science for a changing world

Forecasting Bacteria Levels at Bathing Beaches in Ohio

The U.S. Geological Survey developed models for predicting exceedance of the bathing-water standard for Escherichia coli (E. coli) at three Lake Erie beaches and one inland lake in Ohio. The statistical models were specific to each beach, and the best model for each beach was based on a unique combination of environmental and water-quality variables as explanatory factors. For the Lake Erie beaches, these factors included wave height, number of birds on the beach at the time of sampling, lake-current direction, rainfall, turbidity, and streamflow of a nearby river. For the inland lake, these factors included date, wind direction and speed, number of birds, and rainfall. The prediction error in the models was too large to accurately estimate concentrations of E. coli; however, the models can be used like weather forecasts to predict the probability, given a set of input variables, that the Ohio bathing water standard used to judge swimming safety will be exceeded.

WHY AND HOW ARE BACTERIA LEVELS CURRENTLY MONITORED AT BATHING BEACHES?

People may risk illness from exposure to disease-causing microorganisms (pathogens) at recreational beaches. Twenty-seven percent of the U.S. beaches that responded to a recent beach survey reported at least one advisory or closure during the 2001 swimming season (U.S. Environmental Protection Agency, 2002a). Because monitoring for pathogens is difficult and expensive, beach advisories or closings are issued on the basis of standards for concentrations of indicator organisms. Indicator organisms do not necessarily cause disease, but they are present in feces and therefore indicate the possible presence of pathogenic organisms. The levels of indicator organisms provide a measure of the quality of the recreational water and the risk of illness due to water-contact activities. The State of Ohio uses the indicator bacterium *Escherichia coli* (*E. coli*) to assess recreational water quality (Ohio Environmental Protection Agency, 2002) because it is the indicator recommended by the U.S. Environmental Protection Agency (U.S. Environmental Protection Agency, 1986). *Escherichia coli* is a natural inhabitant of the gastrointestinal tract of warmblooded animals and is direct evidence of fecal contamination from them. For Ohio, one standard used to assess bathing-water quality is the single-sample maximum level for *E. coli* of 235 colonies per 100 milliliters (col/100 mL); this level cannot be exceeded in more than 10 percent of samples collected during any 30-day period (Ohio Environmental Protection Agency, 2002).

Under the current advisory system, users of recreational waters may still be exposed to waterborne pathogens because of delayed notification of monitoring results (U.S. Environmental Protection Agency, 2002a). Current methods to assess concentrations of *E. coli* take at least 18 hours to complete. During this period, *E. coli* concentrations may change dramatically, which may mean that a beach may be posted with an advisory when the bacterial water quality has already returned to safe levels. This also means that an advisory may not be posted on a day when the risk of pathogen exposure is high; for example, after a heavy overnight rain.

WHY WAS THE STUDY DONE?

Agencies that monitor the beaches need tools that can provide quick, reliable indicators of recreational water-quality conditions. Predictive models are one means to provide these rapid assessments (U.S. Environmental Protection Agency, 2002b). In cooperation with local agencies, the U.S. Geological Survey (USGS) addressed this concern in a study at three Lake Erie beaches near Cleveland, Ohio (Edgewater Park, Villa Angela, and Huntington Reservation) and one inland lake beach in Trumbull County, Ohio (Mosquito Lake) (fig. 1). In this study, scientists investigated use of water-quality and environmental variables in beach-specific statistical models to predict *E. coli* concentrations.

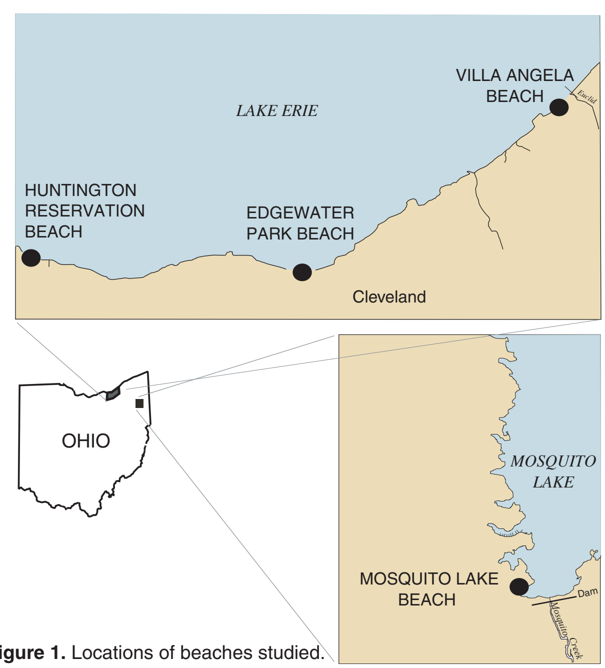

Figure 1. Locations of beaches studied.

