Professional Practice

Eugene Bardach
Editor

Candidates for inclusion in the *Professional Practice* section may be sent to the incoming section editor, Richard P. Nathan, Director, Rockefeller Institute of Government, 411 State Street, Albany, New York 12203, Work: (518) 443-5831, Fax: (518) 443-5834, Email: nathanr@rockinst.org.

WHAT IF…?

Eugene Bardach

Wearing another APPAM hat, that of Program Chair for the 2002 annual research conference, I pronounced that the conference theme would be "What If…?" The call for papers included this paragraph:

> APPAM members do a lot of what might be called "applied social science research." They use scientific methods to describe problems of concern to public policy, to explore their causes, and to estimate the effects of existing policies. This laudable commitment to empiricism, however, often inhibits analysts from taking the next step that is logically required of policy thinking: asking "What if…" questions about the possible effects of various policy options on outcomes that policy-makers desire. These questions are uncomfortably—and unavoidably—speculative. They are also, perhaps equally uncomfortably, normative and multi-dimensional. Dealing with them sometimes requires special methods, such as creating decision trees or running the "Crystal Ball" program or doing break-even analysis. More generally, though, it requires simply a self-conscious effort to pose, and then speculate about, a series of thoughtful "what if…?" questions.

As is (perhaps fortunately) typical of such annual themes, this one was ignored by the great majority of paper and panel proposers. One of the exceptional cases was an entire panel of papers devoted to "What if…?", organized by Jeryl Mumpower and his colleagues centered around the Rockefeller Institute at SUNY Albany. I asked the paper writers to consider condensing and revising their work for a symposium to be published in the *JPAM* Professional Practice section. Here are three of

Journal of Policy Analysis and Management, Vol. 23, No. 4, 889–920 (2004)

Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.20040

them. They deal with diverse subject matters: implementing welfare-to-work policy, destroying stockpiles of chemical weapons, and choosing decision parameters when interpreting mammograms.

USING SIMULATION MODELS TO ADDRESS "WHAT IF" QUESTIONS ABOUT WELFARE REFORM

A.A. Zagonel, J. Rohrbaugh, G.P. Richardson, and D.F. Andersen

BACKGROUND

In 1996, then President Clinton signed into law the Personal Responsibility and Work Opportunity Reconciliation Act, a sweeping piece of legislation designed to "end welfare as we know it." While many of the implications of this law have been made clear with the passage of time, when the law was passed state and local managers faced a thicket of policy design challenges comprising difficult "what if . . ." and "what . . . when" questions. Indeed, many of the most important normative questions did not concern the readily predicted aspects of the reforms, but inherently probabilistic concerns that come with big policy changes. Moreover, since the reform was outside traditional policy experience, policymakers and researchers had no steady intuition to guide them in considering plausible outcomes: What if there is a big recession, while people are facing time limits? What is the effect on states and localities if the behavioral effects of reform are very different from what is expected? This last question was not merely hypothetical: Caseload declines were indeed much steeper than was predicted by pre-reform policy. In New York State, these questions had a special edge since Article 17 of the state constitution mandates that local governments have a constitutional requirement to provide for the indigent and needy. While the federal government can end entitlements after 5 years, New York State and its local governments cannot suspend all benefits.

A coalition of state agencies and county governments set out to use simulation technologies supported by group model-building techniques (Andersen and Richardson, 1997; Richardson and Andersen, 1995; Vennix, 1996; Vennix, Andersen, and Richardson, 1997) to address these many what-if and what-when questions. Technical aspects of this project have been reported previously in the literatures on simulation (Allers et al., 1998; Andersen et al., 2000; Lee, 2001; Lee et al., 1998a; Rogers et al., 1997; Zagonel, 2003) and public affairs (ESR, 1998; Rohrbaugh, 2000; Rohrbaugh and Johnson, 1998). This brief paper uses this project to illustrate one way policy analysis can move from analyzing questions of "what is" to "what if."

What-if Questions Facing Welfare Managers in 1996

Asking what-if questions about future policy effects is intrinsically difficult because traditional policy analysis tools are data based, and data about the future simply do

DOI: 10.1002/pam.20054

not exist. Moreover, traditional policy models are designed and calibrated within a specific policy environment. Although such models are very useful for predicting the likely consequences of incremental policy changes, experiences in many policy domains indicate that these models are much less reliable in assessing very large policy changes that alter the environment on which the models are implicitly based. Thus, what-if analyses must be grounded in the data of what is, but also contain the judgments and guesses of policy experts concerning system structure and behavior in the future. These what-if analyses, simulated in robust nonlinear models that merge hard data with expert judgment, are best compared to intuitive judgments, insights, and "seat-of-the-pants" forecasts that managers and policymakers are often forced to make. They contrast with the precise and data-driven social scientific estimates that emerge from traditional policy analytic work.

Two distinct classes of questions are illustrated in the present case. The first of these are of the form, "What might happen if we were to make such and such a policy change?" For example: What might happen if the county were to invest more heavily in job training and placement services? What if we limited TANF services to only 2 or 3 years? What if we place additional emphasis on monitoring and assessment programs?

A second class of what-if questions is of the form, "What might happen if some scenario not under our control were to change dramatically?" Examples of these questions might be what might happen if (or when) the economy turns for the worse and unemployment begins to rise, or what happens if neighboring counties cut benefits, potentially starting a "race to the bottom" for discretionary welfare benefit levels. These were the kinds of questions county managers facing new state and federal mandates worried about in early 1997 as they faced multiple choices necessary to implement the new Personal Responsibility and Work Opportunity Reconciliation Act.[1]

A Thumbnail Sketch of the New York State Welfare Simulation Project

In January of 1997 a partnership between New York State agencies and three county governments undertook the simulation project we are reporting upon here.[2] The project emerged in four overlapping phases. The first phase involved using a group model-building approach to construct a simulation of the basic TANF system for Cortland County, a rural county located in central New York State (Rogers et al., 1997; Rohrbaugh, 2000; Zagonel, 2003).

The second phase of the project concentrated on formulating the "Safety Net" sectors of the model, exploring the flows and accumulations of clients moving through the system at some future point in time after they had lost TANF eligibility. This aspect of the model-based study is especially important, since the effect of policy on non-TANF safety-net institutions has received less systematic

[1] Policy research since then has answered some of these questions. Presidential addresses by Bane (2001) and Gueron (2003) to the Association for Public Policy Analysis and Management provided insightful reviews. Earlier policy work by Bane and Ellwood (1994), Baum (1991), Gueron and Pauly (1991), and Hollister, Kemper, and Maynard (1984) was strengthened and deepened by subsequent work by Bloom and Michalopoulos (2001), Greenberg, Michalopoulos, and Robins (2001), Grogger, Karoly, and Klerman (2002), Gueron (2002), and Weaver (2000). The work reported here adds to this growing body of scholarship the perspectives and tools of group modeling and policy simulation.

[2] The Commissioners of Social Services in the participating counties: Jane Rogers, Robert Allers, and Irene Lapidez, and the genuine engagement of their management teams drove the success of these projects.

attention than the effects of policy changes on the size of the TANF caseload. After losing eligibility, clients would have to be served somehow by the so-called "safety net," as mandated by Article 17 of the New York State constitution. These safety-net sectors also explored the resource stocks (such as assessment services, job training, and monitoring) that might influence the inflow, outflow, and recidivism flows of clients who had lost TANF eligibility. This portion of the modeling effort imported in whole cloth the TANF model developed in Cortland County, but had it calibrated and joined with the safety-net model developed for Dutchess County, a mid-sized suburban county located in the Hudson valley (Allers et al., 1998; Zagonel, 2003).

The third phase of the project was conducted in conjunction with a network of service providers located in Nassau County, a large and demographically complex county directly adjacent to New York City on Long Island. This portion of the project explored how model structures and approaches developed in smaller and less complex regions applied in the more complex environment in the downstate metropolitan area. Once again, the model was carefully parameterized and calibrated before it was used.

The final phase of the project was aimed at implementing policy insights from the model. Follow-up workshops were held in these counties with broadly based groups of stakeholders, including new participants who were not involved in model development, but who were involved in exploring policy futures, and scenarios, using the simulation model. To facilitate this process, the joined TANF–Safety Net model was wrapped in a graphical user interface that allowed it to be used by managers, policymakers, and laypersons who had not been previously involved in the model construction process (ESR, 1998; Rohrbaugh and Johnson, 1998). Below, we report on the final version of the model embedded in this "management flight simulator" for use by policymakers and laypersons alike.[3]

AN OVERVIEW OF GROUP MODEL BUILDING

Group model building[4] is a specialized form of group decision-support systems that allows a group of managers to interact in facilitated face-to-face structured activities resulting in the creation of a formal computer simulation model of a specific policy system. Typically a management team of ten to twenty persons meets with a facilitation team of three or four modeling and facilitation experts for two to four full working days spread out over three to four months.

The result of these group modeling activities is a formal computer simulation model that reflects a negotiated, consensual view of the "shared mental models" (Senge, 1990) of the managers in the room, tested and tried against existing administrative and time series data whenever possible.[5] The final simulation models that emerge from this process are crossbreeds, sharing much in common with data-

[3] The literature on management and policy-oriented simulation models uses the term "management flight simulators" to refer to this class of models that are tuned for use by lay audiences with easy to manipulate graphical interfaces.

[4] Early work emerged shortly after client-friendly simulation software appeared and was combined with group facilitation; see Reagan-Cirincione et al. (1991), Richardson and Andersen (1995), and Vennix et al. (1992). Developments since can be seen in Andersen and Richardson (1997), Vennix (1996), and Vennix, Andersen, and Richardson (1997).

[5] Zagonel (2002) has explored in some depth the tension that arises within these group modeling efforts as the model strives to meet the dual and sometimes conflicting goals, sometimes serving as a boundary object to facilitate the process of building group consensus, and sometimes serving as a micro-world depicting data and relationships in the policy environment external to the group in the room.

based social scientific research while at the same time being comparable to the rough-and-ready intuitive analyses emerging from backroom conversations.

In addition to working directly with the whole policy team, modelers do extensive work structuring the model and collecting administrative data. These data are used to cross check and verify parameters quickly estimated by the client group and to calibrate the model. The process used to build user confidence in the overall simulation model and its implications involves a linked set of structural and behavioral tests, as described below.

Overall, the final simulation model contained more than 600 active equations, with all of the functions, parameters, or initial conditions being connected to either administrative data from the county, time series, or expert judgments made by administrators on the management team, with all of the assumptions and data documented (Lee et al., 1998b, c; Zagonel, 2003).

"What-if" Insights from the Simulation Model

Because the flight simulator can easily and quickly show the consequences over time of changing any input assumption or parameter, it can be used rapidly to explore a wide range of policy options and scenario changes. For example, managers could double TANF employment services (or halve or eliminate them) to see what might happen. Highly risky and extreme policies could be explored quickly and in a no-risk situation since any possible system "crashes" would only be simulated crashes (without threats to client well-being or career-endangering cost overruns).

