponse in Surveys* (New York: Wiley Press, 1996); King et al. (fn. 25).

with software by Honaker, King, Joseph, and Scheve).[28] The idea of multiple imputation is to fill in each of the missing data with several imputed values, creating several completed data sets (where the observed data are identical for all). Then whatever statistical procedure that would have been used in the absence of missing data is applied to each data set, and there is an easy procedure for combining the results from the different data sets. Unlike listwise deletion, this uses all information in the data and appropriately represents uncertainty by filling in the missing values. Since the information to be imputed under our approach is far less than the amount of data discarded under listwise deletion, imputation tends to be more robust. We encourage methodologists to develop multidimensional neural network models for use in imputing missing data, since one could then improve on the missing value techniques used here.

IV. An Improved Forecasting Model

We now discuss the statistical specification for our improved model, evaluate its forecasting performance, and summarize a variety of unusually stringent additional tests we used to ensure against overfitting. In each case, we compare this model with the corrected version of the task force model.

Statistical Specification

We began with the task force's three-variable model and added from their data set a variable we constructed for the *military population* (a logistic transformation of the fraction of the total population of a country in the military). The logic is based on the "resource" (that is, rather than grievance) component of conflict theory:[29] the larger the fraction of the population that has weapons and is trained in military conflict, the more risk there is that internal dissent may lead to state failure.[30]

We also built a *population density* variable (the log of the number of people per square mile relative to the regional median), under the

[28] James Honaker, Anne Joseph, Gary King, and Kenneth Scheve, "Amelia: A Program for Missing Data" (http://gking.harvard .edu, 2000) .

[29] Ted Robert Gurr, "Why Minorities Rebel: A Global Analysis of Communal Mobilization and Conflict since 1945," *International Political Science Review* 14, no. 2 (1993); James B. Rule, *Theories of Civil Violence* (Berkeley: University of California Press, 1988), 178; Mark Lichbach, *The Rebels' Dilemma* (Ann Arbor: University of Michigan Press, 1995), 4–6.

[30] See also John D. McCarthy and Mayer N. Zald, "Resource Mobilization and Social Movements: A Partial Theory," *American Journal of Sociology* 82 (May 1977); Charles Tilly, *From Mobilization to Revolution* (New York: McGraw-Hill, 1978).

simple theory that internal conflict requires people to be near others who might disagree. Population density is a very crude indicator of physical proximity, which, as Lichbach[31] (and the many references therein) explains, should reduce collective action costs by making the communication of grievances easier and by allowing for repeated interactions and therefore more trust. Collier[32] is also interested in population density but finds the opposite result. Fearon and Laitin[33] find results that support the theory they describe as not robust to specification decisions.

As a measure of the institutionalization of democratic institutions, we also included the task force's measure of *legislative effectiveness* (qualitatively coded as none, largely ineffective, partly effective, and effective). Przeworski et al.[34] argue that parliamentary institutions make a democracy more likely to endure. Legislative effectiveness is an important component of democratization, but it is sufficiently distinctive and divergent from the other components that we control for it separately.

Like the task force's variables, our additions are reasonable choices and widely discussed in the qualitative literature as possible risk factors, but they are not derived from anything approaching an empirically verified formal theoretical model. A sufficiently convincing story can be told about the theoretical expectations for each of these six variables (seven, if we count the two democracy dummies separately), but instead of pretending that our "hypotheses" were constructed ex ante, we prefer to recognize this as an exploratory analysis. Our more modest goal for this stage (that is, in addition to the more difficult goal of producing reliable forecasts) is to identify empirical regularities that may help in building theories rather than to test an existing fully specified theory.

This line of work is still quite valuable from a theoretical perspective by virtue of the substantial evidence it provides against all theories that assign a role to any variable other than the six in the present analysis (from the task force's original set of 1,231). The qualitative literature on the causes of state failure and its various components is far richer than is summarized in these six variables, but unless some case can be made

[31] Lichbach (fn. 29), 158–65.

[32] Paul Collier, "Economic Causes of Civil Conflict and Their Implications for Policy," in Chester A. Crocker, Fen Osler Hampson, and Pamela Aall, eds., *Managing Global Chaos* (Washington, D.C.: U.S. Institute of Peace, 2000), 6.

[33] James Fearon and David Laitin, "Weak States, Rough Terrain, and Large Scale Ethnic Violence since 1945" (Paper presented at the annual meeting of the American Political Science Association, Atlanta, 1999).

[34] Adam Przeworski, Michael Alvarez, J. A. Cheibub, and F. Limongi, "What Makes Democracies Endure," *Journal of Democracy* 7 (January 1996).

that the largest state failure data set ever constructed excludes variables identified in these theories, these theories can be regarded as inconsistent with the data and should be rejected.

Finally, our prior work suggests that we should expect massive interactions and nonlinearities, just as in international conflict data,[35] in part since the effects of the explanatory variables are expected to differ over types of countries and regions and because of the heterogenous definition given for state failure. In contrast, assuming that all or most interactions are absent, as most scholars do (and as the task force did) when they use logit models even with some interactions, is a heroic assumption. The "curse of dimensionality" ensures that the six-dimensional space represented by all the linear and nonlinear interactions of our six explanatory variables, plus the possible nonlinearities in the main effects, is almost incomprehensibly immense.[36] Since no accepted theory can rule them out, and few theories have even addressed the issue, we prefer not to assume knowledge of these interactions, beyond some smoothness in the functional forms, and instead introduce a model capable of estimating what it can from these data. We then use extensive out-of-sample tests (described below) to protect against being fooled by overfitting. We therefore follow this rule when feasible: *when we know something, we assume it; when we don't know, we estimate it.* As long as the estimation process is scientifically disciplined, this approach is superior to making draconian assumptions without empirical knowledge.

We impute the missing data and then use the neural network statistical model described in Beck, King, and Zeng,[37] modified by what are known as "committee methods." A neural network model is parametric and is just like a logistic regression analysis except that the functional form can take on many shapes in addition to the logit model's escalator-

[35] Nathaniel Beck, Gary King, and Langche Zeng, "Improving Quantitative Studies of International Conflict: A Conjecture," *American Political Science Review* 94 (March 2000).