HOW WAS THIS STUDY DONE?

Data were collected during 2000 and 2001 to build on an earlier study in 1997 (Francy and Darner, 1998). Water samples were collected four or five days a week by local agencies at beach study sites during the May through August recreational season and analyzed for concentrations of *E. coli*. The local agencies that collected and analyzed samples were the Northeast Ohio Regional Sewer District, the Cuyahoga County Board of Health, and the Cuyahoga County Sanitary Engineers. Environmental factors that were measured included turbidity, temperature of the water, number of birds on the beach at the time of sampling, wave height, streamflow at a nearby gaging station, wind speed and direction, intensity of ultraviolet radiation, rainfall amounts in the previous 24 or 72 hours, and lake-current direction. Data were compiled and analyzed by the USGS. The factors found to be related to *E. coli* concentrations were used as variables in multiple linear regression (MLR) models to predict concentrations of *E. coli*.

WHAT WERE THE STUDY RESULTS?

What types of variables did the models contain?

The first step in development of predictive models is to identify factors related to *E. coli* concentrations and that may be used in the models as explanatory variables. Variables that were not related or weakly related to concentrations *E. coli*, such as water temperature and ultraviolet intensity, were not used in predictive models. Although the previous day's concentration of *E. coli* was found to be related to the present day's concentration of *E. coli*, it was not used as a variable in the model. Data on the previous day's *E. coli* concentration were not always available and did not improve the predictive ability of any of the models enough to warrant additional work. An example of one of the related variables is turbidity at Huntington Reservation (hereafter "Huntington"). As *E. coli* concentrations increased, turbidity values also increased (fig. 2). However, the scatter of *E. coli* concentrations at a given level of turbidity was considerable. One reason for the scatter is that turbidity does not explain all of the variability in *E. coli* concentrations; other explanatory variables are needed to more fully explain this variability.

Different combinations of variables related to *E. coli* were tested by use of MLR techniques. The explanatory variables used in MLR models that best predicted recreational water quality and reduced multicolinearity (where at least one explanatory variable is related to one or more other explanatory variables) varied from beach to beach (table 1). Wave height was included as an explanatory variable in all of the Lake Erie models. The two models presented for Huntington included wave height, turbidity, and two additional variables and had R^2 values of approximately 0.4. The R^2 value represents that fraction of the variation in *E. coli* concentrations that is explained by the model. The R^2 values for models at the other beaches were not as high as those for Huntington, which means that greater amounts of the variations in *E. coli* concentrations were unexplained at these other beaches. At Mosquito Lake, the variables wind direction, wind speed, date, number of birds on the beach, and rainfall were used to predict *E. coli* concentrations. Some of the variables used in the Mosquito Lake model differed from those used for models at the Lake Erie beaches. This is not surprising because Mosquito Lake is an inland lake, and hydrologic processes affecting the beach at Mosquito Lake differ from those affecting the Lake Erie beaches.

What type of information does each model provide for beach managers?

The probability that *E. coli* concentrations would be equal to or greater than 235 col/100 mL was used as the model output variable because the prediction errors in the models were too large to accurately estimate concentrations of *E. coli*. This approach provides estimated probabilities similar to those in a weather forecast and can be illustrated by means of one of the Huntington models. The variables used in this model are listed in line 2 of table 1. For example, if the following data are measured one morning at Huntington: wave height is 1-3 feet, rainfall amount in the last 24 hours is 0.45 inch, turbidity is 35 Nephelometric Turbidity Units (NTUs), and current direction is easterly, the Huntington model results predict a 40-percent probability that the single-sample bathing-water standard will be equaled or exceeded.

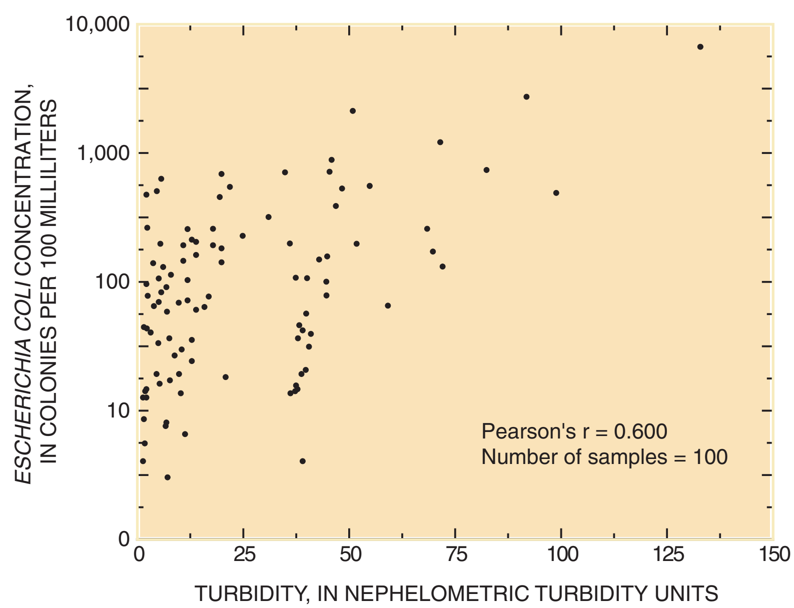

Figure 2. Comparison of turbidity and *Escherichia coli* concentrations for Huntington Reservation, 2000 and 2001.

