EPA 823-R-07-006

REPORT OF THE EXPERTS SCIENTIFIC WORKSHOP ON CRITICAL RESEARCH NEEDS FOR THE DEVELOPMENT OF NEW OR REVISED RECREATIONAL WATER QUALITY CRITERIA

Airlie Center
Warrenton, Virginia
March 26-30, 2007

U.S. Environmental Protection Agency
Office of Water
Office of Research and Development

June 15, 2007

CHAPTER 3
METHODS DEVELOPMENT

Stephen Weisberg, Chair, Southern California Coastal Water Research Project
Alfred Dufour, USEPA
Charles Hagedorn, Virginia Tech
Sharon Kluender, Wisconsin State Laboratory of Hygiene
Erin Lipp, University of Georgia
Robin Oshiro, USEPA

[this page intentionally left blank]

3.1 Introduction

The Methods Development workgroup focused on addressing the following four key questions:

1. What are the attributes and criteria for deciding whether a new method or indicator is ready for adoption by EPA?
2. What kinds of studies are necessary to quantify those attributes?
3. Are there any new indicators/methods for which those studies have been conducted and that are ready for adoption?
4. What studies (or modifications to planned studies) are most critical for EPA to implement in the next 3 years to support adoption of new methods/indicators in a criteria development framework?

A critical starting point for the workgroup members was recognition that the evaluation of methods and/or indicators needs to be considered in context of the Clean Water Act (CWA) applications in which they would be used. The following five primary uses were identified by workgroup members:

- Routine beach monitoring to support public health warning notification systems;
- Routine beach monitoring data to support total maximum daily load (TMDL) decisions;
- Rapid methods to track the progress of a sewage spill as it moves downstream or downcoast to improve the beach closure determinations;
- Compliance assessments conducted at the terminus of National Pollutant Discharge Elimination System (NPDES) discharge pipes; and
- Trends assessments to determine whether water quality conditions at a site are changing over time.

The workgroup focused on the first two applications beacause members felt that they are most relevant to EPA's desire to redefine their current recreational water quality criteria. However, several workgroup members also recognized the relevance of the other applications so a short section is included (see Section 3.7) that illustrates the similarities and differences in the method evaluation process for these other CWA uses.

For water quality notification systems, two principal issues were identified that need to be addressed. The first is that current laboratory measurement methods require up to 24 hours to enumerate indicator bacteria. Contaminated beaches remain open during this processing period, but indicator bacteria may already have returned to acceptable levels by the time laboratory results are available and warning signs are posted. Continued advances and improvements in molecular- and immunological-based techniques provide new opportunities for measuring bacteria more rapidly. Although current (traditional/standard) methods rely on bacterial growth and metabolic activity, these new methods allow direct measurement of cellular attributes, such as genetic material or surface immunological properties. By eliminating the necessity for a lengthy incubation step, some of these methods have the potential to provide results in less than 4 hours, enabling managers to take action to protect public health (i.e., post warnings or close beaches) on the same day that water samples are collected. This assumes that samples can be

processed at the beach or that the time required for transportation to a laboratory is brief. For same day posting to be achieved, the results of the tests also have to be delivered to and evaluated by beach managers in a timely manner.

The second issue workgroup members identified is that present standards used to evaluate recreational water quality data are based on a "one-size-fits-all model," relying on use of a single indicator (e.g., enterococci at marine beaches) and a single standard for all recreational waters. There is growing recognition that enterococci measured on the beach may derive from many sources, including humans, domesticated animals, indigenous wildlife (including shore and migratory birds), and regrowth in sand, sediments, or on biofilms. The health risk to humans varies depending on which of these sources is responsible for the measured enterococci. As such, existing warning systems do not provide an equal level of health risk protection at all beaches. Moreover, the costly cleanup processes associated with the TMDL programs are not necessarily focused on the beaches that represent the greatest public health risk. There are additional concerns that cleanup activities, and associated costs, are being targeted at beaches where enterococci concentrations that exceed standards result from natural sources and processes.

EPA could consider two means of adjusting their criteria framework to address one-size-fits-all concerns. The first adjustment is to develop additional indicators to replace, or to augment in a tiered fashion, the existing enterococci indicator as it is now used at marine beaches (US EPA, 1986). These new indicators would be more specific to human sources and better related to human health risk than the existing indicator.

