Disease Transmission Models for Public Health Decision Making: Toward an Approach for Designing Intervention Strategies for *Schistosomiasis japonica*

Robert C. Spear, Alan Hubbard, Song Liang, and Edmund Seto

Center for Occupational and Environmental Health, School of Public Health, University of California, Berkeley, California, USA

Mathematical models of disease transmission processes can serve as platforms for integration of diverse data, including site-specific information, for the purpose of designing strategies for minimizing transmission. A model describing the transmission of schistosomiasis is adapted to incorporate field data typically developed in disease control efforts in the mountainous regions of Sichuan Province in China, with the object of exploring the feasibility of model-based control strategies. The model is studied using computer simulation methods. Mechanistically based models of this sort typically have a large number of parameters that pose challenges in reducing parametric uncertainty to levels that will produce predictions sufficiently precise to discriminate among competing control options. We describe here an approach to parameter estimation that uses a recently developed statistical procedure called Bayesian melding to sequentially reduce parametric uncertainty as field data are accumulated over several seasons. Preliminary results of applying the approach to a historical data set in southwestern Sichuan are promising. Moreover, technologic advances using the global positioning system, remote sensing, and geographic information systems promise cost-effective improvements in the nature and quality of field data. This, in turn, suggests that the utility of the modeling approach will increase over time. ***Key words:*** **disease transmission, mathematical models, parameter estimation, schistosomiasis.** ***Environ Health Perspect*** **110:907–915 (2002). [Online 12 August 2002]**
http://ehpnet1.niehs.nih.gov/docs/2002/110p907-915spear/abstract.html

In a companion article, Eisenberg et al. (2002) present an approach to the analysis of infectious disease transmission for waterborne pathogens using dynamic models studied via computer simulation techniques. Here we present an application of this approach to designing local control strategies for the parasitic disease schistosomiasis. The schistosomiasis transmission cycle involves mammals and freshwater snail species linked through contact with different forms of the parasite in surface waters. Our work focuses on agricultural villages in the southwestern part of Sichuan Province in China, where schistosomiasis is endemic. The challenge is to determine whether a dynamic modeling approach can be a useful tool in specifying effective intervention strategies. We propose to use the model to integrate general knowledge of the factors controlling transmission of the disease, quantitative data specific to the transmission of schistosomiasis in China, and site-specific data of the sort typically available in these settings.

This report is of work in progress in that our activities to date have been concerned with model formulation and its parameterization, particularly in light of the kind of field data commonly generated in rural China. We have not yet designed and implemented an intervention program. However, much of our work has been devoted to analysis of data from a study that culminated in a successful intervention program carried out by our colleagues at the Sichuan Institute of Parasitic Disease over 1987–1995. Regrettably, that intervention was not sustainable because of recurrent annual costs of drug treatment. This underscores that the search is for an intervention strategy that is not only effective but also sustainable in a local context.

It is important to point out at the outset that we are not designing intervention trials in a traditional epidemiologic context. Our objective is not to determine whether a particular intervention is effective when all other factors are controlled. For schistosomiasis, there is a considerable body of knowledge about the array of methods of controlling transmission that have been employed in different settings. The task is to determine which blend of the subset of feasible interventions should be used in a particular setting and to predict its probable effectiveness in diminishing disease transmission. To accomplish this task, we require a well-informed computer model of schistosomiasis transmission that can be calibrated to local conditions. Eventually, we hope to use the model as a tool for routinely designing the management strategies for the many sites where the disease is endemic.

The Disease

Schistosomiasis is a waterborne parasitic disease that affects 200 million people and poses a threat to 600 million in more than 76 countries (WHO 1993). The disease is caused by infection by parasitic worms of the genus *Schistosoma*. These parasites are transmitted via contact with contaminated water. The life cycle of the schistosome begins with the sexual pairing of adult worms in the blood vessels of the host and the production of copious numbers of eggs, a fraction of which are excreted in feces (or urine in the case of *S. haematobium*). The eggs hatch in water and release a free-swimming miracidium, whose objective in life is to find and penetrate an appropriate snail in which to develop. After a period of asexual reproduction, tailed, free-swimming larvae called cercaria leave the snail and are transported in water, where they actively seek an appropriate vertebrate host. Cercaria penetrate the intact skin of the host, thus infecting it. The parasites subsequently mature into adult worms in the host, where they mate to complete the cycle.

Four to six weeks after schistosome penetration and once worms have migrated to and settled in the mesenteric veins of the vertebrate host, mated adult worms begin to produce eggs. On rare occasions, infected people will experience a severe condition at this time, called Katayama fever, in the Asian form of the disease. The worms themselves cause little or no damage to the body. They are generally undetected by the body's immune system because of the ability of the worm's tegument to attach host proteins to itself as a kind of camouflage. In the long term, it is the eggs that are the real culprits of clinical disease. Eggs are carried off in circulation and are sieved by small blood vessels, especially in the liver and spleen, where the body's immune system attacks them and covers them with fibrotic tissues that accumulate into granulomas.

Long-term infections can lead to development of severe lesions that block blood flow. The resulting increase in blood pressure can in turn direct eggs out of the

Address correspondence to R. Spear, Center for Occupational and Environmental Health, School of Public Health, University of California, 140 Warren Hall #7360, Berkeley, CA 94720-7360 USA. Telephone: (510) 642-0761. Fax: (510) 642-5815. E-mail: spear@uclink4.berkeley.edu

We acknowledge our long-standing collaboration with the Schistosomiasis Department, Sichuan Institute of Parasitic Disease, Chengdu, under the leadership of D. Qiu and X. Gu, as well as with the Schistosomiasis Control Unit of Xichang County.

Financial support for this work was provided in part by the Endemic Disease Office of the Provincial Government of Sichuan and the National Institute of Allergy and Infectious Disease, 1RO1-AI43962.

Received 26 October 2001; accepted 5 March 2002.

abdominal area into other parts of the body, including the lungs and brain. The tissue damage and lesion development caused in these areas can be fatal in severe cases. Symptoms of chronic infection may include general malaise; abdominal pain; headache; enlargement of the liver, spleen, and lymph nodes; and presence of blood, pus, and mucus in the stool. Cirrhosis may develop as lesions accumulate in the liver.

In the continuing absence of a vaccine for schistosomiasis, it is necessary to rely on various environmental and behavioral interventions to diminish risk of infection. Virtually all have been tried in one setting or another. Because water contact is the route of exposure of the vertebrate host, it is possible to identify particularly hazardous aquatic environments and attempt to control access to them by both humans and animals. Alternatively, one may either protect the snails from infection by humans and animals or destroy the snails or their habitat. In China, an ancient and pervasive practice involves mixing human excrement, termed nightsoil, with that of animals, which is then used for crop fertilization. Where animal involvement is marginal, this practice is central to maintaining the infection in the snail population. Hence, the strategy of enhancing sanitation facilities and conditions employed in other regions of the world to date has not been a viable strategy in China. More commonly, the large-scale use of chemotherapy for humans and animals has been used and has the beneficial effect of controlling morbidity while interrupting the egg burden shed into the environment. Although various combinations of these control strategies have been used quite successfully to reduce the incidence of disease in China, as recently as 1995 approximately 865,000 people and 100,000 water buffalo were infected (Chen 1999). The endemic areas in 2000 are shown in Figure 1. Control has been particularly difficult in certain regions, including the mountainous areas of Sichuan, where our work has been focused.

