Why Do Bureaucrats Delay? Lessons from a Stochastic Optimal Stopping Model of Agency Timing, with Applications to the FDA

Daniel P. Carpenter

Why do bureaucratic agencies often exhibit a reluctance to make their decisions quickly?

Why do some agency decisions take longer than others, even when they are made under identical statutes, agency structure, and administrative procedures?

Despite two decades of illuminating research in bureaucratic politics, one of the fundamental powers of the government—*the power to wait,* to delay its response to the requests of citizens—has not been systematically studied. The popular adage that agencies are "slow," whether right or wrong, does not answer either of these questions. It simply restates the first query and cannot answer the crucial issue of variation in the second. The aim of this essay is to develop a modeling framework that will offer some generalizable answers to these questions. I will model the regulatory approval decision as the optimal stopping of a stochastic process, derive a set of predictions about the duration of regulatory review, and conduct tests of the model in duration analyses of drug review times by the U.S. Food and Drug Administration (FDA).

The duration of government decisions, in addition to their content, is one of the central issues of modern democracy. Government agencies display a high degree of variability in the time they take to grant permits, issue licenses, adopt new policies, take criminal and civil cases to trial, and approve new products.

Of these examples, there is no more controversial and seminal case than the approval of new pharmaceutical products by the FDA. The FDA exhibits wide variation in review times for new drug applications (NDAs). From 1993 to 1995, for instance, the FDA approved sixty-seven new chemical entities (NCEs), but the drug Flumadine (marketed by Forest Labs) spent eighty-two months in the NDA review stage, whereas Orlaam (BioDevelopment Corp.) was approved in less than a month (Kaitin and Manocchia 1997, 47). The speed of approval is important for patients and firms alike. As scholars and politicians have agreed, these are nontrivial differences. Drugs that are approved more quickly will be available to patients who need them. And firms that are able to get their drugs approved more quickly will gain important advantages over competitors in product recognition and physician prescribing practices (Olson 1997, 378).

All new drug applications are considered under *the same statute*—the Food, Drug, and Cosmetic Act of 1938, as amended—and *the same administrative procedures.* They are also approved by *the same agency.* The question is obvious, but it cannot be answered well by reference to procedures, structure, or statute: Why are some NDAs approved more quickly than others?

Political science research has until recently eschewed questions of the duration of agency operations. Beginning with the work of Simon (1947) and Bernstein (1955), scholars in institutional political science have studied extensively the behavior of bureaucratic agencies. In the main, this research has taken two forms. The first is the study of political control. During the past two decades, following largely the work of Arnold (1979, 1987), Moe (1982, 1985), Weingast (1983 [with Moran], 1984), and Wood (1988, 1991 [with Waterman]), a number of quantitative studies have concluded that political control of bureaucratic agencies by elected authorities (or "principals") does prevail (see also McFadden 1976; Magat, Krupnick, and Harrington 1986; Rothenberg 1994; Wood and Anderson 1993; Ringquist 1995; and Carpenter 1996; but see Krause 1996a). Moreover, scholars have identified a number of mechanisms by means of which elected authorities can influence agency behavior, including congressional committee oversight (Weingast and Moran 1983), presidential and multi-institutional control (Moe 1995; Hammond and Knott 1996), statute (Moe 1990; Huber and Shipan 2002), "fire alarm" oversight

(McCubbins and Schwartz 1984), interest group monitoring (Spiller 1990; Banks and Weingast 1992), direct citizen contact (Scholz, Twombly, and Headrick, 1991; Brehm and Gates 1997), administrative procedures (McCubbins, Noll, and Weingast 1987; Bawn 1995), appointments (Wood 1988; Ringquist 1995), and budgetary control (Yandle 1988; Carpenter 1996).

A closely related literature, grounded in organization theory, has attempted to analyze the internal behavior of agencies. These studies include analyses of agency adaptation (Bendor and Moe 1985), bureaucratic learning (Bendor 1995; Brehm and Gates 1997), conflict and decision making in hierarchies (Williamson 1975; Miller 1992), budgetary decision making under bounded rationality and hierarchy (Padgett 1980, 1981), and "parallel versus serial" information processing (Bendor 1985; Heimann 1993, 1997).

None of these studies has placed the duration of agency decisions at the center of analysis. It is as if political scientists have assumed that the timing of decisions has nothing to do with "politics" and that the proverbial slowness of government agencies is, for all intents and purposes, neutral or trivial. The aim of this essay is to redress these shortcomings of the bureaucratic politics literature by focusing directly upon the duration of agency decision processes.