Below, we present two simulated policy runs to contrast with the base run. The base run (or "reference" run) shows the model's projection of what would happen to the county's welfare system if no policy changes were made and if there are no external scenario changes (in unemployment, for example).[6] The first policy run, labeled the "middle" policy, simulated a high investment in and emphasis on assessment, monitoring, and job-finding and promotion functions that are traditionally associated with a social services unit. In the simulation the resource stocks associated with TANF employment services were increased, as were TANF assessment and monitoring services (similar investments were made in the Safety Net sectors of the model). The second policy mix was labeled the "edges" policy and it contained a mixture of resource investments that concentrated on the "front" and "back" ends of the system. These investments were in prevention, child support enforcement, and self-sufficiency promotion—at the time investments such as these were typically outside the traditional purview of social services departments and would require extensive collaboration with other government agencies, as well as private and voluntary organizations in the county.

Figure 1 compares the base run to the middle and edges policy packages for one of the key performance indicators for the TANF welfare reform system—total job finding flows from TANF. The X axis is time, measured in years, and the Y axis is the total flow of people out of TANF, measured in people per year. The base case

[6] "Base" policy assumed that all welfare programs were funded at their 1996 levels. The "base" assumption about unemployment was that the economy was at the exact unemployment rate that would cause no growth, but also no decline over the time horizon 1984–1998 (the modeling team "backwards computed" this figure as part of the model-testing and confidence-building phases (see below). This calculation was intended to "hold constant" the very large effects of unemployment on TANF caseloads so that the runs could show "pure" effects of policy changes.

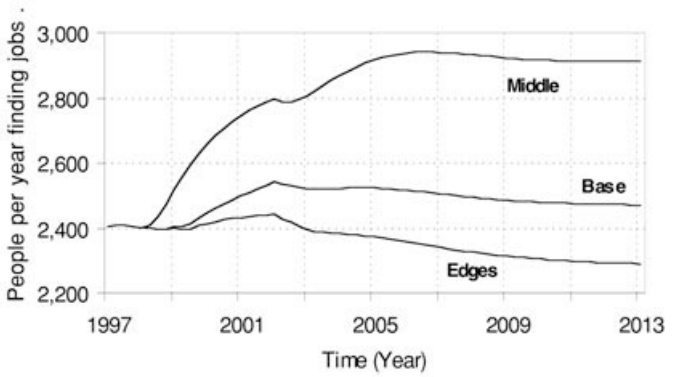


Figure 1. Total job-finding flows from TANF (base vs. middle vs. edges policy packages). The base case provides the initial reference point (2400 people per year in 1997) matching history in the particular county, against which to compare the alternative policy scenarios.

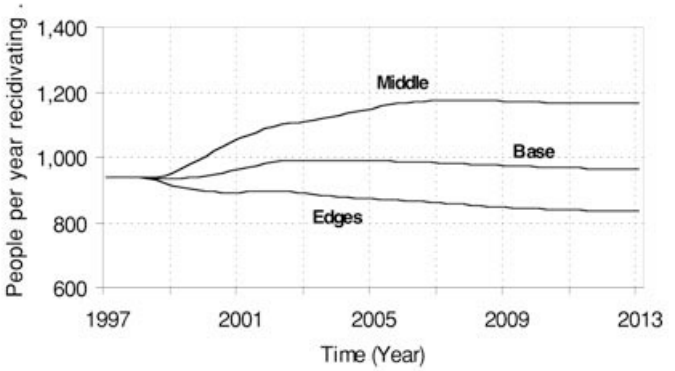


Figure 2. Total recidivism flows back to TANF (base vs. middle vs. edges policy packages). Again, the base case provides the initial reference point matching history in this particular county, against which to compare the alternative policy scenarios.

provides the initial reference point (2400 people per year in 1997) matching history in the particular county, against which to compare alternative policy scenarios. This key indicator, the sum of several welfare client flows emerging from TANF and finding jobs, shows how well each of the policy packages performs in terms of getting TANF clients off the welfare rolls and into employed status.

The base run shows that the 1996 welfare reform without any significant new resource investments starts to move clients off of the welfare rolls into employment. However, significant new investment in the middle policy package greatly accelerates this presumably beneficial trend. Investing in the edges of the system appears in the simulation actually to retard job-finding relative to the base run. These dynamic patterns are an example of the class of interesting and seemingly "coun-

terintuitive" results that often emerge from policy simulations of complex systems (Forrester, 1971; Sterman, 2000). As shown below, such predictions are the entry point for further policy discussion—both to examine the credibility of the result and to understand the mechanisms that lead to counterintuitive but beneficial findings. Does expanded investment in the middle policy create a critical bottleneck in our services? Can complementary investments elsewhere in the system relieve these bottlenecks? Does the relative timing of investments matter?

By focusing on a less commonly articulated performance measure—total recidivism—Figure 2 begins to reveal an important structural insight lurking behind these graphs.[7] The edges policy has the effect of significantly reducing recidivism in the model. Pavetti (1993) and Bane and Ellwood (1994) have shown that many families beginning a new welfare spell are not first-time entrants. Rather, these families are returning to the welfare system after a previous completed spell receiving public aid. Since the total number of persons coming on to TANF at any point in time is the sum of first-time recipients plus recidivists, this reduction in recidivism can dramatically decrease the overall TANF caseload. By contrast, the middle policy actually has the simulated effect of increasing recidivism. Richardson, Andersen, and Wu (2002) have explored this surprising dynamic further. They demonstrated that in the simulation model the high influx of families on TANF into the post-TANF employment support system had the effect of "swamping" these downstream resources, leading to long-term increases in recidivism. Since recidivism is not mandated to be tracked in the federal legislation and is hard to document across the welfare system, the increased TANF caseload could easily be misinterpreted as the result of some external influence such as rising unemployment rather than as a natural, endogenous consequence of the middle policy intended to reduce caseloads.

We learn that by concentrating in different areas of the system, the two policy packages strove to decrease the number of individuals receiving TANF cash aid by acting through two quite different mechanisms—the middle policy "pumped" persons off of TANF while the edges policy tended to shut down the inflow from recidivism. For most of the policy runs examined, the edges policy could do a better job of reducing overall caseloads. Equally important, the policy reduced strain on resources, thus enriching the quality of services. To summarize the mechanism at work here, the middle policy is great at getting people into jobs, but then they lose those jobs and cycle back into the system because there are not enough resources devoted to helping them stay employed. The edges policy lets them trickle more slowly into jobs but then does a better job of keeping them there.

But what about costs? Both the edges and the middle policy packages could be expensive. For example, the simulation projected that the edges policy could increase overall program expenditures by about 10 percent 5 years into its implementation. This contrasted with an approximate 5 percent increase in program expenditures associated with the middle program over the same period. Such large cost considerations could deter local administrators from implementing either of the policy packages given a tight fiscal situation. The simulation model produced detailed cost projections for all policy options, relying on detailed administrative

[7] Note that in order to show the pattern of the dynamics, the scales on Figures 1 and 2 are not zero-based scales and the scales are different. Thus the visual intervals between graphs in Figures 1 and 2 are not comparable.

data available at the state and local level to create the costing equations. Costs were broken out by federal, state, and local shares. A clear cost-performance tradeoff emerges over time: While more costly in the short run, the edges policy eventually reduces costs below the base and middle scenarios in the long run, in a classic worse-before-better pattern.

Whatever the final policy choice, the simulation model provides a "level playing field" for evaluating the what-if implications of multiple policy and scenario changes always using precisely the same agreed-upon set of assumptions and numbers. The model always uses exactly the same mathematically precise logic to make inferences from these assumptions. Of course, these inferences are only as good as the model upon which they are based. Hence, the overall project attempted to pay careful attention to model testing, sensitivity analyses, and calibration. Where does sensitivity analysis fit into this discussion? All models are wrong, but we hope to understand the critical assumptions that would really hurt us.

Model Calibration and Confidence

The class of system dynamics simulation models used for this project does not rely solely on statistical measures of goodness of fit to assess model quality. Rather, the modeling effort uses a battery of tests of model structure and behavior to build user confidence in the model as a policy tool (Forrester and Senge, 1980; Richardson and Pugh, 1981; Sterman, 2000). We relied primarily on five classes of structural and behavioral tests to refine and calibrate the model, thus increasing confidence in the resulting simulations and structural analyses.

First, and perhaps most important, the actual structure of the model emerged directly in front of the group of managers who would ultimately lead decision-making processes. The "theory" of how clients moved through the system and how resources and programs could affect those movements emerged from a series of focused discussions with initial sketches being drawn by hand and later converted to system flow diagrams. This step is also important to bolster the credibility and sense of joint ownership of the models. Managers are naturally (sometimes correctly) dismissive of many complex modeling efforts. They must be included in every way possible to bolster the legitimacy of the effort (Richardson and Pugh, 1981).

Second, the parameters used to run the model were based on administrative data for the involved agencies whenever possible. Members of the management team who were the local "data experts" helped to comb through extensive data sets to arrive at model parameters (Lee et al., 1998b, c). In addition, model output for some cross-section in time could be compared directly to data reported on the monthly or annual administrative data sheets.

Third, some parameters and functional relationships were not directly available in the existing administrative data. Lee et al. (1998a) reported on the data-gathering conferences that were used to capture expert judgments of model parameters and functions that were missing from the administrative data. In these half-day conferences, teams of knowledgeable managers made independent estimates of key parameters and functions and these data were compared and analyzed for consistency and convergence between separate estimates of the same number or effect.

Fourth, sensitivity analyses helped managers gain confidence by showing that the policy insights they were deriving from the study were largely independent of plausible ranges of parameter values. For example, in the flight simulator users could

pick alternative economic or client behavioral assumptions and see the impacts those changes would have on outcomes.

Fifth, running the model and comparing its output to what is known to have actually happened in the past or could be expected in the future is another test for building confidence in the model. Figure 3 shows a simulation run for Nassau County where the model was initialized with data from 1984 and run through 1998 driven by the historic unemployment rate. Plotted on a separate scale are the actual and the simulated caseloads for the same period. The actual caseload shows a trajectory highly similar to the simulated (R^2 = 0.91), although the simulated data show a turning point that lags the actual caseload. The simulated trajectory arises from dynamics internal to the model as driven by historic unemployment.