The second potential adjustment is to adopt a framework similar to that of the World Health Organization (WHO, 2003), in which watershed characterization studies are used to adopt site-specific standards. These site-specific standards would be based on perception of health risk resulting from the types of fecal sources in the watershed and the proximity of those sources to the beach. The Methods workgroup members felt strongly that source identification methods needed to be a key tool in characterizing risk and that further evaluation of source identification methods needed to be conducted if they are to be used in this context.

This chapter is organized around describing the approach that would be used for assessing methods/indicators in the following three contexts: (1) replacement of existing methods with more rapid methods, (2) replacement of existing indicators with those that are more specific to human sources of fecal contamination, and (3) determination of source identification methods that can be used to characterize risk in the development of site-specific standards. Within each section, the adequacy of evaluations of methods/indicators is discussed and the most immediate research activities that would provide the greatest benefit to EPA for modifying monitoring and/or indicators within the next 5 years are highlighted.

3.2 Classes of Indicators

The evaluation of methods is a critical element in bringing new technology to the measurement of water quality. Current evaluation protocols were developed for cultural methods for enumerating bacterial indicators of fecal contamination. The evaluation usually included method

attributes regarding the performance of the method, such as specificity, accuracy, and precision. Further evaluation that addressed how the method performed in and between laboratories included multi-laboratory testing that determined how robust a method might be (i.e., how poorly can the method be performed and still produce useful results?). The question that arises is whether the current protocols for evaluating membrane filter culture-based methods are suitable for evaluating new methods that are being proposed for measuring water quality. Some of the new or alternative methodologies that are available for testing water quality include molecular-based methods, such as quantitative polymerase chain reaction (qPCR), nucleic acid sequence based amplification (NASBA), and transcription-mediated amplification (TMA). These methods amplify nucleic acid sequences to high levels such that they can be easily detected. Other methods use antibodies to which fluorescent compounds are attached. The fluorescent-tagged antibodies then attach to specific microbes and are "counted" in a flowcytometer. The preceding methods "count" dead and live bacteria and thus differ significantly from currently used quantitative cultural methods.

Some recent methods do measure viable microbes in an indirect manner. For example, enzyme-based methods measure substrate utilization employing compounds that fluoresce when metabolized by specific bacteria. Comparison of the fluorescence to a standard curve allows a "count" to be established. Another method measures adenosine triphosphate (ATP) using a bioluminescence measuring instrument to determine the amount of ATP that is produced only from viable bacteria.

In the current context, there are indicators available or in late stages of development that are ready for evaluation to determine if they are appropriate for use in routine beach monitoring. Some can be measured with the technology described above while others can be measured with currently available methodologies.

Leading candidates are indicators and detection methods that can be used to replace current culture-based indicators of water quality (i.e., enterococci and *E. coli*). For instance, nucleic acid sequences from enterococci have been used to measure the density of enterococci in bathing beach water. Some aspects of the performance of this method have been completed. As described above, enterococci have also been quantified using a fiber optic/fluorescent antibody detection method, an enzymatic/substrate method, and a method that measures ATP. None of the latter methods have been evaluated with respect to either their performance characteristics or for robustness. Similarly, molecular-based methods that measure viruses (e.g., adenoviruses) that might replace currently used indicators of recreational water quality have not had their performance evaluated. If these indicators are shown to be effective in their performance, they will be candidates for use in epidemiological studies to determine how well their densities in recreational waters relate to swimmer health.

Another class of microbes and other analytes are related to identifying the source of fecal contamination that might affect beach microbial water quality. Other markers include genes such as the *Esp* gene from enterococci, which might be specifically associated with human feces; male-specific (F+) coliphage that can indicate whether water has been contaminated by humans or animals; and chemical markers such as optical brighteners, caffeine, coprostanol, and urobilin that may be associated with human use or are the end-products of human metabolism (see also

Chapter 2). Optical brighteners are measured quite easily with a spectrofluorometer, while caffeine, coprostanol, and urobilin require more complex instrumentation, such as a high performance liquid chromatography (HPLC) instrument. Measuring genomic markers is less complicated and does not require a thermocycler to perform a PCR test. The varied nature of these source identification markers may require modification of the performance evaluation criteria to accommodate the different characteristics of these source specific analytes. For instance, the range of applicability and practicality may be more important than the accuracy and precision characteristics of these chemical or genomic source identification approaches.