The Benefits and Costs of Waiting

The theoretical core of this chapter may be expressed in two arguments. First, waiting is a way of gaining more information about an uncertain decision that the agency regards as irreversible (or reversible only at cost). In the case of drug approval, the FDA sees its approval decisions as reputationally irreversible even though drugs can be recalled; once a drug has done sufficient harm that it must be recalled, the damage done to the agency's reputation cannot be regained. (Indeed, recall as an agency admission of error may worsen the situation.)

Yet there are costs to waiting as well. Each potential drug has a *political demand* associated with it, a demand that rises in rough proportion to the number of potential citizens who need it (who have the diseases for which it is indicated), the degree of political organization of those citizens, the political power (or rents) of the pharmaceutical firm submitting the drug, and the degree of media and political attention given to

the drug and the disease that it treats. Drugs may also have opponents to their approval, as in the case of the abortion-inducing "morning-after pill," RU–486. The degree to which the supporters of approval are unified, relative to any opposition, may be termed the *cohesion* of the political demand for the drug and may be considered an important part of the drug's political demand. The FDA's experience with AIDS drug approvals (see Epstein 1996) has forcefully demonstrated the political power of organized disease advocates.

The agency's response to these demands is a stochastic dynamic optimum that embodies the trade-off between the value of additional time and the political cost of delay. The model has several interesting and counterintuitive predictions and carries a specific distributional prediction about the form of the hazard to be used in duration analysis.

The essay thus has the advantage, I hope, of linking empirical estimation tightly to an underlying theory of choice, a theory that is "dynamic" in the sense of Bellman optimality.[1] The optimal stopping models considered here should serve as useful tools with which to analyze a range of bureaucratic behaviors as well as to structure duration analyses of bureaucratic and regulatory decision making. In previous uses of duration models in bureaucratic research, these models have not flowed explicitly from an underlying formal model (e.g., Whitford 1998; Balla and Knight 2002; though for a more intuitively based attempt to premise duration models upon an underlying theory, see Gordon 1999).

A central reason to analyze the duration of agency decisions is to redress an unfortunate theoretical and empirical imbalance in the bureaucratic politics literature. Most analyses in bureaucratic politics are either time-series analyses of agency enforcement aggregates (Moe 1982, 1985; Wood 1988; Ringquist 1995; Carpenter 1996; Krause 1996a; Olson 1995, 1997) or discrete choice analyses of agency behavior (McFadden 1976; Arnold 1979; Weingast and Moran 1983). While extensive insight has been generated by these studies, they all but ignore the critical question of the *timing* and *rapidity* of bureaucratic decisions. The speed with which agencies make decisions is an instrument of power over private sector actors and a source of immense agency discretion. Most new drug applications that pass the clinical trial stage are approved by the FDA. The critical question in FDA drug approvals is not *if* but *when.* This is the core issue in FDA reform debates.

Two Caveats: Other Explanations for Delay and the Role of Political Intuition

Before proceeding with the theoretical and empirical analyses, it merits remarking that other mechanisms can generate delay in bureaucratic decision making. An agency may have too many tasks assigned to it by Congress. Delay may be equivalent to procrastination. Delay may be an attempt to pass the buck or dump an unpleasant task upon a successor. Or a certain procedure may simply take more time than another.

While some level of delay in organizational decision making occurs in all contexts, thinking of delay as overload, procrastination, buck passing, or procedural complexity has its limits. For one thing, it would be difficult to model any of these four phenomena in a way that would generate predicted variation in the delay time. If we observe the FDA taking more time with one drug than another, it surely cannot be due to buck passing or procrastination (the FDA will eventually have to decide upon the drug). If overload and the procedural complexity of the drug decision explain drug approval delay, then how can drugs considered by the same organization at the same time and under the same procedures receive different review times? Undoubtedly, all four of these factors are at work, but none of them can be the decisive component of variation.

A second caveat worth remarking here is that some of the hypotheses predicted by the model are relatively intuitive and are not derived entirely from the optimal stopping model itself. These include the hypothesis that a drug's approval time is decreasing with the political clout of the firm submitting it; the hypothesis that approval time is decreasing with the organization of sufferers of the disease for which the drug is intended; and the hypothesis that expected FDA approval time is increasing in the disease-specific order in which the drug enters the market. Still, the stochastic optimal stopping model is useful here because it shows that commonsense observations about FDA drug approval are consistent with dynamic optimality and bureaucratic learning.

The chapter is organized as follows. In the section that follows, I outline the formal model that I will use to model an approval decision. In the next section, I will offer hypotheses and suggest measurements, and I will briefly test one of the model's predictions.

An Optimal Stopping Model of the Approval Decision

The approval of a product or license presents the regulator with a learning problem. The agency must study a new product application (with accompanying "data") and decide when the apparent benefits of the application—whether a product, a proposal, or a permit for new economic activity such as grazing or construction—outweigh the costs or risks associated with its use.