[36] To understand the curse of dimensionality in this context, consider a regression with one continuous dependent variable and a single ten-category explanatory variable. To estimate this regression without assumptions, we need to estimate ten quantities, the mean of *Y* within each of the ten categories of *X* (e.g., the mean starting salary for people with each of ten levels of education). We could easily do this if we had, say, a sample of one hundred observations within each of the ten categories. By contrast, linear regression would summarize these ten numbers with only two, a slope and a constant term, by making the assumption that nothing is being lost. Now suppose we added one more ten-category explanatory variable. The curse of dimensionality is that we need to multiply not add—to estimate one hundred quantities, not merely twenty (graphically, we move from a bar chart to a checkerboard where the height of each square represents dollars of starting salary). An analysis with, say, ten ten-category explanatory variables requires the estimation of ten billion quantities, and summarizing that with a linear regression that has only eleven parameters and maybe even a few (linear) interaction terms is a stunningly strong assumption.

[37] Beck, King, and Zeng (fn. 35).

shaped curve (ignoring the steps!), depending on what the data suggest. We summarize these models in Appendix 2. The idea of committee methods is to run a number of neural network models, varying the number of hidden neurons and prior weights (which together indicate how much smoothness we expect in the functional form). The individual models are then combined, either by weighting them according to some estimate of performance or by simple averaging. Much empirical and some theoretical work indicates that simple averaging, which we use, usually works better because in-sample estimates of performance tend to be highly variable. Bishop[38] proves that committee methods improve out-of-sample performance through variance reduction. Committee methods remove some of the arbitrariness that accompanies the real use of most statistical methods that require choosing one model specification out of a large potential set.[39]

Evaluating Forecasting Performance

We now provide evidence that our new model forecasts better than the task force model, regardless of the costs one assigns to the two types of misclassification. To do this, we use a *Receiver-Operating Characteristic* (ROC) curve.[40] (This obscure terminology comes from signal processing theory, where the receiver [a decision maker in our framework] must decide whether each item in a string of noisy binary data is really a 0 or a 1.) The ROC graph thus has the fraction of 0s correctly predicted plotted vertically and the fraction of 1s corrected predicted plotted horizontally. The key point is that for any value of *C*, a model and data will produce only *one* pair of numbers for the percentage of failures correctly predicted and the percentage of nonfailures correctly predicted. This one pair of numbers appears as *one point* in an ROC graph, such as that in Figure 1. Changing *C* a little at a time over its entire possible range and plotting the corresponding pairs of percentages correctly predicted values on the graph give the complete ROC curve.

[38] Christopher M. Bishop, *Neural Networks for Pattern Recognition* (Oxford: Oxford University Press, 1995), 366.

[39] All members of the committee that constituted our model were based on the same input variables and three numbers: a random number seed for the starting values (which we include here to make it easier to replicate our results), the number of hidden neurons, and the prior standard deviation for the weights. The triples for the members of our committee are 45,3,1; 8,3,2; 908,3,3; 85,3,5; 908,4,2; 35,5,1; 12345,5,5; 768,5,6; 134,5,10; 8,7,3; 9,8,5; 45,8,6; 923,10,1. In general these are all fairly smooth neural network models. We chose this set based on our experience in fitting analyses to similar data and through some preliminary analyses. We expect models that can forecast even better could be developed.

[40] See D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*, rev. ed. (Huntington, N.Y.: Krieger, 1974); C. E. Metz, "Basic Principles of ROC Analysis," *Seminars in Nuclear Medicine* 8 (Spring 1978).

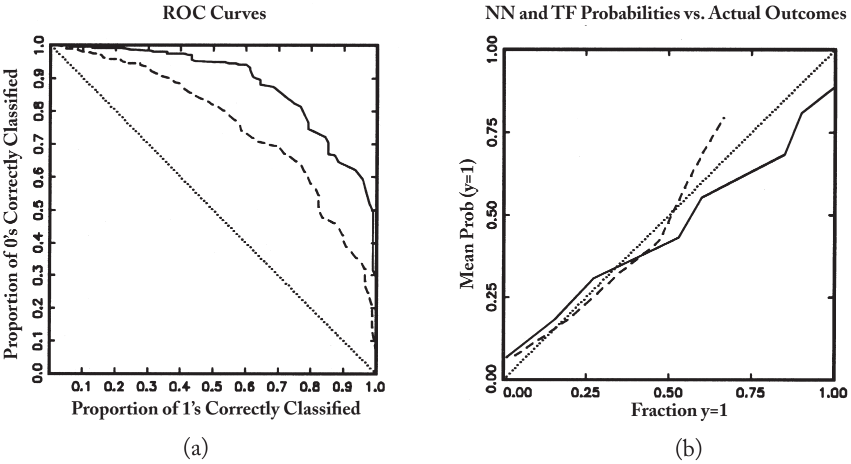

FIGURE 1
IN-SAMPLE CASE-CONTROL MODEL FIT[a]

[a]Graph a is the ROC curve, which evaluates comparative model performance, and graph b evaluates the veracity of the estimated probabilities. In both graphs the dashed line refers to the task force model and the solid line is from the neural network model.

We offer an example of the ROC curve in the left graph in Figure 1 and now discuss how to interpret it. For reference, the diagonal line in this graph is what the ROC curve would be if probability values were selected randomly (from a uniform distribution) and unrelated to the data. The top right point in the graph indicates 100 percent of 1s and 0s correctly predicted. Thus, the farther above the diagonal line a curve drawn on the basis of an empirical model falls, the better is the model performance. On the basis of this type of analysis, a researcher would conclude that one model dominates another if its curve is greater than the other model for every point, that is, for every possible cost of misclassification. In situations where a model performs better in some areas and worse in others, we may need to narrow the range of possible values for *C* in order to choose one model over the other. Out of sample, however, we do not know until after the fact where we are on the curve, that is, what are the percentages of 0s and 1s correctly predicted.

