National Drinking Water Advisory Council Report

on the

CCL Classification Process

to the

U.S. Environmental Protection Agency

May 19, 2004

Foreword

This report presents recommendations from the National Drinking Water Advisory Council (NDWAC) to the Administrator of the U.S. Environmental Protection Agency. The report was prepared by the NDWAC's Work Group on the Contaminant Candidate List (CCL) Classification Process. The Work Group prepared the report for consideration by the NDWAC. After the deliberations, the Council unanimously approved and adopted the NDWAC Work Group's report on the CCL Classification Process with one minor clarification.

The Charge to the CCL Work Group was to evaluate recommendations made by the National Academy of Science's National Research Council, including methodologies, activities and analysis, and to make recommendations for an expanded approach to the CCL listing process for the purpose of protecting public health.

The Work Group's deliberations and recommendations focused on the overall implementation strategy and methodology that EPA should use to develop the CCL. In developing this report the Work Group developed recommendations on a phased adaptive management approach, identified methods to select and validate new classification approaches, and provided a framework for obtaining input from experts and stakeholders.

TABLE OF CONTENTS

TABLE OF FIGURES, EXHIBITS, AND TABLES

Executive Summary

ES.1 Introduction

The Contaminant Candidate List (CCL) Classification Process Work Group (the Work Group) was charged by the National Drinking Water Advisory Committee (NDWAC) with reviewing the National Research Council (NRC) 2001 report, *Classifying Drinking Water Contaminants for Regulatory Consideration.* The Work Group was asked to advise the NDWAC on development and application of the classification approach suggested by the NRC, including evaluating proposed and alternative methodologies. In conducting its review, the Work Group considered the large and growing number of agents that might become candidates for scrutiny in the CCL process, and the rapid expansion of information on these agents. Based on this review, the Work Group drew the following conclusions.

- There is merit in the premise of a three-step selection process as proposed by NRC for both chemical and microbial contaminants:
- *Identify the CCL Universe*
- *Screen the Universe to a Preliminary CCL (PCCL)*
- *Select the CCL from the PCCL*
- The Agency should move forward with the NRC recommendation to develop and evaluate some form of prototype classification approach.
- Expert judgment plays an important role throughout the three-step selection process, particularly in reviewing the prototype model and the new classification approach**.**
- Enhancement of surveillance and nomination processes are essential to assure a full consideration of emerging chemical and microbial contaminants.

The Work Group also identified a number of practical limitations or difficulties in developing and applying the recommended approach, and sought to advise the NDWAC on how these might be addressed.

1.1 Background and Purpose

The Safe Drinking Water Act Amendments of 1996 require the US Environmental Protection Agency (EPA, or the Agency) to publish every five years a list of chemical and microbial contaminants that are known or anticipated to occur in public water systems and that may have adverse health effects, and that, at the time of publication, are not subject to any proposed or promulgated National Primary Drinking Water Standards. The first CCL was published in 1998 and was categorized based on four priority areas in drinking water research (occurrence, health effects, treatment, and analytical methods). On a staggered, second five-year cycle (three-and-a-half years after a CCL is required), EPA is required to evaluate this research together with any already available

information and make a determination for at least five contaminants on whether or not to proceed with the regulatory development process.

The first CCL was developed based upon a review by technical experts of readily available information, and contained 50 chemical and 10 microbial contaminants/groups. EPA completed its first regulatory determination process in July 2003. EPA recognized the need for a more robust and transparent process for identifying and narrowing potential contaminants for future CCLs and requested advice from the NRC on developing such a process. In its 2001 report, the NRC proposed a broader, more comprehensive screening process to assist the EPA drinking water program to identify those contaminants for which further research – and ultimately a decision on whether or not to proceed with a regulatory development process – would be appropriate.

1.2 Charge to the NDWAC Work Group

The NDWAC's Charge to the CCL Work Group is set forth below.

> "Evaluate recommendations made by the National Research Council, including methodologies, activities and analysis, and make recommendations for an expanded approach to the CCL listing process for the purpose of protecting public health.
>
> "This may include, but not be limited to, advice on developing and identifying:
>
> i. Overall implementation strategy
> ii. Classification attributes and criteria (and methodology that ought to be used)
> iii. Pilot projects to validate new classification approaches (including neural network and other prototype classification approaches)
> iv. Demonstration studies that explore the feasibility of the VFAR approach
> v. Risk communication issues
> vi. Additional issues not addressed in the NRC Report"

1.3 Guiding Principles

The Work Group adopted the following principles to guide its work:

- *As public health is the first and foremost consideration, development and maintenance of the CCL should maximize protection of public health, including consideration of sensitive subpopulations.*
- *The CCL process should be built on the best available science, consistent with the goal of protection of public health and development of the CCL in a reasonable time frame.*
- *All aspects of the CCL process should reflect the important role of expert judgment in both establishing procedures and reviewing the results of those procedures.*
- *All aspects of the CCL process should be systematic, open, accessible and available to informed stakeholders, and well-documented so that a knowledgeable reader could understand and reproduce the process of analysis leading to specific decisions made for the PCCL and CCL.*

- *All aspects of the CCL process should apply equal rigor to chemical and microbial agents, consistent with the data available for these two categories.*
- *There should be opportunities for public involvement at all key points in the CCL process, with broad participation by affected parties.*

ES.2 Overview of Recommended CCL Classification Process and Overarching Issues

2.1 Building on the NRC Approach

In reviewing the NRC selection process, the Work Group focused on the following:

- More completely addressing the scope of the CCL "Universe" as described by NRC – with respect to both chemicals and microbes
- Identifying a robust and practical means of screening the Universe to a Preliminary CCL (PCCL)
- Evaluating the application of a prototype classification algorithm to select a CCL from the PCCL
- Ensuring that both chemical and microbial contaminants are adequately and equally considered by the CCL process
- More fully developing the role of expert judgment acknowledged by the NRC but not developed in its report
- Reviewing the NRC's call for transparency throughout the CCL process
- Expanding on the NRC model to explicitly allow for nomination of potential contaminants for consideration
- Expanding on the NRC model by explicitly encouraging the Agency to maintain the CCL process as an ongoing programmatic element, rather than as a protocol that is repeated every five years (This expansion includes the concept of surveillance for data to support the CCL process.)
- Suggesting data and information "hierarchies" that might be used in the process
- Following through on the NRC's recommendation to incorporate consideration of data quality into the CCL process
- Developing a framework for incorporating genomics and proteomics, including the NRC's Virulence Factor Activity Relationship (VFAR) concept, into the CCL process

2.2 Transparency and Public Participation

The CCL process will need to be explained so that the public can generally understand the method used to develop the CCL. Key criteria, data, and assumptions that affect inclusion or exclusion of contaminants ought to be noted, where possible, so that the reader can follow the logic regarding why decisions are made. Decision-makers, stakeholders, and drinking water consumers need to be able to understand why EPA has selected the CCL contaminants and why further research on these

contaminants is a good use of resources. The public will want to know why investment in the methods used to select contaminants and investment in research on certain contaminants is an efficient and effective use of resources that will lead to improved protection of public health. If EPA is transparent in its decision-making, the public will have the rationale needed to understand how the method works and why specific contaminants are or are not on the list.

The CCL Work Group agrees with the NRC that the EPA will need to garner public support to implement the CCL method effectively and efficiently. The Work Group recommends that EPA consider early and ongoing consultation with key stakeholders and outreach to the public as implementation proceeds. Finally, the Work Group agrees with the NRC that the public involvement program needs to be tailored to the public's needs and should start early in the process.

2.3 Overview of the CCL Process

Figure ES.1 diagrams the CCL Classification Process recommended by the NDWAC Work Group. It is a three-step process. The first step consists of the parallel development of the Microbial CCL Universe and the Chemical CCL Universe (which together constitute the "Universe" of agents identified as Step 1 in the diagram below). The second step consists of screening contaminants from the Universe of identified agents to the Preliminary Contaminant Candidate List, or PCCL. The third step is the classification of contaminants on the PCCL to produce the proposed CCL.

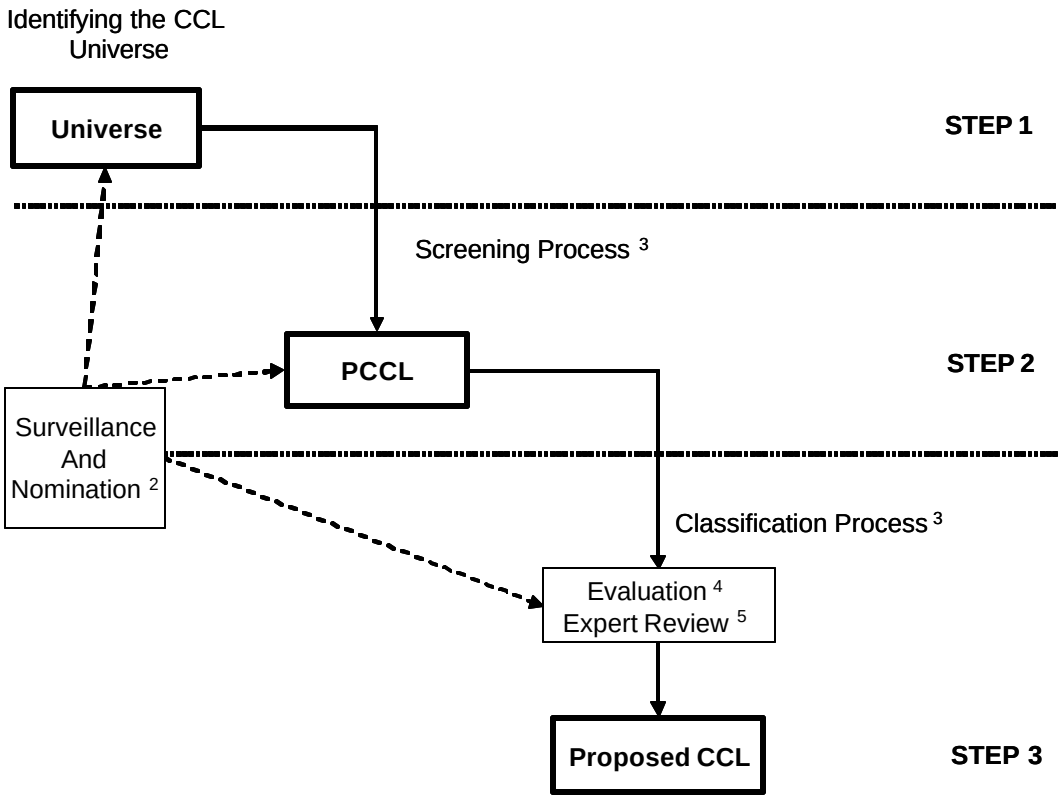
Selection of microbial and chemical contaminants through a single CCL process is mentioned in the NRC recommendations. The Work Group found that, at this point in time, there are still systematic differences in the strengths and weaknesses of the information available for chemical and microbial contaminants.

→ **The Work Group recommends that the procedure for screening and selecting CCL contaminants consist of parallel processes for microbial and chemical contaminants that meet in the formation of a single CCL, but that take best advantage of the information available for each type of contaminant.**

For the third step, classifying contaminants on the PCCL to select contaminants for the CCL, the NRC proposed five attributes, or characteristics of a contaminant that contribute to the likelihood that it could occur in drinking water at levels and frequencies that pose a public health risk.

→ **EPA should proceed initially with using the two health effects attributes and three occurrence attributes described by the NRC as input for the PCCL-to-CCL classification modeling for contaminants.**

Figure ES.1 Overview of CCL Process Recommended by the NDWAC Work Group[1]



Notes:

1. Steps are sequential, as are components of each step, with the exception of surveillance and nomination. This generalized process is applicable to both chemical and microbial contaminants, though the specific execution of particular steps may differ in practice.

2. Surveillance and nomination provide an alternative pathway for entry into the CCL process for new and emerging agents, in particular. Most agents would be nominated to the CCL Universe. Depending on the timing of the nomination and the information available, a contaminant could move onto the PCCL or CCL, if justified.

3. Expert judgment, possibly including external expert consultation, will be important throughout the process, but particularly at key points, such as: reviewing the screening criteria and process from the Universe to the PCCL; assessing the training data set and classification algorithm performance during development of the PCCL to CCL classification step.

4. After implementing the classification process, the prioritized list of contaminants would be evaluated by experts, including a review of the quality of information.

5. The CCL classification process and draft CCL list would undergo a critical Expert Review by EPA and by outside experts before the CCL is proposed.

2.4 Overarching Issues

The Work Group identified several overarching issues that must be considered in developing the CCL process. In addition to the need for transparency and public participation, these overarching issues include:

- *The integration of expert judgment throughout the CCL process*
- *Active surveillance and nomination/evaluation processes for new and emerging agents*
- *Information quality considerations*
- *The application of an adaptive management approach to implementing the CCL process*

The approach to address these overarching issues is intended to be consistent with the Work Group's guiding principles.

Integrating Expert Judgment into the Process. NRC recommendations include provisions for "expert" and "scientific review" in the CCL process but provide little guidance as to what, how, and when such review would be used. Like the NRC panel, the Work Group observed that expert judgment is inherent throughout the development of the CCL process and in implementing that process once it is developed. Critical reviews, involving various types of expert consultation and collaboration, up to and including more formal expert reviews, will be useful at key points in the new, evolving CCL process outlined in Figure ES-1.

→ **There are several key milestones in the CCL process where a critical review would be especially relevant:**

- ***In Step 2, to review the screening criteria and their application to screen agents from the CCL Universe to the PCCL;***
- ***In Step 3, during development of the classification process from the PCCL to the CCL, to assess the training data set(s), assess the performance of the classification algorithm(s) tested, and to determine whether that performance is sufficient to justify immediate use of the algorithm(s) or suggests the need for further development;***
- ***After the classification process is implemented, to evaluate the prioritized list of contaminants, including a review of the quality of information, to provide judgments on the proposed draft listing;***
- ***The CCL classification process and draft list should undergo a formal expert review, including external experts, before the CCL list is proposed.***

Surveillance and Nomination/Evaluation Processes. The Work Group believes that a surveillance process will prove to be an important and necessary component to ensure timely identification of information relevant to new and emerging agents.

→ **The Work Group recommends that EPA establish an active surveillance process to provide identification of new and emerging agents for the CCL.**

→ **The Work Group recommends that EPA develop a nomination and evaluation process for new and emerging agents, to enable agencies and interested stakeholders from the public and private sectors to nominate agents for consideration in the CCL process.**

It is envisioned that surveillance and nomination would be integral components of the CCL process, providing an alternative pathway for entry into that process rather than a separate process. Typically, agents identified by the surveillance process would be nominated for placement in the CCL Universe, not on the CCL. However, depending on the timing of the identification of the new and emerging agents (in relationship to CCL publication schedule), and the nature of the information about them, contaminants could move onto the PCCL, or even onto the proposed CCL through an expert review process, or (if justified) through an accelerated Agency decision-making process.

Information Quality Considerations. It would be expected that, for many of the agents initially selected for consideration in the CCL process, the available data would consist of various types with different characteristics and robustness. The Work Group also recognized that the data or information used to select the CCL will be more detailed and comprehensive than the data used to identify the CCL Universe. Additionally, the CCL process will apply more scrutiny to contaminants when selecting the CCL than when screening the Universe of agents to identify contaminants for the PCCL. To address the variability of the disparate types of data, it is essential that the nature of the data used to support these steps be documented for review in the later steps of the CCL process. This characterization should identify the data sources, the methods used to derive the data, what quality assurance procedures were in place during data gathering, processing, or analysis, and whether the data characterize "demonstrated" or "potential" occurrence or health effects. In selecting the CCL, the nature and type of information should be considered further and in a manner that is consistent with the development of the prototype classification algorithm. The Work Group emphasized that it is important for EPA to develop and document appropriate data quality approaches as part of the adaptive management approach to implementation discussed below. EPA should establish data quality approaches applicable throughout Steps 1 through 3 prior to identifying the CCL Universe.

Adaptive Management Approach. The Work Group proposes an adaptive management approach. Adaptive management principles could be applied in the development, implementation, and refinement of the three-step CCL method, particularly in the initial phases of implementation. This process incorporates systematic and continual integration of design, management, and monitoring, which would enable EPA to make informed adjustments and adaptations. This process incorporates systematic and continual integration of design, management, and monitoring, which would enable EPA to make informed adjustments and adaptations, resulting in an improved method based on experience from the outcomes of successive generations of implementing the Universe-to-CCL approach.

These overarching issues are discussed in more detail in Chapter 2.

ES.3 CCL Classification Approach for Microbial Contaminants

The Work Group evaluated the differences in chemical and biological characteristics of demonstrated and potential water contaminants. The conclusions suggest that identifying the Microbial CCL Universe and screening the set of biological agents to a PCCL can be consistent with the NDWAC's proposed principles for chemicals but will require different data sources and data

elements, and may require more involvement from experts than the approach described for chemical agents and contaminants. The Work Group's recommendations for a classification approach to microbial contaminants are summarized as follows.

→ **The NDWAC Work Group recommends that the Microbial CCL Universe be based on the evaluation of data sources and literature reviews that identify organisms known or suspected to cause human disease.**

→ **The Work Group recommends that the selection of human pathogens for the PCCL start with a Microbial CCL Universe of recognized human pathogens (e.g., the amended Taylor et al. 2001 list), and that those pathogens known to be associated with source water, recreational water, and drinking water be selected for inclusion into the PCCL.**

→ **The Work Group supports the following concepts for EPA's consideration as they develop future CCLs:**

- ***Biological characteristics should be recognized as legitimate criteria for screening pathogens for the PCCL.***
- ***The list of pathogens inhabiting the Microbial CCL Universe should be screened for biological characteristics promoting or mitigating against survival and transmission in water.***

→ **The Work Group recommends that organisms associated with opportunistic infections be excluded from the PCCL unless clinical, epidemiological, or similar information implicates them as the potential or known cause of waterborne disease. The Work Group suggests that EPA increase surveillance for infections caused by these organisms, especially in sensitive subpopulations.**

→ **EPA should review public health surveillance techniques, in conjunction with the Center for Disease Control (CDC), with a view to making those techniques as proactive, robust, and effective as possible in identifying the occurrence of waterborne or watershed disease outbreaks and the organisms associated with those outbreaks.**

The Work Group also evaluated the use and potential of VFARs for the CCL process. Genomics and proteomics are recognized as powerful tools for the elucidation of pathogenic mechanisms but the technology is yet largely unproven for CCL application.

→ **The Work Group recommends that EPA should monitor the data and information that emerge as genomics progresses and integrate them for consideration in the CCL process. The process should be updated and maintained in a continuing process and verified against expert opinion. The Work Group recommends that EPA monitor the progress of genomics and the related technologies and integrate them into the CCL process, as feasible.**

ES.4 CCL Classification Approach for Chemical Contaminants

The recommended process contains three distinct steps: (1) building the Universe of chemical agents; (2) screening contaminants from the Chemical CCL Universe to the PCCL; and, (3) moving contaminants from the PCCL to the CCL.

4.1 Building the Chemical Universe

After review of NRC's recommendations, available data sources, and consideration of the potential scope of the Universe of known chemical agents, the Work Group recommends EPA adopt a principles-based approach, consistent with that described by the NRC.

→ **EPA should use the inclusionary principles as the foundation for identifying the Chemical CCL Universe. These principles are as follows:**

- ***The Chemical CCL Universe should include those agents that have demonstrated or potential occurrence in drinking water; or***
- ***The Chemical CCL Universe should include those agents that have demonstrated or potential adverse health effects.***

The Work Group recommends a strategy of accessing discrete databases to retrieve various, unique sets of records with multiple selection criteria, a process known as a "data source compilation approach." The Work Group further recommends supplementing this iterative information retrieval process with surveillance and nomination processes that provide alternative pathways into the CCL classification process.

4.2 Screening Contaminants from the Universe to the PCCL

The Work Group proposes EPA develop a screening process that relies on widely available data elements that reflect certain aspects of both health effects and occurrence.

→ **The Work Group recommends that the screening criteria and methods be:**

- ***capable of assessing as many of the contaminants in the CCL Universe as possible, even those with limited data;***
- ***as insensitive as possible to data limitations;***
- ***as simple as possible, to require fewer resources and less time;***
- ***capable of identifying those contaminants of greatest significance for further consideration; and,***
- ***to the extent feasible in light of the significant differences in availability of data for chemicals and microbes, as similar as possible to the microbial approach.***

→ **The Work Group recommends that a limited set of data elements that are widely available and that represent important characteristics of health effects and occurrence be used as the basis of the screening to select contaminants from the Chemical CCL Universe.**

Chapter 4 (section 4.2.3) details the Work Group's analysis and recommendations for a "workable approach" to screening the Chemical CCL Universe using widely available data elements for health effects (where the data element with the most "health-protective" value would ultimately be used in the screening process) and for occurrence (where the data element with the greatest occurrence frequencies and highest contamination levels would ultimately be used in the screening process).

→ **The Work Group recommends that the contaminants that are screened to the PCCL be those for which values for data elements for both health effects and occurrence reach a level of concern, based on the screening process, for inclusion on the PCCL. Generally, neither alone would be sufficient under this screening process.**

At the same time, the Work Group recognizes that there are likely to be contaminants that are highly toxic but have low potential for exposure or that have high potential for exposure but do not appear to be highly toxic. The Work Group recommends that EPA use a supplemental assessment to identify such agents that should be further investigated and perhaps should be included on the PCCL.

→ **The Work Group recommends that EPA allow expert judgment to be used to correct mistakes or oversights that will arise from this relatively simple process. It will likely be appropriate to add some number of contaminants to the PCCL that pose a concern but that do not fit the process outlined. The Work Group recognizes that unforeseen circumstances will arise, and recommends that EPA allow for supplemental consideration to address them.**

The Work Group considered a number of other issues specific to the classification of chemical contaminants.

→ **The Work Group recommends that "tags" be used to retain information about the sources of values used in the screening process and that this be done in such a way as to preserve this information for later steps in the process. The tags should identify values derived from models such as QSARs. The tags should also identify what combination of "demonstrated" and "potential" values for health effects and occurrence were used.**

→ **The Work Group recommends that, as the Agency develops approaches to screen chemical agents from the Universe to the PCCL, it should consider a range of options both for using data element values in the screening process and for establishing appropriate screening criteria to select PCCL contaminants. The screening method developed should be practical and transparent, and should efficiently screen the Universe to the PCCL. The method should also employ a level of precision that appropriately characterizes the nature and type of information used. While the Work Group discussed several options and identified their advantages and disadvantages, it did not recommend a single approach.**

ES.5 Moving from the PCCL onto the CCL

The Work Group discussed structured decision approaches to select the CCL from the PCCL. Some of the structured decision approaches discussed, particularly classification algorithms, require as inputs some specific measures of the attributes that characterize a contaminant's known or potential health risks. These specific measures could be either the actual

values reported in the scientific literature (such as water concentration measurements or Reference Dose values), or generated values or "scores" based on the actual values reported in the literature to characterize the attributes. The Work Group discussed several methods to quantify attributes. Each of these quantification methods presents a set of benefits and challenges particular to the method. The Work Group did not develop specific recommendations for quantifying attributes or a preferred structured decision approach. The Work Group does provide a series of general recommendations for the Agency to use as a framework to develop and evaluate the attributes and classification process to select contaminants from the PCCL.

5.1 Quantifying Attributes for Use as Inputs to Classification Models

The Work Group considered two basic approaches to quantifying attributes.

1) Using the actual quantitative value or measurement provided by the data element to quantify the attribute.

2) Scoring attributes, using a set of rules to convert the data element values to either:

   - a normalized numerical score with continuous values allowed within the given scoring range; or
   - a limited set of categorical scores within a given range.

The NDWAC Work Group did not reach a conclusion regarding which approach to quantifying attributes is preferred, and therefore does not make a specific recommendation favoring one over the other. Some of the recommendations that follow refer to aspects of attribute scoring and are therefore relevant where EPA determines that attribute scoring is the preferred approach.