Learning in Continuous Time

The model here is tailored toward the bureaucratic review of a new product application, specifically a new drug application. The framework is, however, generalizable to learning about the guilt or innocence of a suspect (Whitford and Helland forthcoming; Gordon 1999) or the eventual environmental damage of a construction or grazing permit or energy-generating license. For the present, the model imagines product review as a process of page turning through a product application. One by one, the agency observes a series of experiments (in the case of FDA drug approval, clinical trials) in which a drug either harms or does not harm the person taking it. The sequence of binary outcomes—"harm" or "not harm"—becomes the data for the agency's decision. I model the evolution of these observations as a continuous-time Wiener process, more commonly known as "Brownian motion."[2]

Let all drugs be indexed by i, diseases by j, and firms by k. All drugs in the model are characterized by two parameters.[3] First, let $\gamma_{ij} (0 \leq \gamma_{ij} \leq 1)$ be the *curing probability* of the drug (which can be interpreted as the fraction of people with disease j that drug i will cure). I assume that γ_{ij} is fixed and known with certainty throughout the agency's decision process.

Second, let μ_i be the *danger* of the drug, which can be thought of as the expected number of people who will be harmed or killed by the drug over a given interval of time. Normalizing the interval to unity, μ_i may be thought of as the *rate* of harming consumers. The greater the danger of the drug the more its approval will harm the agency's reputation for protecting public safety. I assume throughout that a drug's danger is independent of its curing power, which implies $\text{cov}(\mu_i, \gamma_i) = 0$.

Observed harm in regulatory review evolves according to a Wiener process $X_{it} = X(t)$, a linear function of underlying danger (μ) plus a random component, as follows:

$$X(t) = \mu t + \sigma z(t), \tag{1}$$

where μ and σ are constants and $\sigma > 0$ and where $z(t)$ is a standard normal variable with mean zero and variance t. A more "dangerous" drug—one with higher μ—will yield more harm, but harm will also be affected by the random term $z(t)$. A higher σ will yield a more volatile review—namely, a series of clinical results from which the agency will find it harder to learn what μ is.[4]

Observed harm is therefore a Markov process with independent increments, and the agency learns about μ in a simple Bayesian fashion based upon the stochastic history of $X(t)$. Letting $X(t)$ start arbitrarily at 0, then it is normally distributed with mean μt and variance $\sigma^2 t$. We assume that σ is the same across drugs but that μ differs across them, according to a normal distribution with mean m and variance s.[5] For any drug review of length t and accumulated harm $X(t) = x$, all relevant information for the agency's decision is captured in a pair of "sufficient statistics" (x and t). Bayesian estimates of μ are

$$\text{Posterior Mean} \equiv E_{xt}(\mu) = \hat{\mu} = \frac{m/s + x/\sigma^2}{1/s + t/\sigma^2}, \tag{2a}$$

$$\text{Posterior Variance} \equiv S(t) = \frac{1}{1/s + t/\sigma^2}, \tag{2b}$$

where m is the mean danger for the average drug in the population—or the expected danger of any drug, before the agency knows anything about it—and s is the "prior" (or preexperimental) variance of this danger. Notice that

$$\lim_{t \to \infty} \hat{\mu} = \frac{x}{t} = \mu \quad \text{and} \quad \lim_{t \to \infty} S(t) = 0. \tag{3}$$

The posterior variance $S(t)$ is a crucial parameter. Simply put, $S(t)$ is the best estimate of the agency's uncertainty about the true value of μ. For this reason, at any time *the value of waiting for another moment is an increasing function of* $S(t)$.

Reputation and the Value of Waiting to Approve

The fundamental assumption of the model is that the regulator guards its reputation for protecting citizen welfare (here "safety"). Students of the 1962 Kefauver Amendments argue that the FDA's decision not to

approve the drug thalidomide played a crucial role in the expanded discretion that the agency received under the new laws (Quirk 1980). This reputation is an important political asset—it can be used to generate public support, to achieve delegated authority and discretion from politicians, to protect the agency from political attack, and to recruit and retain valued employees (Carpenter 2000b, 2001).

Yet, although waiting to approve always has *some* value, there is not always a *net benefit* to waiting. In other words, in many cases it is *not* optimal to wait forever. As with most inference problems in politics, the marginal return to more information is decreasing. Voters may not need ten debates or twenty polls to help them sort out the differences between two candidates; often two or three of each will do. A similar logic prevails in regulatory product approval: *the value of waiting declines over time, as more is learned about the drug.*

The Approval Payoff as the Cost of Information

There is another reason why waiting is not always optimal. Put simply, waiting is costly. Patients want a drug for their diseases, and firms that profit from drug sales want entry into potentially lucrative markets. To delay the approval of a drug is to impose costs upon these interests, and when these interests are organized or well publicized they can make it costly for the agency to delay. The case of AIDS offers a lucid example. When ACT-UP (AIDS Coalition to Unleash Power) protestors, dismayed that the FDA might delay approval of lifesaving therapies, demonstrated at agency headquarters in 1988, they embarrassed the agency and prompted a sharp change in policy on AIDS drugs (Epstein 1996).