Schistosomiasis is a disease whose distribution is particularly sensitive to environmental change, most clearly environmental change of human origin. Two events loom that promise major environmental changes: the completion of the Three Gorges Dam on the Yangtze River in China and the increasing probability of global warming. The changes that will be caused by these events promise to have a substantial impact on the distribution and extent of *S. japonicum* in China. Hotez et al. (1997) have written on the impact of the dam, speculating that the effect will generally be to increase both *Oncomelania* snail habitat and human disease transmission. More recently, our colleagues in the Sichuan Institute of Parasitic Disease have completed a 3-year project for the Chinese Ministry of Health, in which they reach a similar conclusion for the environment upstream of the dam, although they forecast that the time frame for fully establishing the disease may extend 50–70 years into the future (SIPD 1998).

Modeling Schistosomiasis Transmission

Mathematical modeling of schistosomiasis transmission began with the work of MacDonald (1965). Considerable further work has occurred since that time (Anderson and May 1991; Woolhouse 1991). Essentially, there have been two approaches: models based on disease prevalence and models based on parasite burden. As discussed more generally by Eisenberg et al. (2002), prevalence models track the number of infected, infectious, or susceptible individuals, whereas models based on parasite burden track the intensity of infection, most commonly mean parasite burden in a population. Because, in the case of schistosomiasis, clinical disease is linked to the duration and intensity of infection, parasite burden models may be preferred (Anderson and May 1991). This is clearly the case for morbidity control programs. Hence, the structure of the model being used in our work follows that of Anderson and May (1985), which tracks mean worm burden in the human population and the mean density of infected snails in the environment.

In our variant of the model, we also track the uninfected snail density as well as divide the human population into risk groups generally defined by occupation and residence. In particular, we employ a connected set of models for each of these risk groups, each of identical structure. To date, we have focused on occupational groups comprising farmers, domestics, students, and others (e.g., teachers and administrators) living in a natural village. Residence is indexed by natural village because residential areas are generally included within the boundaries of the land farmed by the villagers who live there. The rationale for dividing the population by residence and occupational group was originally suggested by an analysis of the prevalence data from villages in Sichuan in which the dominant classification variable was residence, with occupation second, followed by other small subgroups defined by task (Maszle 1998). The residence–occupation classification is also attractive because it defines convenient groups around which intervention can be structured.

The state equations of the model describe the worm burden, w_{ik}, of the ith occupational group living in village k; the average density of infected snails, z_k; and the density of uninfected snails, x_k, in that village. Each of these equations is of similar structure in that the rate of change of each variable with time depends on the difference between the rate at which worms, for example, develop *in vivo* and the rate at which resident worms die. Similarly, uninfected snails reproduce, die, or become infected. The mortality rate of infected snails is higher than that of uninfected snails. However, because the fraction of the snail

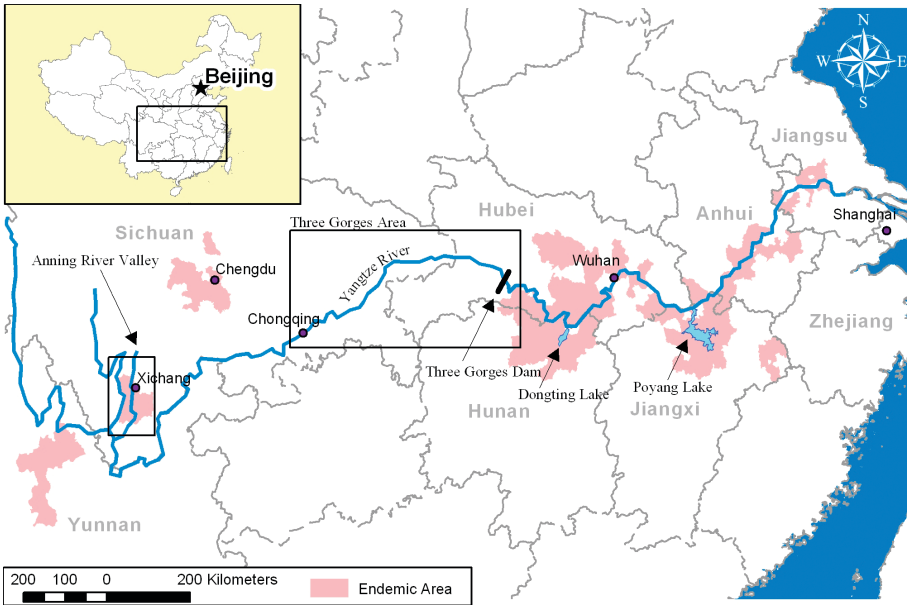

Figure 1. The schistosomiasis endemic areas in China in 2000. The Chuanxing villages are within the Anning River Valley.

population that is infected seldom exceeds 1% of the total population, the rate of decrease of the uninfected snail population is essentially all due to natural mortality.

Although the death rates of worms, infected snails, and uninfected snails are all modeled simply as first-order processes, the processes and rates at which worms develop in humans, the infection process in snails, and the reproduction of uninfected snails in the environment are all complex. That is because these processes depend on environmental variables such as temperature and rainfall, agriculture-related variables such as irrigation water area and fertilizer use, and variables related to the development and maturation of the parasite in snails and in people. The model and its parameters are described in detail elsewhere (Liang et al. 2002); we summarize its structure and parameters here to illustrate how the model serves as a platform for the integration of the quite diverse data bearing on the intensity of disease transmission and the opportunities for its disruption at a local level (Liang et al. 2002).

Figure 2 shows the model and its structural relationships, together with local data inputs. The outputs of the thick-ruled boxes are the state variables w_{ik}, z_k, and x_k, where $w_{ik}(t)$ = mean worm burden in the ith group in environment k; $x_k(t)$ = mean density of uninfected snails (snails/m^2 of habitat); and $z_k(t)$ = mean density of infected snails (snails/m^2 of habitat).

As mentioned above, the common features of the equations are the death processes, denoted by μ, with corresponding subscripts for worms *in vivo*, infected snails, and uninfected or susceptible snails. Also, there are three temperature-dependent developmental delays, from human infection to the maturation of the worm *in vivo*, τ_w; infection of the snail to the time when cercaria are excreted into the environment, τ_z; and the time from snail hatch to adulthood, that is, to infectable status, for the snail, τ_s. The exponential terms depending on the $\mu\tau$ products in each of the equations are the fraction of the developing population that dies before development is complete. Table 1 provides a summary of the model parameters and environmental variables, together with their units, except four parameters associated with snail and sporocyst development, discussed below, which do not appear explicitly in Figure 2.

The cercarial concentration is assumed to be directly proportional to the density of infected snails adjusted by C_{net}, the net import or export from or to adjacent villages. Both $C_k(t)$ and $z_k(t)$ are spatial averages over k. Because the lifetime of both cercaria and miracidia is less than a day, and the time step of the model for simulation purposes is 1 day, these relationships are algebraic. Because the adult snails live above the waterline, it is clear that rainfall is an important mechanism for flushing cercaria into ditches as well as transporting eggs from the field, where they are distributed in nightsoil, into the ditch environment, where the snails live. Hence, $r_{ck}(p)$ is the fraction of the daily production of cercaria that reach the surface water and is a function of the daily precipitation, $p(t)$. The units used for infective cercaria and miracidia are based on the premise that both inhabit surface or near-surface water.