In addition to "simulating history," cross sections of model output were compared to actual data for the same period, and multiple policy and scenario runs drove the model with extreme conditions to assure that it behaved sensibly over a wide range of policy assumptions.[8]

Crafting General Insights from a Specific Model

The model structure elicited from the teams of county managers exhibited surprising dynamics that could be traced to endogenously generated recidivism. This insight had the practical effect of refocusing managerial attention on making downstream investments intended to prevent clients from recidivating back into TANF. Lee (2001) and Richardson, Andersen, and Wu (2002) have suggested that this practical result also opens a deeper set of theoretical questions centering on how to manage multi-stage service delivery systems. Significant investments in upstream

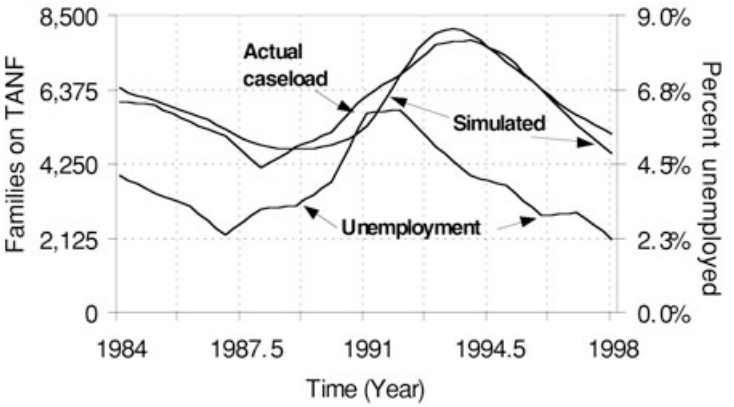


Figure 3. Simulated versus actual caseload in Nassau County (model driven by historic unemployment rate as shown).

[8] Zagonel (2003) presents a critical reappraisal of the model's confidence-building tests, suggesting improvements in model structure, providing additional statistical tests of model fit, and updating behavioral tests with more recent time series data.

capacity (here moving clients off TANF) can speed clients downstream and swamp downstream resources (here straining programs trying to prevent recidivism), leading to overall poor system performance.

This same dynamic is hypothesized to exist within patient versus outpatient mental health services (Huz et al., 1997; Richardson, Andersen, and Wu, 2002) and other structurally similar multi-stage service delivery systems having the potential for recidivism. Hence, as managers address "what if" questions using well-structured simulation models, they both generate practical insights to support future specific initiatives and create more generic insights that deepen our overall understanding of service delivery systems.

WHAT HAPPENED IN THE THREE COUNTIES?

The pattern of pursing implemented results varied considerably in the three counties. The Commissioner of Social Services in Cortland County used the model and its results to plan her investment priorities to implement Welfare Reform for the mid- to long-term (Rogers et al., 1997). After the simulation phase of the project was completed, Commissioner Rodgers convened a series of county-wide workshops involving stakeholders in government and private, not-for-profit organizations. The purpose of these workshops was to agree on a set of ranked investments. As a result of these initiatives, Cortland County decided to spend about $700,000 in investments targeted mainly at edges policies, including a resource center to coordinate community efforts toward diversion, a program of job counselors and case managers to promote self-sufficiency, and an expansion of child care services in the community. The workshop groups reconvened two years later to evaluate qualitatively the success of these investments.

In Dutchess County, Commissioner Allers convened a county-wide workshop to roll out the results of the model-based study. This workshop did not result in direct investments. However, within a year Allers created a public-private task force to design and implement concrete initiatives aimed at the "Edges" policies (Allers et al., 1998; see also ESR, 1998; Rohrbaugh, 2000; and Rohrbaugh and Johnson, 1998).

Implementations in Nassau County were the least extensive of the three sites. Of the three counties, Nassau was the largest and most demographically diverse, situated on Long Island adjacent to New York City. The commissioner convened her direct staff to work with the model but did not involve a wider group of community stakeholders to implement model-based implications.

HOW GROUP MODELING CAN HELP "WHAT IF" ANALYSES

A number of the process features related to building these models contribute to their appeal for front-line managers:

- *Engagement*. Key managers are in the room as the model is evolving, and their own expertise and insights drive all aspect of the analysis.
- *Mental models*. The model building process uses the language and concepts that managers bring to the room with them, making explicit the assumptions and causal mental models managers use to make their decisions.
- *Alignment*. The modeling process benefits from diverse, sometimes competing points of view as stakeholders have a chance to wrestle with causal assumptions in a group context. Often these discussions realign thinking and are among the most valuable portions of the overall group modeling effort.

- *Refutability.* The resulting formal model yields testable propositions, enabling managers to see how well their implicit theories match available data about overall system performance.
- *Empowerment.* Using the model, managers can see how actions under their control can change the future of the system.

Group modeling merges managers' causal and structural thinking with the available data, drawing upon expert judgment to fill in the gaps concerning possible futures. The resulting simulation models provide powerful tools to ground what-if thinking.

A preliminary version of this paper was presented at the 2002 APPAM Research Conference in Dallas, Texas, November 2002. The authors in this fully collaborative effort are presented in reverse alphabetical order. We are especially grateful for the support, guidance, and welfare expertise of Irene Lurie. We thank Tsuey-Ping Lee, Naiyi Hsiao, and Robert Johnson for their contributions to this research. The authors are also grateful for helpful comments received from anonymous reviewers.

A.A. ZAGONEL is at the Critical Infrastructure Assurity Group at Sandia National Laboratories.

J. ROHRBAUGH is Professor of Public Administration and Policy at the Rockefeller College and Dean of International Students at the University at Albany, State University of New York.

G.P. RICHARDSON is Professor of Public Administration, Public Policy, and Information Science at the Rockefeller College and Chair of Public Administration and Policy at the University at Albany, State University of New York.

D.F. ANDERSEN is Professor of Public Administration, Public Policy, and Information Science at the Rockefeller College, University at Albany, State University of New York. .

REFERENCES

Allers, R., Johnson, R., Andersen, D.F., Lee, T.P., Richardson, G.P., Rohrbaugh, J., & Zagonel, A.A. (1998). Group model building to support welfare reform, Part II: Dutchess County. Proceedings of the 16th international conference of the system dynamics society, Québec City, QC, Canada. Albany, NY: System Dynamics Society.

Andersen, D.F., & Richardson, G.P. (1997). Scripts for group model-building. System Dynamics Review, 13(2), 107–129.

Andersen, D.F., Richardson, G.P., Rohrbaugh, J., Zagonel, A.A., & Lee, T.P. (2000). The impact of US welfare reform on states and counties: Group facilitated system dynamics modeling, policy analysis and implementation. GDN 2000, INFORMS Section on Group Decision and Negotiation. University of Strathclyde. Glasgow, Scotland.

Bane, M.J. (2001). Expertise, advocacy and deliberation: Lesson from welfare reform. Journal of Policy Analysis and Management, 20, 191–197.

Bane, M.J., & Ellwood, D.T. (1994). Welfare realities: From rhetoric to reform. Cambridge, MA: Harvard University Press.

Baum, E.B. (1991). When the witch doctors agree: The Family Support Act and social science research. Journal of Policy Analysis and Management, 10, 603–615.

Bloom, D., & Michalopoulos, C. (2001). How welfare and work policies affect employment and income: A synthesis of research. New York: Manpower Demonstration Research Corporation.

ESR (1998). Understanding welfare. Empire State Report, (January), 17.

Forrester, J.W. (1971). Counterintuitive behavior of social systems. Technology Review, 73(3), 52–68.

Forrester, J.W., & Senge, P.M. (1980). Tests for building confidence in system dynamics models. In A.A. Legasto Jr., J.W. Forrester, & J.M. Lyneis (Eds.), System dynamics (pp. 209–228). TIMS Studies in the Management Sciences 14. New York: North-Holland.

Greenberg, D.H., Michalopoulos, C., & Robins, P.K. (2001). A meta-analysis of government-sponsored training programs. Baltimore: University of Maryland.

Grogger, J., Karoly, L.A., & Klerman, J.A. (2002). Consequences of welfare reform: A research synthesis. Santa Monica, CA: Rand Corporation.

Gueron, J.M. (2002). The politics of random assignment. In Mosteller & Boruch (Eds.), Evidence matters: Randomized trials in education research (pp. 15–49). Washington, DC: Brookings Institution Press.

Gueron, J.M. (2003). Fostering research excellence and impacting policy and practice: The welfare reform story. Journal of Policy Analysis and Management, 22(2), 163–174.

Gueron, J.M., & Pauly, E. (1991). From welfare to work. New York: Russell Sage Foundation.

Hollister, R.G. Jr., Kemper, P., & Maynard, R. (Eds.) (1984). The National Supported Work Demonstration. Madison: University of Wisconsin Press.

Huz, S., Andersen, D.F., Richardson, G.P., & Boothroyd, R. (1997). A framework for evaluating systems thinking interventions: Lessons from a pilot test of an experimental approach. System Dynamics Review, 13(2), 149–169.

Lee, T.P. (2001). The dynamics of New York State social welfare reform finance at the county level: A feedback view of system behavior. Proceedings of the 19th International Conference of the System Dynamics Society. Atlanta, Georgia. Albany, NY: System Dynamics Society.

Lee, T.P., Zagonel, A.A., Andersen, D.F., Rohrbaugh, J., & Richardson, G.P. (1998a). A judgment approach to estimating parameters in group model-building: A case study of social welfare reform at Dutchess County. Proceedings of the 16th International Conference of the System Dynamics Society. Québec City, QC, Canada. Albany, NY: System Dynamics Society.

Lee, T.P., Andersen, D.F., Rohrbaugh, J., & Zagonel, A.A. (1998b). Parameter booklet for the Nassau County joined TANF and safety net model (September). Welfare Reform Project. Center for Policy Research, University at Albany.

Lee, T.P., Andersen, D.F., Rohrbaugh, J., & Zagonel, A.A. (1998c). Parameter booklet for the Dutchess County joined TANF and safety net model (April). Welfare Reform Project. Center for Policy Research, University at Albany.

Pavetti, L. (1993). The dynamics of welfare and work: Exploring the process by which women work their way off welfare. Cambridge, MA: Harvard University.

Reagan-Cirincione, P., Schuman, S., Richardson, G.P., & Dorf, S. (1991). Decision modeling: Tools for strategic thinking. Interfaces, 21(6), 52–65.

Richardson, G.P., & Andersen, D.F. (1995). Teamwork in group model-building. System Dynamics Review, 11(2), 113–137.

Richardson, G.P., Andersen, D.F., & Wu, Y.J. (2002). Misattribution in welfare dynamics: The puzzling dynamics of recidivism. Proceedings of the 20th International Conference of the System Dynamics Society. Palermo, Italy. Albany, NY: System Dynamics Society.

Richardson, G.P., & Pugh, A.L. III (1981). Introduction to system dynamics modeling with DYNAMO. Cambridge, MA: Productivity Press.

Rogers, J., Johnson, R., Zagonel, A.A., Rohrbaugh, J., Andersen, D.F., Richardson, G.P., &

Lee, T.P. (1997). Group model-building to support welfare reform in Cortland County. Proceedings of the 15th International Conference of the System Dynamics Society. Istanbul, Turkey. Albany, NY: System Dynamics Society.

Rohrbaugh, J. (2000). The use of system dynamics in decision conferencing: Implementing welfare reform in New York State. In G.D. Garson (Ed.), Handbook of public information systems (pp. 521–533). New York: Marcel Dekker.

Rohrbaugh, J., & Johnson, R. (1998). Welfare reform flies in New York. Government Technology, (June), 58.