In graph a of Figure 1, the dashed line is computed from the task force model estimated with the in-sample case-control data from 1955 to 1990 and evaluated in the same data (which makes it a measure of fit, not of forecasting ability). In addition, the solid line in the graph gives the ROC line from our neural network model. Since this solid line is always above the dashed line, the neural network model dominates the task force model, no matter what normative decision one might make about the costs of misclassification, *C.* Of course, since this graph is both fit and evaluated in the same data, it indicates that the neural network model fits the data better, not that it necessarily forecasts better.

Before moving to an evaluation of out-of-sample performance, we also offer a test of the veracity of the estimated probabilities. A probability that is accurate gives the fraction of times a state with the given characteristics will actually fail. To evaluate these probabilities, we sort estimated probabilities into bins of 0.1 width: [0,0.1), [0.1,0.2),…,[0.9,1]. For observations falling in each bin, we compute the mean predicted probability from the model (which will often be somewhere near the midpoint of each interval), as well as the observed fraction of 1s. If model probabilities are accurate, these two quantities should be close: for example, if the probability of failure is forecast to be 0.2 for a group of states, then about 20 percent of these states should actually fail. We then compare the two in graph b in Figure 1 to check the fit of the model in the training set (and below to evaluate the forecasts in the test set). In this figure both models are fairly close to the 45-degree line, indicating fairly accurate in-sample probabilities. The graph reveals the neural network model to have more informative probabilities (higher values), although these appear to be slightly underestimated (perhaps suggesting that the neural network priors should be adjusted to allow somewhat less smoothness, although we do not follow up on this minor point).

We now consider the more important out-of-sample performance of both models. Figure 2 gives analogous graphs, estimated from the 1955–90 data and evaluated in the 1991–98 data. When we refer to the "task force" model in these graphs, we make the case-control corrections described in Section III (otherwise, the probabilities in graph b would be far worse).

As can be clearly seen, in the ROC graph a, the (solid) line for the neural network model is always above the (dashed) line representing task force model. Thus, for every value of *C,* the neural network model has a higher percentage of 1s correctly predicted and a higher percentage of 0s correctly predicted out of sample. Whatever one's normative preferences, therefore, the neural network model is superior to the (prior-corrected) task force model.

Graph b in Figure 2 indicates that the neural network probabilities are both more accurate (closer to the 45-degree line) and more informative than those for the task force model (because they extend farther

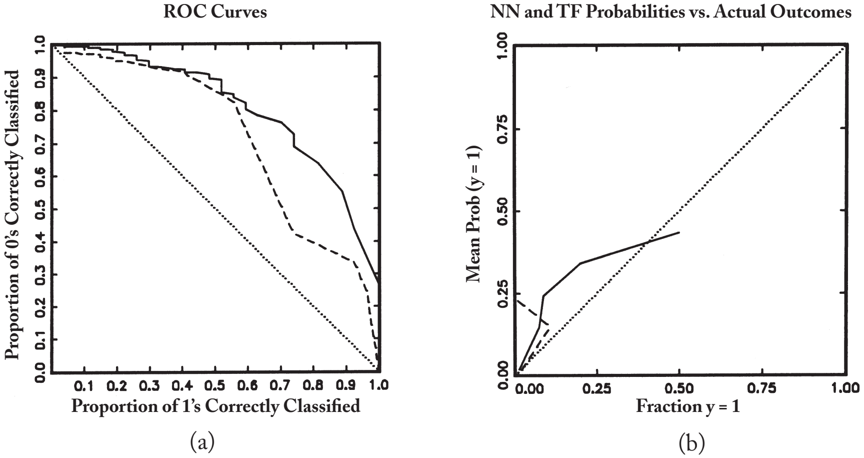

FIGURE 2
OUT-OF-SAMPLE GLOBAL MODEL FORECAST[a]

[a]Graph a is the ROC curve, which evaluates comparative model performance, and graph b evaluates the veracity of the estimated probabilities. In both graphs the dashed line refers to the task force model (although we also prior corrected, as in Section III) and the solid line is from the neural network model.

up the diagonal). The task force figures here are prior corrected, which is why they do not extend as high as in Figure 1 and why they are anywhere near the 45-degree line. Unfortunately, even with the correction, they do fairly poorly and are not accurate. Indeed, even prior correction is insufficient, since the out-of-sample ranking of states in the probability of failure is not preserved in the task force model. The dashed line doubling back on itself means that higher estimated task force probabilities actually correspond to lower actual rates of state failure. The solid line, representing our neural network analysis, is not perfect, but it indicates that our estimated probabilities are at least monotonically related to actual instances of state failure and are usually fairly close to the diagonal equality line. Taken together with the ROC graph, the available evidence indicates that our neural network committee model offers out-of-sample forecasts that are better than the (prior-corrected) predictions of the State Failure Task Force.[41]

[41] Each of the methodological improvements we made to the task force model improved results over the same model without that feature, and all were necessary to generate a model that dominated the (prior-corrected) task force model for any value of *C*. Of course, prior correction alone was sufficient to improve a great deal on the original task force analysis. A rough ranking from most to least important in changing the results is prior correction, neural networks, committee methods, the additional covariates, and multiple imputation for missing data.

Additional Tests to Ensure against Overfitting

We conducted several additional tests to verify our claims to have built a model that can forecast more accurately and to have picked up some piece of the underlying structure. In addition, we make these tests somewhat more difficult by using various subsets of our case-control data for our training sets and subsets of the global data for our test set. Taking all this into account, we can think of no political science modeling exercise that has applied more stringent tests, whether for the purpose of forecasting or for estimating causal effects.[42]

For our first test, we divided the country-years randomly between test and training sets and examined the ROC and probability graphs, as in Figures 1 and 2. In all cases, aside from what would be expected due to random error, our neural network committee approach dominated the (prior-corrected) task force model. We also computed the marginal effect graphs we report in Section VI and found that the causal structure uncovered stays quite stable across the different random subsets.