The miracidial concentration is given by $M_k(t) = r_{ek}(p,\beta_k)E_k(t)$. As with the cercaria, $r_{ek}(p,\beta_k)$ represents the precipitation-dependent fraction of the total daily egg production, $E_k(t)$, which enters the aquatic environment of the kth village and hatch into miracidia in surface waters. As with cercaria, $E_k(t)$ includes a transport term, E_{net}, to or from adjacent villages. Egg burden into the environment also depends on the fraction of the total nightsoil production used for fertilization, β_k, which varies seasonally and by crop demand. Egg excretion by humans, which drives miracidial production in the absence of significant infected animal populations, depends on the worm burden in all occupational groups living in village k; hence the stacked boxes at the top of Figure 2. Egg excretion is estimated from

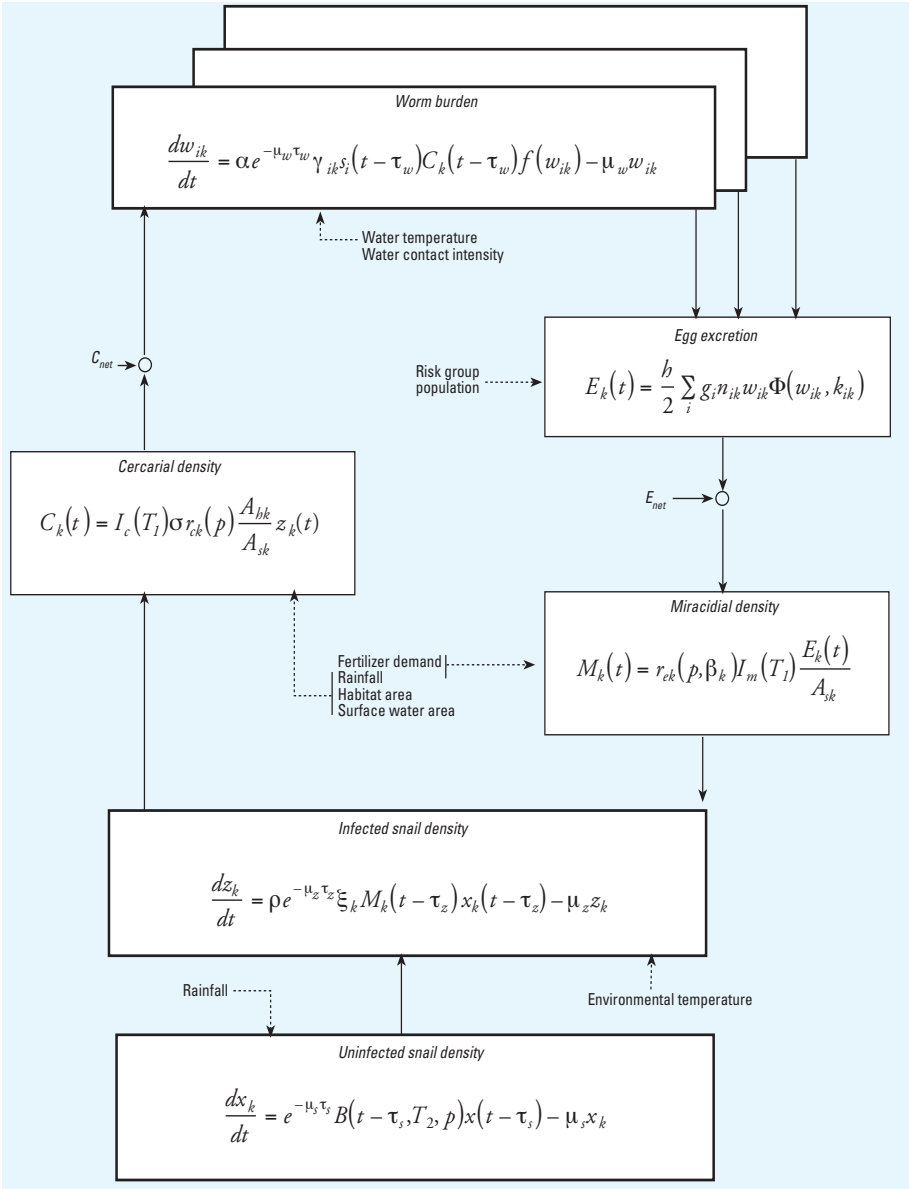

Figure 2. The structure of the transmission model describing group specific worm burden w_{ik} in humans; their egg output, E_k; resulting miracidial density in the irrigation system, M_k; the density of uninfected snails, x_k; infected snails, z_k; and the resulting cercarial density in the irrigation system, C_k. The dashed lines represent local data inputs to the model.

data from two separate tests, the Kato-Katz test, which involves microscopic examination of fecal smears and results in egg counts, and a miracidial hatch test that is sometimes used in China to detect infection. These data are fitted to a statistical model whose parameters k_{ik}, r, and h are estimated from local data and embedded in the mathematical model of Figure 2 (De Vlas et al. 1992)

Temperature enters the model in several additional ways. First, the infectivity of cercaria is known to be temperature dependent and is reflected in the model through the unit-free infectivity function $I_c(T_1)$, where T_1 is water temperature. A similar phenomenon exists for miracidia, which is similarly represented by $I_m(T_1)$. Hence, C_k and M_k are effective concentrations after adjusting for temperature-dependent infectivity. Also, the time delay between infection of the snail by a miracidia and the development of the sporocyst to a point where cercaria begin to be excreted is temperature dependent and represented by the delay time, τ_z. Maszle (1998) developed a degree–day model that specifies τ_z from the local temperature time series, a formulation that we continue to employ. However, the degree–day models for both infected and uninfected *Oncomelania* snails depend on temperature not of the water but of the microenvironment along the ditches above the waterline. The function $B(t-\tau_s, T_2, p)$ is the effective per capita reproduction rate of uninfected snails, which also depends on the microenvironmental temperature T_2 and rainfall, p.

At this point, it is clear that the model is structured to integrate very diverse information both from the field and from laboratory investigations regarding factors influencing the life cycle of snails and the biology of the schistosome. The challenge is to move from structural issues to quantitative forecasts of infection rates in humans and in snails. This requires moving from functional relationships to numbers.

Model Parameters

At this point in our exposition, experienced modelers will be questioning the level of detail of the model presented above and the data that exist to yield realistic parameter estimates. Unquestionably, the success of our approach rests on narrowing the uncertainty in important parameters to a degree that will result in sufficiently narrow ranges of uncertainty in the predicted outputs; the uncertainty in the outputs determines the resolution with which one can compare candidate intervention strategies. Hence, the parameter estimation issue is central. As discussed below, our formal approach to parameter estimation results in explicit information on residual parametric uncertainty at each point of the analysis. As a prelude, however, in Table 1 the parameters of the model are divided into three groups, biologic, measurable, and spatial, based on the nature and extent of data available for their estimation.