The last class of indicators that may be ready for evaluation as indicators of fecal contamination are those that may have been rejected previously, for whatever reason, but should be considered again because of the availability of new information about their occurrence in water or because of new methods for their detection. Other potential indicators may be candidates because they are species within a group indicator, such as the enterococci and clostridia, and individual species may better indicate the quality of a waterbody. It is likely that this class of indicator will fit well into the current paradigm for characterizing microorganisms that might be used for routine recreational water quality monitoring.

3.3 Evaluating New Methods for Existing Indicators

Workgroup members felt that after a method passes defined performance criteria, it must be evaluated for its application as an equivalent (or superior) water quality tool compared to the current assays. An example of when this approach might be used is the transition from culture-based enterococci detection to detection by a rapid (molecular-based) assay.

The workgroup identified two major approaches to conduct this evaluation, (1) determining the relationship to health risks based on epidemiological studies or (2) establishing equivalency to an existing water quality tool.

3.3.1 Health Risk Evaluation

Workgroup members felt that determining the relationship to health risk is the best approach to evaluating a new method. An epidemiological study that can associate human risk with a new method is the preferred approach. The new detection method will ideally show an improved relationship to illness and will therefore be more protective of public health than the current approach that relies on indicator detection. Likewise, if the new method offers other improvements over the existing method (e.g., more rapid, less costly, etc.), then its relationship with human health should be at least as good as the current indicator.

The health risk evaluation should also be used when the target of the new method differs significantly from the current system. For example, a culture-based enterococci assay does not measure the same thing as a polymerase chain reaction (PCR)-based assay, which detects DNA rather than culturable (viable) cells. In these cases, a direct comparison of methods (as described below) may not be appropriate or possible.

3.3.2 Establishing Equivalency between New and Standard Methods

The equivalency validation approach assumes that for methods with similar targets (e.g., viable cells), the performance of the new method can be compared to that of the existing method without the need to determine health risk directly. Given the cost and time involved in large scale epidemiological studies, the equivalency approach can be performed for many new methods. The EPA should determine how dissimilar the method targets can be and still be evaluated by this approach. For example, cellular activity-based assays (e.g., immunomagnetic separation and ATP bioluminescence [IMS/ATP]) and membrane-filtration assays both measure viable cells, even though the end points are different. The workgroup members suggested that this activity-based assay is similar enough to be evaluated through equivalency validation. This level of flexibility is important because of the limited number of epidemiological studies that can be carried out in the near- or long-term.

The current EPA (2003) protocol, *EPA Microbiological Alternate Test Procedure (ATP) Protocol for Drinking Water, Ambient Water, and Wastewater Monitoring Methods*, provides a suitable vehicle for performing these evaluations. The EPA recommends approval of a proposed method if it is similar or better than the approved method (the "gold standard") for 80% of the matrices tested. Currently, only culture-based methods can be included as an alternate test procedure; therefore, consideration should be given by the EPA on the comparability of other methods (as mentioned above).

Along these lines, California has adopted equivalency validation between methods with different targets (i.e., culture-based versus PCR-based). This protocol, *Beta Testing of Rapid Methods for Measuring Beach Water Quality* (SCCWRP, 2007), provides guidelines for comparing between methods. Similar to EPA, this validation compares method performance between multiple sample types and laboratories and also sets acceptable variability between results at 0.5 log (based on within method variability previously reported [Griffith et al., 2006; Noble et al., 2003]). Additionally, the precision should be equal to or better than for the existing methods.

Many workgroup members felt that EPA's protocol is too prescriptive because it disallows applications for methods that are not culture-based. In the future, for example, should the IMS-ATP test be found to have a health risk-based association, EPA should consider allowing its comparison to culture-based methods since both assay for live organisms, albeit not exactly via the same mechanism (membrane-filtration colonies or Most Probable Number [MPN] results versus ATP occurrence). By the same token, workgroup members felt that the California protocol was too relaxed in that genetic methods were compared to culture-based methods for the purposes of acceptance of the former. Because these methods do not measure the same targets (DNA versus membrane-filtration colonies), this was perceived as comparing "apples to oranges" as the criteria for making such comparisons are not yet well established.

3.4 Performance Criteria

Regardless of which of the two evaluation approaches is chosen (health risk-based or method equivalency-based), performance criteria for the method should be completed, and preferably before using the method in an epidemiological study to obtain heath risk-based association data.