→ **Attribute scoring protocols for contaminants should accommodate multiple data sources and a variety of data elements that may be available to score contaminants on the PCCL.**

→ **Attribute scoring across different types of data elements for a given attribute should be consistent and allow for a meaningful comparison among scored PCCL contaminants.**

→ **EPA should systematically refine and improve upon the details of the attributes as more experience is gained, including refinements and improvements in gathering and processing the needed data and information to score the attributes and with respect to using the attribute scores in the selected classification approach. Further refinements may include reducing the number of attributes; other refinements may pertain to the data elements used to score the attributes, the scoring protocols, and the actual attribute scoring process itself.**

→ **EPA should generate and include, along with the actual values or the attribute scores that are generated, descriptive "tags" that provide additional data quality information that may be used by experts reviewing the data, attribute scores and/or the PCCL-to-CCL classification modeling results.**

→ **If attribute scoring is used, the scoring system selected by EPA for each attribute should enable discrimination among contaminants, and there should be sufficient number of scoring categories so that information loss during characterization of contaminants is limited. At the same time, the scoring categories should not be so numerous that they convey a false sense of precision.**

→ **If attribute scoring is used, the scoring protocols should be transparent and straightforward.**

5.2 Overview of Classification Approaches and Work Group Recommendations

→ **The Work Group recommends that EPA pursue development of a prototype classification algorithm (*a posteriori* approach) for selecting contaminants for the next CCL. The Work Group recommends moving forward to develop and test one or more prototype models as tools to be used with expert judgment for decisions on classifying contaminants for future CCLs.**

The Work Group did not have time to evaluate the alternatives and recommend a particular prototype model. It may be useful to have several models that are used in concert to corroborate results. Also, it may be necessary to develop separate models for chemical and microbial contaminants, or models that differentiate chemicals and microbes within the model structure. *The development of any model should be an adaptive process, and should be reviewed by experts, with consideration given to updating the training data set, with each successive CCL cycle.*

→ **The Work Group recommends that the entire model development process be as transparent as possible. The development process should be viewed as iterative, and EPA should involve experts and allow opportunities for meaningful public comment on the evaluation.**

→ **EPA should use another approach for selecting CCL contaminants in the near term (i.e., for CCL3) if there are difficulties in the model development process that cannot be overcome.**

→ **The Work Group recommends that experts should be involved throughout the process of narrowing a PCCL to a CCL, specifically as advisors in the design of an approach, development of a training set, scoring of contaminant attributes, evaluation of algorithm results, and ultimate selection of CCL contaminants.**

5.3 Development of a Training Data Set

There are several issues to consider in the selection of a "training data set" used to inform the decision-making tool, or algorithm. With respect to training data sets, the Work Group makes the following recommendations.

→ **The training data set should consist of contaminants (and corresponding decisions to "list" or "not list" each contaminant) that reflect technically sound, consistent judgments about what should and should not be included on the CCL.**

- → **The training set should include contaminant attribute data that are distributed throughout the attribute space, and the training set should be selected to define the discriminant surface (the function that defines "include" and "exclude" decisions) as precisely as possible.**

- → **The Work Group recommends that EPA maintain transparency and clarity when developing the training data set. To the extent feasible, EPA should document training data set development and communicate its rationale for assigning decisions to training set contaminants.**

- → **The rationale for the number and distribution of training set contaminants should be described. Quantitative rationale should be expressed for the prototype classification approach.**

These are important considerations for determining if the training set and models have been adequately developed to begin processing PCCL contaminants. The rationale should include a description of the methods used for calibration and validation, and measures used to assess goodness of fit, such as misclassification rates.

ES.6 Summary

EPA should proceed with the development of prototype classification methods. The NDWAC Work Group identified several overarching principles that EPA should use in developing a CCL. These include the use of experts at key steps to allow for technical checks on the process, and nomination and surveillance processes that provide an alternative pathway for contaminants to enter the CCL classification process when new information surfaces.

To classify chemicals, the Work Group recommends a three-step process that includes defining and building the Universe of chemical agents, screening from the Chemical CCL Universe to create a PCCL, and developing a CCL from the PCCL. For microbes, a somewhat different but parallel process is recommended, involving identifying a Universe from a list of known human pathogens, reducing this list to a PCCL based on habitat and biological properties indicative of a pathogen's ability to be transmitted via water, and developing a CCL from the PCCL.

In making its recommendations, the Work Group identified a number of practical limitations or difficulties in developing a classification approach. These limitations are outlined in the report, and will require additional work to resolve, however the NDWAC Work Group's assessment concludes that there is merit in the NRC-proposed process for classifying microbes and chemicals, and that the Agency should move forward in pursuing the approach outlined in this report.

Chapter 1

Introduction

1.1 Background on the Contaminant Candidate List and the National Research Council Recommendations

The Safe Drinking Water Act Amendments of 1996 require the US Environmental Protection Agency (EPA) to publish every five years a list of chemical and microbial contaminants that are known or anticipated to occur in public water systems and that may have adverse health effects, and that, at the time of publication, are not subject to any proposed or promulgated National Primary Drinking Water Standards. The first Contaminant Candidate List (CCL) was published in 1998 and was categorized based on four priority areas in drinking water research (occurrence, health effects, treatment, and analytical methods). On a staggered, second five-year cycle (three-and-a-half years after a CCL is required), EPA is required to evaluate this research together with any already available information and make a determination for at least five contaminants on whether or not to proceed with the regulatory development process.

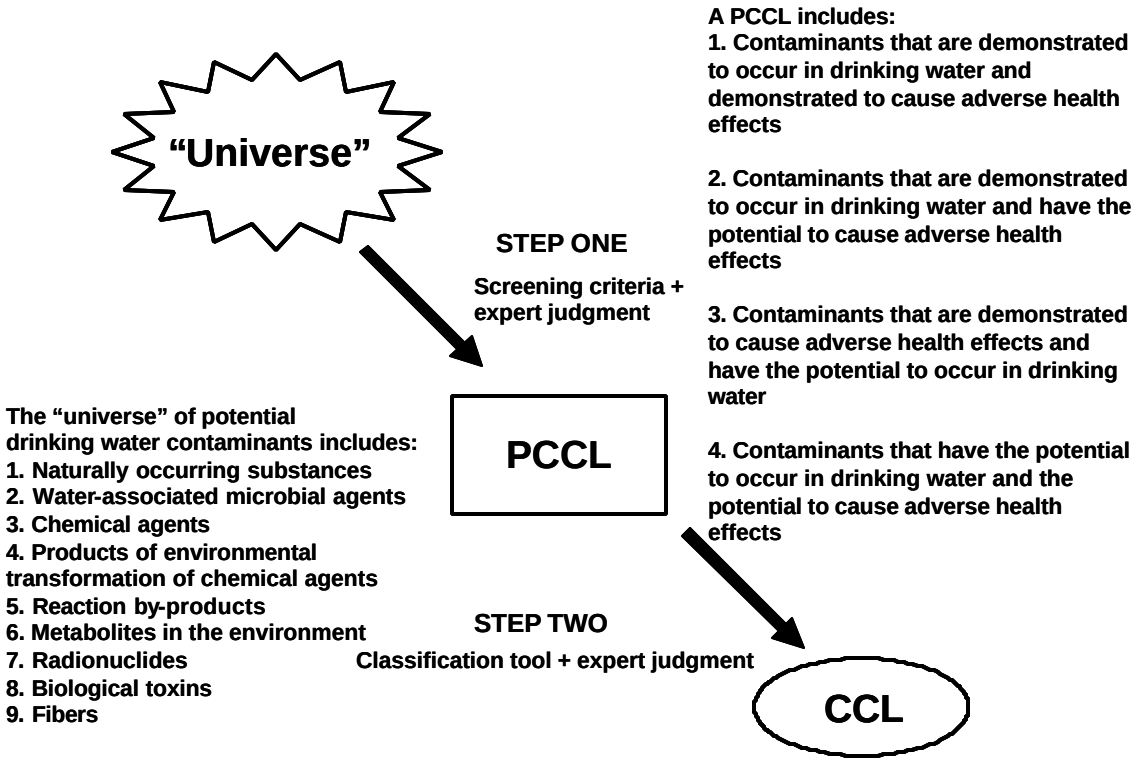
The first CCL was developed based on the review by technical experts of readily available information and contained 50 chemical and 10 microbial contaminants/groups. EPA recognized the need for a more robust and transparent process for identifying and narrowing potential contaminants for future CCLs and requested advice from the National Academy of Sciences National Research Council (NRC) on developing such a process. In its 2001 report, *Classifying Drinking Water Contaminants for Regulatory Consideration,* the NRC proposed a broader, more comprehensive screening process to assist the EPA drinking water program to identify those contaminants for which further research – and ultimately a decision on whether or not to proceed with a regulatory development process – would be appropriate.

1.1.1 The NRC Recommendations

The NRC's major recommendations are summarized in the following excerpts from the Executive Summary (pages 4-6) of the 2001 report.

> *"The committee continues to recommend that EPA develop and use a two-step process for creating future CCLs as illustrated in Figure ES-1 [reproduced below as Figure 1.1]."*

Figure 1.1 - NRC Proposed "Two-Step" CCL Process[1]



> *"The committee also continues to recommend that this two-step process be repeated for each CCL development cycle to account for new data and potential contaminants that inevitably arise over time…*
>
> *The committee recommends that the process for selecting contaminants for future CCL(s) be systematic, scientifically sound, and transparent. The development and implementation of this process should involve sufficiently broad public participation."*

The NRC recommended that a broadly defined universe of potential drinking water contaminants be identified, assessed and culled to a preliminary CCL (PCCL) using simple screening criteria and expert judgment. All the contaminants on the PCCL would then be assessed in more detail using a prototype classification tool, in conjunction with expert judgment, to evaluate the likelihood that they could occur in drinking water at levels and at frequencies that pose a public health risk and move onto

[1] The two steps referred to in the NRC report are 1) screening the universe of potential contaminants to generate a preliminary CCL, or PCCL; and, 2) refining the PCCL to produce a CCL. However, because the NDWAC Work Group elaborated further on the NRC's concept of a "Universe" of potential contaminants and on how to identify its scope and contents, this report generally refers to the NRC approach as a "three-step" process (except where quoting directly from the NRC report).

the CCL. NRC recommendations associated with specific steps in the CCL process are discussed in subsequent chapters of this report, along with the NDWAC Work Group's deliberations and recommendations. A detailed listing of the NRC recommendations is provided in Appendix A.

The NDWAC Work Group deliberated these major recommendations and some of the issues relevant to ensuring the use of best available science and assisting in the transparency and communication of the CCL – issues that should be considered in the evaluations to move a contaminant from a broad Universe of potential drinking water contaminants and onto the CCL.

1.1.2 Charge to NDWAC Work Group

With the NRC recommendations in hand, the Office of Ground Water and Drinking Water turned to the National Drinking Water Advisory Committee (NDWAC) to provide advice on different aspects of the staged approach recommended in the NRC report and to work out how this could be implemented. The NDWAC formed a Work Group on the Contaminant Candidate List Classification Process (Work Group) to evaluate the NRC recommendation and report back to the full NDWAC. The Charge to the CCL Work Group is set forth below.

> "[To] Evaluate recommendations made by the National Research Council, including methodologies, activities and analysis, and making recommendations for an expanded approach to the CCL listing process for the purpose of protecting public health.
>
> "This may include, but not be limited to, advice on developing and identifying:
>
> i. Overall implementation strategy
> ii. Classification attributes and criteria (and methodology that ought to be used)
> iii. Pilot projects to validate new classification approaches (including neural network and other prototype classification approaches)
> iv. Demonstration studies that explore the feasibility of the VFAR[2] approach
> v. Risk communication issues
> vi. Additional issues not addressed in the NRC Report"

1.2 Convening and Membership of the NDWAC CCL Classification Work Group

On June 19, 2002, the Federal Register published a notice announcing the formation of the NDWAC CCL Classification Process Work Group and requesting nominations to the group. During the convening process, several areas of expertise were identified as important for the Work Group, including computer modeling, epidemiology, contaminant occurrence, statistics, toxicology, chemistry, microbiology, risk analysis, risk communication, water system operation, and public health. The convening process sought to identify candidates with expertise in these areas as well as individuals to represent the views of several stakeholder groups, including the water industry, environmentalists, the public health community, rural water systems, and local elected officials. From

[2] Virulence-factor activity relationships

among the candidates identified, EPA and the chair of the National Drinking Water Advisory Council selected individuals to serve as members of the Work Group. Part way through the process two members resigned from the group because of changes in their work obligations. The final membership of the Work Group was as follows:

Dr. Laura Anderko, University of Wisconsin, Milwaukee

Dr. Richard Becker, American Chemistry Council

Dr. Douglas Crawford-Brown, University of North Carolina Chapel Hill

Dr. Michael Dourson, Toxicology Excellence for Risk Assessment

Dr. Alan Elzerman, Clemson University

Dr. Jeff Griffiths, Tufts University

Dr. Wendy Heiger-Bernays, Boston University School of Public Health

Mr. Buck Henderson, Texas Commission on Environmental Quality, Association of State Drinking Water Administrators

Dr. Nancy Kim, New York State Department of Health

Mr. Ephraim King, U.S. Environmental Protection Agency

Ms. Carol Kocheisen, National League of Cities

Mr. Gary Lynch, Park Water Company

Mr. Ken Merry, Tacoma Water of Tacoma Public Utility

Mr. Brian Ramaley, Newport News Waterworks

Dr. Graciela Ramirez-Toro, Centro de Educación, Conservación y Interpretación Ambiental, Interamerican University, Puerto Rico

Dr. Craig A. Stow, University of South Carolina

Dr. O. Colin Stine, University of Maryland, Baltimore

Mr. Ed Thomas, National Rural Water Association

Ms. Lynn Thorp, Clean Water Action

Dr. Daniel Wartenberg, University of Medicine & Dentistry of New Jersey – Robert Wood Johnson Medical School

The Work Group was supported by a team of technical consultants and EPA staff. The technical consultants included Amy D. Kyle, PhD MPH, Consulting scientist, health and environment; Doug Owen, Malcolm Pirnie, Inc.; Jeff Rosen, Perot Systems Environmental Services; Paul Rochelle, Metropolitan Water District of Southern California; and Steve Via, American Water Works Association (AWWA). George Hallberg, JoAnne Shatkin, Frank Letkiewicz, Nelson Moyer, and other staff from the Cadmus Group, Inc. also served on the technical team as contractors to EPA. Facilitation was provided by Abby Arnold, Sara Litke, and other staff from RESOLVE, and the document was edited by Susan Savitt Schwartz.

1.3 NDWAC CCL Classification Process Work Group Guiding Principles

Early in their deliberations, the Work Group adopted the following principles to guide their process:

- *Public health is the first and foremost consideration. Development and maintenance of the CCL should, to the extent possible, maximize protection of public health. Full consideration should be given to sensitive subpopulations.*
- *The CCL process should be built on a foundation of science, and explicitly state and explain the rationale for adoption of assumptions and estimates when these are used in lieu of actual data.*
- *All aspects of the CCL process should be systematic and scientifically sound and should maximize transparency, while acknowledging that expert judgment also will be necessary; when expert judgment is used, it should be clearly identified.*
- *Ultimately, the CCL process should be described and documented to such an extent that a knowledgeable reader could understand the rationale for why a contaminant would be on or off the "Universe," PCCL, or CCL. Ultimately, it should be clear which decisions are based on expert judgment, science, policy considerations, or other considerations.*
- *The CCL process should apply equal rigor to chemical and microbial contaminants from a public health perspective.*
- *The CCL decision-making process must be open, accessible, and available to all informed stakeholders, including the interested general public as well as the professional and scientific community and all directly affected parties.*
- *Consistent with the authority under which the CCL Work Group was formed, the group encourages the opportunity for public involvement throughout the entire process. Broad participation that is representative of the range of affected and interested parties is to be encouraged, thereby incorporating public values, viewpoints, and principles into the process.*
- *As much as possible, the goals and objectives of the CCL process should guide information and data collection. EPA should clearly communicate these goals so that the desired types of data can be identified and developed for future CCLs by sources other than solely EPA. In addition, EPA should articulate the types of data and data elements preferred for developing the CCL and how those data may affect the selection of contaminants for the CCL.*

1.4 Summary of the NDWAC CCL Work Group Deliberation Process

The Work Group met in plenary ten times in Washington, DC: September 18-19, 2002; December 16-17, 2002; February 5-6, 2003; March 27-28, 2003; May 12-13, 2003; July 16-17, 2003; September 17, 2003; November 13-14, 2003; January 22-23, 2004; and March 4-5, 2004. At the first meeting the Work Group heard an overview of the NRC recommendations from NRC committee members who had helped to develop the recommendations. Based on this overview the Work Group began to identify issues to address and formed several activity groups to focus on specific aspects of the NRC proposed CCL process. The Work Group agreed to follow a fairly detailed work plan designed by Work Group members and the technical team. The work plan proposed to address various

issues in parallel. Detailed meeting summaries for each meeting are available on the EPA website [http://www.epa.gov/safewater/ndwacsum.html].

All Work Group meetings were open to the public and announced in the *Federal Register*. At each plenary meeting as well as by conference call and in activity groups, the Work Group reviewed the components of the proposed NRC approach and examined their feasibility through various analyses. For each component of the proposed approach, the remaining chapters of this report summarize the questions considered by the Work Group; the analyses conducted to explore the questions; key points discussed; and the Work Group's recommendations and rationale for the recommendations.

The Work Group was not able to address all 23 recommendations included in the NRC report, (see Appendix A of this report), but rather focused on specific aspects of the NRC recommendations. Also, within the Work Group's deliberations, some topics received extensive discussion and analysis while others were less extensively debated. This uneven consideration of the NRC recommendations should not be construed either as an endorsement or as a refutation of the NRC recommendations that are not specifically addressed. Likewise, very general recommendations tend to reflect one of the following situations:

1) Resolution of some topics will require time and resources on a scale required for the Agency's actual implementation of the CCL process (e.g., development of training data sets). The Work Group's schedule did not allow this level of involvement.

2) Specific guidance was inappropriate, as the Agency's actions will in reality need to reflect the success of intermediate actions toward reaching recommended objectives.

3) Given the information available to the Work Group and the schedule, detailed recommendations were not developed for all the NRC recommendations.

1.5 Role of the CCL in Protecting Public Health and Implications of Inclusion on the PCCL or CCL

To understand the process proposed in these recommendations, it is useful to consider the role of the CCL in the protection of public health and what it means for a contaminant to be placed onto or left off the PCCL or the CCL. The Safe Drinking Water Act (SDWA), as amended in 1996, requires the EPA to publish a list of contaminants that are known or anticipated to occur in public water systems, and which may require regulation under the SDWA [section 1412(b)(1)]. The SDWA, as amended, also specifies that EPA must publish this list of contaminants (Drinking Water Contaminant Candidate List, or CCL) not later than 18 months after the date of enactment (i.e., by February 1998), and publish a new CCL every five years thereafter. The SDWA requires that the list of contaminants include those which, at the time of publication, are not subject to any proposed or promulgated national primary drinking water regulation (NPDWR). The list must be published after consultation with the scientific community, including the Science Advisory Board, after notice and opportunity for public comment, and after consideration of the occurrence database established under section 1445(g). The unregulated contaminants considered for the list must include, but not be limited to, substances

referred to in section 101(14) of the Comprehensive Environmental Response, Compensation, and Liability Act of 1980 (CERCLA), and substances registered under the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA).

Contaminants on the CCL are evaluated to determine what additional data are needed and to identify the next steps for each contaminant. Contaminants requiring additional data on their occurrence, health effects, treatment, or analytical methods become research priorities to develop sufficient data to support a regulatory determination. When sufficient data are available, regulatory determinations evaluate the extent of exposure and potential public health protection to populations via drinking water and whether or not initiation of a regulatory process is appropriate. The Agency may determine that an appropriate action is development of health advisories, NPDW regulations, or no action. The precepts for guiding EPA in making regulatory determinations for a drinking water contaminant are included in Section 1412(b)(1)(A) of SDWA. This section of SDWA requires EPA to consider the following three evaluation criteria prior to making a regulatory decision:

1) potential adverse health effects from the contaminant;

2) occurrence of the contaminant in public water supplies with a frequency and at levels of public health concern; and

3) whether regulation of the contaminant would present a meaningful opportunity for health risk reduction for persons served by public water supplies.

Figure 1.2 illustrates the regulatory process in its entirety and provides a schematic of the CCL process for identifying potential drinking water contaminants in relation to the development of NPDWRs. Below the dashed line are the components deliberated by the Work Group. Further detailed consideration of health effects and opportunities of actual public health risk reduction are assessed through the subsequent steps of the regulatory development process and were not part of the charge to the Work Group (i.e., regulatory determinations, development of MCLGs, and development of MCLs). The Work Group focused on the immediate objective of the CCL process.

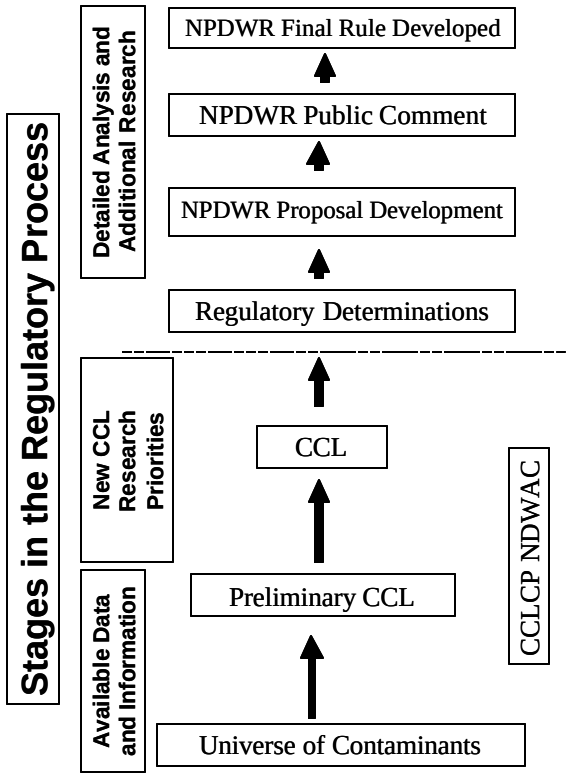


Figure 1.2 - Overview of the Regulatory Process

We note finally that limited societal resources mean that a finite number of contaminants can move through the CCL process, and that uncertainties generally may be large in risk estimates prior to the detailed process of regulatory development. The CCL should identify a set of priority contaminants that pose risk to public health. Simply increasing the size of the CCL may not contribute to improving public health. Listing a contaminant on the CCL means that the assembled data indicate it has properties generally indicative of significant risk and/or suggestive of the need for future research aimed at clarifying those suggestive properties. It also means that, relative to other contaminants on the PCCL, a contaminant on the CCL has evidence that is more indicative or suggestive of risk, and therefore offers a greater possibility of improving public health through allocation of resources to better understand and/or control that risk. The CCL should identify those contaminants whose existing evidence indicates either: 1) that a subsequent risk calculation might produce risks warranting regulatory attention; or, 2) that the contaminant has a high measure of either occurrence or effects, and that subsequent research on the other component might be expected to yield the information needed for an estimate of risk.

The PCCL aids in narrowing the pool of candidates for the CCL. Placing a contaminant on the PCCL means that there are aspects of existing data, not necessarily conclusively validated, that suggest significant risk and warrant resources to clarify this suggestion. It also means that, relative to other agents in the Universe, a contaminant on the PCCL has evidence that is more indicative or suggestive of risk, and is more likely to retain these characteristics after available data and information are assembled, and employed as part of the process leading to the CCL. Placing a contaminant on the PCCL does not mean it is established to pose a significant risk or has characteristics that are fully indicative of significant risk that would warrant concern or justify further research. The PCCL is simply an intermediate resting place for contaminants that will be scrutinized in more detail.

Chapter 2

Overview of Process and Overarching Issues

This chapter presents a brief overview of the CCL Classification Process recommended in this report. It emphasizes the importance of making that process transparent to the public, and highlights the ways in which the NDWAC Work Group's recommendation builds upon that of the NRC. The chapter also presents a discussion of overarching issues – issues that affect or apply to more than one aspect of the CCL process.

- *Transparency and public participation*
- *Integration of expert judgment into the CCL process*
- *Active surveillance and nomination/ evaluation processes for new and emerging agents*
- *Information quality considerations*
- *The use of quantitative structure activity relationship (QSAR) models*
- *The application of an adaptive management approach to implementing the CCL process*

The approach to address these overarching issues is intended to be consistent with the Work Group's guiding principles (discussed in Chapter 1).