Senge, P.M. (1990). The fifth discipline: The art and practice of the learning organization. New York: Doubleday.

Sterman, J.D. (2000). Business dynamics: Systems thinking and modeling for a complex world. Boston, MA: McGraw-Hill.

Vennix, J.A.M. (1996). Group model building: Facilitating team learning using system dynamics. Chichester: Wiley & Sons.

Vennix, J.A.M., Andersen, D.F., & Richardson, G.P. (Eds.). (1997). Group model building. System Dynamics Review, 13(2), 103–106.

Vennix, J.A.M., Andersen, D.F., Richardson, G.P., & Rohrbaugh J. (1992). Model building for group decision-support: Issues and alternatives in knowledge elicitation. European Journal of Operations Research, 59(1), 28–41.

Weaver, R.K. (2000). Ending welfare as we know it. Washington, DC: Brookings Institution Press.

Zagonel, A.A. (2002). Model conceptualization in group model-building: A review of the literature exploring the tension between representing reality and negotiating a social order. Proceedings of the 20th International Conference of the System Dynamics Society. Palermo, Italy. Albany, NY: System Dynamics Society.

Zagonel, A.A. (2003). Using group model-building to inform welfare reform policy making in New York State: A critical look. Proceedings of the 21st International Conference of the System Dynamics Society. New York City. Albany, NY: System Dynamics Society.

FOUR WHAT IFS? EVALUATION OF ALTERNATIVE TECHNOLOGIES FOR THE DESTRUCTION OF CHEMICAL WEAPONS

Sandor Schuman and John Rohrbaugh

The application of "what if" analysis in the evaluation of alternative technologies for the destruction of chemical weapons by the United States Army provides an opportunity to examine four different types of what-if questions.

CONTEXT, ORGANIZATION, AND ANALYTICAL APPROACH

By 1996, to meet existing Congressional mandates (Public Law 99-145, 1985) and then-anticipated international treaty obligations (Chemical Weapons Convention, 1997), the U.S. Army was already in the process of destroying its stockpile of 31,000 tons of chemical weapons at an estimated cost of over 7 billion dollars and with a

DOI: 10.1002/pam.20053

target completion time of December 31, 2004. Having conducted more than 12 years of research and development, the Army had found incineration to be the best available technology for destroying the chemicals. However, because of strong public opposition to incineration, the Army was required to evaluate alternative technologies for destruction of the mustard and nerve gases stored in bulk tanks in the states of Maryland and Indiana (Public Law 102-484, 1992).

Amid concerns about the ability of the program to meet its deadline (Foote, 1994), the Army initiated in 1996 an extensive evaluation of alternative technologies. This evaluation involved a team of 29 subject matter experts selected from Army organizations and outside contractors (U.S. Army Program Manager for Chemical Demilitarization, 1996) who systematically compared incineration—as the baseline technology—and five alternative technologies. The evaluation process involved the subject matter experts working in five teams, each addressing a different subject matter (see Table 1). To ensure thorough and consistent consideration of over 150 specific questions to which each of the competing technology vendors was required to respond, the subject matter experts were supported by an analytical facilitation team comprised of group facilitators and decision analysts headed by the senior author. The analytical facilitation team used one computer projector to display Logical Decisions™ software for the Analytic Hierarchy Process and Multiattribute Utility Analysis, and a second computer and projector to display word processing software that was used for narrative documentation of the decision-making process. In a series of meetings over the course of six weeks, the expert teams were guided by the analytical facilitation staff to systematically integrate the available factual information and their expert judgments regarding each of the technology alternatives (see Figure 1).

THE FOUR "WHAT IFS?"

At least four distinct types of what-if questions can be asked:

1) If we do this, what results?
2) If we are wrong (in 1 above), what results?
3) If we do this and conditions change, what results? and
4) If we anticipate the results above, which results do we prefer?

1. If We Do This, What Results?

Perhaps one of the earliest forms of mathematical modeling in operations research was quantifying the strength of relation between action and consequence (e.g., acceleration of an assembly line by one foot per minute will result in a 10 percent increase in daily output). As Andersen and Rohrbaugh (1992) point out in their discussion of simulation models for urban dynamics, a substantial class of issues is subsumed in this focus on the nature of means-ends relations. Policy analysts often are required to establish the magnitude of a change in Y (elasticity) that can be attributed to a change in X.

To review the six technologies systematically, the anticipated results were described in five major categories or primary criteria: efficacy, safety, environmental impact, schedule, and cost. Note that, with the exception of incineration (an established technology for which empirical results were well established), the five alternative technologies were experimental and at various stages of development.

Table 1. Evaluation of alternative technologies for the destruction of chemical weapons.

Technologies Evaluated:
Incineration (baseline technology)
Neutralization followed by onsite biodegradation
Neutralization followed by offsite post treatment
Molten metal catalytic extraction process
Electrochemical oxidation
High-temperature gas-phase chemical reduction
Subject Matter Expert Teams:
Environmental impact
Process efficacy
Process safety
Schedule
Cost

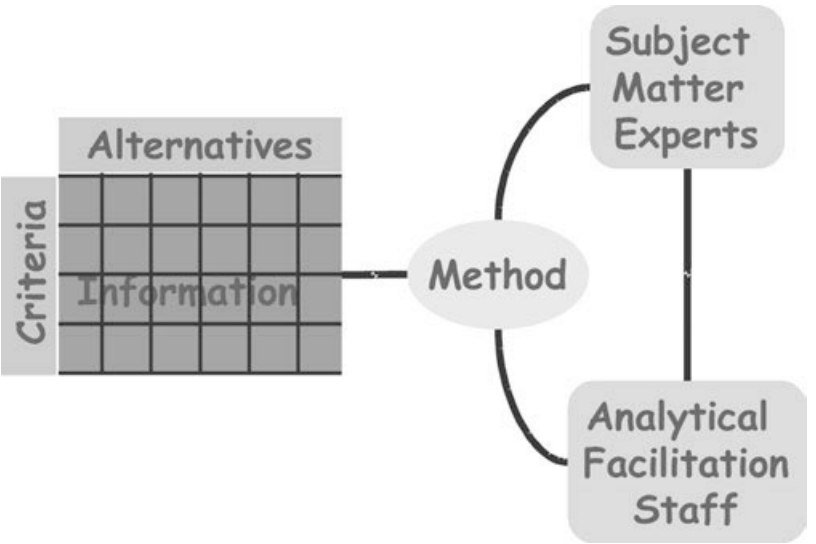


Figure 1. The analytical facilitation staff worked in partnership with the subject matter experts who systematically evaluated the technology alternatives against a uniform set of criteria and integrated this information by building a multi-attribute utility model.

Thus, the anticipated results for these five had to be based on the judgment of the subject matter experts rather than on exhaustive empirical data. All primary criteria had been disaggregated by a previous panel of experts who organized their concerns into 14 factors, 49 subfactors, and more than 150 specific questions used as the basis for measuring the anticipated results of implementing each of the six tech-

nologies. These questions were the basis for soliciting information from vendors regarding the performance of their technologies, so that the subject matter experts could predict and assess the likely results. For example, implementation of high-temperature, gas-phase reduction was viewed to require the greatest number of months to project completion, while the current incineration process, of course, required the fewest.

2. If We Are Wrong (in 1. Above), What Results?

As we know from the most elementary application of inferential statistics, estimates must be accompanied by some form of confidence interval. Operations research developed the use of sensitivity analysis for this reason, that is, to undertake a careful review of the magnitudes of potential error in estimation that could alter a managerial decision.[1] Sensitivity analyses are performed in many ways. One-way sensitivity analysis examines how changing a single parameter alters important outcomes, while *n*-way sensitivity analysis involves more complex exploration of how simultaneous changes in *n* parameters lead to systematic differences in results. In the present case, for example, changes in assumed weather patterns away from a base case may expose vulnerability to low probability failures in safety systems. Even more sophisticated sensitivity analyses can be directed at the elasticities (i.e., the table functions), not just in isolation but also in combination with one or more other alterations in the initial model. Where slight perturbations in the estimates comprising the initial model lead to apparent shifts in preferred alternatives, one is led to investigate the validity and reliability of such estimates more thoroughly.

In the assessment of alternative technologies to be used in chemical demilitarization, these basic forms of sensitivity analysis were employed extensively. Initially, agreement on the ratings for each of the 49 subfactors was accomplished, these ratings were aggregated by factor, then by criterion and, finally, by a single, composite figure for each alternative. In the model used here, the aggregation rules for creating factor scoring from subfactors and overall criterion levels from factors (and the composite figure for each alternative) were expressed in linear and summative form with differential weights (derived from pairwise range comparisons of minimum and maximum scores) attached to the components at each stage. Thus, it was possible to test how specific adjustments in subfactor ratings might affect factor scoring or even overall criterion levels for any of the alternatives under consideration. Sensitivity analyses also were attached to these aggregation rules to assess the extent to which arguable modifications would influence the conclusions to be drawn from this evaluation.

3. If We Do This and Conditions Change, What Results?

Another distinct type of "what if" question identified in operations research pertains to making explicit assumptions about possible states of nature, that is, about what "the future may hold in store for us." Decisionmaking—either under uncertainty or under risk—requires the specification of unfolding events beyond one's control that would affect the outcomes experienced from particular choices; decision trees make these chance forks explicit. In scenario planning, for example, mod-

[1] Sensitivity analysis also might be used to find the range of application for particular policies; however, it was not used in this way in the present case.

eling begins not with the identification of alternatives or criteria but with specially constructed stories about a variety of distinct, plausible changes in the status quo. The purpose of scenario planning is not to predict or control the future but to build an appreciation of how well-designed courses of action might contribute to shaping future conditions and thereby improve their consequences for the decision maker.

In modeling the choice between technologies for the destruction of chemical weapons, the most important what-if question posed about the future concerned uncertainty regarding public opposition that potentially could delay the Resource Conservation and Recovery Act permit process. Under the "conservative" scenario, risk-adjusted estimates were made for subfactors of both the schedule and cost criteria which future public opposition might primarily affect; parallel estimates also were agreed for these subfactors under the base scenario (i.e., future conditions in which substantial permitting delays do not occur). For example, in the base scenario 30 percent contingency costs for delays were included in several cost estimates, but the comparable risk-adjusted cost estimates were placed higher still—sometimes by 50 to 75 percent.

4. If We Anticipate the Results Above, Which Results Do We Prefer?

Another class of issues with great pertinence to modeling is the exercise of judgment about preferred policy outcomes. What effects are intended? How much effect is desired? Are some effects more important than others? The mathematical solution to the classic "knapsack problem" in linear programming offered an analytical method for the allocation of scarce resources but, of course, depended on an explicit statement of the tradeoffs between such contents as water, sandwiches, apples, and cookies while on the hike. Public policy problems are not different: To ascertain what we should do, we must ask what results we prefer. The pertinent what-if question of this type examines whether our objectives will be achieved by our proposed actions.