In short, the information obtained while waiting, though valuable, is not free. To approve a drug is to please the patients who need it and the firms that will profit from its sales. I assume that the approval payoff A for any drug is strictly positive and is a combinatorial and strictly decreasing function of the number of drugs that have already been approved for disease j, as follows:

$$A = \theta(\psi_j, L_j, N_j, \omega_k) = (1 + \lambda_k)\psi_j \gamma_{ij} L_j \prod_{i=0}^{N_{i-1,j}} (1 - \gamma_{ij}), \tag{4}$$

where L_j is disease j's *prevalence,* or the number of persons with disease j, and ψ_j is the *political multiplier* of disease j, a positive parameter.[6] We can

interpret ψ_j as the expected number of citizens, for every citizen afflicted by disease j, who will apply pressure upon the agency or the politicians governing it. The term N_j is the number of marketed drugs that already treat disease j; and λ_k is the political clout of the submitting firm, also strictly positive.

Since the approval payoff is a multiplicative function of its components, it is possible to collapse the different terms of the approval payoff in equation (4) into a single-index parameter. In this case, the function θ can represent the entire payoff.

The Agency's Optimal Policy

The problem facing the agency can be described as the optimal stopping of the process $\hat{\mu}_t$ (see eq. 2), consistent with the following objective:

$$\max Ee^{-\delta(t_{app})}[A - E_{\hat{\mu},t}\int_t^\infty e^{-\delta(y-t)}\alpha\mu^*(y,\omega)dy]$$

$$= Ee^{-\delta(t_{app})}[A - \delta^{-1}\alpha\mu^*(t_{app},\omega)], \tag{5}$$

where δ is the discount factor, t_{app} is a given approval time, μ^* is the agency's estimate of danger at the optimal stopping time (as given in eq. 2), ω denotes an elementary event in the probability space Ω, and y is a variable of integration (a variable that has the same dimensions as time and functions purely as an index in the optimization problem). The parameter α represents the agency's aversion to danger, or its "preferences" with respect to the relative weight it puts upon danger versus the clinical benefits of drugs.[7] In all respects, the agency can be modeled as stopping the process $\alpha\hat{\mu}_t$.

The political cost of approval, then, is simply the reputational loss that accrues to the agency from the observable danger of the drug. Upon the approval of the drug, the agency therefore "pays" the parameter $\alpha\mu$, which can be learned only through preapproval review of the product.

The problem facing the agency, then, is the problem of "stopping" the review only when the payoff of approval exceeds both the utility losses associated with the danger of the drug *and* the value of waiting for more information. In other words, the rational agency does *not* simply approve the drug when its apparent danger is less than the payoff of approval. This is the aspect of the agency's decision that differentiates product approval from standard problems of administrative choice.

It is shown elsewhere (DeGroot 1970) that for sequential decision

problems upon Markov processes of this sort the optimal policy entails the division of the space of possible outcomes $[\hat{\mu}_t, t]$ into two regions—a *continuation region,* where observed values of $\hat{\mu}_t$ suggest waiting, and a *stopping region,* where values of $\hat{\mu}_t$ suggest termination of the review process and approval of the drug. Figure 1 offers a sample division of the space, where $\hat{\mu}_t$ has been replaced by W_t. The policy can be described by a unique partitioning of $[\hat{\mu}_t, t]$ by a "barrier" $\eta(t)$ such that the first departure from the continuation region involves the process $\hat{\mu}_t$ "hitting" the barrier $\eta(t)$ for the first time.

For purposes of the present discussion, I assume the existence of an optimum for the agency's problem. Elsewhere (Carpenter 2000c) I prove the existence of an optimum for this problem, and existence theorems for related problems appear in Jovanovic 1979 and, most important, Miroschnichenko 1963.

The solution to (5) is complex and must satisfy three conditions. First, the agency's policy must be *dynamically optimal.* It must be the case that whatever the agency's initial decision its subsequent choices constitute an optimal policy with respect the subproblem starting at the state that results from the initial actions (Dixit and Pindyck 1994, 100). Second, the optimal barrier $\eta^*(t)$ must satisfy a *value-matching condition* such that the agency is indifferent between taking the approval payoff A and losing the capitalized value of the drug's danger. The value-matching condition says simply that the barrier represents indifference between the matched value of two options. Finally, the optimal stopping of Wiener processes must satisfy a third condition known as the *smooth-pasting* condition (Dixit 1993). The smooth-pasting condition is difficult to explain simply, but it requires that along the approval barrier the time derivative of the approval payoff and the time derivative of the danger losses are equivalent. At the point of indifference between waiting and approving, the time path of the approval payoff cannot differ from the time-path of the danger loss or the agency would be better off either to have approved the drug earlier or to wait.