Definitions

- ***agent:*** *any physical, chemical, or biological substance.**
- ***known agents:*** *physical, chemical, or biological substances that have been identified in the technical literature and adequately characterized to enable a judgment regarding their inclusion in the CCL Universe.**
- ***emerging agents:*** *a subset of known physical, chemical, or biological substances previously evaluated as not requiring inclusion in the CCL Universe, for which new information becomes available which heightens concern and triggers re-evaluation.**
- ***new agents:*** *physical, chemical, or biological substances that are or may be newly-discovered or synthesized, for which little is known about their potential occurrence or adverse health effects.**
- ***contaminant:*** *contaminant is defined similar to agents, as any physical, chemical, or biological substance in water. For this report the Work Group used contaminant to indicate any agent for which data exist that suggests that the agent belongs on the PCCL.*
- ***attributes:*** *characteristics of a contaminant or potential contaminant that contribute to the likelihood that a particular contaminant or related group of contaminants could occur in drinking water at levels and frequencies that pose a public health risk.*

**** See further discussion in Chapter 4.1.3***

2.1 Transparency and Public Participation

Chapter 2 of the NRC report makes clear that, to achieve acceptance, the CCL classification process adopted by the EPA *"needs to be based on sound science, risk perception, social equity, legal mandates to consider the risks of vulnerable populations, and the proper role of transparency and public perception."*

2.1.1 Why Transparency is Important for the CCL

Like the NRC, the CCL Work Group believes that the credibility of the CCL methodology EPA adopts depends on sound science. The method will need to withstand peer review or scientific critique, in which scientists can take the same information and test conditions and achieve comparable

results. Acceptance also will depend on how the method is developed and how transparent – i.e., how clear – it is to the public. The explanation of the CCL process will need to be expressed so that the public can generally understand the method used. This does not necessarily mean the process will be simple or easy to understand.

Decision-makers, stakeholders, and drinking water consumers need to be able to understand why EPA has selected the CCL contaminants and why further research on these contaminants is a good use of resources. The public will want to know why investment in the methods used to select contaminants and investment in research on certain contaminants is an efficient and effective use of resources that will lead to improved protection of public health. If EPA is transparent in its decision-making, the public will have the rationale needed to understand how the method works and why specific contaminants are or are not on the list.

As recommended by the NRC, the CCL Work Group discussed the importance of noting uncertainties in data or information used in the process, as well as uncertainties in the proposed CCL. If EPA is clear about these uncertainties, it will provide decision makers and the public with the tools needed to determine whether they believe EPA has made appropriate contaminant determinations, based upon protection of public health, good science, and occurrence in drinking water**.**

If successful, the CCL classification approach recommended in this report will generate a list of contaminants that enables EPA to concentrate research and other activities on those contaminants that occur or potentially occur in drinking water, and that pose the most concern for public health. This will result in research being targeted as wisely and effectively as possible to support public health protection while addressing a concern of stakeholders and ratepayers that limited resources be spent in a cost-effective manner. By investing in this kind of process up-front, the contaminants of significant concern will be singled out for further study in an open and transparent manner. EPA should use caution when developing this up-front process to assure that resources are wisely invested when implementing the recommendations in this report. This will help EPA allocate limited funds to the contaminants that pose the greatest public health risk with input from stakeholders. The resulting effort will assist in supporting a credible and open *process* so the public knows the rationale for why research is being recommended and can support appropriate listing decisions.

The CCL Work Group endorses the following steps proposed by the NRC to encourage transparency of whatever method EPA adopts (pp. 64-66 of NRC report):

- *One of EPA's major goals in developing future CCLs should be to explain the process sufficiently so that the reader can understand the rationale behind including particular contaminants on the CCL. To achieve this goal would require that transparency be incorporated into the method used in the decision-making process in addition to being an integral component in communicating the details of the decision-making process to the public (NRC 2001, p. 61).*
- *The use of a classification tool needs explanation or rationale.*
- *The method for designing and calibrating the decision-making process must be explained.*

If decision-making for including or excluding a certain contaminant on future CCLs ultimately depends on a combination of the results of a classification tool and EPA judgment, then this relationship must be fully articulated along with the background assumptions and underlying Agency

judgments. Key criteria, data, and assumptions that affect inclusion or exclusion in potentially controversial cases ought to be noted, where possible, so that the reader can follow the logic regarding why decisions were made.

2.1.2 Public Participation

As quoted by the NRC, " '*public participation encompasses a group of procedures designed to consult, involve, and inform the public to allow those affected by a decision to have an input into that decision' (Rowe and Frewer, 2000)"* (p. 66). The NRC also points out that *"a central tenet of public participation is that the public is, in principle, capable of making wise and prudent decisions"* (p 66).

The CCL Work Group agrees with the NRC that EPA will need to garner public support to implement the CCL method effectively and efficiently. Without this, it will be difficult to obtain buy-in from various stakeholder groups. The CCL Work Group's principles on public participation are as follows:

- *The CCL decision-making process must be open, accessible, and available to all stakeholders who are interested.*
- *The CCL Work Group encourages EPA to provide the opportunity for public involvement at key steps along the way. Broad participation that is representative of the range of affected and interested parties should be a priority, thereby considering public health values, viewpoints, and principles.*

The NRC recommended an approach that would lead to scientifically sound policy decisions, informed by technical expertise, that are responsive to stakeholder values and concerns. This Work Group process is a first step in this direction. Because the prototype classification process for developing the CCL is a new approach, the Work Group recommends that EPA develop an outreach program to educate and inform stakeholders about its use. This approach may be a challenge for some to understand. Therefore, in the future, the Work Group recommends that EPA consider early and ongoing consultation with key stakeholders and outreach to the public as implementation proceeds. Finally, the Work Group agrees with the NRC that the public involvement program needs to be tailored to the public's needs and should start early in the process.

2.2 Overview of Recommended CCL Classification Process

In providing an overview of the CCL process, this section of the chapter describes the Work Group's recommendations for:

1) building on the NRC's concept of a three-step[3] CCL classification process;

[3] In its report, the NRC refers to its recommended approach as a "two-step" process, where a Universe of potential contaminants is assumed to exist, the first step is screening that Universe to generate a Preliminary CCL (PCCL) and the second step is refining the PCCL to produce a CCL. However, because the NDWAC Work Group elaborated further on the NRC's "Universe" and how to identify its contents, this report generally refers to the NRC approach as a "three-step" process (unless directly quoting from the NRC report).

2) developing parallel processes for building the microbial and chemical CCL Universes and for classifying the agents that comprise those Universes – first to the Preliminary CCL (PCCL), and then to the CCL;

3) approaching the development of a prototype classification approach; and for

4) incorporating genomic information into the CCL classification process.

2.2.1 Building on the NRC Approach

Having accepted the premise of a three-step selection process as proposed by NRC, the Work Group focused on achieving objectives inherent in the NRC approach. The NRC approach presents a number of logistical and practical hurdles for the EPA. For example, the NRC recommended that the Agency describe the Universe of potential contaminants very broadly. As a result, the information management system and decision criteria employed in the early stages of the CCL process must process tens of thousands of agents – often on the basis of very limited data or information. The Work Group was cognizant of such practical implementation issues and recommended modifications to the NRC approach to address them. Specifically, the Work Group focused on the following objectives.

1) More completely addressing the scope of the CCL "Universe" as described by NRC – with respect to both chemicals and microbes

2) Identifying a robust and practical means of screening the Universe to a Preliminary CCL (PCCL)

3) Evaluating the application of a prototype classification algorithm to select a CCL from the PCCL

4) Ensuring that both chemical and microbial contaminants are adequately and equally considered by the CCL process

5) More fully developing the role of expert judgment acknowledged by the NRC but not developed in its report

6) Reviewing the NRC's call for transparency throughout the CCL process

7) Expanding on the NRC model to explicitly allow for nomination of potential contaminants for consideration

8) Expanding on the NRC model by explicitly encouraging the Agency to maintain the CCL process as an ongoing programmatic element, rather than as a protocol that is repeated every five years. (This expansion includes the concept of surveillance for data to support the CCL process.)

9) Suggesting data and information "hierarchies" that might be used in the process

10) Following through on the NRC's recommendation to incorporate consideration of data quality into the CCL process

11) Developing a framework for incorporating genomics and proteomics, including the NRC's Virulence Factor Activity Relationship (VFAR) concept, into the CCL process

Key areas where the Work Group expanded on the work of the NRC, going beyond the NRC's report and recommendations, include: the role of surveillance and nomination processes, the role of expert input into the CCL process, the issue of data quality considerations, and the concept of an adaptive management approach to implementing the CCL classification process. The Work Group's contributions in each of these areas, along with a discussion of other overarching issues, are presented in Section 2.3. First, however, it will be helpful to step through the recommended CCL process illustrated in Figure 2.1.

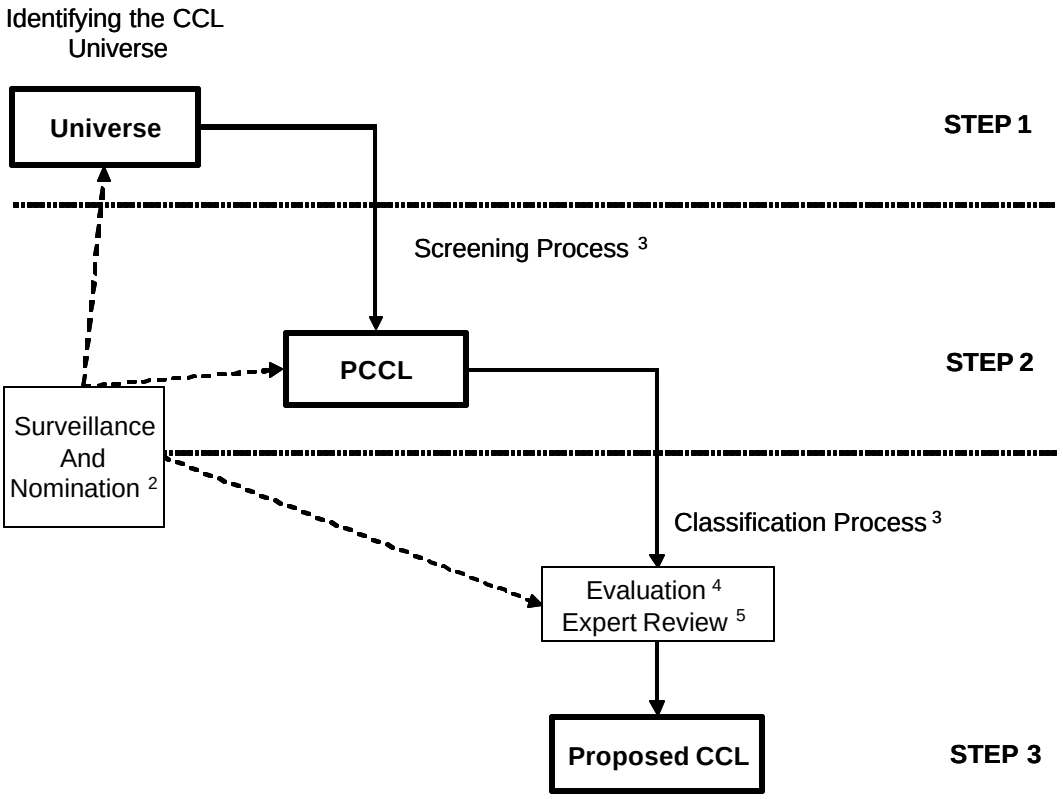
2.2.2 Parallel Processes for Chemical and Microbial Contaminants

Selection of microbial and chemical contaminants through a single CCL process is mentioned in the NRC recommendations. The Work Group found that, at this point in time, there are still systematic differences in the strengths and weaknesses of the information available for chemical and microbial contaminants.

→ **The Work Group recommends that the procedure for screening and selecting the CCL contaminants consist of two parallel processes which meet in the formation of a single CCL, but which take best advantage of the information available for each type of contaminant.**

The general framework of assessment, drawing on information concerning health effects and occurrence, and allowing for consideration of the quality of information available, should be used for both classes of contaminants. The specific recommendations for the microbial and chemical classification processes are discussed in Chapters 3 and 4, respectively.

Figure 2.1 - Overview of NDWAC Work Group Recommended CCL Process[1]



Notes:

1. Steps are sequential, as are components of each step, with the exception of surveillance and nomination. This generalized process is applicable to both chemical and microbial contaminants, though the specific execution of particular steps may differ in practice.

2. Surveillance and nomination provide an alternative pathway for entry into the CCL process for new and emerging agents, in particular. Most nominations would be for agents to the CCL Universe. Depending on the timing of the nomination and the information available, a contaminant could move onto the PCCL or CCL, if justified.

3. Expert judgment, possibly including external expert consultation, will be important throughout the process, but particularly at key points, such as: reviewing the screening criteria and process from the Universe to the PCCL; assessing the training data set and classification algorithm performance during development of the PCCL to CCL classification step.

4. After implementing the classification process, the prioritized list of contaminants would be evaluated by experts, including a review of the quality of information.

5. The CCL classification process and draft CCL list would undergo a critical Expert Review by EPA and by outside experts before the CCL is proposed.

2.2.2.1 Identifying the CCL Universe

Identification of the CCL Universe should follow the principles described by NRC to be inclusive of agents with demonstrated or potential occurrence in drinking water and/or demonstrated or potential adverse health effects. The NDWAC Work Group recommends that the EPA consider adopting a three-stage process to identify the Chemical CCL Universe.

- Identify and retrieve data (including lists of agents) from sources that have data and information about occurrence of contaminants in drinking water or source water or about health effects.
- Identify and retrieve data (including lists of agents) and information from sources that have a link (pathway) to drinking water concerns.
- Use other information sources, such as chemical properties, and models (e.g. QSARs) and surrogate information, to address data gaps.

The Work Group recommends that the Microbial CCL Universe be based on the evaluation of data sources and literature reviews that identify pathogens, i.e., organisms known or suspected to cause human disease. (Those pathogens from this CCL Universe that are known to be associated with source water, recreational water, or drinking water would be selected for inclusion on the PCCL.)

Specific recommendations for microbial and chemical contaminants are described in detail in Chapters 3 and 4, respectively.

A tag indicating the type of data source employed to identify and describe a microbial or chemical agent should be tracked along with information about that agent as it is further processed in the CCL process, as described in section 2.3.4.

2.2.2.2 Screening from the Universe to the PCCL

The Work Group reviewed the NRC's conceptual approach to screening a Universe of agents to a Preliminary Contaminant Candidate List (PCCL). Work Group members agreed with the NRC recommendation that this screened list should receive a higher degree of scrutiny before contaminants are moved to the CCL. To address this intermediate step, the Work Group assessed the likely availability of data about occurrence and adverse health effects of contaminants that may be present in drinking water and concluded that, as noted by the NRC, these data may be very limited. The Work Group considered the advisability of giving priority to contaminants for which more data are available against the interest in taking an inclusive approach. The Work Group also noted that the screening approach should err on the side of allowing contaminants to move forward at this step in the classification process rather than omitting potential drinking water contaminants from further consideration. The Work Group concluded that it would be important for EPA to develop an approach and screening criteria that are:

- *capable of assessing as many of the agents in the Universe as possible, even those with limited data;*
- *as insensitive as possible to data limitations and that treat contaminants with different amounts of data available as similarly as possible;*
- *as simple as possible, to require fewer resources and less time;*

- *capable of identifying those contaminants of greatest significance for further consideration; and,*
- *to the extent feasible in light of the significant differences in availability of data for chemicals and microbes, as similar in approach as possible.*

To develop an approach that would be as simple as possible and allow for the assessment of as many agents from the Universe as possible, the Work Group discussed how to identify the most essential characteristics of agents of concern. The Work Group sought to limit these characteristics to those that are most necessary and informative. These characteristics would become the basis for conducting this initial stage of screening, from the Universe to the PCCL.

The Work Group decided, based on the information available at this time, that a more limited set of the data elements describing health effects and occurrence characteristics would be more effective when defining criteria to select contaminants from the Universe for the PCCL. The attributes used to characterize microbial and chemical contaminants and the data elements used to describe or measure those attributes are discussed in Section 2.2.2.3, below. Specific recommendations for screening contaminants from the Microbial and Chemical CCL Universes are described in Chapters 5 and 6, respectively.

2.2.2.3 Characterizing the PCCL Contaminants

In considering the NRC recommendations, the Work Group agreed that structured decision-making models could be used by EPA, in conjunction with expert judgment, to determine which chemical and microbiological contaminants are most appropriately moved forward from the PCCL to the CCL based on their known or potential health risks. These various models require as inputs some specific measures related to those risks. These specific measures would be quantified attributes developed from either the actual values reported in the scientific literature (such as water concentration measurements or Reference Dose values) or the generated values or "scores" based upon the actual values reported in the literature to characterize attributes. Attributes – or more specifically, either the actual values or scores generated from the actual values for data elements used to characterize those attributes – can serve as the inputs for these models. ***Attributes*** are defined in the context of the CCL classification process as characteristics of a contaminant that contribute to the likelihood that it could occur in drinking water at levels and frequencies that pose a public health risk. The various types of measures or descriptors that may be used as a means for quantifying the attributes are referred to as ***data elements***. The expression "quantifying the attributes" used in this report refers to either the use of actual values reported in the scientific literature or to the generation of values (or "scores") based upon these actual values to characterize the attributes numerically so that they can be used in the classification modeling steps. These quantified values for the attributes can then be applied as inputs to the structured-decision making tools described in Chapter 5 to prioritize the contaminants for moving them from the PCCL to the CCL. In using data elements to quantify attributes, the emphasis at this step in the process should be to have those contaminants posing the greatest public health risk proceed to the CCL.

The NRC indicated that it had spent considerable time deliberating on the number and type of attributes that should be used in the CCL development approach the NRC envisioned. From those deliberations, the NRC developed a set of five specific attributes – two addressing health effects, three addressing occurrence – that they believed constituted a reasonable starting point for EPA to consider.

The NRC envisioned that these five attributes would be applicable to both chemical and microbial contaminants, but recognized that the types of measures and information used to quantify the attributes would differ for these two categories of contaminants. The specific attributes identified by the NRC were:

- ***Potency** and **Severity** as key predictive attributes for health effects*
- ***Prevalence** and **Magnitude** as key predictive attributes for occurrence*
- *And **Persistence/Mobility**, as characteristics that might predict possible occurrence if direct measures of prevalence and magnitude were not available*

For chemical contaminants, the Work Group agreed that EPA should start with the two health effects attributes (potency and severity) and the three occurrence attributes (magnitude, prevalence, and persistence/mobility) generally described by the NRC as input for the PCCL-to-CCL classification modeling. (See text box for general definitions for each of these attributes, as provided by the NRC.)

Exhibit 2.1 - Health Effects and Occurrence Attributes

- ***Potency*** indicates the amount of a contaminant required to cause an adverse health effect; a relative scaling of a dose-response relationship.
- ***Severity*** describes the clinical significance of the most sensitive health end-point; a measure of "How bad is the effect?"
- ***Magnitude*** reflects the concentration or expected concentration of a contaminant relative to a level that causes a perceived health effect.
- ***Prevalence*** describes how commonly the contaminant does or would occur in drinking water.
- ***Persistence/Mobility*** reflects the likelihood that the contaminant would be found in the aquatic environment based solely on physical properties of the contaminant.

For microbial contaminants, the same set of attributes are used, but different kinds and combinations of data elements are required to quantify those attributes, as discussed in Chapter 3.

There are numerous details concerning how many attributes are needed and how they should be characterized and quantified that must be developed in conjunction with the development of the specific classification approach(es) to be used in the process of moving contaminants from the PCCL to the CCL. This is in keeping with the NRC observation that the five attributes discussed in its report were meant to be illustrative and represented a reasonable starting point for EPA's consideration.

2.2.3 Developing a Prototype Classification Approach

The NRC prototype classification approach is a challenging one, both because of the number and difficulty of preparatory steps required and because of the inter-related complexities of the attribute and scoring process. The Work Group was unable, given available time and resources, to actually develop and test a training data set based on the attribute scoring protocols developed by EPA in

support of the Work Group activities. Consequently, the Work Group did not have the opportunity to pilot the NRC recommendation regarding the prototype classification approach.

Despite these limitations, the Work Group did feel that it could offer EPA practical advice on how to proceed in its evaluation of the prototype classification approach.

→` **Specifically, EPA should proceed with the following steps** .

- ***Evaluate a range of performance indicators for the classification approach.***
- ***Proceed to construct the necessary data systems to support a classification approach.***
- ***Prepare training and validation data sets and test the performance of the algorithm against the Agency's performance indicators for the algorithm .***
- ***Employ expert processes to make adjustments in the event the classification approach does not perform adequately.***
- ***Once the classification approach is demonstrated, apply the approach to obtain a draft CCL for expert evaluation and refinement of both the product and the process.***

2.2.4 Incorporating Genomic Information in the CCL Process

Genomics and proteomics are potentially powerful tools for elucidating the pathogenic mechanisms of microorganisms, and thus for understanding individual and population exposure and response to contaminants. At present, the use of these techniques for screening microbes or chemicals in the CCL process is premature; however, the Work Group found considerable merit in the NRC's recommendation for long-term development of VFARs. (See Chapter 3.4 for further discussion and specific recommendations of the Work Group related to VFARs.)

For chemical contaminants, current research in toxicology involves gathering data on the relationship of genomics and chemical mechanisms of action, a growing field called toxicogenomics.

The field of toxicogenomics is rapidly developing information about gene and protein activity in response to chemical exposure. Biological responses following exposure to chemicals are studied at several levels. Presently, toxicologists identify organ system and specific adverse effects by exposing animal and cell models to specific chemicals. There are strain and species differences in these responses. Likewise, in the human population, there are differences in responses. These differences are often related to the genetic make-up of the individuals. A great deal of effort is being focused in this area (e.g., the National Institute of Environmental Health Sciences' National Center for Toxicogenomics). As in microbiology studies, technologies such as DNA microarrays or high-throughput nuclear magnetic resonance (NMR) and protein expression analysis are being used for the assessment of the biological effects from chemicals. In the future, researchers will begin to understand which genes are turned on (or off) in response to specific chemicals. These responses will provide useful information on the degree to which populations are exposed and perhaps begin to identify those populations who may be more susceptible to those exposures.

As these genomic and proteomic techniques are developed and refined, their use should be considered for future CCL development for both microbiological and chemical contaminant evaluation.

→ **As noted in Chapter 3 of this report, the Work Group recommends that EPA should monitor the progress of genomics and related technologies and integrate them into the CCL process, as feasible.**

2.3 Overarching Issues

The remainder of this chapter addresses issues that affect many aspects of the CCL classification process. These overarching issues include the following.

- *Integrating expert judgment into the CCL process*
- *Implementing active surveillance for new and emerging agents*
- *Implementing nomination/evaluation processes for new and emerging agents*
- *Dealing with quality of information in the CCL process*
- *Use of quantitative structure activity relationships (QSARs)*
- *Use of an adaptive management approach to implementation*

2.3.1 Integrating Expert Judgment into the Process

NRC recommendations include provisions for "expert" and "scientific review" in the CCL process but provide little guidance as to what, how, and when such review would be used. Like the NRC panel, the Work Group observed that expert judgment is inherent throughout the development of the CCL process and in implementing that process once it is developed. Critical reviews, involving various types of expert consultation and collaboration, up to and including more formal expert reviews, will be useful at key points in the new, evolving CCL process, as outlined above in Section 2.2 and in Figure 2.1.

→ **There are several key milestones in the CCL process where a critical review would be especially relevant:**

- ***In Step 2, to review the screening criteria and their application to screen agents from the CCL Universe to the PCCL;***
- ***In Step 3, during development of the classification process from the PCCL to the CCL, to assess the training data set(s), assess the performance of the classification algorithm(s) tested, and to determine whether that performance is sufficient to justify immediate use of the algorithm(s) or suggests the need for further development;***
- ***After the classification process is implemented, to evaluate the prioritized list of contaminants, including a review of the quality of information, to provide judgments on the proposed draft listing;***
- ***The CCL classification process and draft list should undergo a formal expert review, including external experts, before the CCL list is proposed.***

Each of these reviews would benefit from a range of relevant expertise from both inside and outside the Agency. There are however, significant time and resource constraints to consider. To best utilize the Agency's limited resources, formal expert review is most critical in evaluating the

classification process and draft CCL. In the Work Group's opinion this formal expert review should involve external experts. This review would consider the performance of the CCL classification algorithm, considering not only what was listed, but also looking selectively at the PCCL to identify inconsistencies or biases in the algorithm's performance, and the application of expert judgment to the prioritized list.