How can these models be used by beach managers to predict recreational water quality?

For the model to be useful to beach managers, the probability that is associated with too great a risk to allow swimming needs to be determined. Establishing a threshold probability provides this information. Computed probabilities that are less than a threshold probability indicate that bacterial water quality is most likely acceptable. Computed probabilities equal to or above the threshold probability indicate that the water quality is most likely not acceptable and that a water-quality advisory may be needed.

Threshold probabilities were established by determining the lowest probability that produced the most correct responses and fewest false negative responses. False negative responses are especially troubling because swimming may be allowed when the bathing-water standard is exceeded. A threshold probability of 32 percent was established for Huntington. This concept is illustrated for the model at Huntington shown in figure 3 (see back page). The data points on the plot are *E. coli* concentrations used to develop the model that were measured during 2000 and 2001 (x-axis) and the computed probabilities determined by the model (y-axis). The plot is divided into four quadrants by a vertical line through 235 col/100 mL on the x-axis and a horizontal line through the threshold probability of 32. The four quadrants in figure 3 are

1. Correct nonexceedance. *E. coli* concentrations did not exceed the standard (were less than 235 col/100 mL), and the predicted probabilities were below the threshold.
2. False positive. *E. coli* concentration did not exceed the standard, but the predicted probabilities were above the threshold.
3. Correct exceedance. *E. coli* concentrations were equal to or exceeded the standard, and the predicted probabilities were above the threshold.
4. False negative. *E. coli* concentrations were equal to or exceeded the standard, but the predicted probabilities were below the threshold.

The threshold probability of 32 gives a total of 88 correct responses (quadrants 1 and 3), 7 false positives (quadrant 2), and 4 false negatives (quadrant 4). Consequently, if a threshold probability of 32 percent was used to dictate posting of an advisory, the correct decision would have been made in 88.9 percent of the cases. In 7.1 percent of the cases, an advisory would have been posted when the standard was not equaled or exceeded, and in 4.0 percent of the cases, an advisory would not have been posted when the standard was exceeded. Threshold probabilities and responses are shown for the other beach models in table 1.

Table 1. Variables, regression statistics, and threshold data for beach models, 1997, 2000, and 2001

[R^2 is the fraction of the variation in the dependent variable that is explained by the model; threshold probability is based on the the Ohio single-sample bathing-water standard for *Escherichia coli*]

Beach	Time period of data used for model development	Variables in model	R^2 of model	Threshold probability	Number of model responses by category (percent)		
					Correct	False positive	False negative
Edgewater	2000 and 2001	Wave height, number of birds[a], lake-current direction, rainfall 24[a]	0.32	45	109 (79.5)	9 (6.6)	20 (14.6)
Huntington	2000 and 2001	Wave height, turbidity, rainfall 24[b], lake-current direction	0.41	32	88 (88.9)	7 (7.1)	4 (4.0)
Huntington	2000 and 2001	Wave height, turbidity, streamflow 7am, number of birds[a]	0.40	30	83 (83.8)	7 (7.1)	9 (9.1)
Mosquito Lake	2000 and 2001	Date, sum sine wind direction[d], wind speed, number of birds, rainfall weighted 72[e]	0.39	29	90 (90.9)	4 (4)	5 (5.1)
Villa Angela	1997, 2000, and 2001	Wave height, log turbidity, rainfall weighted 72[e]	0.32	39	119 (71.2)	24 (14.4)	24 (14.4)

[a] Number of birds on beach at time of sampling.
[b] Rainfall 24 was the amount in inches at Hopkins Airport, Cleveland, Ohio, in the 24-hour period preceding the 9 a.m. sampling.
[c] Streamflow 7am was the streamflow at the Cuyahoga River at Independence, Ohio, at 7 a.m. on the day of sampling.
[d] Sum sine wind direction was the sum of the sines of the instantaneous wind directions determined at 9 a.m. on the day of sampling and on the two previous days.
[e] Rainfall weighted 72 was the amount in inches at Hopkins Airport, Cleveland, Ohio, or Youngstown Airport, Youngstown, Ohio, in the 72-hour period preceeding the 9 a.m. sampling, with the most recent rainfall receiving the greatest weight.