PROPOSITION 1. *The optimal barrier is*

$$\eta(t) = \delta A - \frac{S(t)^2}{2\sigma^2} F_{\hat{\mu}\hat{\mu}}[\eta(t),t]. \tag{6}$$

PROOF. *Carpenter 2002.*

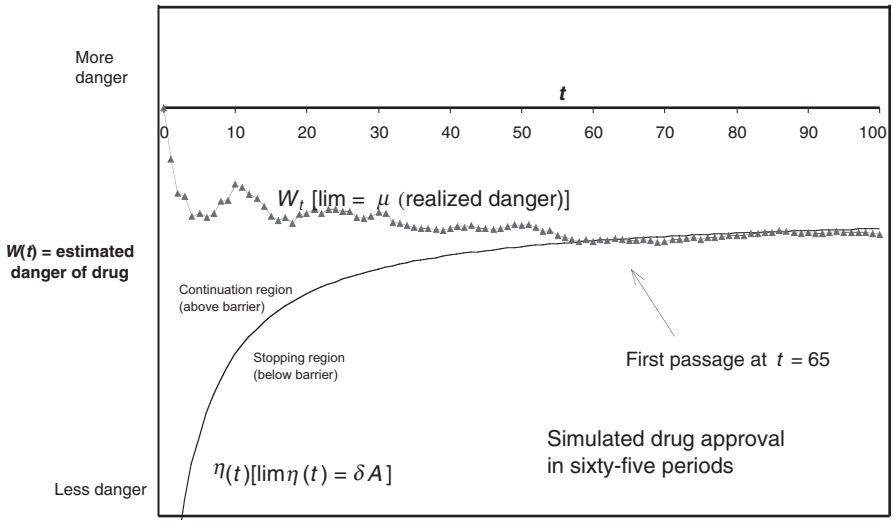

Fig. 1. Simulated first passage of the danger process $W(\mu,t)$ through the agency's approval barrier $\eta(t)$

Here the term $F_{\hat{\mu}\hat{\mu}}$ represents the second derivative of the function F with respect to the estimate $\hat{\mu}$. By equation (3), the limit of $\eta(t)$ is δA, which $\hat{\mu}_t$ approaches from above. In the asymptote, the agency can learn nothing more about the drug because uncertainty about its danger (the posterior variance of μ) is at zero. The benefits of approval (which have remained constant throughout the decision but are fully capitalized in the limit) are equal to δA. Meanwhile, the costs of approval are $(\alpha)\mu$. In the limit, the estimated danger converges to the true danger, so that the drug is approved asymptotically if $\alpha\mu < \delta A$, or $\mu < \delta A$ if we set α to one as we have done above. In finite time, of course, the agency cannot make the decision in such terms. *There is always a value to waiting for more information, regardless of the agency's risk aversion.*

Risk-Neutral Delay and Other Characteristics of the Approval Distribution

The optimal rule for waiting has now been derived. Let $G^*(t)$ be the approval distribution, or the probability of approval under the optimal policy. Then an approximation to $G^*(t)$ can be retrieved from the class of first-passage-time distributions:

$$\Pr[\alpha\hat{\mu}_t < \eta^*(t)] = 2\left\{1 - \Phi\left[\frac{(m - \delta A)}{\sqrt{r(t)}}\right]\right\} = 2\Phi\left[\frac{(\delta A - m)}{\sqrt{r(t)}}\right], \qquad (7)$$

where $\Phi(\cdot)$ is the cumulative standard normal integral and $r(t)$ is the posterior *precision* of the Wiener process, or $r(t) = s - S(t)$. $G^*(t)$ and its density $g^*(t)$ then comprise an *inverse Gaussian* or *Wald variate* (for a derivation see Harrison 1985). Notice that a greater approval payoff increases the approval probability.

The following propositions note interesting properties of the approval distribution.

PROPOSITION 2: DE FACTO REJECTION WITHOUT A REJECTION OPTION. *In expectation, a nonzero fraction of drugs will never receive the agency's approval, as* $\lim_{t\to\infty} G^*(t) < 1$.