Perhaps the most difficult question arises when one seeks to assign specific valuations to desired—or undesired—social outcomes. From whose perspective should these issues be evaluated: Experts? Policymakers? The affected public? This problem is especially vexing when one considers the specific problem of the present case. How should one value the risk of serious health problems that might result from release of chemical warfare agents or improper disposal? How should one value public anxiety about chemical agents and generalized public trauma in the event of unintended release? How should one value the delays and costs associated with politically unfavorable outcomes?

In recommending technology for chemical demilitarization, the expert panel gave considerable attention to the tradeoffs between efficacy, safety, environmental impact, schedule, and cost. This is not just one more set of technical issues, it is a matter of the foundational expression of public values. As in their other assignments of weights to subfactors and to factors, priorities were elicited by repeated pairwise range comparisons of minimum and maximum criterion levels. At this final stage of the evaluation process, the key what-if question pertained to these priorities: What if these numerical expressions of public values are flawed—will we be advocating the wrong technology?

Analyses were undertaken to explore the profile of weights that had been articulated; examples for cost and efficacy criteria are shown in Figure 2. The heavy vertical lines mark the relative priorities of cost and efficacy agreed upon by the expert team (approximately 0.12 and 0.32, respectively). Rising and falling trajectories of

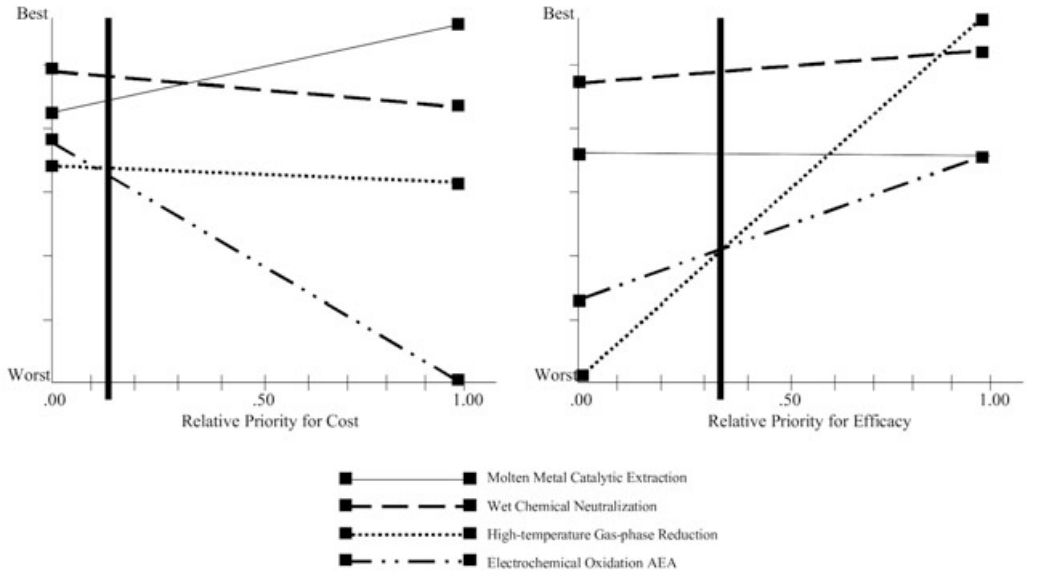


Figure 2. Analytic exploration of the question: If we anticipate the results above, which results do we prefer? The graphs show the change in preference for four technologies as the weight on top-level criteria (cost on the left, efficacy on the right) varies from 0 to 1.

overall acceptability of the four technologies are depicted as the relative priority of each criterion increases from left to right along the horizontal scale. Figure 2 illustrates that, from the standpoint of cost, the preferred technology would change if cost were determined to be a more important criterion; unless efficacy became virtually the only criterion of interest, however, the preferred technology would remain the same.

CONCLUSIONS

The contributions from economics for improved decision making about public policy problems are substantial, since the availability of such analyses can serve to better inform the key stakeholders. Nevertheless, since research designs are partial, data incomplete, and information unreliable, the opportunity for scientific disagreement and scholarly debate (even in a political vacuum) is enormous (Vari and Rohrbaugh, 1996). Yet, even with elaborate research designs, complete data, and fully reliable information, effective public policy making must rely, as well, on the best analytical tools available to us from the history of operations research and the development of management science. We must be prepared to ask—and answer well—at least four types of what-if questions, as in the present case of the Army-sponsored evaluation project.

If we do this, what results? We must be able to have an explicit basis for the pertinent elasticities that we incorporate in our model.

If we are wrong, what results? Sensitivity analyses are essential to establish just how critically dependent any decision may be on certain faulty assumptions about means and ends.

If we do this and conditions change, what results? Wishing for a beneficent future for policy actions does not guarantee one. We must anticipate worse-case, if not worst-case, scenarios and make decisions accordingly.

If we anticipate the results above, which results do we prefer? Rational decisionmaking about a public policy is a process that allocates scarce resources to achieve our most valued objectives. This is an extremely challenging task, not only because priorities may differ across stakeholders, but also because such priorities may be difficult to explicate. We must work especially hard to carefully pack the public policy knapsack.

What actually happened? The Army adopted the expert panel's recommendations to use the neutralization with biodegradation and post-treatment technologies. Army decisionmakers were influenced not only by this thorough what-if analysis, but by the fact that these results were corroborated by two independent evaluations conducted by the National Academy of Sciences and one of the Citizens Advisory Committees. There are two additional points of interest bearing on this result. First, the analytical facilitation team was prepared to perform additional what-if analyses with the Army's decisionmakers to explore different weighting schemes for the top-level criteria. This step was found to be unnecessary, perhaps because the decisionmakers felt that the equal-weighting scheme employed by the expert team was adequate, and that any increase in the weight on cost—while it would change the preferred alternative—would not be publicly acceptable. Second, it is interesting to note that one of the Citizens Advisory Committees relied heavily on a criterion not considered by the expert panel. One of their major concerns was that, once the chemical weapons destruction was completed, operation of the facility would be continued to process other types of toxic chemicals, which would be shipped in from other locations. Consequently an important criterion for them was that the selected technology be uniquely suited to the destruction of these particular chemicals. That is, despite promises from the Army that the plant would be dismantled once the chemical weapons destruction was complete, the committee wanted it to be technologically impossible for the Army to go back on its word.

Supplemented by further testing and development, the Army began construction of the Aberdeen Chemical Agent Disposal Facility disposal facility in Maryland in April 1999, and the Newport Chemical Agent Disposal Facility in Indiana in April, 2000. The Maryland facility started operation in April 2003 (Aberdeen Chemical Agent Disposal Facility, 2003); the Indiana facility was projected to start operation in January 2004, but has since been delayed (Newport Chemical Agent Disposal Facility, 2003).

SANDOR SCHUMAN is Research Associate, Center for Policy Research, Rockefeller College of Public Affairs and Policy, University at Albany, State University of New York.

JOHN ROHRBAUGH is Professor of Public Administration and Policy, Rockefeller College of Public Affairs and Policy, University at Albany, State University of New York.

REFERENCES

Aberdeen Chemical Agent Disposal Facility (2003). APG to begin neutralizing chemical agent stockpile. Chemical Materials Agency (Provisional) Public Outreach and Information Office Press Release, April 22.

Andersen, D.F., & Rohrbaugh, J. (1992). Some conceptual and technical problems in integrating models of judgment with simulations models. IEEE Transactions on Systems, Man, and Cybernetics, 22, 21–34.

Foote, W.G. (1994). The chemical demilitarization program—will it destroy the nation's stockpile of chemical weapons by December 31, 2004? Military Law Review, 146(Fall), 1–93.

Newport Chemical Agent Disposal Facility (2003). News release, November 13.

U.S. Army Program Manager for Chemical Demilitarization (1996). Alternative technology program evaluation report. Aberdeen Proving Ground, MD: U.S. Army.

Vari, A., & Rohrbaugh, J. (1996). Decision conferencing GDSS in environmental policy making: Developing a long-term environmental plan in Hungary. Risk Decision and Policy, 1, 71–89.

DETECTION AND SELECTION DECISIONS IN THE PRACTICE OF SCREENING MAMMOGRAPHY

Thomas R. Stewart and Jeryl L. Mumpower

A large class of problems in society requires detection or selection decisions. Examples include: Who should receive extra scrutiny in airport screening? What personal characteristics, if any, should patrolling police attend to? What blood alcohol levels constitute driving while intoxicated? What thresholds should be used for issuing severe weather warnings or terrorist-related security alerts? At what age, if any, should men routinely receive PSA testing for prostate cancer, and what thresholds should be established for treatment?

In this paper we focus on an important member of this class—the practice of screening mammography. Substantial uncertainty and disagreement persist concerning the value of regular mammogram screenings for women, particularly those between the ages of 40 and 49. For example, the National Breast Cancer Coalition (2003) has concluded that there is insufficient evidence to recommend for or against screening mammography for any age group of women. Conversely, the U.S. Preventive Services Task Force (2002) has recommended screening mammography, with or without clinical breast examination, every 1 to 2 years for women aged 40 and older.[1]

The results of seven controlled clinical trials on the effectiveness of screening mammography remain controversial (e.g., de Koning, 2003; Olsen and Gotzsche, 2001). Critics argue that the studies are flawed and inconclusive. Taken as a whole, data from these trials indicate a 24 percent decline in breast cancer mortality associated with

[1] As an anonymous reviewer noted, the Task Force was rather tepid in its endorsement. Their summary of recommendations concluded: "For women age 40–49, the evidence that screening mammography reduced mortality from breast cancer is weaker, and the absolute benefit of mammography is smaller than it is for older women."

DOI: 10.1002/pam.20055

mammography. Despite the large number of women enrolled in clinical studies of mammography effectiveness, meta-analysis indicates that the estimated protective effects of mammography depend strongly upon the inclusion or exclusion of specific contested studies. Olsen and Gotszche (2001) argue that the principal studies most favorable to mammography display systematic prior differences between the mammography and non-mammography comparison group that undermine study validity. The exclusion of such studies leads to no statistically significant benefit associated with the intervention. Most discouraging, large increases in the proportion of women receiving mammography have not resulted in large declines in breast cancer mortality.

OVERVIEW OF MAMMOGRAPHY SCREENING AND PRACTICE

Figure 1 provides a framework for describing the relations among three domains of decision making about mammography screening: decisions by women and their doctors to obtain screening, decisions by radiologists to recommend biopsy, and decisions by policymaking bodies to recommend and support screening for certain sub-groups of women.

Beginning at the left, some fraction of the eligible women voluntarily decides to be screened.[2] Their decision may be influenced by the recommendations of influential organizations, by their doctor, by their families and friends, and by information they receive about mammography. Recently there has been substantial effort to increase the proportion of women who decide to be screened. Some studies suggest that many women are confused and are not well-informed decisionmakers (Rimer et al., 2002; Schwartz et al., 1997).