Biologic parameters. Biologic parameters are those that might be expected to be relatively invariant across regions of similar ecology and single snail species, for example, the mountainous endemic areas of Sichuan. Among these parameters are the natural death rate of infected snails, μ_z, and parameters that depend principally on human or parasite physiology, for example, the death rate of worms in the body, μ_w. The proportionality constants α and ρ, the egg model parameters r and h, and the developmental delay τ_w, the time between cercarial penetration to the mature worm, are all biologic parameters. The data available for estimating/constraining these biologic parameters consist of published experimental data (e.g., h, r, μ_z, and μ_w) as updated by infection data collected in epidemiologic surveys of the villages. Thus, one of the most important challenges of our approach is to use data integration techniques that give proper weight to each source of data.

Measurable parameters. The measurable parameters are those that can be at least approximated from site-specific data. Clearly, the areas of habitat and surface water are estimable from field surveys (Seto et al. 2001). A more complicated example is the parameter $s_i(t)$, which reflects the intensity of water contact of an average person in occupation i. This quantity varies with season and is intuitively quite important. The parameter $s_i(t)$ can be estimated from monthly time–activity questionnaire data or using the more sophisticated methods of Ross et al. (1998). We assume $s_i(t)$ estimates to be valid across a region for villages engaged in similar agriculture.

Spatial parameters. The spatial parameters are those that can currently be estimated only from site-specific longitudinal data that allow the model to be fit to initial and final values of state variables via simulation studies. That is, presently these parameters cannot be estimated independently of the model. These parameters are γ_{ik} and ξ_k. The notion is that they will both equal unity if cercaria and miracidia are uniformly distributed in the surface water system, but both can be larger or smaller than unity. For example, if cercaria and human water contact patterns have different spatial distributions over the water system, one might expect γ_{ik} to be small, reflecting that the worm burden is less than one might expect from the number of infected snails and the

Table 1. Model parameters, variables, and inputs.

Parameters	Interpretation and units	Values
Biologic		
τ_w	Worm development delay, infection to maturity in humans (days)	20–30
μ_w	Worm mortality rate (per capita per day)	0.000456–0.0014
h	Eggs excreted per worm pair per gram stool	0.728–2.56
r	Aggregation of eggs in stool sample at constant worm burden	0.1–10
γ_w	Density dependence of worm establishment *in vivo*	0.001
τ_z	Sporocyst development delay to cercarial release (days)	60–83
μ_z	Patent and latent snail death rate (snails per snail per day)	0.0042–0.0074
σ	Cercarial production per sporocyst per day	3–26
τ_s	Snail development delay, egg hatch to infectable age (days)	71–115
μ_s	Snail mortality rate (per capita per day)	0.0023–0.01
B_m	Maximum snail reproduction rate (eggs per snail per day)	0.2–1.1
α	Schistosome acquired per cercaria per square meter contact	0.0001–0.5
ρ	Probability of snail infection per miracidium per square meter surface water	0.000001–0.0005
Measurable		
$w_{ik}(0)$	Initial worm burden in the ikth group	Local data
$x_k(0)$	Initial snail density	Local data
$z_k(0)$	Initial infected snail density	Local data
s_i	Average water exposure for the ith group [exposure area (m^2/contact/day)]	Local data
k_{ik}	Worm aggregation parameter	Local data
A_{hk}	Snail habitat area (m^2)	Local data
A_{sk}	Surface water area (m^2)	Local data
β_k	Fraction of nightsoil used for fertilization	Local data
Spatial		
γ_{ik}	Spatial index for the distribution and interaction between human exposure and cercaria	1 (default value)
ξ_k	Spatial index for the distribution and interaction between snails and miracidia	1 (default value)
Inputs		
p	Rainfall (mm/day)	Local data
T_1	Water temperature (°C)	Local data
T_2	Snail microenvironment temperature (°C)	Local data

intensity of water contact. If γ_{ik} is large—that is, if human water contact occurs where infected snails are concentrated—it suggests the converse, together with the implication that an environmental intervention, either focal application of molluscicide or ditch alterations, might have a significant impact on transmission intensity. To specify such an intervention would require spatial data that would allow targeting snail clusters proximate to water contact locations associated with the ith risk group, perhaps a popular clothes-washing site.

As noted above, the challenge is to estimate the various parameters of the model to a degree of precision that will allow discrimination of the effects of various simulated interventions in the presence of residual uncertainty. As outlined by Eisenberg et al. (2002), we begin by assigning to each parameter of the model a distribution function reflecting current uncertainty of its value at each stage of the study. The essence of our strategy is to refine these estimates from new information. There are two dimensions of refining our knowledge, the acquisition and integration of site-specific data and the statistical methods of updating the parameter estimates. Computer simulations are central to both aspects.

Local Data Sources

Our efforts to date have focused on adapting the model to incorporate the nature and extent of site-specific data. In this context, much of our work is based in the Anning River Valley of southwestern Sichuan. Villages were selected as being typical of the environment of about 90% of the population in the Daliang mountainous region. The living and working styles of people in a residential group are usually very similar, and the fields that they farm are usually adjacent to their housing areas. In general, the agriculture typical of the river valley plains does not rely heavily on animal husbandry; hence, the animal populations are relatively small in comparison with the high mountain valley regions, also found in the Daliang region.

Within the Chuanxing township, which is typical of the area, the maximum elevation is 2,010 m in the north, dropping to 1,530 m in the south. The climate is subtropical, with an annual average temperature of 17°C and annual rainfall of about 1,000 mm, over 90% of which falls between the beginning of June and the end of October. The main agricultural products are rice, wheat, garlic, eggplant, and tomatoes, although more diversified vegetable crops and flowers are increasingly common. A complex irrigation system was substantially expanded in the late 1970s. Rainfall and mountain runoff feed the irrigation system in the wet season, and during the dry season, water can be pumped from Qionghai Lake, several kilometers to the south. Since the expansion of the irrigation system, the prevalence of schistosomiasis has increased in the area. In Minhe village, for example, the infection rate was 32% in 1977, 38% in 1978, 39% in 1980, 49% in 1984, and 57% in 1987. An important factor to sustaining the disease cycle in this area is that fertilization practices make extensive use of nightsoil that is moved from residential pit latrines to field storage pits without treatment and with minimal holding times. Snail habitat is principally on the margins of irrigation ditches because they offer year-round moisture and a relatively stable habitat, unlike the farmed areas themselves. A typical ditch network mapped using the Global Positioning System (GPS) is available online (Seto and Liang 2002) or in Seto et al. (2001).

In those regions of China in which schistosomiasis is endemic, there are units organized within county health departments whose focus is on surveillance and control of the disease. They are supported by sections of the provincial health departments and by both research- and surveillance-based activities at the national level that are, in turn, in touch with relevant units of the World Health Organization. This has standardized methods and protocols to differing degrees at the provincial, national, and international levels. In Sichuan, field data that can be collected, given adequate resources, include the following.

Human prevalence. Prevalence surveys may begin with an immunologic screen that, if positive, is followed by examinations of fecal samples. As mentioned above, this might involve a miracidial hatch test and/or Kato-Katz quantitative egg counts for those with positive hatch test. Infection histories are often available for each individual based on these data. If animal populations are significantly involved in transmission, similar techniques are used to determine their infection status.

Snail survey. In our work, this now begins with ditch-mapping procedures using the GPS, as noted above, an aspect of which allows sampling of sites and estimation of snail density and the extent of habitat by location on the ditch network. The fraction of the snail sample that is infected is typically very small; hence, knowledge of the spatial distribution of infected snails is generally very poor. This motivates the cercarial bioassay.