Workgroup members consider the following to be the major parts of performance criteria: repeatability, accuracy, specificity, sensitivity, robustness, range of applicability, and practicality. These performance criteria are summarized below.

Repeatability asks the question: if a test is repeated, will the results be the same? Note, this does not take into account the degree of error with regard to how well the test does at identifying its target (accuracy). For example, if a person is throwing darts at a target, repeatability is the measure of how often the darts hit a specific place. Repeatability does not measure whether one hits the center of the target or not—that is accuracy (see more below). Repeatability is sometimes referred to as precision and can be expressed both on an absolute scale (i.e., standard deviation) and on a relative scale (i.e., relative standard deviation [RSD]). The RSD (sometimes referred to as coefficient of variation) is calculated as the standard deviation divided by the mean, expressed as a percent. For the purpose of summarizing data, both standard deviations and RSDs should be calculated. Generally, RSDs are most appropriate for summarizing precision when variability increases as concentration increases. To provide an indication of the effect of multiple matrices on precision, standard deviations should be calculated separately for each matrix as well as for the method over all matrices. In addition to within and among matrix/matrices for repeatability, it is important to test intra- (within lab) and inter-laboratory (among labs) repeatability to ensure consistency.

Accuracy measures the degree to which the method identifies its target. It is defined as the degree of agreement between an observed value and an accepted reference value. Accuracy includes random error (precision) and systematic error (recovery) that are caused by sampling and analysis. Using the above dart example, this would be the number of times that the dart hits the "bulls-eye."

Specificity includes the false positive and false negative rates. The false positive question asks if the method is significantly more likely or less likely to detect non-target organisms or other sample constituents that would be reported as the target organism by the analyst when compared to the reference method. To assess whether the false positive rates are significant, replicates known to contain non-target organisms that could be falsely identified as the target organism should be analyzed. The determination that the samples do not contain the target organism should be based on a third independent standard method. For example, if the target organism is cultured *E. coli*, the test should be used against, at a minimum, other enterobacteria, and, depending on what the test is, potentially Gram positive organisms as well. If the test is for genetic material, then the primers and probes should be tested against GenBank to look for potential false positives from non-*E. coli* species with the same sequences. Specificity also asks the false negative question regarding whether the new method is significantly more or less likely to exhibit non-detections for samples with the target organism or to exhibit results that are biased low when compared to the reference method. To assess whether the false negative rates are significantly different between methods, replicates known to contain target organisms should be analyzed. As in false positive studies, the determination that the samples do not contain the target organism should be based on a third independent standard method. For example, if the target organism is genetic material from *E. coli*, then a method for culturable *E. coli* can be used.

If the culture method is able to detect *E. coli*, then the genetic method should, in general, also detect *E. coli*.

Estimates of false positive and negative rates as percentages can be calculated as follows:

1. false positive rate = # false positives/(# of true negatives + false positives) × 100%; and
2. false negative rate = # false negatives/(# true positives + false negatives) × 100%.

The sensitivity of a test is the analytical detection limit of the test (the smallest amount detectable using the method). For chemical methods, the sensitivity may be defined as the minimum amount of a particular component that can be determined by a single measurement with a stated confidence level. Generally, these refer to instrument analysis; thus, it is the lowest quantity of a substance that can be distinguished from the absence of that substance (a blank). For microbial methods, sensitivity is the limit of detection of a particular method. In general, methods are not used at this level since confidence around that level is lower and more subject to user error.

The robustness of a test is the degree to which the method can perform in the presence of incorrect inputs or stressed conditions. More simply, how poorly can a method perform and still produce useful results? For example, does the method perform as intended in the hands of a semi-novice user (e.g., a qPCR method performed by a person familiar with molecular-based methods including PCR but not qPCR)? If the test is for cultured microorganisms, can it detect stressed organisms in ambient waters (e.g., the EPA *E. coli* methods have a 2-hour resuscitation step at a lower temperature for stressed organisms)? Robustness is not a measurable attribute per se but must be considered and applied for overall method performance.

The range of applicability should also be considered as it answers the question: is the test reliable on a nationwide basis (e.g., does it work equally well in temperate and tropical climes, in the Great Lakes and other inland waters, etc.), in the presence of inhibitors (e.g., turbidity, alkalinity, organics [humic acids]), and in a variety of matrices (e.g., sewage, septic tanks, urban runoff, agricultural waste, known animal sources)? In general, the range of applicability does not apply to matrices other than the one for which the test was designed; that is, a recreational water quality method should not be expected to perform equally well for sewage sludge. Like robustness, this is not a measurable attribute but must be considered and applied for overall method performance.