We also use what we call the Stanford Test (so named because several Stanford faculty and fellows suggested it at a talk we gave there on a related subject), which combines a simple version of cross-validation with out-of-sample verification. First, all countries are randomly divided into two groups, which we label A and B. The training and test sets are defined by dividing the sample chronologically at 1990, as before. Then country group A in the training set is used to forecast country group B in the test set. Similarly, country group B in the training set is used to forecast country group A in the test set. Hence, the forecast is both out of sample (to a future time) and out of space (to different countries). This is one of the hardest (reasonable) tests that can be constructed to ascertain whether a statistical model has uncovered a stable, causal structure and has not overfit the idiosyncrasies in the data that do not persist. Although difficult, the Stanford Test is reasonable to apply to any analysis aimed at uncovering lawlike causal statements or making genuine policy-relevant forecasts. (Although we do not pursue the possibility beyond the use of the test here, the procedure could be profitably generalized to all possible subsets A and B and formalized to yield sampling probabilities.)

In the present case our test was made more difficult by the fact that there are only twenty-seven events in our test set after 1990 and thus only about half that in country groups A and B. But even with the sam-

[42] We summarize the results of these tests here, rather than presenting detailed accompanying figures, since this would involve including numerous figures for each one presented in this paper.

pling error induced by using only half the data at a time, the forecasting performance, as judged by the ROC and probability graphs, and the causal structure, as judged by our marginal effect plots, were all quite consistent with one another. In virtually all cases our neural network committee approach dominated the (prior-corrected) task force model.

We also examined whether the substantive variables we measured, along with the estimated functional form, were sufficient to pick up the information represented in the country labels. That is, in almost any country-level analysis, a set of regional dummy variables will correlate highly with the outcome variable. The key is determining whether one's substantive variables pick up that variation, making the ad hoc idiosyncrasies of including dummy variables unnecessary. In our case we compared the ROC curve for our model with one where we also added a set of regional dummy variables. Predictably, the model fit the in-sample (training set) much better. However, the key is that the out-of-sample forecasts to our test set were considerably worse. This indicates that, indeed, we were able to "get rid of proper nouns":[43] the dummy variables are unnecessary and so our substantive variables apparently do not exclude any important structural components that correlate with the country names.

We also tried several measures of economic growth, because the literature at least since Huntington[44] has suggested that growth might lead to state failure by empowering middle classes in authoritarian regimes. Unfortunately, we were unable to find evidence in support of this hypothesis or evidence that growth in any way adds forecasting power to our models. We also tried adding the number of years since the last state failure, to model time-series dependence in the data,[45] but we found no evidence to support this variable either, although some of this effect might be represented in existing variables. In addition, we tried a time trend, but like the regional dummies it helped fit the in-sample data better but forecast considerably worse. We also examined many other individual variables from the task force data set, but finding no evidence that they could help in forecasting or would alter our substantive conclusions, we excluded all of them.

[43] Adam Przeworski and Henry Teune, *The Logic of Comparative Social Inquiry* (Malabar, Fla.: Krieger, 1982).

[44] Samuel P. Huntington, *Political Order in Changing Societies* (New Haven: Yale University Press, 1968).

[45] Nathaniel Beck, Jonathan Katz, and Richard Tucker, "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable," *American Journal of Political Science* 42 (October 1998).

The results of these successful validations from test sets defined on the basis of time, random assignment, the Stanford Test, the regional dummies, and other substantive variables cause us to be more confident about our ability to forecast and to believe we have found something approximating causal structure. We could still be wrong, and past performance is no guarantee of future success; but it seems hard to argue that these tests would not increase the chances that our model would hold for new data not yet collected. We think that future research by the task force and others in this and related fields would benefit from applying these procedures.

V. Sample Forecasts

Studying individual forecast probabilities of state failure must be done carefully, following the ideas about decision theory presented in Section IV. A key issue is that a probability that may seem high from one perspective could easily be considered low from another.

Our individual country-level forecasts easily distinguish countries that obviously do not have much risk of failure from those with some risk, something that previous literature has not accomplished. For example, the published task force numbers—the only figures that have been used by policymakers—include a forecast that France would fail in 1960 with an incredibly high 0.29 probability. In contrast, our model forecasts a probability of state failure in France of nearly 0 for almost all years. More difficult is distinguishing among the countries with some risk of failure. For example, with information available in 1964, our model gave a 0.35 probability of Uganda failing in 1966 (which is very high for the probability of failure occurring in just one year). Uganda failed that year. By contrast, our model gave a forecast probability of Kenya failing in that year of less than a third of the probability for Uganda (Kenya did not fail).

Of course, these examples are anecdotes culled from the thousands of predictions that come from our model and that cannot be fully presented in this short article. Instead, we now summarize the numbers by giving, in Table 4, our best and worst forecasts for all post-1990 countries, based on the data from the case-control subset of countries up to 1990. (Thus, our methods could be used to compute much better two-year-ahead forecasts than these by using all the information available; the 1998 forecasts, for example, are based on estimates from more than eight years earlier.) For example, the ten country-years in the top left section of the table are those with the highest forecast probabilities of

Table 4
Best and Worst Predictions

Best Predictions			*Worst Predictions*		
1. Highest Prob (Failure), Failure Observed			*1. Lowest Prob (Failure), Failure Observed*		
Country	*Year*	*Prob.*	*Country*	*Year*	*Prob.*
Senegal	1991	.5307	Mexico	1994	.0137
Kyrgyzstan	1995	.4563	Bosnia-Herzegovina	1992	.0174
Kazakhstan	1995	.4147	Algeria	1991	.0259
Cambodia	1997	.4122	Sierra Leone	1991	.0312
Georgia	1998	.3913	Congo-Kinshasa	1992	.0333
Armenia	1994	.3470	The Gambia	1994	.0405
Guinea-Bissau	1998	.3214	Haiti	1991	.0592
Thailand	1991	.2787	Lesotho	1994	.0627
Zambia	1996	.2348	Tajikistan	1992	.0667
Georgia	1991	.2285	Indonesia	1997	.0781
2. Lowest Prob (Failure), Nonfailure Observed			*2. Highest Prob (Failure), Nonfailure Observed*		
Country	*Year*	*Prob.*	*Country*	*Year*	*Prob.*
Finland	1992	.0014	Peru	1998	.5955
Sweden	1994	.0015	Bangladesh	1996	.4841
Estonia	1997	.0015	Kyrgyzstan	1993	.4781
Finland	1993	.0016	Kyrgyzstan	1994	.4752
Switzerland	1993	.0016	St. Kitts-Nevis	1997	.4570
Switzerland	1994	.0016	Uzbekistan	1991	.4394
Norway	1993	.0017	Bangladesh	1995	.4355
United Kingdom	1994	.0017	Kyrgyzstan	1997	.4187
United Kingdom	1996	.0017	Guinea-Bissau	1996	.4110
United Kingdom	1991	.0017	San Marino	1998	.4006

state failure among post-1990 country-years that experienced actual state failures. Only one of these probabilities is greater than 0.5, and so we would not have necessarily expected any one of the others to fail in the year predicted, unless the costs of misclassifying state failure were worse than the costs of misclassifying nonfailure. However, the results clearly indicate that the set of countries should have produced a number of state failures and so, for policy purposes, these would presumably have been watched carefully. That is, our methods would have made it possible to know that the probability of state failure was very high in each of these cases and likely to occur in the disturbingly high fraction of cases indicated by the given probability.