Cercarial bioassay. Bioassay data are obtained at various locations on the ditch networks at particular times, generally at the height of the mid-summer infection cycle. The procedure involves exposure of one cage of five mice per location to the surface water in the ditch for a total exposure period of 10 hr. Because the exposures are integrated over multiple days, good information is gained on the relative hazard of each location if the timing of the assays is reasonably coincident. This provides a crude measure of the spatial variation in cercarial concentration.

Water contact survey. This collects one-point-in-time questionnaire data, generally early in any site-specific investigation, and it may be the least commonly collected of the items on this list. It allows estimates of the relative intensity of water contact by season and occupation and informs the parameter $s_i(t)$, as discussed above. In Sichuan, stratified random sampling procedures are often used with about 25% population samples.

Agricultural and environmental data. It is becoming increasing clear that the intensity of transmission is closely connected to the types of crops grown and associated fertilizer use demands. Information on these issues is available both from records of the administrative village and from local interviews. Also, air temperature and rainfall data are regionally available. There is also considerable variability in the animal population in these villages as well as in the potential for their participation in the transmission cycle. However, it is clearly much less of an issue in this area than it is on the lower Yangtze River, where water buffalo are centrally involved (Chen and Zheng 1999).

Simulation Strategy

For any risk group, most of the parameters of the model are biologic parameters—parameters that are likely to be relatively invariant over large regions and, ideally, over time. This is desirable because estimates of the range of values for these parameters can be extrapolated from one site to another, as well as narrowed with experience over time. However, for any village, there are three to five occupational classes, all of which are likely to have some level of infection within them, which means that they must be accounted for in calculating the egg burden to the village environment. So even though there is a modest number of occupational or mixed occupational and site-specific parameters such as $s_i(t)$, k_{ik}, and γ_{ik}, they proliferate quickly as the number of groups increases. Whether or not this is a problem depends on the relationship of the number of parameters and the amount of data available to estimate them. For example, adding a new village and its data to the analysis is much less problematic than is subdividing the data from a particular village into a larger number of risk groups.

The issue of grouping or aggregation of the population becomes particularly important when exploring the interaction of one village with the next, either in terms of human infection or that of snails. Zhou et al. (1997) showed that there is significant spatial correlation in the Chuanxing data set for infection-related variables between adjacent villages, but not of such variables as uninfected snail density or human water contact frequency. Hence, we expect that the final definition of risk groups will be based on both infection data and spatial features, the latter related to the clear importance of surface water transport of cercaria and miracidia. In that regard, a separate aspect of our work relates to the use of satellite imagery for the identification of snail habitat as well as other remotely observable landscape data. For example, we have recently acquired high-resolution satellite images (IKONOS) of 20 villages near Qionghai Lake near Xichang to match with all of these other data items in the foregoing list in an attempt to understand the importance of landscape-related risk factors.

Our current modeling work is aimed at using historical data from Chuanxing township to refine the estimates of the biologic parameters available from the literature. Maszle (1998) conducted an extensive review of the relevant literature for this purpose, which forms the first generation of parameter estimates. The second generation will result from the analysis of two cross-sectional surveys in Chuanxing, the first in 1987 and the second in 1989. In 1987, a complete prevalence survey was conducted in all 12 villages, together with snail and time–activity surveys. This analysis collapsed the 12 villages and five occupations into seven relatively homogeneous groups. This means of forming the initial risk groups was based on the premise that, in 1987, the population was in a state of dynamic equilibrium with the environment insofar as there had been no widespread chemotherapy in humans before that time. The first-generation parameter estimates and the 1987 data comprise the initial conditions for the simulation study.

After the 1987 prevalence surveys, all infected individuals were treated with praziquantel, the drug of choice for treatment of schistosomiasis. This compound kills the adult worms *in vivo* in a single course of treatment with about 95% effectiveness. There was also some use of molluscicide before the 1988 infection season, but subsequent snail surveys showed this to have been ineffective. The second prevalence survey, carried out in 1989, provides a target outcome for model calibration, as detailed below.

The form of the simulation studies that we are using for this analysis is similar to the form we have used in previous studies (Eisenberg et al. 1996; Grieb et al. 1999; Spear and Hornberger 1980; Spear et al. 1991). In short, given the model and the first generation of parameter distributions, it is possible to run sets of Monte Carlo simulations, each resulting in a predicted outcome at the time of the 1989 cross-sectional survey. The missing element is the specification of the criteria for assessing which, if any, of these outcomes matches the actual outcome observed in the field. This is not a simple matter. Although we have reasonable data on disease prevalence and worm burden based on the egg excretion data, we have very limited knowledge of the infected snail density. However, there are some cercarial bioassay data that might allow a crude rank ordering of infected snail density by village. Because of the considerable uncertainty in assessing goodness of fit in problems of this sort, in the past we have used binary criteria for assessing the degree to which the model captures the principal characteristics of the field data. The criteria are generally specified as a number of conditions on the output, which in this case include disease prevalence, a worm burden that lies within an acceptable range of values about the 1989 field estimates, and an upper bound on the density of infected snails based on the density of uninfected snails from surveys in each village. Hence, the infection-related outcomes pertain to each risk group, and the snail-related outcome, to each village.

The end result of a set of Monte Carlo runs is then n_g good simulations (passes) and n_b bad ones (not passes). Associated with each run is the parameter vector that gave rise to a simulation that met the goodness-of-fit criteria or that did not. These vectors contain the information from which the second generation of parameter distributions is estimated. Through this calibration procedure, parameter uncertainty can be reduced and the model made specific to the set of villages and risk groups under study. As more data from future research become available, additional pass/not-pass criteria can be defined, and the model parameters can be better refined, resulting in third-generation, fourth-generation, and so on, parameter estimates. However, after the second-generation parameter estimates, we have a calibrated model, which can be used for prediction and the evaluation of intervention strategies.

Although we have had considerable practical success with the pass/not-pass methodology, it transpires that it is a special case of a Bayesian approach recently proposed by Poole and Raftery (2000) that they call Bayesian melding. We believe that their generalization and formalization of the means of dealing with parametric uncertainty, when applied to deterministic models of the sort typified by the schistosomiasis model, expand the potential of the overall approach that we have proposed above. Because the issue of parameter estimation lies at the practical core of our work, we now outline the Bayesian melding approach.

Parameter Estimation and Uncertainty Analysis: Bayesian Melding

In the approach we outline above, the models are used for two purposes. Ultimately, they are used to perform virtual field experiments to compare the performance of competing intervention strategies. We refer to this goal as prediction. However, before prediction can be performed, it is important that model parameters have been well estimated. Although we have based parameter estimates on the best available literature and field data, residual uncertainty still exists, which we would like to reduce. We can refine these parameter estimates by statistically comparing the output predicted by the model with observed outcomes. In essence, we must reconcile what we know to be reasonable outputs of the model with outputs that are induced by running the model on what we also believe to be reasonable input parameters. In fact, for some input parameters for which we have no prior information—notably, the spatial parameters—the model and output data provide the only means for estimation. Because this situation requires several disparate sources of data to be integrated in estimating the parameters of interest, a Bayesian approach lends itself to estimation in this complicated and somewhat messy situation. An iterative process allows for the comparison of information from past experience with future data to further refine estimation and calibration of the disease models.