Workgroup members felt that practicality should also be considered when considering a method. This issue is largely addressed in Chapter 7 (Implementation Realities workgroup). However, four main issues were considered important enough to be mentioned here—capital cost, training cost, per sample cost, and additional sampling requirements. Capital costs include the upfront costs such as equipment purchase and the actual space required for the test. For example, when performing genetic testing, aside from the equipment needed (e.g., platform [specific machine], laminar flow hoods, dedicated pipettors), space is needed, ideally in separate rooms, for reagent preparation (material not containing any genetic materials). Space is also needed for the two types of sample preparation, those containing high target sequence DNA concentrations such as DNA standards and calibrator samples, and those containing expected low target sequence DNA concentrations (e.g., filter blanks and water samples)—the latter of which should also be in

separate laminar flow hoods. Training costs are those incurred prior to routine testing so that the user can perform the test within the performance criteria of the test; these may include participation in a workshop for hands-on experience or completing a training module. The other two issues regard routine use of the test. A high per sample cost may become an issue if a large volume of tests need to be completed on a routine basis. Additional sampling is generally an effort that results from rapid testing. For example, if an early morning sample yields, after 4 hours, a positive result resulting in beach closure, it may then lead to additional sampling to determine if the beach still needs to be closed in mid-afternoon. It should be noted that many laboratories (at least in California) do not object to capital or training costs, but take issue with a high per sample cost or with additional sampling requirements.

3.5 Evaluation Process for Alternative (New) Indicators

Currently, recreational water quality is assessed with a single indicator with a single threshold (i.e., a "one-size-fits-all" approach). Under consideration is the implementation of alternative indicator(s) that are better associated with human health risk than the enterococci. These alternative indicators could theoretically replace the current standard but still be used in a one-size-fits-all approach or could be targeted for specific applications (e.g., one indicator may be best associated with risk in tropical marine waters, another in temperate marine waters, and another in freshwaters). Regardless of the final implementation, any new proposed indicator will need to be vetted through performance based standards.

The system of approving an alternative indicator will follow the same process as outlined for the assessment of any indicator or method, although there will be key differences.

- Any proposed indicator and/or method should be evaluated for the following performance characteristics:
  - repeatability (i.e., precision);
  - accuracy;
  - sensitivity;
  - specificity (false positive/false negative);
  - robustness;
  - range of applicability; and
  - practicality.
- *After* performance characteristics have been demonstrated and the indicator and associated method has been *determined to have adequate performance*, it then should be evaluated for its use and application in a water quality criteria, including:
  - relationships to health risks must be established based on epidemiological studies covering an array of beach types and/or geographic areas; and
  - because of lack of comparable standards, a new indicator cannot be evaluated based upon equivalency to an existing method.

This approach would establish the basis for alternative (new) indicators, and leads into the possibility that such indicators could also serve in a role as source identifiers.

3.6 Evaluating Source Identification Methods – Proficiency and Evaluation

When bacterial levels in recreational waters exceed adopted State Water Quality Standard, the potential risk to the public health requires local authorities to post advisories or close swimming areas, risking significant losses in local revenue. The goal of microbial source tracking (MST), as applied to U.S. waters, is to accurately identify the contributors and, if possible, the relative proportions of fecal pollution from all potential sources, or at least the major contributors. Proper use of MST can assist watershed managers in implementation of best management practices (BMPs) that can reduce fecal inputs, thereby limiting or reducing public health risk.

Two major classes of microbe-based and one class of chemical-based MST methods are currently being developed and utilized in surface waters across the world (Blanch et al., 2006; Stoeckel and Harwood, 2007). Although there has been significant progress in the MST field over the past decade, variability among performance measurements and validation approaches in laboratory and field studies has led to a body of literature that is very difficult to interpret, both for scientists and for end users (Stewart et al., 2003; Stoeckel et al., 2004). This section lists and defines/describes performance characteristics that should be uniformly applied across MST studies, although selection of which criteria from the following list to use will vary somewhat based on the target. All methods and MST projects need to include some considerations for representative sampling, sampling frequency, sample volumes required, and the number and choice of source categories. Although the use of a toolbox approach has been important in MST studies, there is a desire to develop an appropriate tiered approach to avoid costs and time from using multiple methods simultaneously. Within the MST community, and largely as a result of the method comparison studies, library-independent methods are currently the priority, while chemical-based methods appear to be desirable for rapid screening and presence-absence tests (with perhaps quantification in the future). Library-based methods still have a role in MST, but only in those circumstances where detailed information is needed, such as many TMDL-based studies.