For another example, our worst (that is, highest probability) predictions among states that did not fail post-1990 are given in the bottom right of Table 4. If policy analysts had used our methods, they would have expected to see some state failures among this group. Of course, the high probabilities without observed failures do not necessarily indicate a problem with our model, since the probabilities overall are accurate (that is, 30 percent of states with 0.3 probabilities really do fail 30 percent of the time). If they were accurate for these countries, nonfailure would be perfectly consistent with the model's predictions for any *one* country at any one time, although for sets of countries the probabilities ought to be realized in actual failures as predicted.

VI. Exploring Empirical Regularities

In this section we discuss a variety of empirical regularities about state failure uncovered by our analyses. These regularities are descriptive features of the underlying causal structure. The tests given in Section IV indicate that these empirical regularities are stable and predictable features of the world and account for some of our success at forecasting. They are not necessarily equivalent to causal effects, which require additional assumptions about counterfactuals, the validity of which neither we nor the task force explores in any detail. For example, that people with more education make more money is an empirical regularity. The claim that any one person, or people on average, *would have* made more money if, ceteris paribus, they had received more education is a causal claim. Causal claims are more difficult to substantiate because they involve counterfactuals for which no direct evidence exists.[46] Although different from causal effects, empirical regularities are still very valuable components of knowledge, since any theories of state failure would need to be consistent with them. Similarly, any causal story would need to account for these verified facts about the world.

Our procedure for summarizing the empirical regularities involves using marginal effect graphs—plots of the probability of state failure by one explanatory variable (at a time), while holding constant the set of control variables at different values to see how the relationships change. In addition, each marginal effect graph itself portrays an interaction between democracy and another variable.

Figure 3 presents some of these marginal effect graphs. In each graph the predicted probability of state failure, computed from our neural network committee model, is plotted vertically. One of the explanatory variables is plotted horizontally in each graph, with three

[46] See King, Keohane, and Verba (fn. 11).

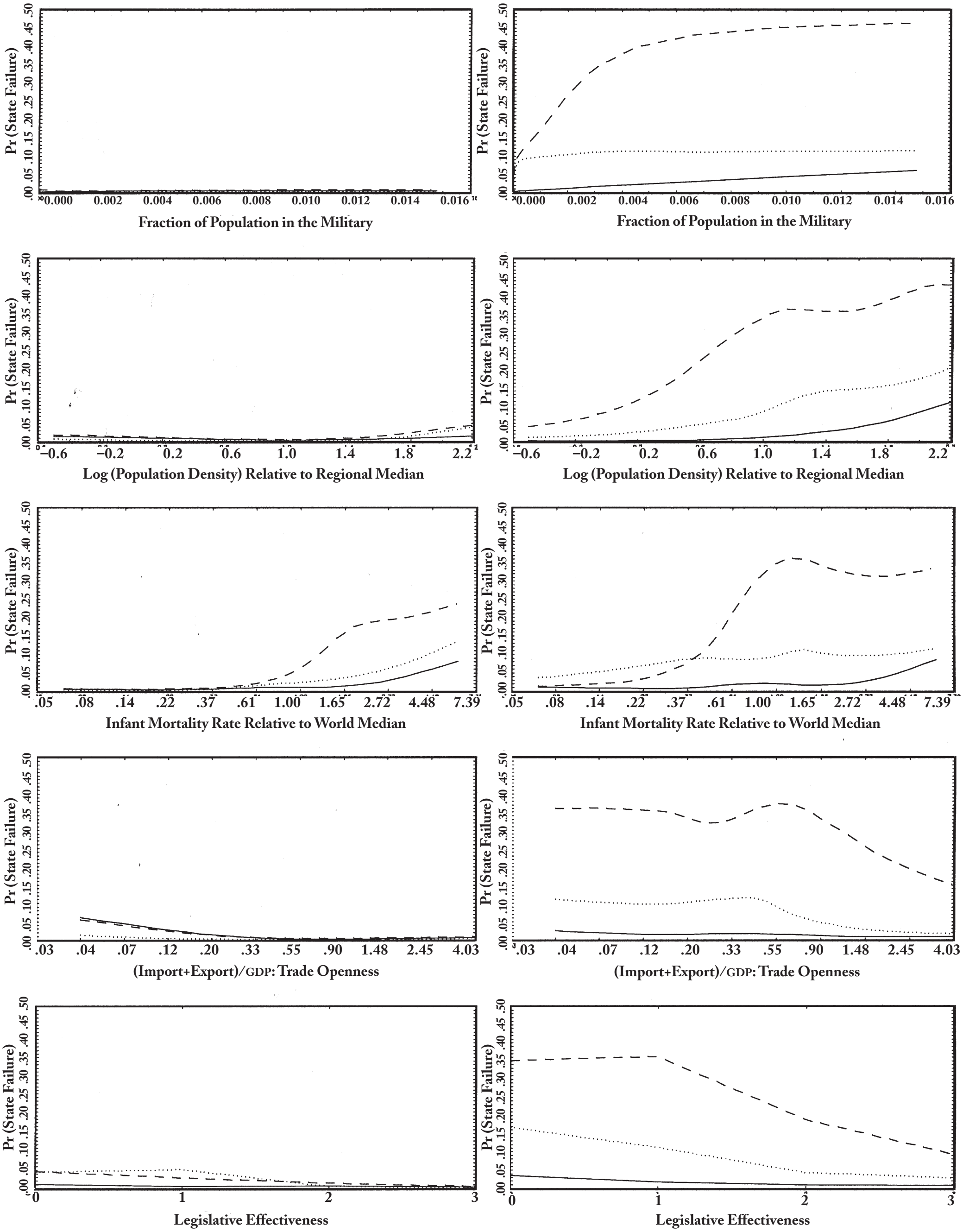

FIGURE 3

MARGINAL EFFECT PLOTS[a]