SKETCH OF PROOF. *The proof comes from examination of* $G^*(t)$ *in (7). Because* $\lim_{t\to\infty} r(t) = s < \infty$ *(eq. 3), the term in brackets in* $G^*(t)$ *never reaches unity.*

Proposition 2 implies that the probability of an endless review is strictly positive. Statistically, proposition 2 is sufficient to identify $G^*(t)$ as a defective distribution.

In two ways, proposition 1 coheres with the FDA's actual behavior. It is trivially consistent, of course, with the fact that not all drug submissions are approved. Yet it is also consistent with the fact that *the FDA never formally rejects a drug.* It simply deems a drug "not approvable," and nothing prevents the producing company from submitting more data about the drug if it seeks approval at a later date (GAO 1995).

A central result of the model is that drug approval should be least likely precisely when some firms and patients most desire it: when a drug has completed early clinical testing and is submitted for review.

PROPOSITION 3: THE SCARCITY OF QUICK APPROVAL. *For any drug* $i\ [t_{app}(i) < \infty]$ *such that an optimal approval time exists, the approval hazard* $\theta(t) = g^*(t)/[1 - G^*(t)]$ *has the following two properties:*

(1) $\theta(t = 0) = 0$,
(2) $\forall t, t < (A - \delta\mu)^2$, $\lim_{t\to 0} \theta(t) = 0$.

PROOF. *Result (1) obtains by the property of* $g^*(t)$ *such that* $g^*(0) = 0$. *Result (2) follows from the unique mode of* $g^*(t)$ *at* $(A - \delta\mu)^2$, *below which* $g^*(t)$ *tends to zero faster than* $G^*(t)$.

PROPOSITION 4: NONMONOTONICITY OF THE APPROVAL HAZARD. *The hazard* $\theta(t) = g(t)/[1 - G(t)]$ *is nonmonotonic.*

SKETCH OF PROOF. *By proposition 3, result (1),* $\theta(t = 0) = 0$. *By proposition 2,* $\theta(t)$ *must return to zero in the asymptote.*

Propositions 2 through 4 have important theoretical and empirical implications. First, we now have a generalizable explanation for agency delay, one derived under the assumption of risk neutrality and without assuming anything about the "slowness" of the agency. The reason for bureaucratic delay, the model suggests, is not sloth or slack but uncertainty reduction: *the information to be gained by waiting is most valuable at the beginning of a decision problem.* Even when A is large, the hazard of approval will be near zero at the beginning of bureaucratic review.

Second, the model shows that waiting can be a de facto form of rejection. Even where no formal rejection option exists, long-term waiting increasingly implies a zero probability of eventual approval under the optimum policy.

Finally, the model offers some important lessons for hazard estimation. The likelihood employed should be inverse Gaussian, and, more significantly, the hazard function should be nonmonotonic. This rules out the simple exponential distribution as a hazard specification, but it *also rules out the most commonly used parametric distribution of political science duration analysis: the Weibull distribution.*

Hypotheses

With the characterization of a dynamic stochastic optimum in proposition 1 and the theoretical results in propositions 2 through 4, the following hypotheses can be posed.

H1: Below a unique mode, the approval hazard $\theta(t)$ tends to zero.
H2: $\theta(t)$ is an increasing function of the prevalence L_j of the disease j treated by the drug i.
H3: $\theta(t)$ is an increasing function of the political influence and publicity of the patients with the disease treated by the drug, or ψ_j.
H4: $\theta(t)$ is an increasing function of the political clout ω_k of the firm k.
H5: If the agency's preferences over danger versus availability are affected by elected authorities (as captured in α), then $\theta(t)$ is decreasing in α.

The hypotheses advanced here can be tested using a maximum likelihood duration model. Specifically, the model predicts that the duration of product approval for an agency faced with the problem in (5) will approve products according to an *inverse-Gaussian distribution.* The present framework therefore offers a more theoretically driven approach to the estimation of duration models in political science than is customarily the case (Carpenter 2002). From a strictly scientific point of view, an important advantage of stochastic process models is that they allow the functional form of statistical tests to be derived directly from the underlying model of behavior (Padgett 1980; Carpenter 1996, 2002).

A Nonparametric Test of H1

The first hypothesis can be tested using nonparametric statistics calculated from duration data that are unaccompanied by covariates. Before doing so, it will be useful to describe the FDA approval process and the approval time data that result from it.

Drug review by the FDA occurs only after a drug completes phase I clinical trials (testing for information on drug safety), phase II trials (testing for effectiveness and safety), and phase III trials (verification of safety and effectiveness and testing for adverse effects). On average, these trials consume six years of drug development. This is more than half of the time required to bring a new drug to market (Dranove and Meltzer 1994). After completion of phase III clinical trials,[8] the sponsoring company submits a new drug application to the FDA. The FDA's Center for Drug Evaluation and Research (CDER) then reviews the NDA by assigning it to a review team composed of doctors, pharmacologists, chemists, statisticians, microbiologists, and regulatory affairs experts. The most novel of these drugs, and the drugs that occasion the greatest controversy over approval and delay, are called new chemical entities. The dependent variable in the following analyses is the length of the NDA review stage (in months) for an NCE.