In the fortunate circumstance that a detailed time series on one of the outcome variables (such as disease prevalence) has been recorded, least-squares or maximum likelihood techniques can be used to estimate the parameters (Eisenberg et al. 2002). Such detailed data are most commonly available in acute, outbreak situations. These procedures have been also adopted to account for prior information on the parameters by using a traditional Bayesian approach (Raftery et al. 1995), where one can define both priors on the inputs and likelihood on the outputs, given the model, input parameters, and data. The data associated with schistosomiasis, however, are different, in

that disease prevalence in humans is measured annually at most and typically not every year. In addition, there will be information on other state variables, such as the density and infection rate of snails, also measured at most annually. Finally, there is also expert opinion regarding plausible trajectories of the state variables, and these are typically a small subset of the possible trajectories that the model can produce and that match the sparse data. Technically, the information available can be translated into prior information on the input parameters and prior information on the output state variables. As discussed above, the pass/not-pass method is one that identifies parameter sets consistent with reasonable output state-variable time series. Others have introduced metrics to define distance between model outputs and expected characteristics of these model outputs, such as frequency and magnitude of oscillations (Kendall et al. 1999). The Bayesian melding approach unifies procedures that combine prior information regarding both input parameters and model output to further constrain the acceptable solution space of the input parameters.

In Bayesian melding, two priors on the output are compared. One prior is based on literature or field data as to what is reasonable output. The other prior on the output is induced by running the model on valid prior information on input parameters. These two output priors are "melded" together and inverted to the input parameter space, thereby refining the estimate of the input parameters. In detail, and using Poole and Raftery's terminology, let

$$M:\theta \rightarrow \phi,\ \theta \in \Theta \subseteq \Re^n,\ \phi \in \Phi \subseteq \Re^p, \quad [1]$$

where M is the deterministic model that relates an n-vector of input parameters, θ, to a p-vector of outputs:

$$\phi = M(\theta). \quad [2]$$

Using an example drawn from our model, one component of θ might be the input h, the number of eggs/worm pair/stool quantity, and a component of ϕ might be the output $w_{ik}(t)$, the worm burden in the ith village, kth risk group at time t. Define the posterior, joint model of inputs and outputs to be

$$\pi(\theta,\phi) \propto \left\{ \begin{matrix} p[\theta, M(\theta)] \text{ if } \phi = M(\theta) \\ 0 \text{ otherwise} \end{matrix} \right\}, \quad [3]$$

where $p[\theta,M(\theta)]$ is the premodel joint distribution. One can think of $p[\theta,M(\theta)]$ as containing the statistical information and relationships among the parameters and state variables before considering how these inputs determine the outputs, ϕ. The post-model joint distribution, $\pi(\theta,\phi)$, however, only puts mass on input/output combinations consistent with the model, so it is a rescaled version of $p[\theta,M(\theta)]$ with the mass of the impossible input/output combinations set to 0. In the schistosomiasis model, for instance, if the prior $p[\theta,M(\theta)]$ allows positive probability on the combination $h = a$ and $w_{ik}(t) = b$, whereas the structure of the model suggests such a combination could never exist (although each value is acceptable in different combinations), then the posterior puts mass 0 on (a,b).The interesting distribution with respect to the estimation of input parameters is the marginal posterior of the inputs, or,

$$\pi(\theta) \propto p[\theta, M(\theta)]. \quad [4]$$

As discussed above, we will ignore the possibility of having sufficient data to construct meaningful likelihood on the outputs. Thus, one can typically write the prior as

$$p(\theta,\phi) \propto q_1(\theta) q_2(\phi), \quad [5]$$

where q_1 and q_2 are the prior distributions for θ and ϕ, respectively. Because $\phi = M(\theta)$, the model M and the prior q_1 induce another, independent prior on ϕ, say, $q_I(\phi)$. Bayesian melding is a method for reconciling these two priors on the output, resulting in a single prior, say, $\tilde{q}(\phi)$. The final step is to invert $\tilde{q}(\phi)$ to get a marginal, posterior distribution on the input parameters, $\pi(\theta)$. The overall approach is illustrated at the top of Figure 3. The distillation of the above technical discussion is that Bayesian melding

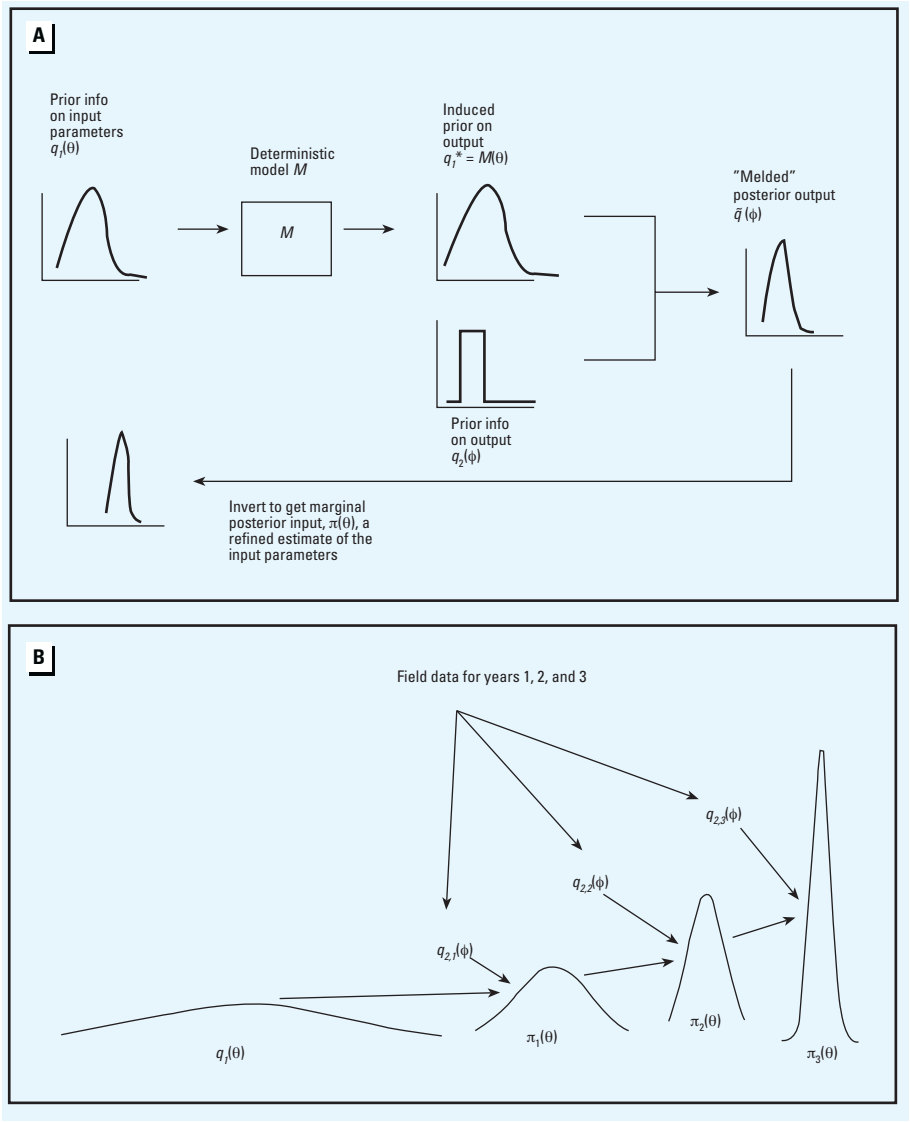

Figure 3. Illustration of the Bayesian melding procedure, where melding the prior knowledge of what the model output should be and what the model actually produces results in a refined knowledge about the parameters (*A*); schematic diagram of repeating use of Bayesian melding to update the posterior input parameter distributions when data are collected from three field seasons and converted to output priors (*B*).