In emphasizing this final review in the CCL process, we do not intend to diminish the importance of expert or critical review in the earlier steps of the CCL process. Expert involvement can be particularly valuable for the Agency as it develops and implements an entirely new approach for the CCL classification process. Inclusion of expert review early in the CCL process offers assurance that the final product, the proposed CCL, will be technically sound and scientifically defensible. This will afford the Agency opportunities to spot problems early and make timely and efficient adjustments. Another benefit may be increased credibility with the stakeholder community, as expert review provides technical checks on the process as it evolves, rather than solely relying on the comments from stakeholders and interested parties during the proposal and final Federal Register publication at the end of the CCL process.

As implied, critical review and expert involvement can take many forms, from reviews internal to EPA, less formal technical consultation with experts external to the Agency and stakeholders, to formal external review. In particular, the Work Group does not envision applications of the more rigorous external peer-review type activity during the earlier stages of the CCL Classification Process. It is also important to note that the Work Group recognizes that these reviews should be integrated into the overall CCL process so that concurrent activities and overall progress toward proposal of the CCL can occur in a timely fashion. Further, the Work Group noted that involvement of the same experts in review of the various steps in the process may afford both logistical and technical advantages.

2.3.2 Implementation of an Active Surveillance Process for New and Emerging Agents

The Work Group recognized that it can take considerable time for information to be generated about contaminants before they appear in many data sources and can be captured in the mainstream CCL process. In fact, it is likely that any such broad process will not, for example, be able to quickly reflect outbreak investigations that may identify new and emerging contaminants. Hence, the Work Group discussed the need for active surveillance and nomination processes to provide an alternative pathway for entry into the CCL process. The Work Group believes that a surveillance process will prove to be an important and necessary component to ensure timely identification of information relevant to new and emerging contaminants. Such relevant information may include recent epidemiological or toxicological studies, new information related to sensitive subpopulations, or new investigations of occurrence or exposures. The Work Group recognizes that the surveillance and nomination processes are key areas where expert judgment would provide input to the CCL process.

→ **The Work Group recommends that EPA establish an active surveillance process to provide identification of new and emerging agents for the CCL.**

This process of identification should be an integral part of EPA's CCL process. The burden of identifying new and emerging problems should not be solely on the public. While the recommended surveillance process has not been characterized in depth, the following aspects should be considered.

- *Implementation of a proactive process to survey or obtain information from institutions and organizations that might be expected to observe or generate new information about occurrence or health effects of potential agents or contaminants. These could include federal, state and local health departments, environmental agencies, drinking water utilities, and research institutions. This would include an ongoing process for communication with these institutions.*
- *Identification of key published data sources (or criteria for their selection) based upon consistency with the inclusionary principles (discussed in detail in Chapter 4), and updated with adequate frequency to provide the most current information available on potential agents.*
- *A means for identifying new information from recent updates of data sources to minimize redundant searching.*
- *A review process that is technically sound and logistically practical.*
- *A means for documenting the process and any decisions reached (transparency).*

Further discussion and specific recommendations related to surveillance and nominations of microbial and chemical contaminants are presented in Chapters 3 and 4, respectively.

2.3.2.1 Surveillance Activities

The Work Group recognizes that EPA has considerable ongoing activity that potentially relates to surveillance, ranging from ongoing literature reviews, to attendance at professional meetings, sponsorship of special meetings and special sessions at meetings dealing with drinking water issues, communications with researchers in the field, and liaisons with foreign institutions. EPA's Office of Water maintains linkage among Offices within EPA (e.g., Office of Pesticide Programs) and with other agencies and organizations that play a key role in the surveillance process (e.g., Centers for Disease Control and Prevention, US Geological Survey). Many offices within EPA are being called upon to conduct surveillance activities, so coordination must be a key component. Examples of EPA activities are noted below (see box) as types of activities that the Work Group recognizes as beneficial. These activities may need to be expanded and their linkage to the CCL may need to be strengthened.

Exhibit 2.2 - EPA Activities Relevant to the Surveillance Process

- *The Office of Ground Water and Drinking Water (OGWDW) is working with the Office of Wetlands, Oceans, and Watersheds (OWOW) to strengthen linkage between ambient water (i.e., source waters) contaminant concerns and criteria (Clean Water Act) and the drinking water program.*
- *The Office of Science and Technology (OST) ecological health (i.e., ambient/source waters) team and its human health (i.e., drinking water) team collaborate with the Office of Prevention, Pesticides, and Toxic Substances (OPPTS) to identify new and emerging contaminants of concern.*
- *OGWDW, OST, and Office of Pesticide Programs (in OPPTS) coordinate on cross-cutting scientific issues related to pesticides (e.g., share data on health effects; coordinate activities on risk analysis and occurrence studies).*

- *OST maintains a relationship with State and Regional risk assessors through the Federal-State Toxicology and Risk Analysis Committee (FSTRAC). It is often through the interactions with this group that EPA becomes aware of local contamination problems and emerging contaminants. FSTRAC meetings are managed by OST and are held twice per year.*
- *Office of Research and Development (ORD) staff conduct research and annual reviews of drinking water issues. (For example: Richardson, S.D. 2003. Water Analysis: Emerging Contaminants and Current Issues. Analytical Chemistry. 75(12):2831-2857); Daughton, C., and Ternes, T., Pharmaceuticals and Personal Care Products in the Environment: Agents of Subtle Change? Environmental Health Perspectives. Volume 107, Supplement 6, December 1999; Birnbaum L.S. and D. F. Staskal. 2003. Brominated Flame Retardants: Cause for Concern? doi:10.1289/ehp.6559 (available at **http://dx.doi.org/** 17 October 2003).)*
- *The EPA Drinking Water Hotline receives reports of contamination incidents that are reviewed as part of the CCL process and that may not appear in other data sources.*
- *The Centers for Disease Control and Prevention (CDC), Council of State and Territorial Epidemiologists (CSTE), and EPA maintain a collaborative surveillance system for the occurrence and causes of waterborne-disease outbreaks. "Surveillance for Waterborne-Disease Outbreaks" is published biannually in the Morbidity and Mortality Weekly Report (MMWR). This collaboration is clearly recognized as part of the process to identify possible new or emerging waterborne microbial contaminants (see Chapter 5).*
- *Various offices within EPA interact with other offices or programs within the National Institute of Health, such as the National Toxicology Program (NTP), National Cancer Institute (NCI), and offices/centers of the National Institute of Environmental Health Sciences (e.g., Center for the Evaluation of Risks to Human Reproduction).*
- *EPA and ATSDR assess the presence and nature of contaminants and health hazards at Superfund sites and may conduct public health assessments at RCRA sites.*
- *The USGS Water Resources programs have established formal liaison coordination with OGWDW (and OPP and OWOW) for sharing information and program coordination for the National Water Quality Assessment (NAWQA) program and the Toxics Substances Hydrology Program, including the National Reconnaissance of Emerging Contaminants.*
- *OGWDW/OW has been working to strengthen interaction with and information review from its foreign counterparts, particularly in Canada, the EU (including WHO), Japan, and Latin America.*
- *OGWDW and other OW offices have official linkages or liaisons for information sharing with various groups on the front lines of water quality issues such as the Association of State Drinking Water Administrators (ASDWA), Association of State and Interstate Water Pollution Control Administrators (ASIWPCA), National Water Quality Monitoring Council (NWQMC), Ground Water Protection Council (GWPC), the Association of State and Territorial Health Officials (ASTHO), among others.*
- *OGWDW and other OW staff participate in support and review of research on water contaminant issues with the American Water Works Association Research Foundation (AWWARF) and the Water Environment Research Foundation (WERF).*

- *EPA staff often directly participate in meetings with various groups, such as those mentioned above, as well as American Water Works Association (e.g., Water Quality Technical Conference), American Chemical Society (ACS), Society of Environmental Toxicology and Chemistry (SETAC), Society for Risk Analysis (SRA), American Society for Microbiology (ASM), American Water Resources Association (AWRA), National Ground Water Association (NGWA), Society of Toxicology (SOT) and the American Public Health Association (APHA).*

2.3.2.2 Primary Source Literature Review

Another component of surveillance is review of the primary research literature to identify, and provide information for new or emerging agents. Many "text" and bibliographic sources of information were identified in the Work Group's efforts to identify databases/data sources for the Universe. Work Group discussions recognized that bibliographic sources could not likely be part of the more automated process of identifying the CCL Universe. As noted in reports to the Work Group, EPA has begun to develop automated data extraction tools (software/programs) that would be able to partly automate data collection from some important text sources (e.g., Developmental and Reproductive Toxicology (DART), part of the National Library of Medicine). As noted in other reviews, bibliographic sources may be used to fill in data gaps for contaminants identified in the CCL process. Various search engines can be employed; past discussions have identified key sources such as PubMed, TOXLINE, CCRIS (Chemical Carcinogenesis Research Information System), GENE-TOX, DART/ETIC (Developmental and Reproductive Toxicology/Environmental Teratology Information Center), and ISI Web of Science. These sources can readily be searched to locate possible information to fill data gaps on identified agents; however, searching the literature for information on new and emerging agents must be part of a surveillance process because of the largely manual effort that is required for more detailed review to assess pertinent literature. For example, preliminary studies might be identified in bibliographic sources (e.g., epidemiological studies) related to emerging issues or agents of interest. Such studies will nearly always need to be evaluated using expert judgment to assess if the results can be utilized based upon sound scientific information.

The most up-to-date, "emerging" information may well come from information presented at professional conferences. Surveillance of meeting proceedings requires yet a different level of effort. This might be most efficiently handled through enhanced relationships with professional societies and organizations, as discussed below.

2.3.2.3 Additional Surveillance Activities and Recommendations

While there are many activities and mechanisms in place that can contribute to the surveillance process, for the explicit needs of the CCL process these may need to be strengthened and new activities may need to be initiated. In many of the cooperative efforts noted, EPA may need to explicitly outline the needs for the CCL process to ensure that there is adequate consideration and communication.

At least a few professional organizations have been forming committees and sponsoring forums to focus on emerging water quality issues. The EPA CCL staff will need to ensure close links to such groups, perhaps through joint sponsorship of regular meetings, workshops or conferences. In a similar manner, EPA may need to help stimulate similar focus groups within other organizations. This can be

accomplished in part with formal communications and requests to professional and interest organizations. To facilitate appropriate interest, EPA might undertake other actions, as needed, such as:

- EPA might designate a formal liaison with outside groups to coordinate efforts in emerging drinking water quality issues, or at least specify an EPA point-of-contact for a specific organization(s).
- EPA might set up workshops, or more narrowly focused meetings with appropriate professional organizations and stakeholders, on emerging or problematic contaminant groups such as microbiological contaminants or personal care products and pharmaceuticals.
- EPA might strengthen communications and review of unique state level programs that are working to identify and even monitor for new and emerging contaminants, or conducting special health studies related to water-borne contaminants (e.g., California, Iowa, New Jersey, and New York, among others).
- EPA might work with journals or publication groups to standardize key words to facilitate data gathering (as noted in Chapter 3 for microbiological surveillance).

→ **In particular, the Work Group recommends that EPA institute a regularly scheduled (e.g., biennial) conference on "Emerging Issues in Drinking Water" as part of their research for the CCL process, where stakeholders and professional groups could present their findings and concerns on emerging and new agents.**

With the myriad groups that can be involved, the Work Group suggests that this could be a particularly efficient mechanism. This would also provide a particularly visible and transparent component for gathering stakeholder input. EPA's sponsorship of such meetings and EPA's presentation of research and data needs for CCL consideration would also act to stimulate and structure needed research both within EPA and by various interest groups for the future of the program. All of these activities should serve to provide "nominations" of agents to add to the CCL Universe.

2.3.3 Implementation of a Nomination and Evaluation Process for New and Emerging Agents

It is envisioned that surveillance and nomination would be integral components of the CCL process and not a separate process. As such, surveillance and nominations typically would provide an alternative pathway for entry for an agent into the CCL evaluative process. In other words, agents identified would typically be considered for placement in the CCL Universe, not on the CCL. However, depending on the timing of the identification of the new and emerging agents (in relationship to CCL publication schedule), and on the nature of the information about them, nominated contaminants could move onto the PCCL – or even onto the proposed CCL through an expert review process, or (if justified) through an accelerated Agency decision-making process (as further discussed below in section 2.3.3.2).

→ **The Work Group recommends that EPA develop a nomination and evaluation process for new and emerging agents, to enable agencies and interested stakeholders from the public and private sectors to nominate agents for consideration in the CCL process.**

As noted, all of the surveillance activities should serve to provide "nominations" of agents to add to the CCL Universe. It is important to note that nominations from the surveillance process can occur both before and after the PCCL stage (see Figure 2.1). Where they occur after the classification approach, nominations can still be considered as part of the expert review process prior to publishing a proposed CCL. In addition, there is the opportunity as part of the formal process of public comment and response on the proposed CCL that can include nominations. This existing process commences when the Agency, in establishing a new CCL, issues a proposal in the *Federal Register* with request for public comment. The request for public comment includes not only the opportunity to comment on what the Agency has proposed (the CCL), but also the opportunity to nominate additional potential contaminants to be considered for the CCL. In this process, EPA reviews comments and nominations, and responds with its decision, as part of establishing the final list.

However, the NDWAC Work Group recommends that additional opportunities to nominate agents should be available during the CCL evaluative process in advance of, and distinct from, the formal comment period on the Agency's proposed CCL. The NDWAC Work Group recommends that throughout the CCL process (for example as part of an "Emerging Issues" conference as discussed above), suggestions from stakeholders for agents to be considered would be provided to EPA, and that, if appropriate, these nominees could be added to the CCL Universe. (As with all parts of the CCL development, documentation of how and where the agent was identified would be part of the process.) Additional components to be considered for the nomination process are outlined below.

2.3.3.1 Additional Considerations for the Nomination Process

The Work Group also suggests that EPA develop additional components to the nomination and evaluation process.

→ **Although the nomination and evaluation process would require further specification by EPA, the Work Group recommends that the nomination process consider the following elements:**

- A communications strategy to identify and engage prospective stakeholders
- Recommendations for systematic communications with stakeholders
- Development of a consistent and transparent evaluation process by EPA, to include:

a) information and documentation requirements (i.e., new and emerging agents or potential contaminants should not just short-circuit the evaluative process);

b) an evaluation process that the nominated agents must undergo (i.e., for a new agent to go directly to the PCCL or CCL), it must present appropriate occurrence and health effects information as other PCCL or CCL contaminants;

c) a means for confirming that the information offered has not previously been considered;

d) a process and criteria for taking appropriate action for those found to have merit, and;

e) a means for documenting the process and any decisions reached.

2.3.3.2 Accelerated Listing Process

As new agents are identified, or as new information becomes available, there may be justification to accelerate their passage to the CCL Universe, from the Universe to the PCCL, or from the PCCL to the CCL. EPA could, if the data warrant, consider these contaminants on an accelerated basis. The Work Group recommends EPA develop a formal accelerated ("fast track") process and ensure that the process is communicated before, or at the time the Agency requests nominations from the public. The process should be open and transparent and be consistent with the overall CCL screening and evaluation procedures. The accelerated process should also consider the elements outlined in section 2.3.3.1, above.

2.3.4 Information Quality Considerations

2.3.4.1 NRC Discussion and Recommendations

In its 2001 report on classifying drinking water contaminants for regulatory consideration, the NRC addresses[4] some of the difficulties and challenges that EPA will face in applying data and information to any classification designed to sort a very large number of chemical and microbiological contaminants into exclusive categories: On or Off the PCCL; On or Off the CCL. The NRC recognized that EPA would likely encounter many challenges in implementing a classification scheme where imperfect or incomplete data must be used to determine whether a specific chemical or microbiological organism may or may not pose an existing or potential threat to consumers of public drinking water.

The NRC did not, however, make specific recommendations in its 2001 report as to how EPA should address or resolve issues related to the quality of the information used in the CCL development process. NRC refers to section 1412(b)(3)(A) of the SDWA Amendments which addresses the use of science in decision-making under this statute, specifying that EPA shall "use the best available, peer-reviewed science and supporting studies conducted in accordance with sound and objective scientific practices; and data collected by accepted methods or best available methods (if the reliability of the method and the nature of the decision justifies the use of the data)."

2.3.4.2 Work Group Considerations and Recommendations on Information Quality

The Work Group also considered the issues related to ensuring the use of the best available information and methods with respect to the data sources to be accessed, the data elements to be extracted from those sources, and the processes to be applied using those data elements to screen or classify a very large number of contaminants in the CCL Universe to reduce it to the relatively smaller numbers on the PCCL and then the CCL. The Work Group recognized that EPA's process should explicitly address compliance with Agency data quality guidelines and the Information Quality Act. The Work Group also recognizes, however, that the Agency must have some flexibility in the data

[4] In the section titled "The Nature of the Task."

quality guidelines to fully embrace the inclusionary principles. Work Group members noted that contaminants considered in the early stages of the CCL process will not necessarily be robustly characterized, and the data available for some of those contaminants will consist of different types of data. The Work Group also recognized that the data or information used to select the CCL will be more detailed and comprehensive than the data or information used to identify the CCL Universe.

To address the variability of the disparate types of data, and to ensure transparency, all steps in the CCL process should document information about the data sources (e.g., what quality assurance procedures were in place during data gathering, processing, or analysis). Additionally, the CCL process should apply more scrutiny to contaminants when selecting the CCL than when screening contaminants from the Universe to the PCCL. The nature of the data used to support these steps should be documented for review in the later steps of the CCL process.

Different data quality approaches can be established commensurate with the purpose for which the data will be used (e.g., screening from the CCL Universe to the PCCL versus classifying from the PCCL to the CCL). This is a priority-setting process that does not require the same detailed analysis as a rulemaking process, and therefore data quality considerations should recognize this difference. The Work Group noted a related key consideration in the CCL process should be that, in general, false negatives should be avoided when going from the Universe to the PCCL and false positives should be avoided when going from the PCCL to the CCL. It is important for EPA to develop and document appropriate data quality approaches as part of the process of implementing the adaptive management approach discussed below (in section 2.3.6). EPA should establish data quality approaches for use in each step of the CCL classification process prior to identifying the CCL Universe.

→ **The Work Group, therefore, recommends that information quality be considered in the CCL process.**

This recommendation raises two questions.

1) How is information quality to be summarized at any stage in the process from building the CCL Universe to selecting the CCL itself?

2) What are experts, or algorithms, to do with this information quality summary at each stage?

The answers to these questions must reflect that an assessment of information quality or uncertainty about some "best estimate" of a numerical value (such as the exposure or potency for an agent) can be resource-intensive, often requiring more resources than does determining the "best estimate" value itself.

→ **As an overall recommendation, the Work Group recommends that EPA collect and consider the "best available" data sources and data elements without restrictions or screening-out of information based on any minimum quality criteria developed in advance.**

The Work Group also offers the following specific recommendations regarding the consideration of information quality at the major stages of the CCL development process.

1) ***Establishing the CCL Universe:*** It will be possible to "tag" the agent with a reference to the quality of the data source or other information used to assign that agent to the Universe. This indicator would refer to considerations of the quality of the information source, and not be specific to information on the agent itself obtained from that source. Since the quality of information on different agents in the same information source can vary, it is recommended that this "tag" not be used for screening agents out of the Universe, but only to provide an indication of the general reliability of the source of information that should be considered at later stages.

2) ***Going from the Universe to the PCCL:*** Even at this stage of the process, it will not be feasible to perform an information quality analysis specific to a contaminant. It will be possible, however, to provide a richer "tag" for each contaminant. For example, the "tag" might document whether a measured value or a QSAR estimation was used for a screening element. Chapter 4 discusses this in further detail. Since the "tag" still does not reflect a full analysis at this stage, it is not necessary to use the "tag" to screen contaminants off the PCCL.

3) ***Going from the PCCL to the CCL:*** The Work Group recommends that EPA (and expert reviewers) consider the information quality "tag" more fully at this stage than at earlier stages of the process. EPA should consider developing information quality tags for the PCCL entries and using those tags explicitly in developing the classification algorithm and using it to create the CCL. One possibility is to include information quality as a sixth candidate attribute in developing the classification algorithm. If, however, a final algorithm is selected that does not include the information quality attribute, then explicit consideration of the "tags" should occur during expert review after the algorithm has been applied to the PCCL, but before the CCL is published.

Work Group members agreed that the list of contaminants selected for the CCL should undergo an expert review. Members noted that documenting the nature and type of information by assigning a "tag" for consideration at this step allows this information to be used in the final analysis for the listing decision. By fully documenting the information used in the process, the review of the information used, and the decisions made to develop the CCL can be conducted in an open and transparent manner.

More specifically, the Work Group discussed using the "tag" as part of the expert review process. For example, the review process could allow contaminants to move from the PCCL to the CCL only if the "tag" indicated sufficiently high reliability of the evidence supporting inclusion of a contaminant on the CCL. This would prevent the CCL from being populated with a number of contaminants that would, upon further review, be rejected both for regulation and for further research. The disadvantage of this approach is that it could require resources to support a judgment of the qualitative, expert-based judgment, of the quality of information for individual contaminants that are candidates for movement from the PCCL to the CCL. Alternatively, considering the nature and type of information used to select contaminants after the draft CCL listing may be useful in determining whether a contaminant remains on the draft CCL and in establishing priorities for regulatory determination.

2.3.5 Use of Quantitative Structure Activity Relationships (QSARs)

2.3.5.1 Introduction

As part of its consideration of options for including potential drinking water contaminants that lack applicable empirical data on health effects or occurrence in the CCL process, the Work Group was presented material on the use of Quantitative Structure Activity Relationship (QSAR) models and the output of those models. The Work Group recognized that the health effects and occurrence-related properties information generated by QSAR models could potentially be used both in screening contaminants to develop the PCCL from the Universe, and as data elements for attributes to develop the CCL from the PCCL. *The focus of this section is whether or not the use of QSAR models appears to be a reasonable approach to generating information about less-well characterized chemicals, that could be used in either or both of these steps in the CCL process.* The present discussion does not, however, specifically address the use of QSAR-generated data in the screening or classification steps. More specific consideration of the use of QSAR-generated data in those steps is addressed in Chapters 4 and 5.

The Work Group agreed that use of QSAR data for agents for which EPA does not have data is a potential tool.

However, a few members raised questions about the proprietary software currently available to develop QSAR data. These Work Group members noted that most of the widely-used models use proprietary algorithms that are not available for independent review. For reasons related to transparency, ethics and validity, these Work Group members could not recommend EPA use QSAR models because of the proprietary nature of these models.

Other members suggested that the value of the use of these models is important. While recognizing this concern about the need for transparency, they supported a recommendation that the Agency should use proprietary QSAR models. These Work Group members also noted that proprietary QSAR applications or computational algorithms were independently reviewed in their development, even though the proprietary models are not available for subsequent reviews.

This section provides what the Work Group learned about QSAR models. The Work Group did not reach consensus on a recommendation for QSAR models. Therefore two different recommendations are included below.

2.3.5.2 Background on QSARs

The Work Group sought to learn enough about QSAR models, their requirements, and their limitations to see if they could arrive at some general conclusions and recommendations regarding a potential role for QSARs in the CCL process. The Work Group did not attempt to conduct a comprehensive analysis of the suitability of all potentially applicable QSAR models or the output that they can generate, but rather sought to develop sufficient information based on a limited number of representative QSAR models to inform the assessment of their applicability for producing the type of information that could be used in the CCL Classification process.

Specifically, the commercially available QSAR application called TOPKAT (The Open Practical Knowledge Acquisition Toolkit) was used to predict the rat chronic oral Lowest Observable Adverse Effect Level (LOAEL), and the QSAR model package developed by EPA and Syracuse Research Corporation called Estimation Program Interface Suite (EPI Suite) was used to predict solubility and aerobic biodegradation information. A set of approximately 700 chemicals was used to test these models. Some of the chemicals that were evaluated also had empirical information for the properties predicted by the QSAR models, and were used largely to get a sense of how reliable the QSAR predictions were. Other chemicals not having empirical data for the QSAR model outputs were evaluated to provide some insight into the potential difficulties of applying the models to substances that are less-well characterized by actual measurements.

Technical reports and presentations documenting the QSAR model evaluation process were prepared by the technical team supporting the Work Group. The information presented, together with discussions of that information by the Work Group, led to the two different recommendations presented in the following section. (See Appendix B for summary of technical reports on QSAR model evaluation.)

2.3.5.3 Conclusions, Recommendations, and Rationale

→ **The full Work Group recommends that EPA explore use of QSAR models to those (agents or contaminants) for which EPA does not have data.**

While QSARs might be a valuable tool, a few Work Group members could not recommend use of the QSAR models because they cannot be properly tested and evaluated. These members suggest that, if EPA chooses to use QSAR models, the Agency should develop and use only fully transparent software that is available for independent review.