Next, screening takes place, and the radiologist's decisions are either positive (biopsy recommended) or negative (normal follow-up).[3] At this stage, the base rate of women with malignancies, the accuracy of interpretation of mammograms, and the thresholds for choosing between a positive or negative finding are important considerations.

There are four possible outcomes of the radiologist's decision. If the preliminary diagnosis is positive, a biopsy is obtained; the biopsy result may then indicate a malignant (true positive) or benign (false positive) condition. Data show that there are about three false positives for every true positive. If the finding is negative, there is no biopsy. While most of the women who are not recommended for biopsy do not have a malignant mass (true negative), some do (false negative).

Each of the outcomes has associated costs and benefits. A true positive may mean earlier identification and treatment of cancer, possibly resulting in a better outcome. Mammographers want to avoid false negatives because undetected and untreated cancer is the worst possible outcome. Because of the high uncertainty, avoiding false negatives results in many false positives (biopsies that turn out to be negative). The appropriate tradeoff between false positives and false negatives is an important factor in the mammography debate.

Cost-effectiveness analyses (e.g., Salzmann, Kerlikowske, and Phillips, 1997) use information about uncertainty, costs, and effectiveness of various outcomes to evaluate alternative options, and these analyses influence recommendations made by organizations such as the National Cancer Society and the National Breast Cancer

[2] This is simplified, of course; women actually have a wider range of options for detecting breast cancer.
[3] Again, this is a simplification because radiologists have other options as well. We use the term "radiologist" rather than "mammographer" because not all radiologists who read mammograms are board certified mammographers.

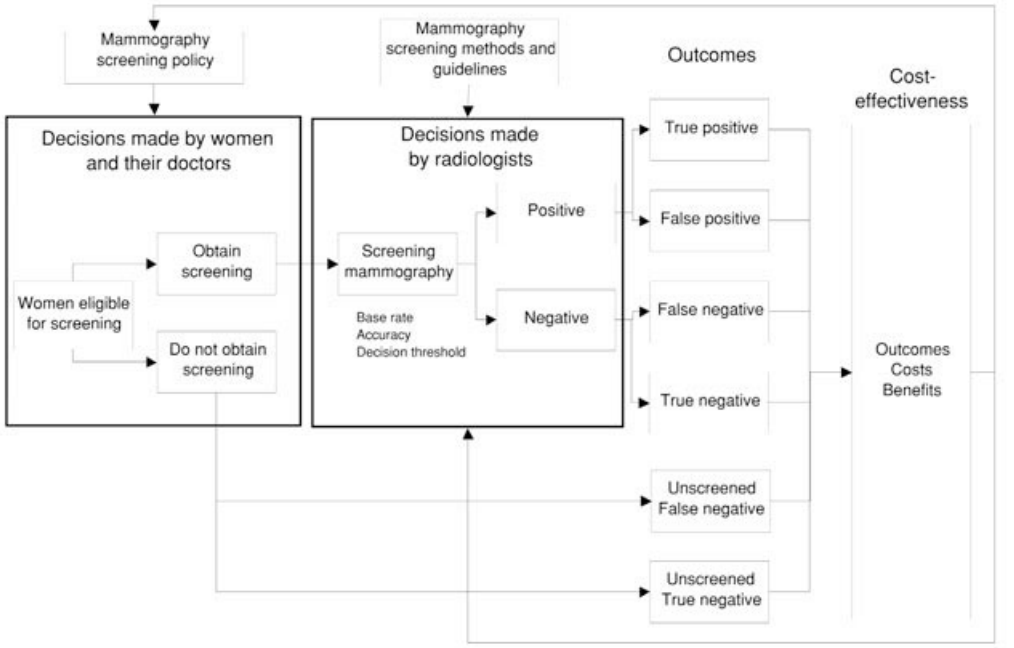


Figure 1. Schematic view of mammography screening.

Coalition and advisory bodies such as the U.S. Preventive Services Task Force. Cost effectiveness analyses, along with other information, influence the doctor's recommendations regarding mammography for their patients, as well as the decisions that radiologists make.

The focus of this paper is the decisions made by radiologists in their practice. These decisions not only occupy a central location in Figure 1, they are central to all issues surrounding screening mammography. While cost-effectiveness analyses and policy debates often assume that the accuracy of radiologists' judgments is fixed, research results strongly suggest that the interpretation of mammograms is less accurate than it could be and that there is wide variation among radiologists regarding the appropriate tradeoff between false positive and false negatives.

Addressing two critical "what if" questions can help inform the crucial decision makers: women, their doctors, and radiologists:

1. What if radiologists selected more or fewer women for biopsy?
2. What if interpretation of mammographic images was more accurate?

A MODEL FOR ADDRESSING DETECTION AND SELECTION PROBLEMS

Selection and detection problems often involve high-consequence decisions under substantial uncertainty. Decisionmakers face two kinds of errors, each with potentially serious consequences: false positives (falsely selecting a case) and false negatives (missing a correct selection). Since uncertainty creates an inevitable tradeoff between these two errors, the decision strategy must be based on values.

Analysis alone cannot solve the problem, but it exposes the nature of the tradeoffs involved and can assure that decisionmakers consider all the possible outcomes. For related discussions see, for example, Hammond, Harvey, and Hastie (1992); Mumpower, Nath, and Stewart (2002); and Swets, Dawes, and Monahan (2000). The critical elements of these problems are base rate, uncertainty, decision threshold, and outcomes.

Base Rate

Among a group of women who have screening mammography, what proportion has malignancies? The base rate depends on age and on the time period considered. The relevant base rate for screening decisions is the probability of a malignancy among undiagnosed women at the time of screening (rather than the lifetime probability of breast cancer). The relevant base rate is for a period of 1 to about 5 years, depending on the screening interval. For any given screening technology, the base rate is critical in shaping the positive predictive value (proportion of true positives to all positives) and the negative predictive value (proportion of true negatives to false negatives) of the screening. All else being equal, low base rates imply low positive predictive values. Even very specific tests yield a large number of false negatives when compared with the small number of true detected cases of breast cancer within the population.

It is difficult to obtain an accurate base rate for a specific population, but we know it is low. Annual incidence of breast cancer is 0.001 to 0.006, depending on age and other factors (Ries et al., 2003).

Uncertainty

Uncertainty is introduced by the inaccuracy of screening mammography. Images vary in quality depending on breast density and other factors, and features that can suggest cancer are often small or difficult to distinguish from normal tissue. Radiologists must therefore make judgments when they evaluate mammograms. Mushlin, Kouides, and Shapiro (1998) conducted a meta-analysis of published studies on the accuracy of screening mammography. They reported true positive rates (correct positive results/total number of patients with cancer), ranging from 0.83 to 0.95; the medical decision-making literature typically refers to the true positive rate as the sensitivity of a test. False positive rates (incorrect positive results/total number of patients without cancer) were reported from 0.009 to 0.065; the medical decision-making literature typically refers to one minus the false positive rate as the specificity of a test. A review by Woolf (2001) found that mammograms miss 0.12 to 0.37 of cancers, depending on the type and stage of cancer, the density of breast tissue, and other factors.

A study by Beam, Layde, and Sullivan (1996) of 108 radiologists from 50 centers found a range of at least 0.53 (0.47 to 1.00) among radiologists in screening sensitivity (the ability to detect cancer when it is in fact present). ROC curve areas (A_z—a measure of accuracy for which a value of 1 indicates perfect accuracy and 0.5 indicates accuracy no better than chance) ranged from 0.74 to 0.95.

Decision Threshold

After examining mammographic images, the radiologist forms an opinion of "suspicion" or "likelihood" of malignancy on some continuum. In other words, the radiologist makes a continuous judgment, but the decision options are discrete (e.g., normal follow-up, return for screening in three months, obtain biopsy). Some way of converting continuous judgments to discrete actions is needed. Generally, it is

assumed that this conversion involves one or more thresholds. If the judgment is higher than some threshold, then biopsy is recommended. Otherwise, no biopsy is recommended. This model is described in detail by Pauker and Kassirer (1980). The threshold may not be determined or even stated explicitly, but it is implicit in the radiologist's decisions.

In effect, the radiologist has to decide what level of suspicion warrants action. The threshold determines the selection rate, that is, the number of mammograms that are selected for biopsy or more frequent screening. A lower threshold means that more mammograms are selected. Furthermore, as is demonstrated below, the threshold influences the relative numbers of false positives and false negatives. Hence, thresholds should reflect the relative costs of these errors.

Different mammographers have different thresholds. Beam, Layde, and Sullivan (1996) found a range of at least 0.45 among mammographers in the proportion of women without breast cancer who were recommended for biopsy. Berg et al. (2000, 2002) found substantial disagreement both about features and about management decisions, and Elmore et al. (2002) found a wide range of false positive rates among mammographers. It is unclear whether this variation indicates differences in values, differences in base rates, differences in accuracy, use of inappropriate thresholds, or a combination of these factors.

Outcomes

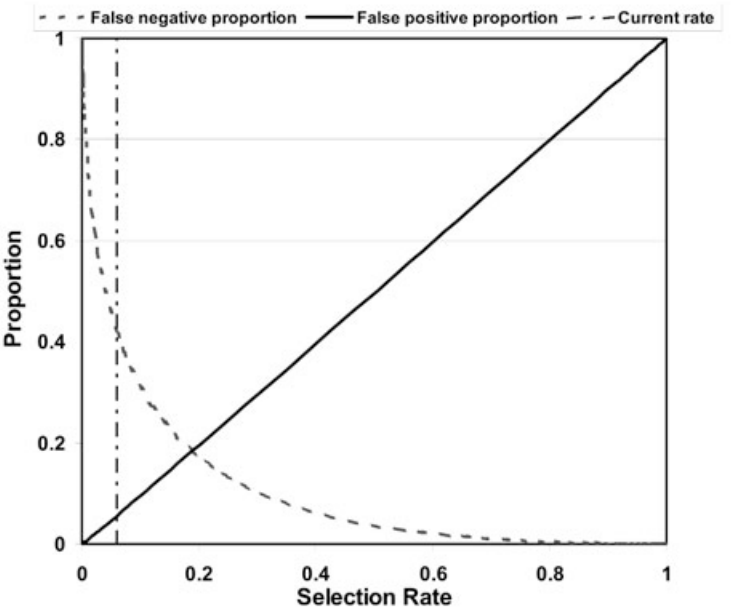
Screening mammography is a detection problem with a low base rate, high uncertainty, and varying decision thresholds determined by physicians. These interact to determine the outcomes. The four possible outcomes of screening mammography were shown in Figure 1.

False positives—defined as women who are referred for biopsies or other additional diagnostic treatments but who are later found not to have tumors—are clearly the most frequent error. After 10 years, 50 percent of women who have annual screening could have a false positive result (Elmore et al., 1998). More than 80 percent of women who have suspicious mammograms have no cancer (National Breast Cancer Coalition, 2003). This leads to many negative biopsies and possibly to over-treatment, because, for example, a positive biopsy indicates the presence of cancer, but does not mean it will spread. False negatives are much less frequent, but potentially more serious, because they may result in an aggressive malignancy going untreated. Both false positives and false negatives depend on the selection rate, but they change in different directions. Figure 2 illustrates the tradeoff between false positives and false negatives.