[a] Each plot gives the predicted probability of state failure vertically by one of the explanatory variables horizontally. Variables not named in each graph are held constant at the medians of the G7 countries for the left column and for state-years with failures for the right column. In each graph, the solid line is autocratic states, the dotted line is full democracies, and the dashed line includes partial democracies.

lines drawn corresponding to full democracies (dotted), partial democracies (dashed), and autocracies (solid). The first column holds control variables (those not named on each graph) constant at the median value of the G7 countries (except for the U.S., which is not included in the task force data). The second column holds constant control variables at the median for all countries with state failures. The graphs with controls held constant at the global median are similar to those in the second column, and so we do not present these. In addition, although in general with flexible models like neural networks holding constant, the control values at even slightly different values can result in very different looking marginal effect plots, such is less likely to be the case with our models since our priors impose smoothness on the functional form. (We exclude confidence intervals from these graphs for visual clarity, but all are sufficiently narrow to make the general interpretations from these lines reliable.)

The striking difference between the first and second columns of this graph shows that our model picks up the obvious differences between major industrial democracies and other countries as interactions among substantive variables, rather than as indicator variables with country labels. Of all the variables examined in the first column, only very high levels of infant mortality and, to some degree, very low levels of trade openness and ineffective legislatures increase the probability of state failure to any significant level. The infant mortality result especially is fairly striking: only states with governments that are sufficiently competent to keep infant mortality below the global median have comparatively low probabilities of state failure; other countries, even when they are alike in all other measured respects to the G7, have substantial probabilities of failure (as high as 0.25).

The second column represents time effects for countries that have values of their control variables like the median of those with state failures. The fact that these graphs are so similar to ones drawn on the basis of the global medians (not shown) indicates the stability and general applicability of the large effects witnessed there. Apparently, the main differences in the size and nature of the effects across the sample of countries is between the G7 and everywhere else. An obvious effect outside the G7 is the similar effect of democracy: for most of the ranges of all of the variables, partial democracies have a higher risk of failure than either full democracies or autocracies. Most often, autocracies have even lower risks than full democracies, although the differences here are mostly very small. We now interpret this effect along with that for legislative effectiveness in the bottom right corner of the figure. The story is similar to accounts such as Hibbs, Gurr, and Muller and Weede;[47] what our data add to these is the qualification that it does not apply to countries when holding constant other variables at the level of the G7. That is, it is not only that the G7 countries are different or that they have some unmeasured characteristics that account for this difference; it is instead that when the control variables take on values similar to those in the G7, even for non-G7 countries, this usual story about democracy is greatly reduced.

One version of this story is as follows.[48] Autocracies are countries that have repressive central governments. Since dissent is nearly impossible, little conflict of any kind, much less state failure, is observed. Such differences between the government and the people as may exist are not expressed. In contrast, in partial democracies differences between the people and the government are publicly revealed to some degree. However, because the institutions of democracy in partial democracies are not capable of adjusting government policy quickly enough, public attitudes can move far from the position of the government and stay there for a sustained period of time. In some cases, when the government cannot control this public conflict, the risk of state failure increases. We measure finer gradations of degrees of democracy in the present framework with the legislature effectiveness measure, which indicates the degree to which the legislature is channeling political dissent through electoral politics. The more this process is institutionalized, the lower the risks of state failure. Countries that are classified as full democracies typically have high legislative effectiveness as well as many of the other related attributes. Since public attitudes and government policy in these countries cannot stray far from one another for long, the risks of state failure drop to low levels, even though there are exceptions. (Clearly this is a very different story from that told by the results and arguments of the democratic peace literature, where the more democracy the better the chances of reducing conflict between nations.)[49]

[47] Douglas A. Hibbs Jr., *Mass Political Violence: A Cross-National Causal Analysis* (New York: Wiley, 1973); Ted Robert Gurr, "Persistence and Change in Political Systems, 1800–1971," *American Political Science Review* 68 (December 1974); Edward N. Muller and Erich Weede, "Cross-National Variation in Political Violence: A Rational Action Approach," *Journal of Conflict Resolution* 34 (December 1990).

[48] See, for example, Huntington (fn. 44); Ted Robert Gurr, "Why Minorities Rebel: A Global Analysis of Communal Mobilization and Conflict since 1945," *International Political Science Review* 14, no. 2 (1993); R. J. Rummell, "Democracy, Power, Genocide, and Mass Murder," *Journal of Conflict Resolution* 39 (March 1995).

[49] See also Edward Mansfield and Jack Snyder, "Democratization and War," *Foreign Affairs* 74 (1995).

Just as having a gun in a household is a risk factor for murder and suicide, the first graph in the second column of Figure 3 indicates that larger military populations are a risk factor for state failure. It is not clear whether this is because governments that fear failure increase the size of their military in anticipation or whether larger fractions of the population belonging to the military means that there are more potential dissenters with access to weapons. Either way (and in all likelihood, both factors are operating), this result may provide additional motivation for the ongoing movement of the international community toward concepts of "human security."[50] Note also that most of the conditional effect of this variable occurs when the military fraction of the population is relatively small, with the effect still increasing but at a much slower rate thereafter. Although this may suggest strategies for designing aid efforts that can achieve the largest effect for each dollar spent, further research would be required first.

The second row of the second column of Figure 3 shows that the risk of failure increases gradually with population density. This is a fairly conventional result: people need to interact before they start fighting, but it could of course be some other characteristic of densely populated nations that is operating.