I have consciously restricted the sample in two ways. First, I analyze only NCEs, while the FDA screens hundreds of other drugs annually. These include both "generic drugs" (simple copies of drugs whose patent protection has expired) and "supplemental" applications, which occur when a company seeks to market its drug for a disease other than the one for which it was originally submitted. Because the review process is very

different for these applications, I leave them to another study. Second, the data (and the model) ignore the clinical trial stage of drug development. I have chosen not to analyze it here because the length of the clinical trial stage depends heavily upon the speed of the company and its clinical testing regime.[9]

Figure 2 displays a plot of the estimated hazard function along with a "confidence function" plot (the dotted line), which represents the value of two standard errors of the hazard estimate from zero. The hazard function is estimated using the nonparametric Kaplan-Meier method. In other words, when the estimated hazard function crosses the dotted line in figure 2, we can say as statisticians that the expected hazard differs from zero at a statistical significance level of $p < 0.05$.

Figure 2 shows that the expected hazard cannot be judged to be statistically distinguishable from zero until ln (review time) equals 1.7, or somewhere between seven and nine months into the review process. This is an interesting result because the 1962 Kefauver Amendments to the Food, Drug, and Cosmetic Act state that the FDA shall approve all drugs within six months of their submission. In other words, figure 2 suggests that for the entire period of legal review, plus one to three months, the expected instantaneous likelihood of approval for a drug is zero.

Yet the model also offers some logic as to why this is the case. The pattern in figure 2 is the result not of bureaucratic sloth nor of mismanagement of the review process. The most compelling theoretical explanation is that the value of waiting for information is greatest at the outset of a product approval process. Because the value of waiting is so high, there is a rational scarcity of quick approval. It is important to note, again, that this result is derived under risk neutrality—the agency neither prefers nor avoids gambles between high- and low-variance drugs—though the agency is always danger averse.

It is also worth noting that figure 2 essentially supports proposition 4 of the model, which predicted the nonmonotonicity of the hazard. Since almost one-half of submitted drugs are not approved (GAO 1995; this is an empirical fact that does not require a statistical test), the empirical hazard function must eventually fall to zero. All that is required for a demonstration of nonmonotonicity, then, is evidence that the hazard starts at zero in addition to ending at zero, which figure 2 shows. (Note that inclusion of nonapproved drugs in the sample would not change this result since the hazard can rise only if a drug is approved.)

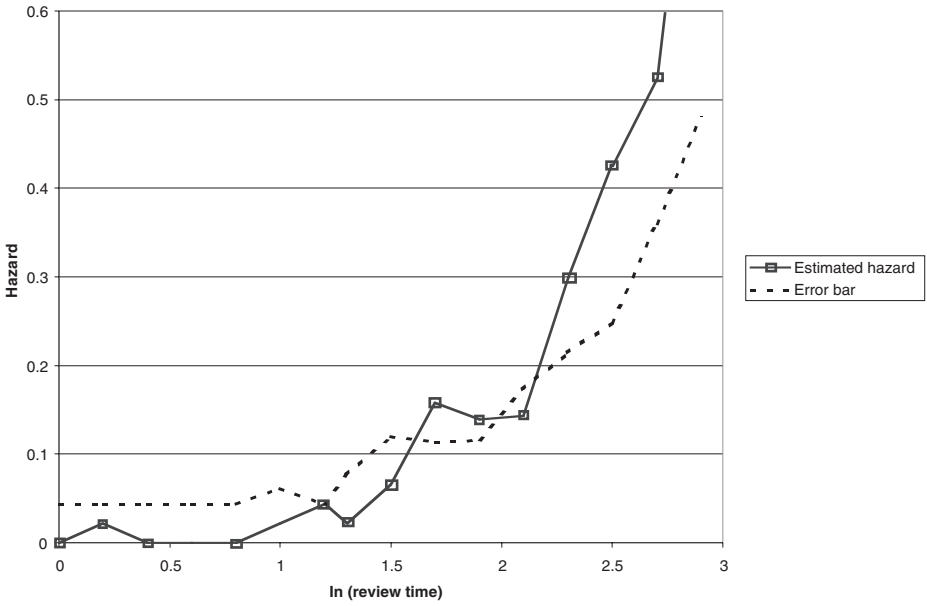

Fig. 2. Nonparametric estimates of the approval hazard, 1985–95 (first eighteen months of NDA review)

Conclusion

While instructive, the theoretical and empirical analyses presented here could doubtless be improved and extended in fascinating ways. Theoretically, it would be interesting to incorporate a budget into the model and allow the agency to bestow resources explicitly upon drugs. A more interesting issue, to my mind, concerns the organizational factors that affect information processing. The FDA has long been hampered by high turnover, particularly when talented reviewers leave for the lucrative pharmaceutical sector. What happens if the agency loses information over time? What happens if agency decision makers are forgetful or, as is more likely the case, that organizational turnover results in a loss of "institutional memory"? Work is currently under way in this direction (Mullainathan 1998; Carpenter 2000d).