takes existing information on the input parameters in the form of a prior distribution, $q_1(\theta)$, and new information/expert opinion on the outputs, $q_2(\phi)$, and constructs new, refined information on the parameter inputs, $\pi(\theta)$. Note that the pass/not-pass procedure simply places a uniform prior on a subspace of Φ, $q_2(\phi) \propto I(\phi \in \Phi^S)$, where Φ^S is the acceptance region of Φ. Unlike traditional Bayesian analysis, Bayesian melding requires priors on the inputs, priors on the outputs, and a somewhat arbitrary specification of how $q_2(\phi)$ and $q_1(\phi)$ are to be melded (relative weights given to each distribution) to get $\tilde{q}(\phi)$.

Bayesian melding can be applied iteratively, adding new field data over time, to progressively refine parameter estimates. Using our field research in China, we begin with a field study to construct $q_{2,1}(\phi)$ (subscript 1 is for first field season) and combine with existing prior information on the input parameters, $q_1(\theta)$, using Bayesian melding to get, first, posterior estimate of the input parameters, say, $\pi_1(\theta)$. In the next field season, we again collect new data to get $q_{2,2}(\phi)$ and use as input prior last years distribution, $\pi_1(\theta)$, now as a prior input parameter distribution to get a new postmodel distribution, $\pi_2(\theta)$. For every new collection of field data, this process is repeated and, ideally, the postmodel distribution becomes "tighter" as more information is collected. The bottom part of Figure 3 shows schematically how this process works.

As more is known about the transmission process for a particular disease, the more potentially complicated one can make the mathematical models. Increasing complexity typically involves the addition of more input parameters, such as the addition of occupation as a risk factor in the models described above. Another consequence is that the models have a greater variety of possible patterns in the output state variables. It also increases the potential for noninvertibility of the models—the potential that different sets of input parameters will result in the identical output series, or,

$$M(\theta_1) = M(\theta_2) = \phi, \quad \theta_1 \neq \theta_2. \qquad [6]$$

What this implies is that strong (possibly nonlinear) relationships among some of the parameters will characterize the joint posterior distribution. Strong relationships among some parameters imply that the available data on the output series cannot identify these parameters but only linear or nonlinear composites of them. Finding these sets of parameters either can help identify parameters that need independent data or may suggest new ways to parameterize these models that can reduce unnecessary redundancy.

Most procedures that attempt to estimate the marginal posterior, $\pi(\theta)$, do so not by finding the distribution directly but by methods that generate random samples from the underlying distribution of interest, one example being the pass/no-pass method. The result is repeated draws from $\pi(\theta)$. So, finding linear and nonlinear relationships among the parameters is an exercise in multivariate density estimation and principal components analysis. Tree-based density estimation is a nonparametric multivariate density estimation technique particularly well suited to estimation of high-dimensional data (Spear et al. 1994). In addition, newly developed nonlinear principal components (Bakshi and Utojo 1998) permit the discovery of strong nonlinear relationships among the parameters that conventional principal components methods would not discover.

Uncertainty Analysis

Consider now that a posterior distribution of the model parameters has been estimated. Estimating the posterior distribution, $\pi(\theta)$, provides an estimate of the uncertainty of input parameters. It does not, however, measure the cost of this uncertainty, where cost is defined to be the uncertainty of the predicted output series, ϕ. As discussed above, repeated simulations of the model based on random draws from the parameter distribution can be used to evaluate model sensitivity to parameters. More generally, these simulations can be used to evaluate the cost of sensitivity on output uncertainty, a process called uncertainty analysis (Morgan et al. 1990). Note that uncertainty in the output is a function of both uncertainty in the inputs and the sensitivity of the outputs to changes in the inputs. For instance, estimation of $\pi(\theta)$ might imply that a particular parameter—say, θ^*—is very poorly estimated (marginal distribution of θ^* has large variance), but if ϕ is insensitive to changes in this parameter, the uncertainty of θ^* does not propagate into uncertainty in the prediction. To direct research efforts in the future to construct more reliable models, we should identify parameters that have the biggest cost with respect to their uncertainty. As mentioned above, for instance, how does the uncertainty/inaccuracy in measuring local rainfall translate into uncertainty on the model predictions of the prevalence of schistosomiasis in humans?

Uncertainty analysis should be done on groups of correlated parameters as discovered using the methods described above. One method to find the parameters that contribute the most to uncertainty in the outputs is to fix all but one parameter or group of correlated parameters and record the variance in the output when the remaining parameters are allowed to vary according to $\pi(\theta)$. By comparing the results, one can target field research to better identify those parameters that contribute most to uncertainty of the output.

Concluding Example

An example of the general strategy outlined above is based on the historical data from the Chuanxing villages. The most highly infected risk group was the residents of a village where 81.9% of all villagers above 5 years of age were infected in 1987. All villagers were treated with praziquantel in October 1987 and August 1988, which reduced the infection rate to 19.3% in 1989. In simulating the infection experience of these villagers over this interval, we used an earlier version of the model without the spatial parameters γ_{ik} and ξ_k and the rainfall effects parameters $r(p)$. All of the prior parameter distributions of the model were

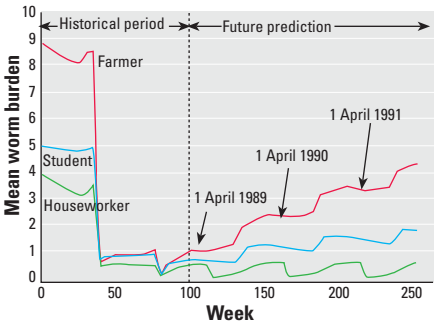

Figure 4. Simulation of the effect of targeted treatment on each risk group using the model calibrated for 1987 and 1989 data from the Xichang study area, where targeted treatment was explored. Simulated treatment was given to the student group every April for 3 years but no treatment for the farmer and domestic-worker groups.

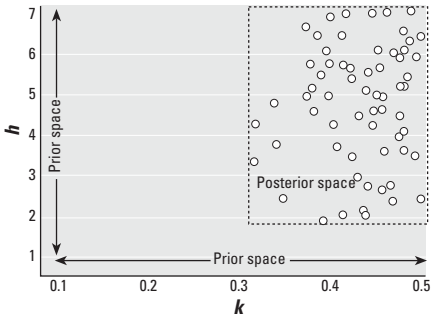

Figure 5. An example from Xichang data of how the posterior parameter space for model parameters can be reduced through simulation and Bayesian melding. Model parameters *h* and *k* were originally sampled uniformly between 1 and 7 and between 0.1 and 0.5, respectively (the prior space). The subset of simulations that pass the calibration criteria (the posterior space), shown as dots, are constrained to the upper right corner of the parameter space, showing that uncertainty in these variables is reduced substantially.

defined to be uniform, some based on data from the literature and some based on local data as outlined above, and Monte Carlo simulations were performed. The calibration trajectories on the left side of Figure 4 are from those simulations that were judged to be consistent with the field data for three risk groups from the 1987 infection surveys. As discussed above, criteria for a good simulation were based on a close match to the 1989 prevalence data, an upper bound on infected snails, and a lower bound on the egg excretion values, given the 1987 initial conditions. All of the trajectories shown in Figure 4 met the criteria, which illustrates that more than one set of parameters can be consistent with the calibration conditions. However, only 62 of the 10,000 simulations met these criteria, so the posterior parameter space has been significantly constrained.