3.6.1 Library-independent Methods (also Reported as Sample-level Classification)

Examples (not comprehensive) include both molecular approaches (*Bacteroidales*, *E. coli* toxin, *Enterococcus Esp* gene, direct measurement of source-specific viruses (polyoma, adenoviruses, enteroviruses, phages, etc.) and microbe-based approaches (*Clostridium perfringens* [alternative indicator], source-related clostridia, source-related enterococci, sorbitol fermenting bifidobacteria [human], *Rhodococcus copraphilus* [grazing animals], human-specific bacteriophages, phage typing, etc.).

Method evaluation includes the following eight performance criteria:

1. Accuracy is defined as the true positive or success rate—if a method identified the presence of the target in 98 out of 100 blind samples, the accuracy would be 98%;
2. Rates of false negatives and false positives of the target are used to describe specificity;
3. The analytical detection limit of the test is used to describe sensitivity;
4. The level of target-host specificity and the range of target-host distribution;
5. Efficiency of recovery of the target from different environments;

6. The reproducibility of analytical results, both inter- and intra-laboratory;
7. The suitability of marker detection (and/or quantification) to meet study-specific objectives; and
8. Detection of several of the above, especially #4 and #5, can be referred to as robustness.

3.6.2 Non-microbial Methods (also Called Chemical Methods)

Examples of non-microbial indicators include, but are not limited to, optical brighteners, host-derived DNA (e.g., eukaryotic mitochondrial DNA), fecal sterols/stanols, and source-specific fecal compounds such as caffeine and pharmaceuticals for humans.

The performance criteria in numbers 1 through 8 above, excluding #4 and #5, apply to non-microbial methods. For chemicals, the analytical detection limit of the test is usually applied to describe both sensitivity (#3) and the efficiency of recovery of the target from different environments (#5).

3.6.3 Library-based Methods (also Called Isolate Matching)

Examples of library-based methods include but are not limited to both molecular approaches (pulsed-field gel electrophoresis [PFGE], ribotyping, PCR with different primer sets, etc.) and phenotype-based approaches (antibiotic resistance analysis [ARA], biochemical, etc.).

The performance criteria in numbers 1 to 3 from library-independent methods (above) are applicable for library-based methods. In addition, the following four criteria apply:

1. Jackknife (also reported as holdout or cross-validation) analysis and the pulled-sample test (recently described as internal proficiency) should be done on each and every library (Stoeckel and Harwood 2007);
2. Library should shave clones removed to reduce redundancy, based on the precision of the typing method;
3. External proficiency or blind tests to determine both size and representativeness of the library should be done as the library is developed; and
4. The benefit-over-random statistic should be used when accuracy is determined, and should be performed on both the library and the external proficiency (or blind) set.

3.7 Modifications to the Evaluation Process When Indicators are used for Other Applications

Indicators are used in many different contexts. Routine beach monitoring, the most time-critical use of indicator bacteria is described extensively in other chapters of these proceedings. This section briefly addresses other (secondary) uses of indicators. Another use of indicators is as an early warning system that would provide evidence for an imminent human health risk, such as a sewage spill. They can also provide evidence of returning to acceptable ambient water quality conditions as designated by the criteria. It is important that the methods be highly specific and robust. Because of the potential for illness in exposed populations, it is extremely important that

this use of an indicator be associated with great specificity and robustness. Specificity in this case refers to the ability of a method to detect an indicator with certainty that the indicator is not giving a false positive response (i.e., an organism or analyte that responds similarly to the target organisms, but is not the target organism). Similarly, target microbes that do not provide a positive response are indicated as false negatives and too many of these could result in a false sense of security that would be highly unacceptable from a public health perspective. Robustness in this case means that the method can be abused and still function properly. Methods of this type are usually used under extreme conditions where the correct result must be obtained in a very short time period.