→ **These members suggest, if EPA chooses to use QSAR models, from an ethical, transparency and validity viewpoint, only fully transparent QSAR models that are available for independent review should be used. If nonproprietary software is not available, QSAR models should not be used.**

Other Work Group members offer a general recommendation (noted below) on the use of QSAR models, along with several considerations to guide the Agency.

→ **These members suggest, based upon a review of the historical use of QSAR in priority setting processes and the limited investigation of QSAR models performed, EPA should pursue using QSAR models, or existing information that has been generated by them, in the CCL development process.**

If EPA does proceed to use QSAR models, the Work Group offers the following general considerations regarding the Agency's investment of time and resources in these tools.

- While it cannot be determined from the Work Group's assessment how successful the use of data generated by QSAR models will be in expanding the range of chemicals that can be included in the CCL process, it does appear that some of these models can provide sufficiently useful information and should, therefore, be included as a potential tool for EPA in developing the CCL.

- Consistent with the overall quality assurance procedures that EPA should apply in the CCL process, the Agency should limit its consideration of QSARs to generally accepted, peer-reviewed QSAR models that are validated, adequately documented, and perform with well-described precision and accuracy.
- The Agency should follow the Office of Research and Development (ORD) framework and recognize that available QSAR model predictions reflect the limitations and biases of the data sets and methods used to develop those models.
- In constructing a sound framework for integrating QSAR predictions into the CCL process, the Agency should be careful to employ QSAR predictions in a scientifically rigorous manner cognizant of the tool's limitations.
- The Drinking Water program should utilize internal Federal office expertise to the extent possible in the selection and application of QSAR models.
- Additionally, outside experts with relevant expertise should review on an ongoing basis development of the Agency's approach to QSAR models and rules by which QSAR predicted values are applied in the CCL process.
- Therefore, the Work Group recommends that when reliable empirical observations are available, QSAR generated values should not be used.

2.3.6 Use of an Adaptive Management Approach to Implementation

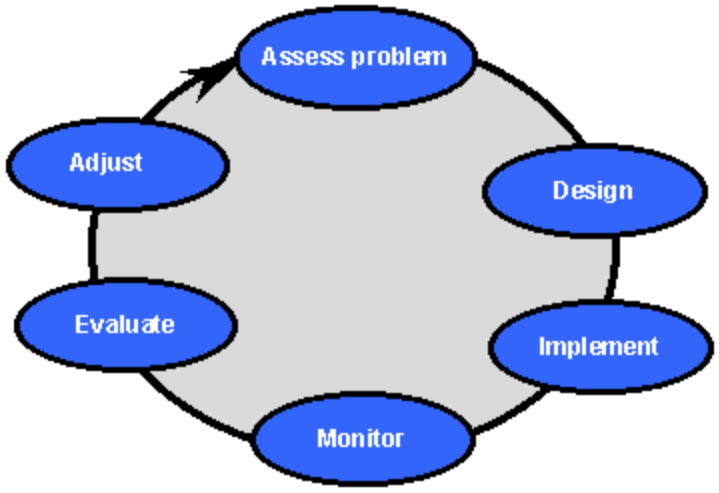
In development of the new CCL classification process, an adaptive management approach could provide a method of evaluating progress at milestones during development, implementation and review of the CCL process. By applying an adaptive management approach as the CCL process is developed and applied, the Agency will be able to determine: a) whether decisions made in the process have provided adequate results and the needed information; and, b) what modifications need to be made to the CCL process in the current or successive CCL cycles. Adaptive management principles could be applied in the development, implementation, and refinement of the three-step CCL process, particularly in the initial phases of implementation, and provide a framework to refine the process. As new information identified in the adaptive management framework becomes available, the Agency could use that information to evaluate and refine the process for current or future CCLs. (See Figure 2.2, below.)

Adaptive management is a well-established concept in environmental management. The idea arose from the recognition that environmental management decisions will always be made with uncertainty about the precise outcome of the alternative actions being considered. This inherent uncertainty may partially be addressed by further research, but often additional study delays action, and occurs on scales that incompletely capture the dynamics of the system that will be affected by the management actions. Thus, the best way to reduce uncertainty is to take action, to treat such management actions as an experiment, to monitor the outcome of such experiments, and thus to learn by doing.

Under an adaptive management approach, reducing uncertainty is an important goal in implementation of each generation of the method. This process incorporates systematic and continual integration of design, management, and monitoring, which would enable EPA to make informed

adjustments and adaptations, resulting in an improved method based on experience from the outcomes of successive generations of implementing the Universe-to-CCL approach.

Figure 2.2 - Diagram Schematic of an Adaptive Management Process



Elements each step in the process should consider:

- *Evaluative criteria for each phase;*
- *Adaptive learning process;*
- *Characterizing data quality;*
- *Transparency;*
- *Use of expert judgment /review.*

Source: British Columbia Ministry of Forests, Forest Practices Branch, http://www.for.gov.bc.ca/hfp/amhome/Amdefs.htm, August 9, 2000.

While adaptive management stresses the need for practical action in the face of uncertainty, it also emphasizes the need to tailor management decisions to the nature and quality of information available at any moment in the process. With little information, some policies with minimal potential for negative consequences ("no regrets") may be in order. More or better information may justify policies with (for example) greater economic costs. As information becomes better established, progressively more explicit decisions, with more serious consequences, are justified.

Concepts of adaptive management are a consistent theme in both the NRC and NDWAC recommendations. (See the bulleted list in text box.) Both reports stress the need to iteratively test and refine the CCL methods, rather than simply waiting until the methods are perfected before applying them in decisions on the CCL. In this regard, the present report emphasizes features that are well described in the context of adaptive management: (1) identify an approach, (2) define evaluative criteria (factors to evaluate), (3) iteratively implement the approach, (4) transparently assess evaluative criteria and (5) make changes to improve performance of the approach. Adaptive management also recognizes the utility of comparing alternative approaches to the creation of the CCL (e.g. different *a posteriori* methods, or an approach rooted more in facilitated discourse than in *a posteriori* methods), and the need to select the approach best suited to the quality of the information and performance available. Perhaps most importantly, adaptive management integrates interim evaluations into the overall approach so that change can take place as information becomes available.

This type of management approach is similar to those used in businesses and complex organizations dedicated to continuous improvement or high performance and should be familiar to most modern managers. This application to environmental systems (in this case, contaminants to be considered for further research and regulatory determinations) is only an extension or adaptation of those design-measure-feedback-redesign business models.

Chapter 3

CCL Classification Approach for Microbial Contaminants

This chapter identifies the challenges presented by the data and information available for microbial agents and provides the rationale from the Work Group's discussion to address those challenges. Section 3.1 addresses developing a Universe of microbial agents. Section 3.2 discusses recommendations to screen the Microbial CCL Universe to the PCCL. Section 3.3 introduces a discussion of protocols and considerations to evaluate microbial contaminants on the PCCL to select the CCL microbes. Section 3.4 presents the Work Group's discussion and recommendations on the use of genomics and proteomics, specifically virulence-factor activity relationships (VFARs), in the CCL classification process. Each section reviews the NRC recommendations and presents the NDWAC Work Group's recommendations for developing a CCL classification approach to microbial contaminants.

Chemicals and microbes exert their toxicological or pathological effects following exposure via ingestion, inhalation, or dermal contact, depending upon the specific agent- and host-dependent variables. However, chemicals and microorganisms behave in markedly different ways in the environment and within the human host. The methods and information used to characterize these two types of agents also vary. Chemical agents tend to be characterized by toxicological and occurrence data that, if not measured, can be modeled or estimated. The adverse health effects of microbial agents tend to be characterized by clinical and epidemiological data. Estimating the occurrence or potential occurrence of microbes can be based on the biological characteristics of the microorganism, but there are few analytical methods available for making such assessments and the information used to characterize microorganisms is not readily modeled or estimated. The differences in chemical and biological characteristics of demonstrated and potential water contaminants suggest that, while identifying the Microbial CCL Universe from the total microbial universe of microorganisms and screening a subset of biological agents from the Microbial CCL Universe to a PCCL can be consistent with the NDWAC's suggested principles for chemicals, at this time the approach to microbial agents and contaminants will require different data sources and data elements, and may require more involvement from experts than the approach described for chemical agents and contaminants.

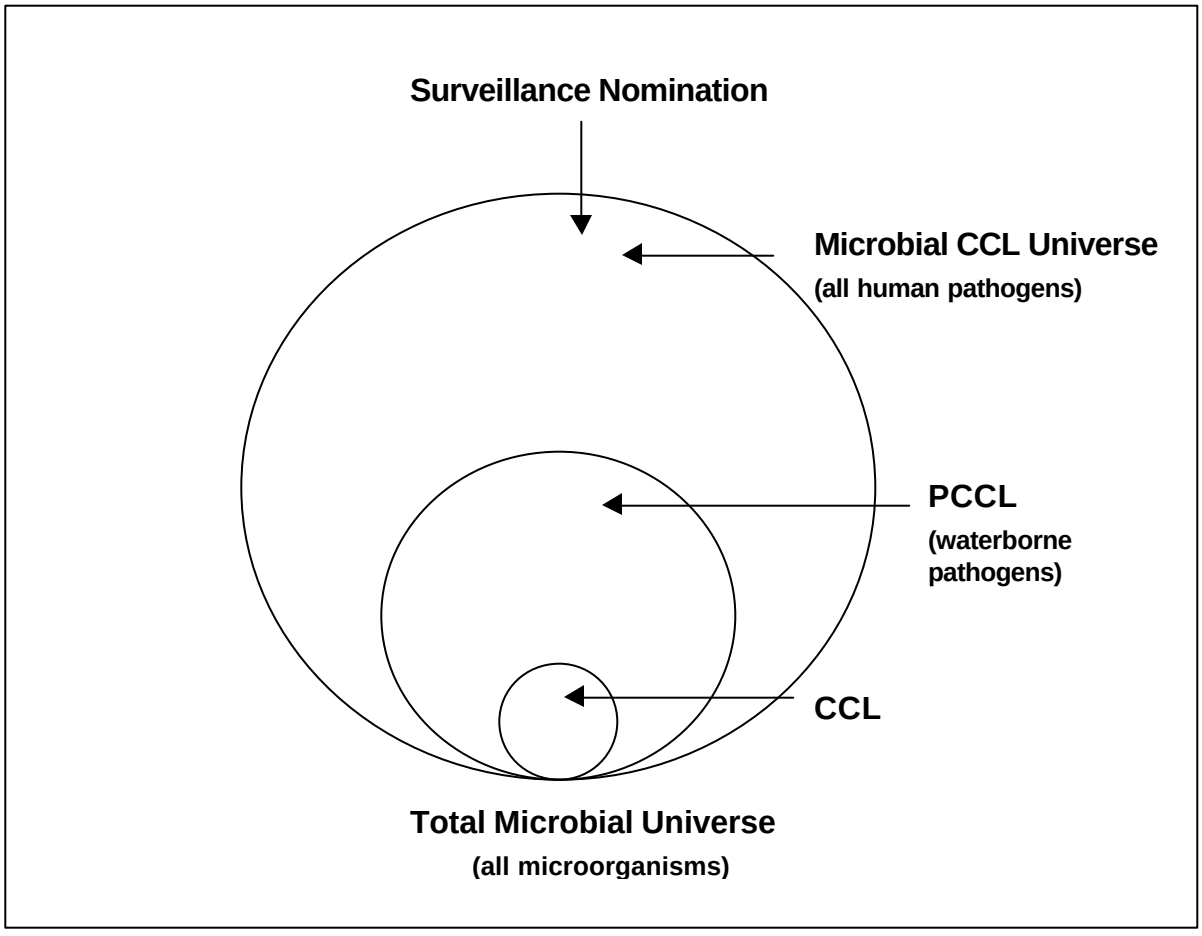
The Work Group has developed the following steps to select microbial contaminants for the CCL.

- *The total microbial universe may consist of all microorganisms.*
- *The Microbial CCL Universe may consist of all human pathogens (i.e., organisms known to cause disease in humans).*
- *The PCCL may consist of all organisms in the Microbial CCL Universe that may plausibly occur in and be transmitted by drinking water.*

- *Microbes in the Microbial CCL Universe may be evaluated against screening criteria to determine the plausibility for water-related transmission (occurrence). Pathogens that are known to cause waterborne disease (health effects) are placed on the PCCL.*
- *Surveillance and nomination provide an alternative pathway for entry into the CCL process for new and emerging microbial agents.*
- *EPA should continue to develop the process for selecting organisms on the PCCL for the CCL.*

These steps are discussed in the following sections of this chapter. A schematic representation of the recommended microbial CCL Classification Process is shown in Figure 3.1.

Figure 3.1 - A Microbial CCL Classification Process



Note that this process differs from NRC recommendations by defining the Microbial CCL Universe as microorganisms known to cause human disease. Microorganisms demonstrating the potential to cause human disease may be added to the Microbial CCL Universe when surveillance demonstrates adverse health effects or by nomination, based upon available data and information. The subset of human pathogens that may plausibly survive in and be transmitted by drinking water comprise the PCCL, and a subset of microorganisms on the PCCL that meet attribute scoring criteria (see Appendix D) are

placed on the CCL. Whereas the NRC recommended including both contaminants that have the potential to cause adverse health effects and those with the potential to occur in drinking water in the Universe of Potential Drinking Water Contaminants (see Venn diagram in Chapter 4.2), the Work Group decided upon a more stringent definition of the Microbial CCL Universe, since using the NRC's criteria would place large numbers of microorganism in the Microbial CCL Universe whose biological properties would prevent them from surviving in drinking water or causing human disease.

3.1 Identifying the Microbial CCL Universe

The Universe of known microorganisms includes bacteria, viruses, protozoa, algae, and fungi. Some microbes from each of these categories are pathogenic to humans, or produce toxins causing human disease. Pathogens that cause gastrointestinal disease (enteric pathogens) are shed in feces. Examples include *Salmonella*, *Shigella*, *Cryptosporidium*, *Giardia,* and noroviruses. Enteric pathogens of humans can be discharged into water by sewage treatment plants, septic tanks, storm sewer flows, runoff events after a rainfall, and other processes. Runoff from animal feeding operations and the fecal contribution of feral animals and migratory waterfowl also have the potential to introduce microorganisms, including human pathogens, into the aqueous environment. Other microbes are natural inhabitants of the soil and environmental waters, and are well-adapted to the low nutrient level and cool water temperatures of the ambient environment. Some aquatic microbes may cause disease in humans under certain circumstances, especially in individuals with a weakened immune system or other major underlying conditions that facilitate infection resulting in disease. Pathogens causing opportunistic infections include *Pseudomonas aeruginosa*, *Legionella pneumophila,* and the *Mycobacterium avium* complex (MAC). Many of the microorganisms in these two categories have not been identified. New microorganisms, including pathogens, are constantly emerging via evolutionary processes.

3.1.1 NRC Recommendations for the Microbial CCL Universe

The NRC recommended general guidelines for defining the Microbial CCL Universe of potential drinking water contaminants as those microorganisms that are known or have the potential to occur in drinking water, and those microorganisms that are known or have the potential to cause human disease from exposure to drinking water by ingestion, inhalation, or dermal contact. These guidelines recognize that knowledge of microbial occurrence and health effects is incomplete, and they provide latitude for inclusion of new and emerging pathogens as they are recognized. The NRC recommendation did not limit the boundary of the Microbial CCL Universe, but suggested that microorganisms could be added to that Universe based upon expert knowledge and new information about their occurrence and health effects. (The NRC also recommended the inclusion of biological toxins in the CCL Universe, but because these are produced and released in the ambient environment, they are addressed as chemical – not microbial – agents and contaminants.)

The NRC (2001) view of the Microbial CCL Universe included agents that occur naturally in water, agents associated with human feces, agents associated with human and animal feces, agents associated with human and animal urine, and agents associated with water treatment systems and distribution systems, together with biological toxins. Table 3.1 illustrates these categories and provides examples of microorganisms for construction of the Microbial CCL Universe of potential drinking water contaminants.

Table 3.1 - Categories and Examples of the NRC-Proposed Microbial CCL Universe

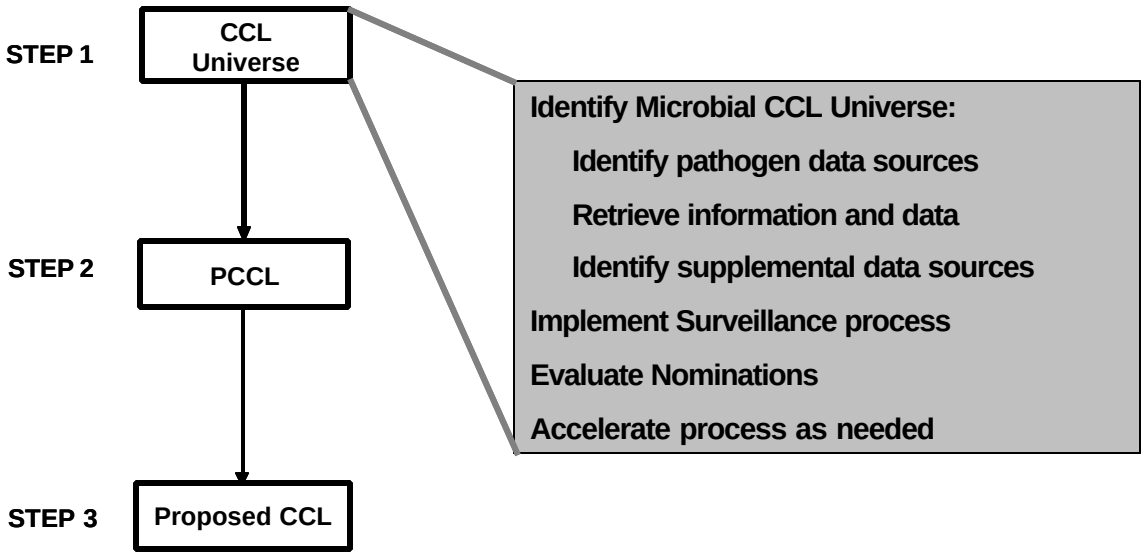
Category	Examples[1]
Naturally occurring agents in water	*Legionella*, toxigenic algae
Agents associated with human feces	Enteroviruses, coxsackie B viruses, rotavirus
Agents associated with human and animal feces	Enteric protozoa and bacteria
Agents associated with human and animal urine	Nanobacteria, microsporidia
Agents associated with water treatment and distribution systems	Biofilms organisms, *e.g. Mycobacterium avium-intracellulare*
Biological toxins	endotoxin, aflatoxin

[1] Some examples can belong to more than one category of contaminants.

3.1.2 Defining the Microbial CCL Universe

Figure 3.2 lists the actions the Work Group recommends EPA implement to identify the Microbial CCL Universe, and locates this step in the CCL process. The recommendations are discussed below.

Fig. 3.2 - Microbial CCL Universe



The Microbial CCL Universe may be framed after thoughtful consideration of all possible occurrences of aquatic microorganisms that may be present in all source waters (ground water, surface water, and marine waters where appropriate), water treatment plants, water distribution systems, plumbing, and recreational water venues supplied by treated drinking water. Microorganisms of primary concern in water treatment and delivery are those that cause human disease and are shed in feces. Pathogens associated with septic waste and sewage may contaminate ground water and surface waters, thereby posing a public health risk. *Salmonella*, *Shigella*, *Cryptosporidium*, *Giardia*, noroviruses, hepatitis A virus, and enteroviruses are examples of pathogens that are shed in feces and that may contaminate water, resulting in sporadic cases of illness or waterborne disease outbreaks. Many microorganisms causing water-related diseases in humans are not of fecal origin, but occur as natural inhabitants of the aquatic environment. *Legionella pneumophila, Aeromonas hydrophila, Pseudomonas aeruginosa,* and many other microorganisms associated with water-related opportunistic infections have their natural habitat in water.

Hundreds of microorganisms are known to be pathogenic, causing infectious diseases in humans, while thousands of microorganisms that may be present in the environment have the potential to cause infrequent opportunistic infections in humans under unusual circumstances of exposure and host susceptibility. Conversely, thousands of microorganisms found in the aquatic environment or in domestic water distribution systems are not known to cause adverse health effects in humans, regardless of their number or route of exposure. The diversity of the Microbial CCL Universe, and the wide range in host susceptibility of human populations, make it difficult to characterize microbial occurrence and the potential for adverse health effects for research and possible regulation.

Health and occurrence data may be more readily available for chemicals than for microbial agents. Existing health effects databases for microbes are based upon case reports from public health surveillance programs and epidemiological investigation of water-borne disease outbreaks. Existing microbial occurrence databases are based upon indicator monitoring, except for data acquired during epidemiological investigation of water-borne disease outbreaks, limited academic research studies on pathogen occurrence, and occasional regulatory information collection requirements (e.g., the Information Collection Rule (ICR) that required selected public utilities to gather information on *Giardia*, *Cryptosporidium,* enteroviruses, total coliforms and fecal coliforms). Limited sources of tabular data on occurrence and health effects of microorganisms are available on the Internet or elsewhere, however the content is frequently incomplete and the quality of data is variable. Because of the lack of pathogen occurrence data in readily accessible form, keyword searches of bibliographic databases of primary literature, conference proceedings, technical reports, monographs, and reference books will be required to adequately populate the Microbial CCL Universe. NRC (1999b) recognized the limitation of existing occurrence and health effects information sources for microbial contaminants, and suggested that expert judgment would remain an important component of the CCL process. Until a unified database of microbial information is available, the process of initial selection of microorganisms for the Microbial CCL Universe and subsequent iterations to move them through the CCL Classification process to the CCL will rely heavily upon expert judgment.

The term “Microbial CCL Universe” implies a subset of contaminants from a universe of all microorganisms. Because of the number of contaminants to be considered in the NRC recommendations, the Work Group discussed building an inclusionary CCL Universe by following the basic NRC principles and selectively combining data elements from data sources into a Microbial

CCL Universe. (This "data source compilation" approach is described in detail in Chapter 4.1.) Construction of a Microbial CCL Universe is envisioned to entail selective compilation of existing data sources into an inclusive and unified data set of known contaminant parameters, from many well-characterized sources of data and information for contaminants recommended by NRC. While several comprehensive sources of data and information have been developed for chemical occurrence and health effects, few equivalent data sources exist for microbes. Thus, the approach for selection of microbial contaminants for the Microbial CCL Universe may of necessity incorporate alternatives, based upon NRC guidelines, by using *information* from a variety of qualitative sources, including surrogate monitoring, modeling, primary literature review, and expert judgment.

→ **The NDWAC Work Group recommends that the Microbial CCL Universe be based on the evaluation of data sources and literature reviews that identify organisms known or suspected to cause human disease.**

A survey of the primary literature was conducted as an example of this approach. Appendix A from Taylor et al. 2001[5] was used as an illustrative starting point for the Microbial CCL Universe. This list includes 1,415 recognized bacterial, viral, parasitic, and fungal pathogens. This article represents an attempt to identify all the known human pathogens through a search of published literature. However, some human pathogens do not appear on the list as a result of recent emergence or taxonomy and nomenclature changes. Additions to the Taylor list have been proposed, and the Work Group suggested a mechanism for adding organisms to the Microbial CCL Universe through surveillance and literature review. Therefore, organisms that are known to cause water-related disease would be included in the Microbial CCL Universe by definition.

Because construction of the Microbial CCL Universe is constrained by limitations of readily available data, the Work Group recognizes the need for a nomination process to provide a means of adding new and emerging pathogens to the Microbial CCL Universe (see Chapter 2.3.3). Advances in genomics and proteomics offer the possibility that molecular techniques, such as the VFAR approach discussed in Section 3.4, may one day provide objective screening capability for selection of microbes for the Microbial CCL Universe.

3.1.2.1 Human Pathogens as the Basis for the Microbial CCL Universe

The Work Group discussed how the Microbial CCL Universe might be identified according to the principle-based, iterative approach (see Chapter 4.1.2) – giving full consideration to the differences between chemicals and microbes, and recognizing the limitations of equivalent data. Members also expressed concern over the blanket inclusion of all microbes with the potential to occur in water, or all microbes with the potential to cause disease, based on the current limited state of microbial occurrence and health data. Members believed that the biological properties of microorganisms controlling population diversity and dynamics should be considered in defining the Microbial CCL Universe. Admission of microorganisms to the Microbial CCL Universe is based upon a proven ability to cause disease in humans; thus autotrophs, thermophiles or other environmental microorganisms that may occur in water are excluded from the Microbial CCL Universe because their biological properties make it implausible that they could cause human disease.