Empirically, an obvious shortcoming of the hazard rate analysis is the absence of drug applications that are rejected or withdrawn. As I have noted, the proprietary character of rejected drug applications renders this

problem far more difficult than one of simply collecting more data. Efforts are currently under way to collect a partial sample of rejected or withdrawn applications, though I note that only a partial sample will be possible and that theoretically substantive and interesting results can be drawn from analyzing approvals themselves. Review times conditioned upon approval, again, lie at the center of the political controversy over FDA regulation.

As a concluding note, I call attention to the fact that optimal stopping offers a different way of thinking about the dynamic behavior of bureaucracies. Most research on bureaucratic dynamics has been associated with the time-series tradition, which includes models of autoregressive and adaptive behavior (Moe 1982, 1985; Wood 1988; Wood and Waterman 1991; Wood and Anderson 1993; Carpenter 1996; Krause 1996b). These models have demonstrated the dynamics of bureaucratic behavior for aggregated decisions made over time. The contribution of the present study, I hope, is that it instructs students of public bureaucracy about the dynamics of *single decisions* in which time itself has value and is an object of choice.

Notes

An earlier version of this essay was presented at National Public Management meetings, December 3, 1999, Texas A&M University. For helpful comments on earlier and different versions, I thank Greg Adams, Frank Baumgartner, John Brehm, Michael Chernew, Scott Gates, Rick Hall, Thomas Hammond, William Keech, George Krause, Ken Meier, Mary Olson, Ken Shotts, Michael Ting, and Andrew Whitford.

1. Under Bellman optimality (or dynamic optimality), it must be the case that the choices at a given period t (or for a given infinitesimal dt) must maximize the agent's utility function for this increment of time and the (assumed) optimal path of all future decisions. Put in more crude and intuitive terms, dynamically optimal actors ignore sunk costs and care only about an optimal (step-by-step) sequence of choices in the future.

2. As Dixit (1993) shows, the Wiener process representation can also be derived by taking limits toward zero on the discrete walk of the binary outcomes harm and not harm. Thus, a discrete time process becomes, in the limit,

a continuous time process; accordingly, all of the results described in this essay may be replicated in a discrete time framework. I use continuous time for its greater mathematical generality and its convenience in the Brownian motion case.

3. I assume an exogenous industry production process in which the agency expects drugs to be submitted at a constant rate over time. Drugs are assumed to treat one disease and one disease only. See the conclusion for planned extensions of the model in which these assumptions will be relaxed.

4. Equivalently, infinitesimal movements in X are given by $dX = \mu dt + \sigma dz$, where z is a standardized Weiner variable whose increment dz has zero mean and variance dt (Harrison 1985).

5. I make the assumption of constant variance only to simplify the exposition of the model.

6. The parameter can exceed unity because citizens other than those directly afflicted by a disease may lobby for drug approval. Those afflicted may have relatives, friends, or other allies in organized patient associations and the media who wish to see the drug approved.

7. Although it does not affect the mathematics of the model, I assume that the agency's utility is linear in danger (μ) and not harm [$X(t)$]. Other functional forms are possible. If the agency's reputation has the character of a durable asset subject to erosion, then perhaps the regulator's utility is dependent not upon μ but upon $\exp(\mu)$. This would render the stochastic process a "geometric" Brownian motion (Dixit 1993). In much of the discussion that follows, I set α to unity without loss of generality. The model sidesteps the question of political control. One can assume that α is a function purely of the preferences of the agency's enacting coalition, or one can assume that the agency's preferences are independent. The model's predictions are the same under either regime.

8. This is a highly cursory summary of the drug approval process. See <www.fda.gov/cder> for more information. Drugs with an "accelerated approval" or "fast-track" status may be submitted for NDA approval before phase II trials are fully completed.

9. Dranove and Meltzer (1994) observe that the length of the Investigational New Drug (IND) stage is positive correlated with the length of NDA review. The speed of NDA review, while doubtless affected by drug firms, is more easily represented as a decision variable. I am currently working on a formal model of product approval with strategic firm submissions (see Carpenter and Ting 2001).