Another use of indicators is for compliance monitoring purposes, such as monitoring sewage treatment effluent for EPA's NPDES Program. Important characteristics for indicators used for such compliance monitoring are precision and specificity. The precision is necessary because sewage treatment plants would receive a fine(s) if limits of the permit are exceeded. The specificity, both false negative and false positive responses, are important for the same reason mentioned above and may influence the way beaches are managed.

Trend assessments are used to determine whether water quality conditions at a site are changing with time. The most important characteristic is precision that contributes to the ability to detect small changes over time (i.e., whether the water quality is decreasing or improving over time). If the water quality decreases then bathing may no longer be allowed. Conversely, if the water quality improves sufficiently then bathing may be re-allowed.

3.8 Research Needs

Several lines of research should be pursued in order to implement improved methods for (1) rapid detection of current water quality indicators, (2) implementing alternative indicators that are more protective of public that the current indicators, or (3) determining source (human or nonhuman) at beaches. This set of research priorities is based upon the current state of available methods and the projected feasibility of implementation in near-term (1 to 3 years) and mid-term (2 to 5 years) or longer timeframes. Although these are listed in priority order, the workgroup members felt that they largely expand on efforts that EPA or its potential partners have already initiated and all are achievable in the next 3 years. Appendix G summarizes currently planned measurements for use in the upcoming Doheny and Malibu Beach (California) epidemiology study.

1. Systematic evaluation of performance criteria for library-independent source identification methods (for use in source characterization [i.e., human versus nonhuman fecal contamination] and in MST) (timeline: 1 to 2 years).

   Workgroup members felt that EPA should fast-track studies to evaluate the performance criteria of source-specific microbial targets.

   A series of controlled trials representing a variety of geographical areas should be conducted to evaluate promising methods. Studies should include samples spiked with known source fecal matter from multiple hosts as well as environmental samples

collected from areas with known dominant sources of fecal contamination. Samples should be assayed by the test methods in several laboratories using blinded controls. These protocols would be similar to those used in the Griffith et al. (2003) studies that EPA co-sponsored approximately 5 years ago, but which need to be updated as new methods have developed and existing methods refined.

Although there are many potential methods that could be included in such studies, the workgroup members identified the following as the most important:

1. enterococci *Esp* gene;
2. *E. coli* virulence genes;
3. human enteric viruses (molecular detection);
    a. DNA-based – adenoviruses and polyomavirus;
    b. RNA-based – enterovirus and norovirus;
4. *Methanobrevibacter smithii* (*nifH* gene);
5. *Clostridium perfringens*;
6. coliphage; and
7. *Bacteroides* human-specific markers.

The last two methods are also being planned for use in EPA's upcoming (2007) health risk (epidemiological) study. The workgroup members felt that the coliphage and *Bacteroides* methods are more advanced than the others and endorses their inclusion in source identification studies.

In coordination with trials over various geographic areas, candidate methods should also be evaluated from the perspective of persistence of genetic or chemical or microbial targets in both primary and secondary habitats (sediments) over longer time periods (multi-year). Although this may be a longer term goal, eventually all methods that appear to be suitable for use regulatory or management-level decisions will need such to be examined over time periods sufficiently long so that there is confidence that the desired targets do not change, or that changes can be captured and dealt with if they do occur.

2. Evaluation of chemical indicators for human sewage (timeline: 2 to 3 years).

Several possible chemical markers of sewage have been reported and have the potential to be used in a rapid to real-time assessment of source. Coordinated studies to evaluate the performance criteria over multiple labs are needed to implement these assays.

The following analytes should be included in near term evaluation studies:

1. optical brighteners;
2. coprostanol; and
3. caffeine.

At least one multi-laboratroy evaluation study of optical brighteners is currently being developed by individual investigators (Hartel et al., 2007).

3. Continued evaluation of rapid assays for the detection of enterococci in human health risk (epidemiological) studies (timeline: 1 year and beyond).

   Rapid detection of current water quality indicators are proposed to allow same day evaluation of water quality. To implement these assays, continued evaluation of the health risk relationship is needed. For qPCR (*Enterococcus*), more epidemiological studies from a range of beach types are needed before implementation. Additionally, other rapid assays for enterococci have been developed and should be evaluated in upcoming and future epidemiological studies.

   Methods under consideration for enterococci detection include the following:

   Immediate (timeline: 1 to 2 years):

   - qPCR (detection of DNA); and
   - TMA (detection of RNA).

   Mid-term (timeline: >2 years; require additional performance evaluation):

   - IMS/ATP (detection of activity);
   - RAPTOR™ (antibody-based detection)[5]; and
   - enzymatic detection.