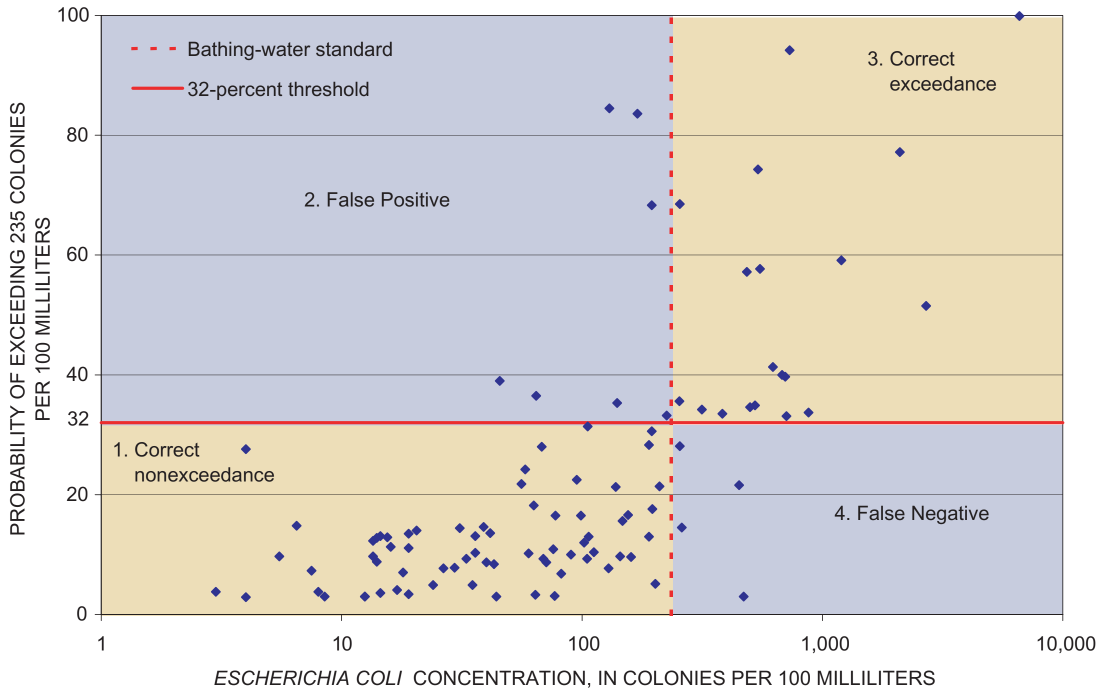

Figure 3. Establishment of the threshold probability based on the single-sample bathing–water standard of 235 colonies per 100 milliliters and the 2000-2001 model for Huntington Reservation, Cleveland, Ohio. (Samples were collected from May through August in 2000 and 2001.)

SUGGESTIONS FOR FUTURE RESEARCH

Additional research could include testing the MLR models developed for Edgewater Park, Villa Angela, Huntington, and Mosquito Lake in subsequent years and compare the models' ability to predict recreational water quality to results from the current method—using antecedent *E. coli* concentrations. If, over time, the probability-based models predict impairment of recreational water quality as well as or better than current methods, beach managers may consider using the models to aid or direct decisions on posting beach advisories. Results could then be provided within 2 hours of data collection, and the beachgoer would have access to more timely information on current water-quality conditions before leaving for the beach.

Donna S. Francy and Robert A. Darner

This study was performed by the USGS in cooperation with the Ohio Water Development Authority, Northeast Ohio Regional Sewer District, Ohio Lake Erie Office, Cuyahoga County Board of Health, Cuyahoga County Sanitary Engineers, and the Cuyahoga River Community Planning Organization.

Information on technical reports and hydrologic data related to the above study can be obtained from

Information Officer
U.S. Geological Survey
6480 Doubletree Avenue
Columbus, OH 43229-1111
(614) 430-7700
http://oh.water.usgs.gov

REFERENCES

Francy, D.S., and Darner, R.A., 1998, Factors affecting *Escherichia coli* concentrations at Lake Erie public bathing beaches: U.S. Geological Survey Water-Resources Investigations Report 98-4241, 41 p.

Ohio Environmental Protection Agency, 2002, Water use definitions and statewide criteria: Ohio Administrative Code, Chap. 3745-1-07, July 21, 2002, p. 9 and 26.

U.S. Environmental Protection Agency, 1986, Ambient water-quality criteria for bacteria—1986: Washington, D.C., Office of Water, EPA-A440/5-84-002, 18 p.

U.S. Environmental Protection Agency, 2002a, EPA's BEACH watch program: 2001 swimming seasons: Washington, D.C., U.S. Office of Water, EPA-823-F-02-008.

U.S. Environmental Protection Agency, 2002b, National beach guidance and required performance criteria for grants: Washington, D.C., Office of Water, EPA-823-B-02-004.