This is an example based on the data presented in Table 1. Different base rates and different levels of uncertainty will result in different graphs, but the basic relations remain the same.

As the selection rate (the percentage of screened women who are diagnosed as positive) increases, the false negative proportion drops, but the false positive proportion increases. By increasing the selection rate, a certain number of false negatives are eliminated, at a cost of creating a number of false positives.[4]

[4] The present analysis relies on a key assumption that mammograms can be ranked by mammographers in terms of their suspiciousness, so that increasing the selection rate implies that the mammographer will refer an increasing proportion of the "next-most suspicious" mammograms for biopsy or other diagnostic treatment. Decisions about whether, for whom, and how often mammography should be deployed will result in changes in the frequency and relative proportions of the four possible outcome cells.



This is an example based on the data presented in Table 1. Different base rates and different levels of uncertainty will result in different graphs, but the basic relations remain the same.

Figure 2. Proportion of error as a function of selection rate.

Table 1. Results for women under 50.

		Decision		
		Negative	Positive	Total
Biopsy result	Malignant	36	50	86
	Non-malignant	9,360	554	9,914
	Total	9,396	604	10,000

Based on Kolb, Lichy, and Newhouse (2002). Sample size adjusted to 10,000 for ease of interpretation.

The rate at which false positives are substituted for true positives at a given selection rate is called the marginal substitution rate. The optimal selection rate is found when the marginal substitution rate equals the ratio of the benefit of a true positive to the cost of a false positive. The optimal selection rate depends on the benefits of true positives (e.g., the health benefits of early detection and treatment) and true negatives (e.g., the psychological benefits of correct disease-free diagnoses) and the costs of false positives (e.g., costly and painful biopsies and other diagnostic treatments of healthy women) and false negatives (e.g., all the costs associated with failure to treat non-detected tumors).

THE PRACTICE OF MAMMOGRAPHY: DATA FROM ONE MAMMOGRAPHER

During a study that lasted almost 6 years (January 15, 1995 to September 30, 2000), Kolb, Lichy, and Newhouse (2002) recorded data on 27,825 screenings of 11,130 asymptomatic women (women were screened more than once) at New York's Columbia–Presbyterian Medical Center. All were patients of one doctor (Kolb). The patients underwent mammography, physical exam, and ultrasound. Presence of breast cancer was determined by biopsy. Absence was determined by negative results for all screenings. Table 1 is based on 5826 screenings of women under 50.

From Table 1, we can calculate a base rate of 0.0086 and a selection rate of 0.0604. The false positive proportion is 554/9914 = 0.056 and the false negative proportion is 36/86 = 0.419. The vertical line in Figure 2 roughly represents these data. Note that although the false negative proportion is higher than for false positives, there are many more negatives than positives (due to the low base rate), resulting in over 15 false positives for every false negative.

The imputed marginal rate of substitution for this case is 33.2.[5] A small increase in the selection rate would result in about 33 additional false positives for every true positive added. If this ratio is unacceptable, then the selection rate should not be increased and consideration might be given to decreasing it. This doctor is behaving as if 33 were the ideal substitution rate. This would be optimal if the benefit of an additional true positive were 33 times greater than the cost of a false positive. We will not argue whether this benefit-cost ratio is correct or incorrect, because that is not a technical matter. The substitution rate is, however, a meaningful number, and informed social policy requires consideration of the substitution rate in the practice of mammography.

Mammography for women aged 50 or older is more accurate than for younger women (due to reduced breast density), and they have a higher base rate. This can result in a decrease in both false positives and false negatives, as shown in Table 2.

Table 3 compares the results for women under 50 and over 50, clearly showing why it is easier to recommend screening mammography for the over-50 group. Even with fewer women selected for biopsy, both the false positive and false negative proportions are substantially lower for the older study sample.

Table 3 raises a question for this mammographer: Why should the marginal substitution rate be lower for women 50 and older than for women under 50? If the selection rate for women 50 and older were increased to 0.036, the marginal substitution rate would be approximately equal to that for women under 50 (33.2) and the false negative proportion would drop to 0.135, although the false positive proportion would increase to 0.029. This is not to say that different marginal substitution rates for different age groups are necessarily inappropriate. For instance, more aggressive tumors or more life years at risk for younger women might justify more aggressive diagnostic practice for younger women. The point is that the substitution rate is almost certainly the result of intuitive processes. If these results were found to generalize to a larger sample of mammographers, health care consumers should decide, as a personal matter, and society should decide, as a policy matter, what tradeoffs are acceptable.

[5] This calculation is based on signal detection theory and assumes that the distributions of the mammographer's judgments for both malignant and non-malignant cases are normal and both have the same standard deviation. For details of the calculation, contact the authors.

Table 2. Results for women 50 years of age or older.

		Decision		
		Negative	Positive	Total
Biopsy result	Malignant	15	74	89
	Non-malignant	9,704	207	9,911
	Total	9,719	281	10,000

Based on Kolb, Lichy, and Newhouse (2002). Sample size adjusted to 10,000 for ease of interpretation.

Table 3. Comparison of women under 50 and those 50 years or older.

Age	Base rate	A_z	Selection Rate	False Positive Proportion	False Negative Proportion	Imputed Marginal Substitution Rate
<50 years	0.0086	0.90	0.0604	0.0559	0.4186	33.2
≥ 50 years	0.0089	0.98	0.0281	0.0209	0.1685	21.9

"WHAT-IF" QUESTIONS

Table 4 examines two critical what-if questions regarding the practice of mammography:

1. What if the decision threshold/selection rate were changed? This is under the control of the radiologist, yet has important policy implications and implications for women and their families that are rarely discussed.
2. What if the accuracy of screening mammography were increased? Ways of increasing the accuracy of mammography are discussed below.

Rows 1 and 2 address the consequences of lowering the threshold (Question 1). In this case, the number of false positives doubles while the false negatives are reduced by one-third. The substitution rate doubles. Comparing rows 1 and 3 shows what happens when the threshold is increased. False positives decrease but false negatives increase, and the substitution rate decreases. Comparing row 2 with 4 and row 3 with 5 shows the effect of increasing the accuracy of mammography (Question 2). Both false positives and false negatives decrease. Rows 4 and 5 again illustrate the effect of raising the threshold. False positives decrease but false negatives increase. The substitution rate drops substantially.

As Table 4 illustrates, changing the threshold alone only trades one kind of error for another, and the marginal substitution rate quantifies that tradeoff. The analysis presents decisionmakers with a difficult dilemma: What is the appropriate substitution rate? Surveys might be one source of guidance. For example, a survey by Schwartz et al. (2000) found that 63 percent of the women sampled would accept

Table 4. Comparison of mammographer performance under various what-if conditions for women under 50.

Age	Base rate	A_z	Selection Rate	False Positive (out of 10,000)	False Negative (out of 10,000)	Imputed Marginal Substitution Rate
1. Kolb et al. data	0.0086	0.90	0.0604	554	36	33.2
2. Lower threshold	0.0086	0.90	0.1200	1138	24	66.9
3. Higher threshold	0.0086	0.90	0.0500	453	39	27.9
4. Lower threshold and more accurate	0.0086	0.95	0.1200	1125	11	103.9
5. Higher threshold and more accurate	0.0086	0.95	0.0500	437	23	32.6

500 false positives to save one life. (This is not the same as our substitution rate because a true positive does not necessarily result in a life saved.)

Current practice might be another guide. Humphrey et al. (2002) calculated that "over 10 years of biennial screening among 40-year-old women invited to be screened, approximately 400 women would have false-positive results on mammography and 100 women would undergo biopsy or fine-needle aspiration for each death from breast cancer prevented" (p. 356). It is noteworthy that these results are consistent with the preferences expressed in the survey described above. Determination of the substitution rate should be a matter for policy debate.

Explicit recognition and discussion of the tradeoff between false positives and false negatives could give women greater control over the outcomes of mammography. Screening might be tailored to the values and risk preferences of specific patients or at least patient subgroups. If patients in group X are more disturbed by biopsy than patients in group Y, one might want to adopt more aggressive screening policies in group Y.[6]

Increasing the accuracy of mammography reduces both kinds of error. There is no need to consider tradeoffs in choosing between rows 2 and 3 or between rows 1 and 4 above. In decision analytic terms, row 3 "dominates" row 2; that is, row 3 is better in all respects. Similarly, row 4 dominates row 1.

A number of studies have investigated ways of improving the accuracy of mammographers' judgments. For example, Getty et al. (1988) found that by using a simple checklist that directed their attention to 12 important features in a mammogram, general radiologists could match the accuracy of experienced mammographers.

Accuracy may increase with the volume of mammograms read. Esserman et al. (2002) compared results in the United States with other countries, such as the United Kingdom and Sweden, that have high-volume, centralized screening programs. They found that for a fixed level of specificity, sensitivity was higher for U.K. radiologists than for U.S. radiologists, and the sensitivity of U.S. radiologists declined as the volume of mammograms read declined. They concluded that "reader volume is an important determinant of mammogram sensitivity and specificity." In an international study of mammography, Elmore et al. (2003) found that

[6] Thanks to an anonymous reviewer for this insight.

North American radiologists judged 2 to 4 percent more mammograms as abnormal than in other countries, without an increase in the number of cancers detected. A recent U.S.–U.K. comparison (Smith-Bindman et al., 2003) found similar results. Beam, Conant, and Sickles (2003), on the other hand, did not find this effect.

Several studies of computer aids for decision making in mammography have found promising results both for improving accuracy and reducing variability among radiologists (Chan et al., 1999; Floyd, Lo, and Tourassi, 2000; Haque, Mital, and Srinivasan, 2002; Jiang et al., 2001; Roque and Andre, 2002). There is also a strong likelihood that digital mammography will replace film in the future, and some evidence this will result in improved accuracy (Fischer et al., 2002; Nawano, 1995). Further, double reading of mammograms is a generally recognized method for improving accuracy (Beam, Conant, and Sickles, 2003; Karssemeijer et al., 2003; Kopans, 2000; Kwek et al., 2003; Liston and Dall, 2003). Kleit and Ruiz (2003) found that availability of previous mammogram reduced false positives by 50 percent.

Although improving accuracy reduces both kinds of error, there is still a tradeoff. The costs involved in methods for increasing accuracy must be justified by the benefits of reduced errors. Through a more complete analysis of the type illustrated here, it can be shown that methods for increasing the accuracy of mammography, even a little, will produce substantial benefits. Further analysis would be needed to determine whether the benefits justify the costs, but some of the methods described above, such as the use of a checklist, would seem to have low cost.

FURTHER EXTENSIONS

The method of analysis we have described is not new. Signal detection theory has been applied to many detection and selection problems, including mammography. Typically, such studies use the area under the ROC curve (A_z) as a measure of accuracy that is independent of the decision threshold. What we have illustrated, and are proposing, differs from current practice in two ways.