Our experience has shown that only a small subset of simulations are "good," generally ≤ 1% of the total. If we compare the posterior space with the prior space for two model parameters, h and k (Figure 5), we can see begin to understand how parameter distributions are refined in multivariate space through our methodology, with acceptable parameter combinations lying in the upper right quadrant of the prior space, substantially reducing the uncertainty in these two parameters. Here, the situation with just two parameters can be easily understood. However, one generally observes that, among parameter sets resulting in good simulations, some parameters range over almost their entire prior distribution range. These two observations suggest that the second- and subsequent-generation parameter spaces are very complex and that more sophisticated multivariate techniques must be used to explore the posterior parameter spaces.

Now that we have calibrated the model to fit our village, to illustrate the exploration of control strategies, we used the 62 parameter vectors to forecast the effects of a sustained intervention among the three groups commencing in 1989, during which only students were administered chemotherapy annually. These results are shown on the right side of Figure 4. Clearly, if this village were isolated such that no cercaria or eggs were imported from neighboring villages, the results suggest that annual mass chemotherapy would be effective in maintaining a low worm burden and low prevalence of infection among students, but that little benefit would be afforded the farmers or domestic workers. In this particular case, the variability about the mean trajectories shown in Figure 4 is modest, largely because of the extent of data available from the Chuanxing studies. We cannot expect that to be the case in forecasting studies based on current surveillance data, and much of our current work focuses on identifying key parameters about which improved estimates would have a significant impact on the variability of forecasts of control effectiveness.

Although such examples raise a variety of questions about the model and its parameterization, the strategic point, we hope, is clear. If we can capture the principal elements of the disease cycle in the structure of the model and demonstrate that it can be parameterized from local data at reasonable cost and with acceptable residual uncertainty, the notion of designing interventions using the model becomes an attractive and, ideally, an effective management tool.

References and Notes

Anderson RM, May RM. 1985. Helminth infections of humans: mathematical models, population dynamics, and control. Adv Parasitol 24:1–101.

———. 1991. Infectious Diseases of Humans: Dynamics and Control. New York:Oxford University Press.

Bakshi BR, Utojo U. 1998. Unification of neural and statistical modeling methods that combine inputs by linear projection. Comput Chem Eng 22:1859–1878.

Chen MG. 1999. Progress in schistosomiasis control in China. Chin Med J 112:930–933.

Chen MG, Zheng F. 1999. Schistosomiasis control in China. Parasitol Int 48:11–19.

De Vlas SJ, Gryseels B, Van Oortmarssen GJ, Polderman AM, Habbema JDF. 1992. A model for variations in single and repeated egg counts in *Schistosoma mansoni* infections. Parasitol 104:451–460.

Eisenberg JNS, Seto E, Olivieri A, Spear R. 1996. Quantifying water pathogen risk in an epidemiological framework. Risk Anal 16:549–563.

Eisenberg JNS, Brookhart MA, Rice G, Brown M, Colford JM. 2002. Disease transmission models for public health decision making: analysis of epidemic and endemic conditions caused by waterborne pathogens. Environ Health Perspect 110:783–790 (2002).

Grieb TM, Hudson RJM, Shang N, Spear RC, Gherini SA, Goldstein RA. 1999. Examination of model uncertainty and parameter interaction in a global carbon cycling model (GLOCO). Environ Int 25:787–803.

Hotez PJ, Feng Z, Xu LQ, Chen MG, Xiao SH, Liu SX, et al. 1997. Emerging and reemerging helminthiases and the public health of China. Emerg Infect Dis 3:303–310.

Kendall BE, Briggs CJ, Murdoch WW, Turchin P, Ellner SP, McCauley E, et al. 1999. Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. Ecology (Washington, DC) 80:1789–1805.

Liang S, Maszle D, Spear RC. 2002. A quantitative framework for a multi-group model of *Schistosomiasis japonicum* transmission dynamics and control in Sichuan, China. Acta Trop 82:263–277.

MacDonald G. 1965. The dynamics of helmingth infections, with special reference to schistosomes. Trans R Soc Trop Med Hyg 59:489–506.

Maszle DR. 1998. Dynamic Modeling for the Control of Schistosomiasis in China in Light of Parametric Uncertainty [PhD thesis]. Berkeley, CA:University of California, Berkeley/San Francisco.

Morgan MG, Henrion M, Small M. 1990. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge, UK:Cambridge University Press.

Poole D, Raftery AE. 2000. Inference for deterministic simulation models: the Bayesian melding approach. J Am Stat Assoc 95:1244–1255.

Raftery AE, Givens GH, Zeh JE. 1995. Inference from a deterministic population-dynamics model for bowhead whales. J Am Stat Assoc 90:402–416.

Ross AGP, Yuesheng L, Sleigh AC, Williams GM, Hartel GF, Forsyth SJ, et al. 1998. Measuring exposure to *S. japonicum* in China. I. Activity diaries to assess water contact and comparison to other measures. Acta Trop 71:213–228.

Seto E, Liang S. 2002. Schistosomiasis in China. Available: http://ehs.sph.berkeley.edu/china [accessed 28 July 2002].

Seto E, Liang S, Qiu D, Gu X, Spear RC. 2001. A protocol for geographically randomized snail surveys in schistosomiasis fieldwork using the global positioning system. Am J Trop Med Hyg 64:98–99.

SIPD. 1998. The Impact of Environmental Change on the Occurrence of Schistosomiasis in the Three Gorges Area. Technical Report 94-8-11. Chengdu:Sichuan Institute of Parasitic Disease.

Spear RC, Bois FY, Woodruff T, Auslander D, Parker J, Selvin S. 1991. Modeling benzene pharmacokinetics across three sets of animal data—parametric sensitivity and risk implications. Risk Anal 11:641–654.

Spear RC, Grieb T, Shang N. 1994. Parameter uncertainty and interaction in complex environmental models. Water Resour Res 30:3159–3169.

Spear RC, Hornberger G. 1980. Eutrophication in Peel Inlet. II. Identification of critical uncertainties via generalized sensitivity analysis. Water Res 14:43–49.

WHO. 1993. The Control of Schistosomiasis. Second Report of the WHO Expert Committee. Technical Report Series 830. Geneva:World Health Organization.

Woolhouse MEJ. 1991. On the application of mathematical models of schistosome transmission dynamics. I. Natural transmission. Acta Trop 49:241–270.

Zhou Y, Maszle DR, Gong P, Spear RC, Gu XG. 1997. GIS-based spatial models of schistosomiasis infection. Geol Info Sci 2:51–57.