---

[5]Taylor, Latham and Woolhouse. 2001. Risk factors for human disease emergence (Appendix A). Phil. Trans. R. Soc. Lond. B 256:983-98

3.1.2.2 Ensuring Inclusiveness of the Microbial CCL Universe

Considering the scope and diversity of microorganisms in the universe of potential water contaminants, and the relatively few comprehensive data sources on their occurrence and health effects, identification of the Microbial CCL Universe will rely on expert knowledge and keyword searches of available bibliographical databases to ensure inclusiveness, while maintaining perspective on the practical likelihood of water contamination and disease transmission.

The Work Group believes that adopting a published list of known human pathogens as the basis for the Microbial CCL Universe is practical and transparent in practice. However, the limitation of this approach in capturing suspected pathogens with the potential to occur in water requires development of a process to identify new information on emerging pathogens. The surveillance and nomination processes described in Section 2.3 ensure that the Microbial CCL Universe remains current.

3.2 Microbial CCL Universe to PCCL

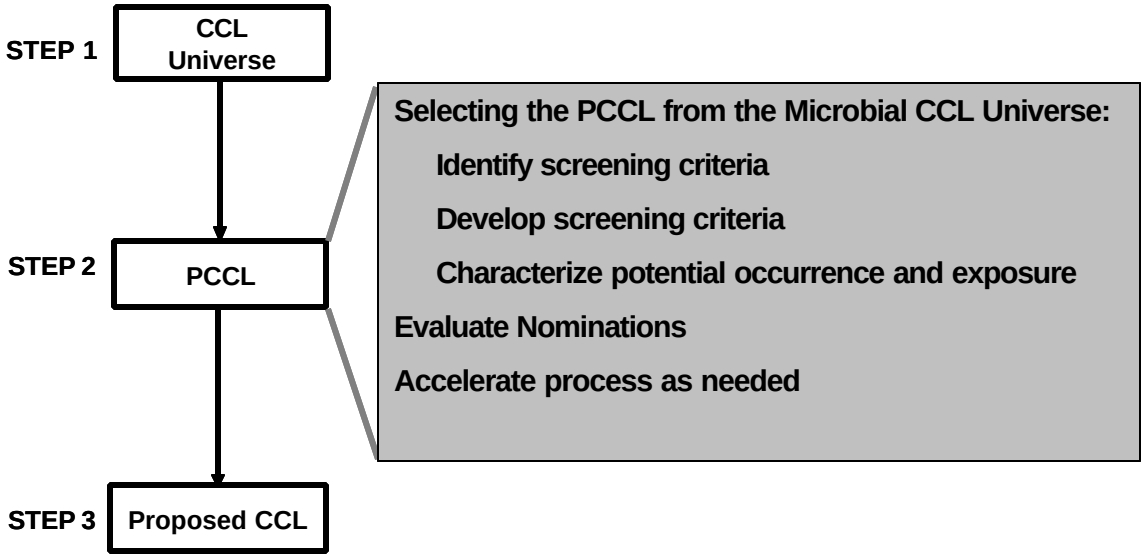
3.2.1 NRC Recommendations for the PCCL

The NRC recommends screening the CCL Universe based on evaluations of occurrence and adverse health effects. Occurrence information includes demonstrated or potential contaminants of ambient or finished water. This concept is illustrated by the list of waters considered by the NRC and the Work Group, e.g. tap water, distribution water, finished water, source water, and watersheds. Having thus broadly defined the CCL Universe based on known or potential occurrence or known or potential health effects, the NRC suggested using the characteristics of occurrence in water and pathogenicity to selectively screen microbial contaminants in the Microbial CCL Universe for inclusion in the PCCL. This screening process would be supplemented by expert judgment.

3.2.2 Screening Microbes for the PCCL

Figure 3.3 identifies the Work Group's recommendations for actions the EPA should develop to screen the Microbial CCL Universe.

Figure 3.3 - Screening Contaminants from the Microbial CCL Universe to the PCCL



The composition of the Microbial CCL Universe and PCCL recommended by the NDWAC Work Group differs slightly from NRC recommendations. Only microorganisms demonstrated to cause human disease would inhabit the Microbial CCL Universe. Because most human pathogens do not occur in water, or lack biological characteristics that permit their survival in water, it is plausible to limit the Microbial PCCL to those human pathogens that may be transmitted by water, and only human pathogens with the potential to occur in water would comprise the PCCL. (See Figure 3.1, above.) A mechanism to prioritize and reduce this list of microorganism for evaluation and ranking is needed, and the Work Group's rationale for accomplishing this selection process centered on two questions:

- *What are the biological characteristics of pathogens that determine the potential for their occurrence in drinking water?*
- *How may a screening process be constructed to identify pathogens for evaluation and possible addition to the PCCL?*

Work Group members recognized that criteria are necessary to selectively identify pathogens to include on the PCCL. The Work Group suggests adoption of a rule-based selection process for moving pathogens from the Microbial CCL Universe to the PCCL. These principles may not be sufficient by themselves, and expert judgment may be needed.

→ **The Work Group recommends that the selection of human pathogens for the PCCL start with a Microbial CCL Universe of recognized human pathogens (e.g., the amended Taylor et al. 2001 list), and that those pathogens known to be associated with source water, recreational water, and drinking water be selected for inclusion into the PCCL.**

The resulting Microbial PCCL should be based on natural habitat and biological characteristics that indicate a pathogen's ability to be transmitted via water. The members identified a simple key to identify organisms that should move to the PCCL (See 3.2.3). Microorganisms having the potential to cause human disease, but not yet demonstrated to do so could be added to the Microbial CCL Universe by identification of genomic or proteomic elements suggestive of virulence as this technology develops (See 3.4). Newly recognized microbes associated with waterborne disease would be added to the PCCL as a result of public health surveillance processes already in place, or by an expert or stakeholder nomination process. This process results in a realistic Microbial CCL Universe and PCCL. The NRC acknowledged that practical limitations (i.e., genomic and proteomic data availability) would constrain the development process, and this Work Group proposal attempts to be consistent with NRC principles while acknowledging those limitations and proposing reasonable and creative solutions.

3.2.3 Screening Based Upon Biological Properties

The Work Group applied further selective principles to restrict the number of microbes on the PCCL to those meeting plausibility criteria in addition to occurrence. Examples of the proposed screening principles that would exclude pathogens from the PCCL are shown in the table below.

Table 3.2 - Proposed Screening Principles to Exclude Pathogens from the PCCL

Proposed Screening Principles
Obligate anaerobes (microorganisms that cannot survive in oxygenated environments)
Obligate intracellular pathogens (environmental survival in water implausible)
Pathogens transmitted exclusively by direct or indirect contact with blood or body fluids (including sexually transmitted diseases)
Pathogens transmitted exclusively by insect vectors
Normal human intestinal, skin, or mucous membrane flora (except when documented to cause water-related disease)
Pathogens transmitted exclusively by respiratory secretions
Pathogens transmitted exclusively by animal bites
Pathogens of animals that are not known to occur in humans (limited host range)
Pathogens causing rare occurrences of disease not associated with water-related transmission

Several pathogens are transmitted by multiple transmission routes, and they would have to be evaluated individually for plausibility of drinking water transmission by ingestion, inhalation, or dermal contact. For example, aerosol transmission of pathogens such as *Mycobacterium* spp. and *Legionella* spp. places them on the PCCL. Respiratory pathogens must be evaluated individually for

plausibility for water transmission. Ten species of the genus *Bacillus* appear in Taylor et al. Appendix A, while only two species are characteristically associated with human illness, and neither of these species represents a significant risk by drinking water transmission. These examples illustrate that, as the Agency develops the CCL classification process, it should refine screening criteria to better identify microbial contaminants that pose risk through drinking water transmission.

→ **The Work Group supports the following concepts for EPA's consideration as they develop future CCLs:**

- ***Biological characteristics should be recognized as legitimate criteria for screening pathogens for the PCCL.***
- ***The list of pathogens inhabiting the Microbial CCL Universe should be screened for biological characteristics promoting or mitigating against survival and transmission in water.***
- ***Genera may be categorically excluded as long as provisions are made for selective exemption of single species of a genus, e.g.* Bacillus anthracis.**

3.2.4 Pathogens Associated with Opportunistic Infections

Many organisms in the Microbial CCL Universe might be included because of their implication in a very few cases of disease, perhaps only a single case. Some of these organisms, such as *Erysipelothrix rhusiopathiae* or *Pantoea agglomerans*, can be found in water.

→ **The Work Group recommends that organisms associated with opportunistic infections be excluded from the PCCL unless clinical, epidemiological, or similar other information implicates them as the potential or known cause of waterborne disease. The Work Group suggests that EPA increase surveillance for infections caused by these organisms, especially in sensitive subpopulations.**

This increase in surveillance, in the Work Group's view, should balance these organisms' exclusion from the PCCL. These organisms would be selected for exclusion by a consensus of expert opinion. Opportunistic pathogens that cause a higher incidence of disease and are normal inhabitants of water (e.g., *Mycobacterium avium* complex and *Pseudomonas aeruginosa*) would not be excluded from the PCCL using this screening procedure.

3.2.5 Alternative Pathways for Adding Pathogens to the Microbial CCL Universe and the PCCL (Surveillance and Nomination)

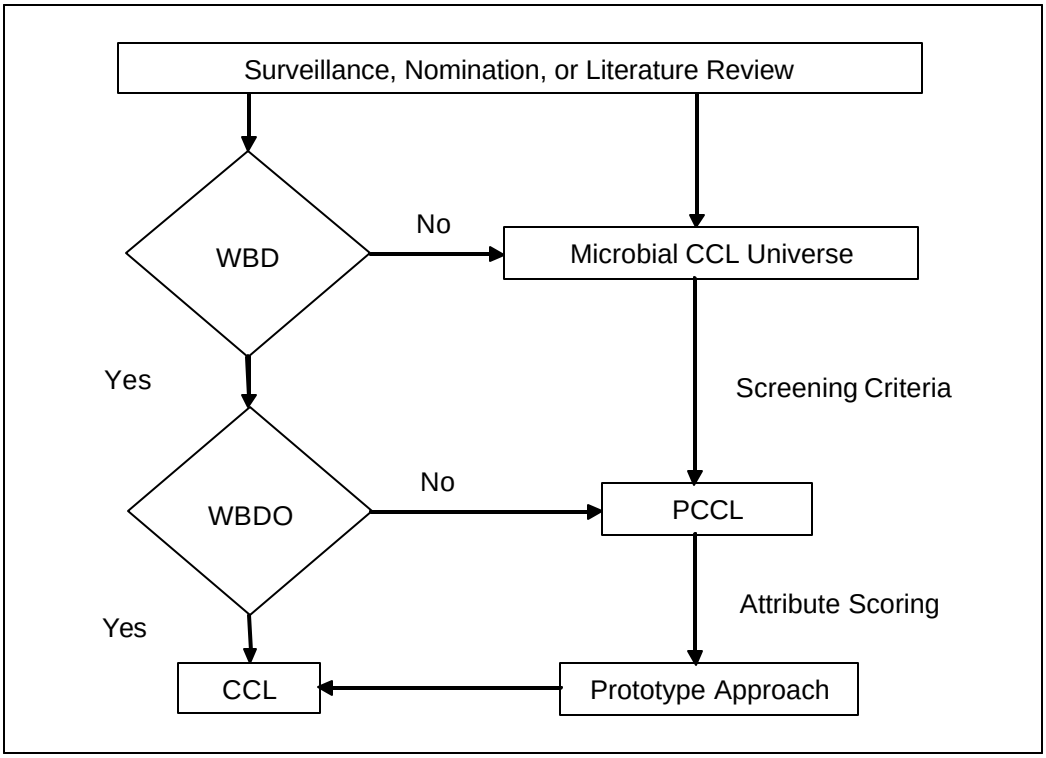
→ **The NDWAC Work Group recommends the development of procedures to include some microbes in the PCCL or CCL outside of the defined process for microbes.**

The dynamic nature of pathogen emergence necessitates use of surveillance data to develop a Microbial CCL Universe that is identified from current information. Other mechanisms for placement of organisms into the Microbial CCL Universe, onto the PCCL, or directly onto the CCL include genomic evidence of pathogen potential, recognition of new water-related waterborne disease agents, identification of a waterborne disease outbreak by an organism not previously known to cause such outbreaks, or nomination of organisms by experts based upon epidemiological data.

Figure 3.4 presents an example of an alternative pathway for expeditiously incorporating emerging pathogens into the CCL classification process. This alternate pathway shows the different types of information to be considered and how a pathogen may be incorporated at different stages in the CCL process. This approach provides a means of integrating pathogens identified through public health surveillance programs, identified by using VFARs, or nominated by experts or stakeholders into the CCL process.

Figure 3.4

Alternative Pathways for Introducing Pathogens to the CCL Classification Process



Emerging pathogens recognized through surveillance or nomination are evaluated for their potential to cause waterborne diseases (WBDs), based upon their biological characteristics and epidemiology. Pathogens recognized to cause waterborne disease are next evaluated for their involvement in recognized waterborne disease outbreaks (WBDOs). Pathogens causing waterborne disease outbreaks may be placed directly on to the CCL. Emerging pathogens with no evidence suggesting their involvement as agents of waterborne disease are placed in the Microbial CCL Universe for further observation. Pathogens causing WBDs but not recognized to cause waterborne disease outbreaks are placed on the PCCL.

The selection of pathogens from the PCCL for inclusion on the CCL is accomplished by scoring attributes as recommended by the NRC (see section 3.3, below).

3.3 Use of Attributes to Classify Microbial Contaminants

3.3.1 NRC Recommendations for Classifying Microbial Contaminants to the CCL

This step in the CCL classification process is intended to reduce the Microbial PCCL to a list of priority pathogens for the CCL. The Work Group concurs with the NRC and recommends that EPA consider a "prototype classification" algorithm (discussed in Chapter 5) to classify contaminants using attributes that characterize occurrence and adverse health effects. This step is dependent upon identification of available data (i.e., what is known about occurrence and health effects of potential contaminants) and quantifying the attributes for use in the prototype classification algorithm. Expert judgment is considered an important component of this step, as it is in the overall process. NRC further recommended that a single approach be developed for selecting chemical and microbial contaminants, requiring the development of predictive measures of pathogen occurrence and virulence. (This is further discussed in Chapter 5.)

Using a prototype algorithm to classify contaminants on the PCCL for consideration for inclusion on the CCL, the NRC selected five attributes to represent the contaminant's ability to cause health effects or potential to occur in water. The Work Group has adopted the health effects attributes of **potency** and **severity**, and the occurrence attributes of **prevalence**, **persistence/mobility,** and **magnitude**, as starting points for evaluating and ranking agents as recommended by the NRC. (See Chapter 2, section 2.2.2.3, for the NRC's definitions of these attributes.)

3.3.2 Use of Attributes for Characterizing and Ranking PCCL Microbes

The Microbial PCCL consists of human pathogens that are documented to be or may be transmitted by drinking water; however the occurrence and health effects of these pathogens range from rare and life-threatening to common and self-limiting. One way to prioritize pathogens for placement on the CCL is to evaluate them for attributes as described above. To quantify attributes, it is necessary to use data elements for occurrence and health affects that are most appropriate for microbes. Components reviewed by the Work Group for constructing scoring protocols are shown in Appendix D.

The terms **potency**, **severity, prevalence** and **persistence/mobility** and **magnitude** most clearly relate to chemical risk assessment models and practices. The Work Group recognizes that these attributes, the protocols used to characterize contaminants for each attribute, and the data and information used will need to be considered in context to describe the multiplicity of factors involved in infective processes in humans. Pathogen occurrence, infectivity, and host susceptibility and immune response determine the outcome of the host-pathogen relationship. An understanding of these terms in the context of host-pathogen relationship is a prerequisite to assessments of microbial health effects.

For example, the *potency* attribute characterizes the amount of a contaminant required to cause an adverse health effect: i.e., an infective dose for a susceptible host. A potential pathogenic microorganism must be viable, infective, and virulent. The pathogen-host relationships determine the course of disease in the host, which relates most closely to the term *severity* from the recommended attributes. The outcome of the pathogen-host interaction is manifested in a spectrum of disease ranging from asymptomatic infection to death of the host. The pathogenicity of the microbe, the mode

of transmission, and the population susceptibility determine the *magnitude* of the health effects. Magnitude, in a microbiological context, was defined as the extent to which the pathogen can cause disease outbreaks or significant numbers of individual cases above the endemic burden of disease in the population.

To support the CCL process, EPA assembled a Microbial Sub-group comprised of microbiologists and risk assessors in the Office of Ground Water and Drinking Water, Office of Science and Technology and the Office of Research and Development to develop draft attribute scoring protocols based on the NRC recommendations. The draft protocols are provided in Appendix D of this report. The elements considered for each of the attribute protocols needed to take into account the data that were available and the expert judgment required to score each attribute. As with the development of the Microbial CCL Universe and screening criteria of microbes from the Universe to the PCCL, the available information was found in the primary literature and not in developed databases.

3.3.3 Developing Draft Protocols to Quantify Attributes

The Work Group had the opportunity during its deliberations to review and evaluate the draft set of specific attribute protocols for microbial contaminants that were developed by EPA. In addition, some Work Group and technical support team members worked closely with EPA staff during its deliberations to develop some specific technical guidelines to quantify attributes for the microbiological contaminants. The Work Group provides general recommendations to develop and evaluate attributes in Chapter 5. The remainder of this section summarizes the discussion specific to microbes and the five attributes. These discussions identified issues EPA should consider in refining the attributes and methods to quantify attributes for microbes and the CCL CP, and are presented below.

Potency. Health effects attributes include potency and severity. Potency is defined by the NRC as the amount of a contaminant that is needed to cause illness. For microbes the infective dose is the most useful marker of potency, however the infective dose is not known for many pathogens. Microbiologists frequently speak in terms of the minimum infective dose, but the terms LD_{50} and lethal dose apply only to animal studies or *in vitro* cell culture assays. Some pathogens cannot be grown in the laboratory and their infective dose can only be estimated. In the future, quantitative virulence-factor activity relationships may become available for determining the relative potency of a pathogen. The draft attribute characterization for potency scoring is constructed in a manner to allow for absent data elements, while admitting the use of available information. The data elements that should be considered for potency include knowledge of water-related disease, the class of pathogen (i.e., bacteria, viruses, protozoa), the burden of disease in the population, the infective dose of the pathogen, the likelihood of fecal or urinary shedding in humans and animals, and the presence of genomic sequences conferring virulence. The attribute score was derived by categorizing the pathogens according to a hierarchical scheme that started with data likely to be known for all organisms such as knowledge of waterborne disease, and then subcategorized using data less likely to be known like morbidity and infective dose. Each layer of subcategorization provides increased resolution for the score. However, even those with minimal amounts of data will receive a score commensurate with what is known. A proposed system for scoring potency is shown in Appendix D, Table D2.

Severity. NRC defines severity as the seriousness of the health effect, and suggests severity be based on "the most sensitive health endpoint for a particular contaminant, and considering vulnerable subpopulations; ... [and] should be based, when feasible, on plausible exposures via drinking water." For microbial agents, severity may be defined in terms of colonization, infection, immune response, disease, sequella, or death. The host-pathogen relationship is variable and dynamic. This continuum may be unrecognizable at various stages. The most sensitive endpoint indicative of host-pathogen interaction is an immune response, however this is not a practical end point for assessment of health effects, since immunodeficient populations may be infected without eliciting an immune response. While chemical health effects may be immediate or cumulative, microbiological health effects may be unapparent for an extended time, depending upon the incubation period of the pathogen, and the manifestation of disease. The data elements for scoring severity include recognition of significant morbidity and mortality, the location and intensity of infective processes, the extent of contagion, the amount of time lost to illness, the extent to which medical intervention is required for recovery, and chronic manifestations or disabilities associated with the disease.

A central issue with severity scoring is whether to score on acute manifestations of disease in normal populations, or to score the worst possible outcome in the most sensitive population. Because most frank pathogens are capable of killing some segment of the population, using worst possible outcome in the most sensitive host inflates and clusters scores. The initial severity scoring tables were constructed to use median outcome in normal populations, with case fatality rate and patient population classification and percentage of patients in the population classifications as weighting factors. One such approach applied the attribute characteristics to the population for which the most data and information were available, then recalculated scores to acknowledge special circumstances and to apply additional stringency. This proposed system applied worst case scoring criteria for healthy and sensitive sub-populations, thereby driving many pathogens to maximal scores.

In an effort to overcome the complexities and limitations of a scoring system using case fatality rates and population-based weighting factors, the Work Group proposed a series of questions carefully constructed so that a 'yes' answer indicated significance while a 'no' answer did not. This approach was useful in sequentially determining the cumulative data elements contributing to severity of disease for both normal and sensitive sub-populations. Examples of the questions include:

- *Does the organism cause significant morbidity (> 1,000/year) in the U.S.?*
- *Does the illness require medical intervention for resolution?*
- *Does the organism cause mild disease in normal populations, but severe disease in individuals with predisposing conditions?*
- *Does the organism cause pneumonia, meningitis, hepatitis, encephalitis, endocarditis, or other severe manifestations of illness?*

The more questions answered affirmatively the higher (more severe) the score. (See Appendix D, Table D3.)

Prevalence. For the occurrence attributes, NRC defines prevalence as, "How commonly does or would a contaminant occur in drinking water?" Prevalence may be determined using six of the seven measures proposed by NRC in the PCCL screening criteria for demonstrated or potential occurrence. In order of preference, these are: demonstrated occurrence in (1) tap water, (2) distribution systems,

(3) finished water of water treatment plants, and (4) source water used for supplying drinking water; and, if no information is available to demonstrate occurrence in water, (5) observations in watersheds/ aquifers, or (6) historical contaminant release data. It should be emphasized that prevalence involves the consideration of both geographical (spatial) and temporal ranges of occurrence. Most pathogen occurrence data are based upon indicator monitoring, hence they become surrogate information, not pathogen occurrence data. Pathogen occurrence data come from epidemiological investigations following outbreaks, research studies on pathogen distribution, and detection method evaluations. There is little pathogen outbreak-occurrence information and even less pathogen data regarding environmental and drinking water occurrence.

The Work Group developed a conceptual framework for prevalence, based upon actual detection in drinking water, actual detection in source water, potential for zoonotic transmission through water contamination, and potential for zoonotic agents to infect humans (host range). As with potency, a simple scheme of hierarchical categories was tested, with the first category dividing pathogens by their presence or absence in drinking or source water and subcategorization according to any known estimate of the frequency. These hierarchical categories are the basis of Appendix D, Table D4. Prevalence scoring using these criteria proved to be more straightforward than other attributes, primarily because occurrence data are either available or not available, limiting the number of criteria in the scoring system.

Persistence/mobility. NRC used a persistence/mobility attribute as a surrogate for potential occurrence when information is unavailable for a contaminant regarding its demonstrated occurrence in water. For microorganisms, the following three characteristics pertain to their persistence and/or mobility: high potential for amplification under ambient conditions; sedimentation velocities and absorption capabilities; and, death or the ability to produce non-culturable or resistant states (e.g. spores and cysts). When a contaminant already has data on demonstrated occurrence in water, and thus information for the prevalence and magnitude attributes, those attributes will take precedence over persistence/mobility.

Persistence implies steady state occurrence or amplification of microorganisms in water. This occurs in surface water by production of resistant forms such as spores, cysts, oocysts; by colonization of other life forms serving as a reservoir; through symbiotic relationships with amoebae; by adsorption to particles; or by production of quiescent forms such as viable but non-culturable bacteria. In water treatment plants and distribution systems, persistence is associated with colonization of infrastructure, e.g. production of biofilms. Organisms that amplify (grow and multiply) are given higher scores than organisms that produce resistant forms but do not amplify in water. This scoring scale may overemphasize relatively innocuous organisms that produce biofilms but rarely or never cause disease in humans.

Data elements for scoring persistence-mobility include survival time in water under ambient conditions, ability to amplify, ability to produce resistant forms, relationship to particles, and potential for symbiotic relationships enhancing survival. The persistence-mobility scoring table (Appendix D, Table D5) emphasizes non-turbid waters (i.e. ground water and treated drinking water), but the Work Group believes that all source water should be included in scoring. Amplification frequently occurs in surface source water where there are large amounts of available nutrients, whereas the assimilable organic carbon is limited in ground water and treated water, slowing or restricting amplification.