The third row illustrates the effects of infant mortality in partial democracies sharply increasing the risks of state failure and then leveling off. Infant mortality is a often a good proxy for the competence of the state and it clearly picks this up here. When the infant mortality rate is equal to the world median or worse (a value of 1 or greater on the horizontal axis), the marginal probability of state failure is approximately constant. However, in those countries where infant mortality is lower (to the left of 1)—and this is often a direct result of governmental intervention—the probability of state failure drops precipitously. Infant mortality is lower in most Western democracies, but there are numerous exceptions, most of which support the role of this variable as a measure of state competence.[51]

[50] For example, Lincoln C. Chen, "Human Security: Concepts and Approaches," in T. Matsumae and Lincoln C. Chen, eds., *Common Security in Asia: New Concepts of Human Security* (Tokyo: Tokyo University Press, 1995); Gary King and Christopher Murray, "Rethinking Human Security," *Political Science Quarterly* (forthcoming), preprint available at http://gking.harvard.edu.

[51] For example, infant mortality in Cuba, where the government is so involved that the minister of health chairs a separate meeting to investigate the case of each infant who dies in the country, is lower even than the U.S. Some exceptions would include cases where natural disasters overwhelm the resources of even the most conscientious governments. Other exceptions include states where targeted interventions, such as by the World Health Organization, reduce infant mortality without as much help from the government involved (Ghana may be an example of this).

The task force's choice of a trade openness variable reflects the optimistic hypothesis that foreign policies designed to open markets might also reduce the incidence of state failure. We have found little direct support for this idea in the literature, but a number of related variables are relevant. For example, O'Kane[52] uses export specialization (the percentage of total export revenue derived from the largest single source in a country) to predict the probability of a coup. Her theory is that overspecialization leads to income instability, which in turn causes frustration, low levels of development, and reduced governmental legitimacy.[53] Export specialization is correlated with trade openness since countries that derive most of their GDP from a single export also receive most of their consumer goods and a lot of their food from abroad and so have high scores on the trade openness variable. This is consistent with our results, to which we now turn.

We find that high levels of trade openness are associated with the probability of state failure, but with virtually all of the effect occurring in nations already having extremely high degrees of openness. This nonlinear trade openness result reveals significant biases in the task force's logit results in Table 1: over most of the range there is no relationship between trade openness and state failure. What the logit model is picking up (and incorrectly attributing to the entire range) is the effect of outlier countries, the lower probability of state failure in countries with larger trading businesses than local economies (that is, when the horizontal axis is to the right of 1). These otherwise unexpectedly wealthy countries include the oil-exporting states and states like Singapore that occupy the "middleman" trading positions.[54]

We reemphasize that these empirical regularities are marginal effect plots; they are obviously not the results of randomized experiments or of an observational study where we know the explanatory variables are exogenous. They are "partial" (or "direct") causal effects that control for some of the consequences of the causal variables being studied. Our tests indicate that stable structures underlie these data, no matter how we look at them, but whether these should be relied on to draw causal inferences or change foreign aid policy must be the subject of additional study. Since foreign aid decisions are easier to identify and their moti-

[52] Rosemary O'Kane, *The Likelihood of Coups* (Averbury: Aldershot, 1987).

[53] For a contrary view, see Collier (fn. 32).

[54] It is also worth noting that this result unfortunately blurs the distinction between variable-based explanations, which we prefer, and country-based "proper noun" stories. More generally, this result also suggests, as is the case, that logistic regression results are more sensitive to outliers than are neural networks. Whereas an outlier can throw off the entire logistic curve, neural networks, by contrast, will usually map outliers separately and localize their effects to small pieces of the functional form.

vations could be (and have been) the subject of much analysis, we have a chance of ascertaining where foreign aid dollars would have their largest effects.

VII. Concluding Remarks

Although we find problems in the work of the State Failure Task Force that lead to overly large forecasts, exaggerated assessments of forecasting performance, and biased causal inferences, we wish to emphasize that it has nonetheless accomplished a great deal. The members have merged their deep knowledge of state failure to create a remarkable, carefully documented data set. The many millions of dollars invested in creating these data far exceed the resources spent on any other data set on state failure, indeed, on almost all other data sets in the discipline. The data set codifies numerous qualitative insights and knowledge from a diverse variety of area studies and other experts brought in to add their expertise to individual variables. The result is that it is now possible to test numerous theories systematically, many for the first time. Even their favored model, when appropriately corrected, performs quite well. We were able to outperform even corrected versions of their forecasting model, but only by introducing new methods to build on what the task force had already produced.

We have no doubt that further work in data, methodological innovation, theory, or area studies would produce better forecasts than even those offered here. We hope this line of research also helps the academic community and policymakers to understand, forecast, and address the critical problem of state failure.

As discussed in the introduction, the task force's definition of state failure is highly heterogeneous, and so progress would likely be made by breaking down this variable into its component parts. We conclude here with a related issue and suggestion for further research. Although all forecasts reported here are "real" in the sense of being tested out of sample, with data only available at least two years prior to the potential event, the logical status of each of the variables in the task force definitions is not entirely what was intended. In some sense, the explanatory variables (infant mortality, partial democracy, legislative [in]effectiveness, and so on) are really indirect indicators that the state has *already* failed, whereas their heterogeneous dependent variables (genocide, disruptive regime transitions, and revolutionary wars aimed at displacing the regime) are not really measures of state failure but instead are indicators of some of the disastrous *consequences* of state failure. Studies intending to forecast and explain the consequences of state failure are obviously important,[55] but so would be a study that tried to predict and explain the *onset* of state failure—the collapse of the central authority structures of the state. To do this would require a different strategy for data collection than that pursued by the task force and a more tailored, operational definition of state failure and the institutionalization, legitimacy, and authority of the state.

Appendix 1: State Failure Data

The task force data set we received includes 1,231 variables for 195 distinct countries between 1955 and 1998. Compiled from many sources, the data set provides information on political, sociological, economic, cultural, religious, educational, demographic, environmental, and public health characteristics of each country with a population greater than half a million. The data set contains information from commercial sources and from sources in the public domain, and some of the data were created or corrected by the task force; none of the data were classified.

While the task force has made older versions of its dependent variable available on the web, it has until now prevented others from accessing the rest of its extensive data collection and (for parts) has not indicated from which commercial source it can be purchased and precisely how to merge the different sources to re-create their data set. Upon publication of this article, the full task force data set to which we had access will be available at http://gking.harvard.edu. The task force continues to update and improve its data, and we encourage it to release subsequent versions so that the scholarly community may continue to benefit from its efforts and further improve forecasts of state failure.