4. Evaluation of alternate indicator candidates in human health risk (epidemiological) studies (timeline: 1 year and beyond).

   Potential alternate indicators (i.e., to replace enterococci and *E. coli*) that have already been vetted for performance criteria should be included in any future epidemiological studies of recreational waters to determine their relationship with health risk.

   The following indicators should be evaluated within the next two years:

   - *Bacteriodales* human specific markers; and
   - F+ coliphage (antibody).

   Other candidates indicators should be added for evaluation as they meet required performance criteria (as listed above)

5. Optimization of sampling, recovery, and processing methods for efficient concentration, processing and detection of rapid, alternative or host specific indicators (Time line: 1 year and beyond).

[5] http://www.resrchintl.com/raptor-detection-system.html

Additional methods need to be optimized for source specific microbial targets. Studies should address issues such as optimization of sample volume, processing/concentration methods, and extraction/purification methods (especially for targets expected to occur at low numbers in the environment).

Furthermore, research addressing straightforward techniques to enumerate *Enterococcus faecium* and *faecalis*, rather than the larger *Enterococcus* group that is presently measured, are needed as the individual species are more likely to be associated with human sewage/feces. Performance-based criteria tests are also needed for these species.

References

Blanch, AR; Belanche-Munoz, L; Bonjoch, X; Ebdon, J; Gantzer, C; Lucena, F; Ottoson, J; Kourtis, C; Iversen, A; Kuhn, I; Moce, L; Muniesa, M; Schwartzbrod, J; Skraber, S; Papageorgiou, GT; Taylor, H; Wallis, J; Jofre, J. 2006. Integrated analysis of established and novel microbial and chemical methods for microbial source tracking. *Applied and Environmental Microbiology* 72(9): 5915-5926.

Griffith, JF; Weisberg, SB; McGee, CD. 2003. Evaluation of microbial source tracking methods using mixed fecal sources in aqueous test samples. *Journal of Water and Health* 1: 141-151.

Hartel, PG; Hagedorn, C; McDonald, JL; Fisher, JA; Saluta, MA; Dickerson, JW, Jr.; Gentit, LC; Smith, SL; Mantripragada, NS; Ritter, KJ; Belcher, CN. 2007. Exposing water samples to ultraviolet light improves fluorometry for detecting human fecal contamination. *Water Research:* In press.

Noble, RT; Weisberg, SB; Leecaster, MK; McGee, CD; Ritter, K; Walker, KO; Vainik, PM. 2003. Comparison of beach bacterial water quality indicator measurement methods. Environmental Monitoring and Assessment: 81: 301-312.

SCCWRP (Southern California Coastal Water Research Project). 2007. *Beta Testing of Rapid Methods for Measuring Beach Water Quality*. Technical Report 506. Available at: ftp://ftp.sccwrp.org/pub/download/PDFs/506_beta_testing.pdf.

Stewart, JR; Ellender, RD; Gooch, JA; Jiang, S; Myoda, SP; Weisberg, SB. 2003. Recommendations for microbial source tracking: Lessons learned from a methods comparison study. *Journal of Water and Health* 1: 225-31.

Stoeckel, DM; Mathes, MV; Hyer, KE; Hagedorn, C; Kator, H; Lukasik, J; O'Brien, TL; Fenger, TW; Samadpour, M; Strickler, KM; Wiggins, BA. 2004. Comparison of seven protocols to identify fecal contamination sources using *Escherichia coli*. *Environmental Science and Technology* 38: 6109-6117.

Stoeckel, DM; Harwood, VJ. 2007. Performance, design and analysis in microbial source tracking studies. *Applied and Environmental Microbiology* 73(8): 2405-2415.

US EPA (U.S. Environmental Protection Agency). 1986. *Ambient Water Quality Criteria for Bacteria – 1986*. EPA440/5-84-002. Washington, DC: US EPA.

US EPA. 2003. *EPA Microbiological Alternate Test Procedure (ATP) Protocol for Drinking Water, Ambient Water, and Wastewater Monitoring Methods – Guidance*. EPA-821-B-03-004. Washington, DC: US EPA.

WHO (World Health Organization). 2003. *Guidelines for Safe Recreational Water Environments. Volume 1 Coastal and Fresh Waters.* Geneva, Switzerland: WHO.