First, performance data that include all four possible outcomes (Tables 1 and 2) can be used to estimate both accuracy and decision threshold. Such data are not routinely collected. Mammography regulations developed by the Food and Drug Administration under the Mammography Quality Standards Act (<http://www.fda.gov/cdrh/mammography/>) require audits based on only positive decisions (the decision to recommend biopsy). Therefore, data on false negatives and true negatives are not routinely available. There are good practical reasons for this. If a woman does not have a biopsy, follow-up is difficult. However, with improvements in medical record keeping, and the advent of electronic records, long-term follow-up becomes more feasible.

Second, while it is generally recognized that decision thresholds vary over time and among individuals, accuracy is often regarded as fixed. We have argued that accuracy can also be changed, and that interventions that address radiologists' judgment are likely to be relatively cheap and effective. It is important to conduct what-if analysis of both accuracy and decision threshold.

Finally, the approach that we have introduced here can be readily extended to address what-if questions related to mammography policy that do not focus narrowly on the practice of radiologists and mammographers. We can illustrate with two obvious extensions; many more examples could be generated:

1. What if the base rate for those who seek screening mammography were changed? The base rate might be changed in various ways, including recom-

mending mammography only for older women, recommending mammography less frequently, and recommending mammography only for women with a family history of breast cancer.

2. What if we could reduce the costs of false positive results or increase the benefits of true positive diagnoses? If we could make biopsy less distasteful or costly, if we could explain the issues better to anxious women, or find other means to reduce the costs of false positive diagnoses, the low predictive value of screening mammography would exact lesser costs from women and their families.[7] Improved treatment protocols that led to better prognoses with fewer noxious and debilitating side effects might likewise change the decision calculus.

The present example, involving screening mammography, illustrates clearly that the facts rarely if ever "speak for themselves" in policy problems that involve detection or selection decisions. Translating scientific or policy analytic studies so that they are made relevant to practice requires application of analytic techniques that are able to pose and answer counter-factual, what-if questions in meaningful ways.

We are grateful to Christine Muller for assistance with the literature review for this paper and to an anonymous reviewer for many helpful suggestions.

THOMAS R. STEWART is Research Professor at the Center for Policy Research, Rockefeller College of Public Affairs and Policy, University at Albany, State University of New York.

JERYL L. MUMPOWER is Interim Provost at the University at Albany, State University of New York. He is also Professor of Public Administration and Policy and Faculty Associate at the Center for Policy Research, Rockefeller College of Public Affairs and Policy, University at Albany.

REFERENCES

Beam, C.A., Conant, E.F., & Sickles, E.A. (2003). Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. Journal of the National Cancer Institute, 95(4), 282–290.

Beam, C.A., Layde, P.M., & Sullivan, D.C. (1996). Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. Archives of Internal Medicine, 156(2), 209–213.

Berg, W.A., Campassi, C., Langenberg, P., & Sexton, M.J. (2000). Breast Imaging Reporting and Data System: Inter- and intraobserver variability in feature analysis and final assessment. American Journal of Roentgenology, 174(6), 1769–1777.

Berg, W.A., D'Orsi, C.J., Jackson, V.P., Bassett, L.W., Beam, C.A., Lewis, R.S., & Crewson, P.E. (2002). Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? Radiology, 224(3), 871–880.

Chan, H.P., Sahiner, B., Helvie, M.A., Petrick, N., Roubidoux, M.A., Wilson, T.E., Adler, D.D., Paramagul, C., Newman, J.S., & Sanjay-Gopal, S. (1999). Improvement of radiologists'

[7] Thanks to an anonymous reviewer for this insight.

characterization of mammographic masses by using computer-aided diagnosis: An ROC study. Radiology, 212(3), 817–827.

de Koning, H.J. (2003). Mammographic screening: Evidence from randomised controlled trials. Annals of Oncology, 14(8), 1185–1189.

Elmore, J.G., Barton, M.B., Moceri, V.M., Polk, S., Arena, P.J., & Fletcher, S.W. (1998). Ten-year risk of false positive screening mammograms and clinical breast examinations. New England Journal of Medicine, 338(16), 1089–1096.

Elmore, J.G., Miglioretti, D.L., Reisch, L.M., Barton, M. B., Kreuter, W., Christiansen, C.L., & Fletcher, S.W. (2002). Screening mammograms by community radiologists: Variability in false-positive rates. Journal of the National Cancer Institute, 94(18), 1373–1380.

Elmore, J.G., Nakano, C.Y., Koepsell, T.D., Desnick, L.M., D'Orsi, C.J., & Ransohoff, D.F. (2003). International variation in screening mammography interpretations in community-based programs. Journal of the National Cancer Institute, 95(18), 1384–1393.

Esserman, L., Cowley, H., Eberle, C., Kirkpatrick, A., Chang, S., Berbaum, K., & Gale, A. (2002). Improving the accuracy of mammography: Volume and outcome relationships. Journal of the National Cancer Institute, 94(5), 369–375.

Fischer, U., Baum, F., Obenauer, S., Luftner-Nagel, S., von Heyden, D., Vosshenrich, R., & Grabbe, E. (2002). Comparative study in patients with microcalcifications: Full-field digital mammography vs screen-film mammography. European Radiology, 12(11), 2679–2683.

Floyd, C.E. Jr., Lo, J.Y., & Tourassi, G.D. (2000). Case-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions. American Journal of Roentgenology, 175(5), 1347–1352.

Getty, D.J., Pickett, R.M., D'Orsi, C.J., & Swets, J.A. (1988). Enhanced interpretation of diagnostic images. Investigations in Radiology, 23, 240–252.

Hammond, K.R., Harvey, L.O., & Hastie, R. (1992). Making better use of scientific knowledge: Separating truth from justice. Psychological Science, 3(2), 80–87.

Haque, S., Mital, D., & Srinivasan, S. (2002). Advances in biomedical informatics for the management of cancer. Annals of the New York Academy of Science, 980, 287–297.

Humphrey, L.L., Helfand, M., Chan, B.K., & Woolf, S.H. (2002). Breast cancer screening: A summary of the evidence for the U.S. Preventive Services Task Force. Annals of Internal Medicine, 137(5 Part 1), 347–360.

Jiang, Y., Nishikawa, R.M., Schmidt, R.A., Toledano, A.Y., & Doi, K. (2001). Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. Radiology, 220(3), 787–794.

Karssemeijer, N., Otten, J.D., Verbeek, A.L., Groenewoud, J.H., de Koning, H.J., Hendriks, J.H., & Holland, R. (2003). Computer-aided detection versus independent double reading of masses on mammograms. Radiology, 227(1), 192–200.

Kleit, A.N., & Ruiz, J.F. (2003). False positive mammograms and detection controlled estimation. Health Services Research, 38(4), 1207–1228.

Kolb, T.M., Lichy, J., & Newhouse, J.H. (2002). Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: An analysis of 27,825 patient evaluations. Radiology, 225(1), 165–175.

Kopans, D.B. (2000). Double reading. Radiology Clinics of North America, 38(4), 719–724.

Kwek, B.H., Lau, T.N., Ng, F.C., & Gao, F. (2003). Non-consensual double reading in the Singapore Breast Screening Project: Benefits and limitations. Annals of the Academy of Medicine, Singapore, 32(4), 438–441.

Liston, J.C., & Dall, B.J. (2003). Can the NHS Breast Screening Programme afford not to double read screening mammograms? Clinical Radiology, 58(6), 474–477.

Mumpower, J.L., Nath, R., and Stewart, T.R. (2002). Affirmative action, duality of error, and

the consequences of mispredicting the academic performance of African-American college applicants. Journal of Policy Analysis and Management, 21(1), 63–77.

Mushlin, A.I., Kouides, R.W., & Shapiro, D.E. (1998). Estimating the accuracy of screening mammography: A meta-analysis. American Journal of Preventive Medicine, 14(2), 143–153.

National Breast Cancer Coalition (2003). Position statement on screening mammography, March 2003. Available at: <http://www.natlbcc.org/bin/index.asp?strid=560&depid=9&btnid=1>; accessed June 7, 2004.

Nawano, S. (1995). Evaluation of digital mammography in diagnosis of breast cancer. Journal of Digital Imaging, 8(1 Suppl 1), 67–69.

Olsen, O., & Gotzsche, P.C. (2001). Cochrane review on screening for breast cancer with mammography. Lancet, 358(9290), 1340–1342.

Pauker, S.G., & Kassirer, J.P. (1980). The threshold approach to clinical decision making. New England Journal of Medicine, 302, 1109-1117.

Ries, L.A.G., Eisner, M.P., Kosary, C.L., Hankey, B.F., Miller, B.A., Clegg, L., Mariotto, A., Fay, M.P., Feuer, E.J., & Edwards, B.K. (Eds.). (2003). SEER cancer statistics review, 1975–2000, National Cancer Institute. Bethesda, MD. Available at: <http://seer.cancer.gov/csr/1975_2000>; accessed June 7, 2004.

Rimer, B.K., Halabi, S., Sugg Skinner, C., Lipkus, I.M., Strigo, T.S., Kaplan, E.B., & Samsa, G.P. (2002). Effects of a mammography decision-making intervention at 12 and 24 months. American Journal of Preventive Medicine, 22(4), 247–257.

Roque, A.C., & Andre, T.C. (2002). Mammography and computerized decision systems: A review. Annals of the New York Academy of Science, 980, 83–94.

Salzmann, P., Kerlikowske, K., & Phillips, K. (1997). Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age. Annals of Internal Medicine, 127(11), 955–965.

Schwartz, L.M., Woloshin, S., Black, W.C., & Welch, H.G. (1997). The role of numeracy in understanding the benefit of screening mammography. Annals of Internal Medicine, 127, 966–972.

Schwartz, L.M., Woloshin, S., Sox, H.C., Fischhoff, B., & Welch, H.G. (2000). US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: Cross sectional survey. British Medical Journal, 320(7250), 1635–1640.

Smith-Bindman, R., Chu, P.W., Miglioretti, D.L., Sickles, E.A., Blanks, R., Ballard-Barbash, R., Bobo, J.K., Lee, N.C., Wallis, M.G., Patnick, J., & Kerlikowske, K. (2003). Comparison of screening mammography in the United States and the United Kingdom. Journal of the American Medical Association, 290(16), 2129–2137.

Swets, J.A., Dawes, R.M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. Psychological Science in the Public Interest, 1(1), 1–26.

U.S. Preventive Services Task Force (2002). Guide to clinical preventive services, 3rd edition: Periodic updates: Screening for breast cancer. Available at: <http://www.ahcpr.gov/clinic/uspstf/uspsbrca.htm>; accessed June 7, 2004.

Woolf, S.H. (2001). The accuracy and effectiveness of routine population screening with mammography, prostate-specific antigen, and prenatal ultrasound: A review of published scientific evidence. International Journal of Technology Assessment Health Care, 17(3), 275–304.