Persistence of bacteria that amplify under environmental conditions is highly variable, and the extent to which they persist and move is largely a function of their population density.

Persistence of bacteria in drinking water is frequently related to the ability of bacteria to produce biofilms, which promote growth of heterogenous bacterial populations while protecting them from disinfection. (Biofilms are dynamic populations of bacteria that slough and serve as a steady-state source of bacteria.) Persistence of bacteria in the environment is determined by physical conditions such as temperature and pH, availability of nutrients, presence of predators, and the ability of microorganisms to form capsules, slime layers, spores, cysts, or other resting forms. Mobility of bacteria in water is passively dependent upon hydraulic flow, which may suspend bacteria adsorbed to particulate material and sheer microcolonies from biofilms. Because of the number and unpredictability of these variables, it may be inappropriate to equate persistence-mobility of organisms in surface waters with persistence-mobility in non-turbid waters such as ground water or treated drinking water.

Mobility is not limited to chemicals, since microorganisms move though the aqueous environment and in distribution system water actively (motility) and passively (adsorbed to particulates, in symbiotic relationship with amoebae, and by hydrostatic flow). Organisms percolate through soil layers to contaminate ground water. Viruses are particularly mobile because of their extremely small size and their relatively long survival times in the environment. Because mobility is associated with the hydrodynamics of distribution systems, presence of biofilms, presence of particulates, and opportunity for symbiotic relationships, it is considered together with persistence for scoring purposes.

Magnitude. NRC defines magnitude as "the concentration or expected concentration of a contaminant relative to a level that causes a perceived health effect" (NRC 2001). For characterizing the attribute of magnitude, ideally two data elements are needed: the concentration of a contaminant in water, and the concentration associated with an adverse health effect. NRC recommended the use of a median water concentration in combination with a measure of potency, if available. Magnitude, in a microbiological context, implies delivery (persistence-mobility) of an infective dose (potency) to the customer's tap with resulting illness. The Work Group scored magnitude according to the number and frequency of waterborne disease outbreaks reported in the U. S. and around the world, pathogen distribution, and biological properties determining pathogen distribution. A draft scoring table is shown in Appendix D, Table D6.

The microbial contaminants considered for preliminary scoring exercises were drawn mostly from the current CCL. A set of seven microbes was used by the EPA Microbiology Sub-Group, and a set of eleven microbes on the current CCL plus *Pseudomonas aeruginosa* was used for a scoring workshop sponsored by AWWA in November 2003. These contaminants were not sufficiently representative of the range of pathogens likely to occur on the PCCL to adequately test the validity of the scoring algorithms. However they did provide participants with examples to test the scoring protocols and provide suggestions to refine the scoring protocols.

The Work Group noted that attribute scoring using the EPA attribute scoring algorithms requires expert knowledge based upon text-based literature to assign scores. Preliminary scoring exercises conducted by different individuals produced different scores, and the same individual scoring organisms on different days may produce slightly different scores as a function of the basic

assumptions entertained at the time. This variability suggests that scoring exercises should be conducted by several experts and the results combined to arrive at final scores and rankings. The Work Group also noted that the scoring will need to document assumptions and data or information used to score the attributes. The present algorithms do not lend themselves to automated scoring and will require expert judgment and interpretation of text based sources. Nevertheless, the scoring algorithms while considering a broad range of available information are relatively simple, thus scoring can be performed easily and updated as necessary. Even in the absence of many bits of information a reasonable attribute score can be determined, and as additional data become available the scores can be refined. The simplicity is appropriate given the triage nature of the CCL, and makes the process readily transparent.

The scoring algorithms proposed result from a lack of tabular data in organized databases. They are based upon premises relating to health effects and occurrence of pathogens, supported by text-based resource materials and expert knowledge. While existing genomic databases may eventually facilitate a more objective approach for selecting genomic sequences associated with virulence of microbes, databases containing health effects and occurrence data elements are only now being considered. It is unlikely that a unified, searchable database of relevant data elements will be available for selection of microbes for the CCL for several years. Meanwhile, expert processes will be required to conduct attribute scoring, to evaluate the validity of scoring results, and to determine the threshold for placing agents on the CCL.

The Work Group recognizes that the preliminary exercises using this scoring approach have not attempted to reconcile scores to produce a composite result for each pathogen in the test data set, thus the plausibility of resulting pathogen rankings has not been evaluated fully. Likewise, no attempt has been made to date to evaluate and rank combined chemical and microbial scores resulting from attribute scoring exercises.

3.4 Applications of Genomics to the CCL Classification Process

The final section of this chapter summarizes the NRC recommendations regarding application of virulence-factor activity relationships (VFARs) to the CCL Classification Process, describes potential applications of functional genomics and proteomics in the context of the CCL process to interested stakeholders, and outlines possible short- and long-term options for further deliberation.

3.4.1 NRC Recommendation on Genomics

The NRC recommended use of VFARs for predicting the virulence of waterborne organisms as a companion approach to quantitative structure-activity relationships (QSARs) for chemicals. Rapid advances in bioinformatics, functional genomics and proteomics, together with development of powerful molecular analytical tools such as polymerase chain reaction (PCR) and microarrays (bio-chips), provide the technology to screen microorganisms at the genetic level even when their genomes have not been fully sequenced. Theoretically, genetic elements coding for surface proteins, toxins, attachment factors, invasion factors, or other virulence descriptors that are shared by microbial pathogens can be identified and related to behavioral traits mediating severity, potency and persistence. Thus, VFARs may be used to detect potential pathogens, and to rank or score attributes pertaining to occurrence and adverse health effects.

3.4.2 Potential Applications of Genomics

Bacteria sense environmental conditions, and respond to host exposure by turning on genes that enhance environmental survival or their ability to invade host cells and cause disease. Polysaccharides contained in the bacterial cell envelope and elaborated into the immediate cell environment (capsule and slime layers), attachment mechanisms, symbiotic relationships with other microorganisms, and induction of a quiescent state all facilitate environmental survival, while activation of adherence, invasion, toxin production, and various secretory genes facilitate pathogenesis. Shared nucleic acid sequences (conserved regions of chromosomes) within gene clusters associated with virulence (pathogenicity islands) may be related to the NRC attributes of severity and potency. Likewise, shared nucleic acid sequences within those gene clusters associated with survival in the environment may be related to the NRC attributes of prevalence and persistence.

The genetic basis of these responses to environmental and host stimuli can be targeted for construction of VFAR gene databases to screen for the presence of VFAR genes in other organisms sharing virulence- or survival-related gene sequences. These VFAR sequences may be used to rank or score the NRC attributes pertaining to occurrence and adverse health effects, and they may serve as primer sequences for PCR for molecular detection of virulence genes in unrecognized pathogens, for direct detection of pathogens in environmental samples, and for construction of microarrays containing hundreds of VFAR gene sequences for rapid screening of microorganisms for their pathogenic potential. These genomic applications may eventually be sensitive enough for detection of non-culturable microorganisms, and for the direct detection of pathogens in environmental samples. For pathogens known to cause waterborne outbreaks, occurrence data alone may be sufficient for inclusion on the CCL.

It is theoretically possible to select large numbers of genes associated with virulence and to use these genes to screen the microbial universe for selection of potential pathogens for the PCCL. By using suites of functionally related genes associated with bacterial pathogenicity islands, it may be possible to further select and prioritize potential pathogens from the PCCL for inclusion on the CCL, based upon genetic function to enhance survival, manifested as potency, and mediating severity of disease. A categorical scoring system might be constructed, based upon the number of VFAR genes identified and upon possession of functional suites of genes assigned to each NRC attribute.

Microarray technology is developing rapidly, together with knowledge of the molecular basis of virulence. Microarrays have been used to detect viruses in cell cultures and clinical specimens, thereby demonstrating the feasibility of the technology for pathogen detection. Similarly, microarrays may be constructed to screen the microbial universe for potential pathogens for inclusion on the PCCL, and to further prioritize PCCL organisms for inclusion on the CCL. An extension of this technology would be the application of proteomics by constructing microarrays to detect gene products associated with persistence or pathogenesis.

3.4.3 Challenges to Use of Genomics

Microbial genomes exhibit considerable plasticity, with frequent acquisition and loss of genetic elements. The presence of multiple mobile genetic elements (e.g. bacteriophage, plasmids, transposons, insertion sequences, etc.), together with the relative frequency of chromosomal recombinations, results in highly dynamic genomes that confound predictability. Presence of virulence

factor genes does not automatically result in expression of gene products; indeed, some genes (toxins, pili, etc.) are controlled by multiple transcription regulators. The presence of enzymes that hydrolyze nucleic acids (nucleases) in the environment, and substances in environmental samples that interfere with PCR reactions (PCR inhibitors resulting in matrix interference) mitigate against detection of identifiable free nucleic acid sequences in environmental samples. Pathogens are typically present in the environment at concentrations below the detection limit of PCR, and culture enrichment techniques are necessary before PCR may proceed successfully. Finally, validation of PCR methods is problematic, thereby restricting its application.

An inherent limitation of genomics and proteomics is that they only recognize known gene functions. Spontaneous mutations cannot be predicted, and heretofore unrecognized gene functions will not be included in virulence factor screens. Available genomic information on microorganisms is variable. While many viral genomes have been sequenced, relatively few bacterial, and even fewer protozoan genomes are known. Sequences deposited in GenBank or other genomic databases are frequently incomplete, and accompanying annotations describing gene function frequently are speculative. Sequence quality is highly variable, and no mechanism exists to assure data quality. Because genomic databases are constructed using known microorganisms, genes, and fragmentary sequences, the VFAR approach has limited predictive value for anticipation of pathogen emergence at this time. Currently, the only means of recognizing emerging pathogens is after they have caused outbreaks, significant morbidity in a population, or serious outcomes in a few cases.

3.4.4 Pilot Projects

Initial explorations conducted for the Work Group using genomic databases to identify VFAR genes were based upon virulence mechanism keyword searches. These searches identified variable numbers of sequences for potential waterborne pathogens, but the vast number of unrelated sequence matches, and the scant number of whole bacterial genomes precluded use of these sequences for screening purposes. Genomic database searches based upon known virulence gene sequences published in peer-reviewed literature detected shared sequences among bacteria, but revealed little information about gene regulation and expression. These gene sequences have potential in selection of microorganisms from the universe for inclusion on the PCCL. Genomic database searches based upon mobile genetic elements, e.g., plasmids and pathogenicity islands, revealed multiple virulence-associated sequences that were widely shared and transferred among bacteria. These suites of genes offer promise for selecting organisms from the PCCL to the CCL by using them to score NRC attributes.

A pilot project constructed a web-based database compiling information on organisms, outbreaks, and genomic data on waterborne pathogens that can be used to identify potential pathogens for inclusion on the PCCL and prioritization of pathogens on the basis of potential virulence for inclusion on the CCL. This database relies upon sequences deposited in GenBank or other genomic databases, together with occurrence and epidemiological data on individual pathogens. This web database is not expected to have predictive value for emerging pathogens.

Another pilot project was devoted to whole genome alignments of viruses and bacteria, to identify conserved sequences that may be used to screen potential pathogens for virulence potential. The approach depends upon the availability of whole genome sequences of the pathogens of interest. Once unique VFAR gene screening sequences are identified, they could be used to screen other

potential pathogens by sequence alignment using custom databases, or by constructing microarrays based upon genomic or proteomic technologies.

3.4.5 Recommendations for the Use of Genomics in the CCL Process

Genomics and proteomics represent powerful tools for elucidation of pathogenic mechanisms of microorganisms; however, there are serious limitations to this technology that affect its application to the CCL Classification Process.

- The technology is largely unproven for the desired applications.
- The technology may not be available in a robust form for use in the next CCL.

Despite these limitations, the Work Group recommends two steps that can be taken to apply genomics to the CCLCP process.

→ **Select known virulence genes of gastrointestinal pathogens to identify data for screening unknown organisms.**

These data are based upon published sequences deposited in genomic databases, and can be used to screen potential pathogens as their sequences become known, to construct microarrays for screening potential pathogens, and ultimately for construction of PCR enhanced microarrays for direct detection of potential pathogens from environmental samples. These data may be used for selection of potential pathogens from the universe for the PCCL.

→ **Select clusters of known virulence genes contained within pathogenicity islands of chromosomes or contained in mobile genetic elements that code for major mechanisms of pathogenicity, e.g. adhesion, invasion, toxin production, etc.**

These suites of genes contain both core function and regulatory control for gene expression, and they are known to confer virulence when transferred to previously non-pathogenic bacteria. Selected pathogenicity islands including genes responsible for attachment pili, protein secretory systems, and toxin production may be used to rank and score attributes to facilitate selection of PCCL organisms for the CCL.

Both of these processes may be implemented in the near term using published literature and genomic databases. However, the microarray applications may be delayed until technical and financial limitations are resolved.

Both genomics and proteomics are developing rapidly and it is probable that microarrays or other evolving technology will be available to facilitate selection of potential pathogens for the PCCL and the CCL by 2010. Meanwhile, expert judgment, based upon published literature, epidemiological investigations, and public health surveillance, remains the key approach for selecting potential pathogens for the CCL. Expertise in bioinformatics and the molecular mechanisms of microbial pathogenesis are desirable additions to microbiology, epidemiology, and water treatment expertise of expert review panels.

To use VFARs to identify pathogens for inclusion in the CCL process, a wide variety of information needs to be integrated. These types of information usually are not found in a single

database but rather in a number of databases. The components of the information that need to be integrated are identified in the NRC report. In many cases the NRC report says these things can be done, but it does not specify the detailed logistics of doing this. The logistics of locating and integrating these different types of information should be explored in detail.

→ **To ease the process of automating microbial CCL processes in the future, the Work Group has identified a number of practical measures that EPA should implement in the short-term:**

- ***EPA should find or define an approach to evaluate data incrementally and in a manner that will readily allow application of search and match techniques to approximate the process QSARs use in eliciting structural similarity (and inferentially, similarity of effect) between structures in known and unknown organisms or genomic fragments.***
- ***EPA should find or provide a physical repository (ie, data warehouse) for that material.***
- ***EPA should monitor the progress of genomics and the related technologies and integrate them into the CCL process.***
- ***EPA should monitor the data and information that emerge as genomics progresses and integrate them for consideration in the CCL process, using an automated process to the extent possible. The process should be updated and maintained in a continuing process and verified against expert opinion.***
- ***EPA should review public health surveillance techniques, in conjunction with the Center for Disease Control (CDC), with a view to making those techniques as proactive, robust and effective as possible in identifying the occurrence of waterborne or watershed disease outbreaks and the organisms associated with those outbreaks.***

Chapter 4

CCL Classification Approach for Chemical Contaminants

The purpose of this chapter is to review the NRC recommendations and present the NDWAC Work Group recommendations for the CCL Classification Process for chemicals. Section 4.1 offers principles and a process for identifying a Chemical CCL Universe that is as inclusive as possible with respect to potential drinking water contaminants. Section 4.2 discusses recommendations for selecting a PCCL from the Chemical CCL Universe and provides potential screening approaches. Section 4.3 addresses the attributes used to characterize chemical drinking water contaminants, and the use of different types of information and data to quantify those attributes for further decision-making.

4.1 Building the Chemical CCL Universe

4.1.1 Summary of NRC Recommendations

NRC (1999b and 2001) noted that an "ideal CCL development process" would identify the entire Universe of potential contaminants and use data-driven screening processes to reduce the agents under consideration to a CCL listing only those contaminants with a high probability that they need to be regulated. However, the NRC recognized that currently the process is far from ideal – no comprehensive list of potential drinking water contaminants yet exists; health effects, occurrence, and other related data for the vast majority of potential contaminants are highly variable, poor or nonexistent; and EPA's resources are constrained (NRC 2001).

The NRC focus on a Universe of agents with "demonstrated or potential occurrence and/or demonstrated or potential human health effects" is illustrated by examples of the kinds and classes of agents recommended for consideration (see Tables 3-1 and 3-5 of the NRC report). The NRC also suggested various data sources that should be reviewed in identifying CCL candidates (Tables 3-2 through 3-4, NRC, 2001). The NRC examples can be grouped into five subject areas as follows:

1) chemical agent groupings (e.g., "pesticides," "gas additives," "military munitions," "pharmaceuticals") and types of microbes ("agents associated with human and animal feces");

2) transformation products (e.g., "reaction and combustion byproducts");

3) naturally-occurring substances (geochemical contaminants, radionuclides);

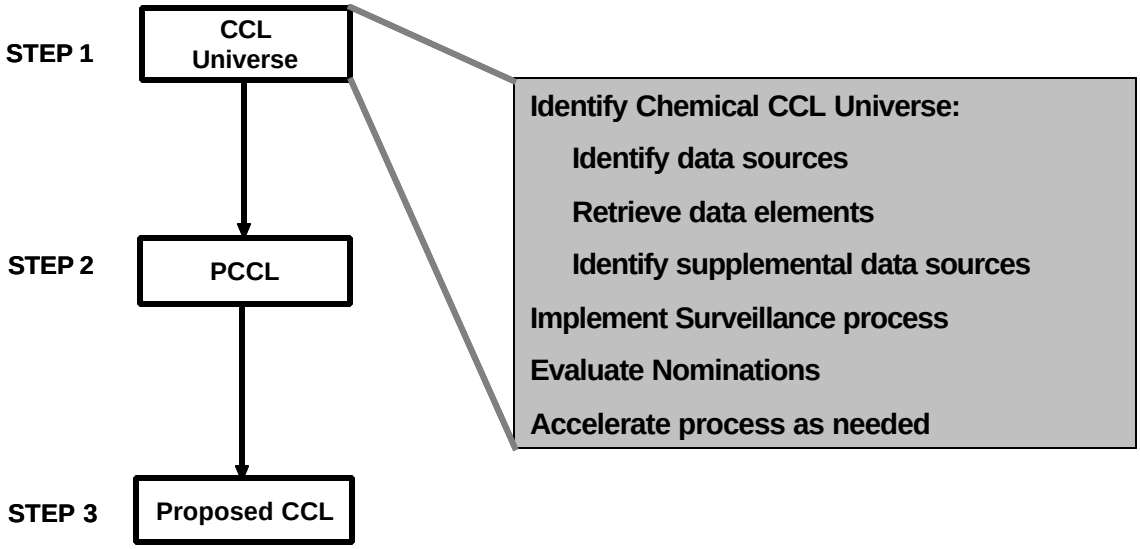
4) biologically-active agents (e.g., "enzyme inhibitors," "hormonally active compounds"); and

5) chemicals with potential to enter drinking water (e.g., "compounds widely applied to land," "constituents found in a landfill leachate," industrial discharges).

These five groupings may provide useful insights in extending the context of NRC's examples when identifying a CCL Universe that is consistent with the NDWAC's inclusionary, principles-based approach discussed below.

Figure 4.1 lists the recommended actions EPA should develop to identify the Chemical CCL Universe, and locates this step within the three-step CCL classification process.

Figure 4.1 - Detailed Overview of Step 1 (Chemical CCL Universe)



4.1.2 Overall Recommendations for Identifying the Chemical CCL Universe

→ **EPA should adopt a principles-based approach consistent with that described by the NRC.**

After review of NRC's recommendations, available data sources, and consideration of the potential scope of the Universe of known chemical agents, the Work Group recommends EPA adopt a principles-based approach, consistent with that described by the NRC. The goal of this approach is to be inclusive of agents with demonstrated or potential occurrence in drinking water and of agents with demonstrated or potential health effects.

→ **EPA should use the inclusionary principles as the foundation for identifying the Chemical CCL Universe. These principles are as follows:**

- ***The Chemical CCL Universe should include those agents that have demonstrated or potential occurrence in drinking water; or***
- ***The Chemical CCL Universe should include those agents that have demonstrated or potential adverse health effects.***

This principles-based approach provides a process for defining the Chemical CCL Universe on the basis of a set of fundamental premises regarding the nature of the agents that should be considered. If an agent meets either of the principles identified above, it is sufficient to place the agent in the Chemical CCL Universe.

The Work Group concluded that a principles-based approach would be most consistent with NRC's recommendations, as it could: a) incorporate the NRC's recommendations for including agents with demonstrated or potential occurrence in drinking water and those with demonstrated or potential health effects; b) provide a framework to include (versus exclude) agents at the earliest stage of the Chemical CCL Universe identification process; c) not limit the number and types of agents or data sources that could be considered for inclusion in the Chemical CCL Universe, now or in the future; and d) be implemented using a data source compilation process.

4.1.3 Specific Work Group Recommendations

The approach recommended by the Work Group is inclusionary with respect to agents that are not robustly characterized, and considers them at an early stage in the CCL selection or classification process. The approach does not limit the number and types of agents or data sources that can be considered for inclusion in the Chemical CCL Universe, and yet it acknowledges that for new and emerging agents, relevant data may not be readily available. Therefore, the Work Group has also included recommendations for surveillance and nomination processes as an integral part of the recommended overall process to identify agents that may need additional research and data collection to provide a means of characterizing potential harmful exposure in drinking water for these new and emerging agents. (See also Chapter 2, sections 2.3.2 and 2.3.3.)

4.1.3.1 Data Source Compilation Approach

→ **EPA should identify agents for consideration in the CCL Universe using a "data source compilation" approach, which is the process of accessing discrete data sources to retrieve various, unique sets of records with multiple selection criteria.**

The Work Group conducted a review of the number and types of known agents and available data sources, and identified alternative approaches as part of its deliberations during development of recommendations for building the Chemical CCL Universe. In discussion papers that the Work Group considered, the numbers and types of known, new and emerging agents were characterized within a hierarchy of available data sources. (See for example, Discussion Drafts for the NDWAC CCL Work Group: "Dimensioning the Chemical Universe," January 13, 2003; and "Top-Down Versus Bottom-Up Database Approaches for Defining the CCL Universe," January 22, 2003.) As a result of Work Group discussions, two approaches were identified for further consideration. One approach was the process of reducing a large array of data sources to relevant subsets of records ("reducing data sources approach"). The second approach was the process of accessing discrete data sources to retrieve various, unique sets of records with multiple selection criteria (the "data source compilation approach"). The advantages and disadvantages of two alternative approaches were discussed by the Work Group and are summarized in Tables 4.1 and 4.2, below.

Table 4.1 - Advantages and Disadvantages of the Data Source Compilation Approach

Advantages	Disadvantages
1. Relevance. Records are pre-screened for inclusion in discrete databases on the basis of key attributes. 2. More robust search capabilities. Discrete databases are typically designed for specialized searches. 3. More data per record. Economical 4. Logistical benefits. Potentially less cost per record, for publicly available databases. 5. Modular approach possible; can merge or recombine multiple databases if elements are consistent.	1. Biases. Screening criteria may not coincide with user's goals. 2. Subjective interpretations of data elements may skew results. 3. Compounds with known issues/data more likely to be included than emerging contaminants. 4. Fewer records. 5. Synonyms, homologues and mixture difficulties. Omissions and redundancies possible. 6. Certain discrete databases proprietary, accessible only by subscription that could hinder transparency. 7. Database incompatibilities. Nomenclature and search fields vary among databases. 8. Weak link issue. Recombined databases are only as current and accurate as least robust sub-database.

Table 4.2 – Advantages and Disadvantages of the Reducing Data Sources Approach

Advantages	Disadvantages
1. Comprehensive scope. Large databases represent the most complete list of known universe of chemicals. 2. Less bias introduced. Elements are included based on broader criteria. 3. Data currency and consistency. Large databases such as CHEMLIST are expanded frequently with new compounds, in a consistent format. 4. Unique substance identifiers. Can reduce inconsistencies in nomenclature.	1. Logistical impracticality. High costs are involved in searching large databases, with fees based on retrieval (e.g., $1 per substance retrieved in CAS; $3 with physical-chemical properties search). 2. Fewer data. Generally, the larger the database, the less data elements per contaminant. 3. Search constraints. Large, general databases contain fewer searchable fields than databases designed for particular purposes. 4. Missing elements. Only known compounds/microbes can be listed; oversights still possible (e.g., emerging contaminants, metabolites). 5. Lack of relevance. Large databases may contain elements not relevant to CCL attributes (e.g., nucleotide sequences, compounds in scant volumes, insoluble compounds). 6. Moving target. Large database searches may not be reproducible as data expand. 7. Large databases costly to maintain, and to update historical entries (e.g., compounds no longer in commercial use, removed from regulatory lists, etc). 8. Cross-referencing hurdles. Unique identifiers (except CASRN) may not be compatible with those in other databases. 9. Synonyms, homologues and mixture difficulties. Omissions and redundancies possible.