This appendix lists some of the representative variables for interested readers. Due to space limitations, it is impossible to give detailed information on the variables (for example, measurement) here, and the original data set should be consulted for that purpose. A complete list of data sources appears in the task force report.[56]

The data set includes such variables as country code/name, year, geographic region, case-control indicators; information on leaders (name, education level/discipline, etc.); characteristics of the ruling elite; ethnic/

[55] See Robert Melson, *Revolution and Genocide: The Origins of the Armenian Genocide and the Holocaust* (Chicago: University of Chicago Press, 1992); Matt Krain, "State Sponsored Mass Murder: The Onset and Severity of Genocides and Politicides," *Journal of Conflict Resolution* 41 (June 1997); Benjamin Valentino, "Final Solutions: The Causes of Genocide and Mass Killing," *Security Studies* 9 (Spring 2000).

[56] Esty et al. (fn. 1, 1998).

linguistic/religious groups; population (sizes, composition, density); income, GDP/GNP (various measures), consumption, government expenditures, money supply, price indexes, foreign investment, imports/exports, foreign aid, exchange rates, interest rates; labor force (sizes, composition, occupational distribution); democracy measures, defense expenditures, military expenditures, assassinations, coups, general strikes, guerrilla warfare, government crises, purges, riots, revolutions, antigovernment demonstrations, seats in the legislature, legislature size, legislative effectiveness, competitiveness of nominating process, party coalitions, party legitimacy, type of regime, major constitutional changes, cabinet size, Freedom House political rights index, Freedom House civil liberties index, historical conflict information, regime durability; school enrollment, illiteracy rate, pupil-teacher ratios, public education expenditures; safe water access, natural disasters, cropland area, crop yield, fertilizer consumption, rail mileage, forest and woodland information; physicians per capita, mortality rate, life expectancy, fertility rate, information on AIDS, standard of living index, birthrate, death rate; crime statistics; newspaper circulation; modernization and postmodernization indexes, cultural zone, and a subjective well-being score.

Appendix 2: A Brief Introduction to Neural Network Models

Binary outcome variables are modeled with a Bernoulli probability distribution, which means simply that the outcome variable Y takes on the value $Y = 1$ (for state failure) with probability π and $Y = 0$ (no failure) with probability $1-\pi$. The only assumptions so far are that the definition of the two categories are mutually exclusive and exhaustive, both of which hold automatically because of the definition of state failure.

The question to be answered by statistical analysis is what the probability of state failure (π) is for possible configurations of values of the explanatory variables, x. In logistic regression, π is assumed to vary with x according to a logistic function, $\pi = 1/(1 + e^{-\alpha-x\beta})$. That is, if we plotted x horizontally and π vertically, the logistic function would look like an escalator with π approaching 0 at the bottom and 1 at the top. The direction (from top left to bottom right or top right to bottom left) and steepness of the escalator-shaped curve are determined by the values of the parameters α and β.

Choosing a logistic regression model is precisely the choice of a specific *family* of possible curves, specific *members* of which are determined by the values of α and β. Logistic regression analysis then involves using data to find the single member of the family (indexed by the values of α and β) that best fits the data. The procedure for choosing a member of a family, known as maximum likelihood estimation,[57] is well known and optimal in many relevant respects. The parameters α and β are of no direct interest, but they can be used to compute quantities of interest such as the probability of state failure.

The neural network model we use is directly analogous to logistic regression: it too specifies a (parametric) family of curves and then uses data to choose the particular member of the family by setting values of the parameters that best fit the data. However, instead of a curve that looks like one of many possible escalators, the possible shapes are now far more numerous and diverse (escalators, roller coasters, linear, quadratic, cubic, and many others). Figure 3 gives some examples of the possible shapes these functions can take. The neural network family of curves also includes members with considerably more intricate interaction effects (that is, when the effect of one explanatory variable depends on the values of others), as evidenced by the differences in the two columns of Figure 3.

To understand the mathematical form, note that the logistic form can be simplified and written as $\pi = \text{logit}(\alpha + x\beta)$ or even $\pi = \text{logit}(\text{linear}(x))$. The specific mathematical form of the neural network we use is then

$$\pi_i = \text{logit}\,[\gamma_0 + \gamma_1 \text{logit}(X_i\theta_{(1)}) + \gamma_2\, \text{logit}(X_i\theta_{(2)}) + \ldots + \gamma_M \text{logit}(X_i\theta_{(M)})], \quad (1)$$

or more simply $\pi_i = \text{logit}(\text{linear}(\text{logit}(\text{linear}(X_i))))$, where M (which is known as the number of "hidden neurons," but of course nothing is hidden and M has no necessary relationship to neurons in the brain), the γs and the θs are adjustable parameters that index particular members of the family. They can be estimated from the data via the same maximum likelihood procedure as for logit. Like logit, the parameters estimated are not of direct interest, but they can be used to compute forecast probabilities and other quantities of interest.

The advantage of neural networks is that they assume less as part of the choice of family and leave the data to guide us more. Indeed, the model in equation 1 can approximate arbitrarily closely any relationship between π and x, given a large enough value for M. In practice, neural networks have been found to approximate a larger range of functional forms with fewer parameters than other relevant possibilities.

The potential disadvantage of neural networks is that they are so

[57] Gary King, *Unifying Political Methodology* (Ann Arbor: University of Michigan Press, 1989).

flexible that they can fit idiosyncrasies in the data rather than structural features that persist, although we guard against this with the methods offered in Section IV. In only rare situations will a neural network model be outperformed by a logistic model. Indeed, choosing a logistic model over a neural network model would be appropriate mainly if there existed strong substantive theory that ruled out all members of the neural network family of curves other than the logistic. In fact, however, we know of no such substantive theory of state failure that even addresses these issues and so there exists little justification for logit over neural networks. See King, Beck, and Zeng[58] for more details along the same lines and Bishop[59] for a complete presentation of neural networks.

[58] Beck, King, and Zeng (fn. 35).

[59] Bishop (fn. 38).