The Work Group agreed that although the "reducing data sources" approach would include emerging and new agents to some degree, it would present significant challenges in developing a manageable Chemical CCL Universe of agents. The "reducing data sources" approach seemed more difficult, because it would include very large numbers of agents with little relevance to the CCL (e.g., the Chemical Abstract Services lists more than 41 million protein and nucleic acid sequences). The Work Group noted that such agents would likely have no health or occurrence data or information, and the "reducing data sources" approach would therefore likely require a significant effort to review a considerable volume of irrelevant records.

The Work Group agreed that the "data source compilation approach" was logistically favorable for identifying the Universe of known agents likely to affect drinking water, even though such an approach may have some disadvantages in identifying new and emerging agents. Therefore, consistent with the inclusionary principles, the Work Group agreed to recommend the "data source compilation" approach coupled with surveillance and nomination processes (described in 4.1.3.2,

below) to consider new and emerging agents as part of an integrated overall process for defining the Chemical CCL Universe. The data source compilation with surveillance and nominations approach should be more efficient at producing a Chemical CCL Universe relevant to the CCL. Furthermore, this approach is more compatible with the Work Group recommendation that data sources for the Chemical CCL Universe be identified on the basis of multiple selection criteria (further outlined in sections 4.1.3.3 and 4.1.3.4, below). The overall approach is envisioned by the Work Group to have sufficient breadth to include known as well as new and emerging agents, consistent with the NRC's recommendations.

4.1.3.2 Supplemental Surveillance and Nomination Processes

→ **The Work Group recommends the "data source compilation approach" be supplemented with a combination of surveillance and nomination processes to provide timely identification of new and emerging agents.**

It is envisioned that surveillance and nomination would be integral components of the CCL process and not separate processes. As such, surveillance and nominations typically would provide an alternative pathway for an agent to enter into the CCL evaluative process. This approach addresses the inclusionary principles by identifying agents through the surveillance process that may be potential drinking water contaminants, but have data gaps. These agents may be identified and placed in the CCL Universe. Chapter 2 (sections 2.3.2 and 2.3.3) provides a more complete overview of the surveillance and nomination processes and recommendations.

4.1.3.3 An Integrated Process for Addressing Known, New, and Emerging Agents

→ **The Work Group recommends that EPA consider adopting a integrated process for building the Chemical CCL Universe, to include the following:**

- ***identification of a Chemical CCL Universe with known agents;***
- ***implementation of a surveillance process for new and emerging agents;***
- ***implementation of a nomination process for new and emerging agents; and,***
- ***adoption of an accelerated process for agents as needed.***

The Work Group recognized that, conceptually, the principles-based approach defined above has adequate breadth to encompass the full range of known, new and emerging agents. For implementation purposes, however, the approach must be tailored to each of the three types of candidate agents/contaminants defined in Chapter 2. For these discussions, the three types of agents are defined as follows.

- ***Known agents*** *are physical, chemical, or biological substances that have been identified in the technical literature and adequately characterized (e.g., occurrence or health effects) to enable a judgment regarding their inclusion in the Chemical CCL Universe. These are CCL candidates, which, by definition, can be identified through analysis of existing data sources. Potential data sources of known agents have been identified for consideration in the Chemical CCL Universe, and the list continues to expand. (Analyses performed for the Work*

Group show how the data sources identified to date, numbering over 200, relate to the examples cited in the NRC's Tables 3-1 and 3-5.)

- ***New agents** are physical, chemical or biological substances that are or may be newly discovered or synthesized, for which little is known about their potential occurrence or adverse health effects. Identification of new agents is challenging in several respects. The Work Group's analysis illustrated that the rate of synthesis and discovery of new agents is prodigious. For example, an average of approximately 4,000 substances are assigned CAS registry numbers daily. The majority of these substances have little data beyond name and structure, however. Most are composed of chemical sequences of biological macromolecules and proteomic sequences, and are not true candidates for the Chemical CCL Universe. On the other hand, some new agents do move into commercial production rapidly now, and these may need to be identified as agents for the CCL because of their potential to contaminate water in the future. The "data source compilation approach" alone, using data sources of known agents, would not likely capture many of these substances.*
- ***Emerging agents** are a subset of known physical, chemical, or biological substances previously evaluated as not requiring inclusion in the Chemical CCL Universe, for which information becomes available that heightens concern and triggers re-evaluation. This group contains agents that were either: a) not included in the CCL Universe; or b) agents for which new information becomes available that may heighten concerns and trigger additional review.*

Identifying the Chemical CCL Universe with Known Agents. Data sources would be identified that provide relevant information about known agents that may be potential drinking water contaminants. Data from these sources would be accessed, using the data source compilation approach, to identify agents for the Chemical CCL Universe and to retrieve novel sets of records with multiple criteria. Recommendations 4.1.3.4 and 4.1.3.5, below, provide further discussion of components of implementing this approach.

Surveillance Process for New and Emerging Agents. The Work Group recommends that EPA establish a surveillance process to provide identification of new and emerging agents. For details of these recommendations and the surveillance process refer to Chapter 2. In short, EPA's surveillance process should include: implementation of a proactive, ongoing process of communication with stakeholder organizations to obtain information; enhanced coordination within EPA and with other agencies; and strengthening the linkage of ongoing activities (ranging from literature reviews to liaisons with professional organizations) to the needs of the CCL process. In particular, the Work Group recommends that EPA institute a regularly scheduled conference (e.g., biennial) on "Emerging Issues in Drinking Water" as part of their research for the CCL process. Such a forum could provide an efficient and transparent mechanism for stakeholders and professional groups to provide their findings and concerns for emerging and new agents (see Chapter 2).

Nomination and Evaluation Process for New and Emerging Agents. The Work Group also recommends that EPA develop a nomination and evaluation process for new and emerging agents, to enable agencies and interested stakeholders from public and private sectors to nominate potential contaminants for consideration in the CCL process. As noted, all of the surveillance activities should serve to provide "nominations" of agents to add to the Chemical CCL Universe. However, as noted in Chapter 2, the Work Group recommends that other opportunities for adding potential contaminants

should be available during the CCL process. Key elements that would require further specification by EPA include a proactive communications strategy (as noted for surveillance) and information and documentation requirements for the evaluation process. (See Chapter 2 for more detailed discussion.)

Accelerated Process. As new agents are identified, or as new information becomes available, there may be justification to accelerate their passage to the Chemical CCL Universe, from the Universe to the PCCL, or from the PCCL to the CCL. EPA could, if the data warrant, consider these contaminants on an accelerated basis. The Work Group recommends EPA develop a formal accelerated ("fast track") process and ensure that the process is communicated before, or at the time the Agency requests nominations from the public. The process should be open and transparent and be consistent with the overall CCL screening and evaluation procedures (See Chapter 2.3.2.2 for discussion.)

4.1.3.4 Chemical CCL Universe Identification Process for Retrieving Information and Data

→ **The NDWAC Work Group recommends that the EPA adopt a three-stage process to identify the Chemical CCL Universe:**

- ***Identify and retrieve data (including lists of agents) from sources that have data and information about occurrence of contaminants in drinking water or source water or about health effects.***
- ***Identify and retrieve information (including lists of agents) from sources that have a link (pathway) to drinking water.***
- ***Use other information sources, such as chemical properties, to address data gaps. Use models or surrogate information to estimate potential occurrence or health effects.***

To identify known agents it is necessary to apply the data inclusion principles to actual data sources. The following guidelines are proposed for identifying data sources that would be used for building the Chemical CCL Universe.

- *If agents in a data source have a reasonable pathway (as identified by NRC) to drinking water sources, the data source should be used for the Chemical CCL Universe.*
- *If a data source contains information in a medium (e. g., sediment) that has potential to transport to water, the data source should be used for the Chemical CCL Universe.*
- *If there are multiple data sources for an agent, all the data sources will be used for the Chemical CCL Universe, as needed.*
- *It may be acceptable to model or estimate values for data elements that cannot be obtained from data sources (i.e., to fill data gaps).*
- *For data sources that contain a mix of information about known agents, it is appropriate and necessary to select only the information that meet the occurrence and health effects principles appropriate for the Chemical CCL Universe (e.g., for data sources from OSHA or NIOSH, data about ergonomic hazards would be screened out).*
- *Bibliographic data sources will primarily be used in the Surveillance and Nomination Processes for new and emerging agents (and to fill data gaps for known contaminants). (However, because of the lack of database-type sources for microbiological contaminants, it*

may be necessary to use bibliographic sources and primary literature to compile data and information for microbes for the immediate future CCL needs. This could be true of other specific issues, perhaps including studies documenting outbreaks. See Chapter 3 for microbe discussion.)

The proposed process would start with sources that contain data about measured concentrations or verified presence of known agents in drinking water or about human health effects. In addition, other sources may also provide data or information to aid in the assessment. In this regard, the Work Group recognized that EPA would need to use expert judgment to assess relevant information. (See also Transparency discussion in Chapter 2.1.) The multi-step three-stage process is explained in more detail below.

Stage (1) Identify and retrieve data (including lists of agents) from sources that have data and information about occurrence of agents in drinking water or source water or about health effects. These data sources would form the first entries to the list of agents in the Chemical CCL Universe. Also, these sources would begin to populate the relevant data elements required for the process of screening from the Chemical CCL Universe to the PCCL, for those agents. Additional data elements would be retrieved, at the appropriate stage in the CCL process, to meet the requirements of the PCCL to CCL classification process.

The Work Group recommends giving equal but separate consideration to occurrence and health effects information. (See Table 4.3 for examples of occurrence and health effects data sources.)

Table 4.3 - Examples of Occurrence and Health Effects Data Sources

Occurrence
▪ National Contaminant Occurrence Database (NCOD)
▪ EPA's (developing) Unregulated Contaminant Monitoring Regulation database (which will become part of the NCOD)
▪ USGS's National Water Quality Assessment program
Health Effects
▪ ATSDR Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) Priority Lists
▪ EPA's Health Advisory Tables
▪ EPA's Integrated Risk Information System
▪ Agency for Toxic Substances and Disease Registry (ATSDR) Minimal Risk Levels (MRLs)
▪ California's Office of Environmental Health Hazard Assessment (CA OEHHA) Toxicity Criteria Database (cancer values only)
▪ World Health Organization (WHO) Drinking Water Quality Guidelines
▪ World Health Organization's Classification of Pesticides by Hazard (CPH) Database
▪ The International Agency for Research on Cancer (IARC) lists of carcinogens
▪ TERA's International Toxicity Estimates for Risk
▪ OSHA or NIOSH data on hazards of agents based on occupational exposures (chemical and microbial agents only)

These sources may be redundant for identifying agents to add to the Chemical CCL Universe (i.e., they will identify and contain data about many of the same contaminants), but each is expected to add unique data elements. New agents encountered while compiling data from these sources will be added to the Chemical CCL Universe. For example, in accessing 23 data rich sources to identify an example Chemical CCL Universe and for evaluating a process to screen contaminants from that Universe to the PCCL, about half (8,750) of the sum of contaminants from all sources (17,891) were unique contaminants.

Stage (2) Identify and retrieve data (including lists of agents) and information from sources that have a reasonable link (pathway) to drinking water concerns. This could include reviewing sources such as the following:

- *High Production Volume (HPV) Chemical Lists*
- *Toxics Release Inventory (TRI)*
- *High Production Volume Master Summary Table*
- *FDA's Generally Recognized as Safe (GRAS) Notices*

- *National Sediment Inventory (NSI). (Contaminants included in the NSI would be compared with the list of contaminants comprising the Chemical CCL Universe. Any contaminants on the NSI that were not already in the Chemical CCL Universe would then be added to the listings in the Chemical CCL Universe. Additional data might be included if deemed relevant to occurrence.)*
- *Lists of pharmaceuticals and personal care products.*

A similar approach might be used with air-deposition data sources. Some data sources with purely ecological endpoints (e.g., AQUIRE) may or may not be appropriate and will require further expert review to assess if relevant information is available.

Stage (3) In a third stage, other information sources would be used to fill data gaps. For some contaminants it may be necessary to use surrogate information, or to model or estimate potential occurrence or health effect end points.

There are two types of data gaps; those for which information has not been generated and those for which the information is available but has not been accessed. This distinction is important as EPA endeavors to fill data gaps cost-efficiently through an iterative approach and considers available options to address this need.

As an example, for some agents added from lists without chemical characteristics data, the Chemical Abstract Service (CAS) Scientific and Technical Network (STN) databases could be used as a supplementary source to fill information gaps for needed data elements, such as solubility. QSAR modeling is an example of information that might be used to estimate water solubility.

EPA's Office for Prevention, Pesticides, and Toxic Substances (OPPTS) routinely uses quantitative structure activity relationships (QSARs) (e.g., from models such as EPIWIN) to fill data gaps for new chemicals as part of the Pre-Manufacturing Notification (PMN) program (under the Toxic Substances Control Act). (See Chapter 4.5 for further discussion of QSARs.)

The Chemical CCL Universe is not finished until all three stages are completed. In addition, the Surveillance and Nomination processes may also add agents to the Chemical CCL Universe.

4.1.3.5 An Approach to Retrieving Data and Evaluating Data Sources

→ **The Chemical CCL Universe should be identified using an adaptive approach to retrieving data and evaluating data sources and data elements for use in the screening and classification steps.**

The three-stage process recommended in Section 4.1.3.4 provides a starting point for retrieving data and information, evaluating data sources and data elements for use in the screening and classification steps. This process can be repeated as needed to obtain additional information.

This injects an important measure of manageability into the identification of the Chemical CCL Universe by combining the inclusionary principles with what can be accomplished given limited resources. This will make it possible to efficiently and cost-effectively prioritize occurrence and health effects data sources, avoid or remove redundancies in data, and provide interim evaluations to

determine how the process is working. The Work Group recognizes that to proceed with the CCL in a timely manner, the Agency should develop and implement the CCL schedule to include the optimal degree of iteration in developing the Chemical CCL Universe, considering the time and effort required to conduct all of the necessary steps to meet the overall CCL schedule.

4.1.3.6 Data Quality Principles Compatible with Inclusionary Principles

→ **For the Chemical CCL Universe, the CCL NDWAC Work Group recommends that EPA establish data quality approaches that do not require such a high threshold that they would be contrary to the inclusionary principles.**

Chapter 2.3.4 provides a detailed discussion of various issues related to data characterization and quality. A few key points pertinent to the Chemical CCL Universe are reiterated here. Section 1412(b)(3)(A) of the Safe Drinking Water Act Amendments specifies that EPA shall "use the best available, peer-reviewed science and supporting studies conducted in accordance with sound and objective scientific practices; and data collected by accepted methods or best available methods (if the reliability of the method and the nature of the decision justifies the use of the data)." The Work Group recognizes that there is a desire by all for use of the highest quality scientific data in developing and implementing environmental policies. However, it is fundamental for the CCL process that, in identifying the Chemical CCL Universe, a wide net be cast to comply with the inclusionary principles. Therefore, on principle, the Work Group supports the use of high quality data, but recommends that EPA establish data quality approaches that do not place too high a bar, which would be contrary to the inclusionary principles.

At a minimum, a description of the origin of the data must be available, including nominal information that reflects what is known about data quality, which could include:

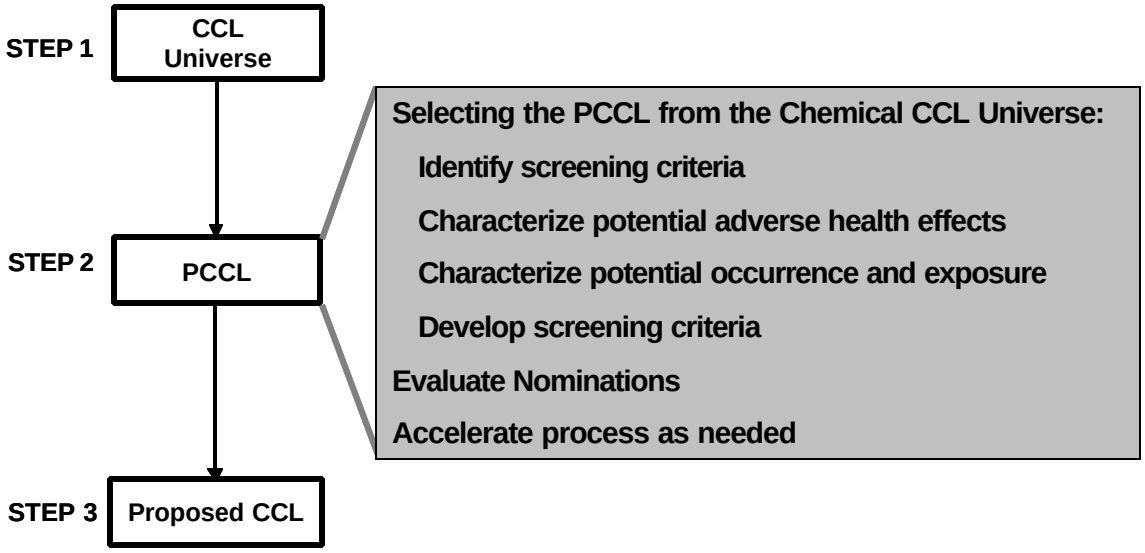
- *contact name;*
- *description of the data elements;*
- *how the data were obtained; and,*
- *meaningfulness and relevance of the data.*

4.2 Process and Criteria for Screening Agents from the Chemical CCL Universe to the PCCL

The previous section described how a Universe of chemical agents with potential or actual occurrence in drinking water or potential or actual capacity to cause health effects in humans would be constructed. This section presents the Work Group's recommendations on how to select from among the agents included in the Chemical CCL Universe those that should be listed on the Preliminary Contaminant Candidate List (PCCL). This intermediate step between the Universe and the CCL would provide a much smaller set of contaminants for more thorough assessment for the CCL.

Figure 4.2 summarizes the actions EPA should implement to screen the Chemical CCL Universe to select the PCCL, and locates these actions within the CCL process.

Figure 4.2 - Selecting the PCCL from the Chemical CCL Universe



4.2.1 Summary of the NRC Recommendations

The NRC recommended that a systematic, transparent, and scientifically sound process – combining expert judgment with well-conceived screening criteria that could be rapidly and routinely applied to a large Universe of agents – be developed to select contaminants from the Universe for the PCCL. The NRC intended that the process be more inclusive than that used for the 1998 CCL and specifically that it avoid excluding contaminants simply because of a lack of data about their occurrence in drinking water.

As shown by the shaded areas denoted by Roman numerals in Figure 4.3 (from page 82 of the NRC report), the NRC recommended that the PCCL include contaminants with "demonstrated" or "potential" occurrence in drinking water and "demonstrated" or "potential" capacity to produce adverse health effects in humans. The shaded areas of the diagram represent the NRC's priority ranking of contaminants for inclusion on the PCCL.

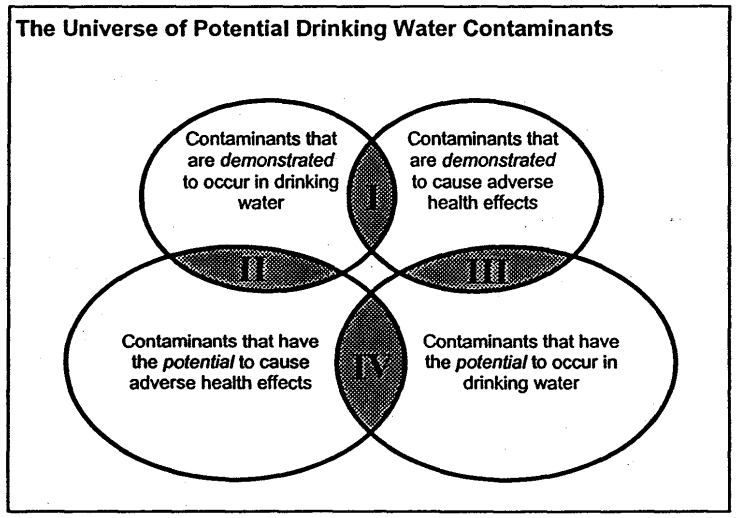
Figure 4.3 - NRC's Diagram of the CCL Universe

I Contaminants for which both health effects and occurrence are "demonstrated" (highest priority)

II Demonstrated health effects and potential for occurrence

III Demonstrated occurrence and potential for health effects

IV Potential for health effects and potential for occurrence (lowest priority)



The NRC concluded that screening criteria would need to distinguish between the effects that would be considered to be "demonstrated" compared to "potential," but also that it would be important to include contaminants for which data were limited. The NRC identified the need to develop screening criteria as a key step. While the NRC did not develop such criteria, it did identify data elements and metrics that could be used in the screening. The NRC discussed comparison of values observed in water to values of concern for health effects and recommended consideration of severity of health effects in the development of the PCCL. For potency, the data elements and metrics identified by the NRC included data from human and animal studies and models. The NRC indicated that data from human and whole animal studies should be considered to indicate demonstrated health effects and that data from other toxicological studies and experiments be considered to indicate potential for health effects. For occurrence, the NRC identified observations in tap water, distribution systems, or finished water to represent demonstrated occurrence. Data about source water and watershed production or release of chemical agents, and physical properties (including persistence and mobility in aquatic systems) would represent a potential for occurrence. The NRC also indicated that, to screen initially for inclusion on the PCCL, aqueous solubility could be used as the sole metric.

4.2.2 Principles for Selecting Agents for a PCCL from the Chemical CCL Universe

The Work Group reviewed the NRC's proposed approach in light of the principles adopted for the process as whole and findings by consultants to the Work Group about the likely availability of data about demonstrated occurrence of chemical agents in drinking water.

An analysis presented to the Work Group in July 2003 found that, as predicted by the NRC, data about demonstrated health effects and demonstrated occurrence were available for relatively few

agents.[6] To achieve the principles of inclusiveness and to develop a more systematic process for assessment, the Work Group concluded that it would be important to develop an approach that would be capable of assessing a large number of agents in the Chemical CCL Universe and that would treat agents with different amounts of data available as similarly as possible. The Work Group developed principles to support this.

The Work Group proposes EPA develop a screening process that relies on widely available data elements that reflect certain aspects of health effects and occurrence. The screening process is to be designed so that values for data elements reflecting both health effects and occurrence would reach a level of concern for an agent to be screened through to the PCCL. This is a key distinction from the development of the Chemical CCL Universe, in which an agent is to be included if there are data suggesting that either health effects OR occurrence may be of concern.

→ **The Work Group recommends that the screening criteria and methods be:**

- ***capable of assessing as many of the contaminants in the CCL Universe as possible, even those with limited data;***
- ***as insensitive as possible to data limitations;***
- ***as simple as possible, to require fewer resources and less time;***
- ***capable of identifying those contaminants of greatest significance for further consideration; and,***
- ***to the extent feasible in light of the significant differences in availability of data for chemicals and microbes, as similar as possible to the microbial approach.***

4.2.3 Workable Approach to Screening Using Widely Available Data Elements

The Work Group sought to develop a process for screening the large number of contaminants in the Chemical CCL Universe for the PCCL.

To develop an approach that would be as simple as possible and allow for the assessment of the largest possible number of agents from the Chemical CCL Universe, the Work Group discussed how to identify the most essential characteristics to be used as the basis of selection of data elements for the screening process. The Work Group concluded that it is not necessary to use data elements that reflect all five of the attributes defined by the NRC. The Work Group recognizes that subsequent steps in the process, including the classification from the PCCL to the CCL, are more likely to use data elements that reflect all five of the attributes.

During its investigations, and drawing upon work performed by its technical consultants, the Work Group found that data about demonstrated occurrence of contaminants in drinking water would be available for fewer than 3% of the agents likely to be included in the Universe (*Example CCL Universe Data Set*, July 2003 presentation to the Work Group). Consequently, the Work Group sought

[6] *Example CCL Universe Data Set: Progress and Recommendations*, Presentation to NDWAC CCL Work Group Washington, DC, July 15, 2003

to identify data elements that would be informative for screening but also available for as many of the contaminants in the Universe as possible.

The Work Group also sought to identify data elements for health effects that would be most informative for screening and available for as high a percentage of the agents in the Universe as possible.

So, for both occurrence and health effects, the Work Group identified data elements that they thought would reflect the most important characteristics of contaminants for screening and that would be most widely available. In doing this, the Work Group did not intend to preclude or restrict EPA from considering other types of data that may come to its attention or become available. The intent was to provide a workable approach to screening that would not be hamstrung by foreseeable limits in the availability of data.

→ **The Work Group recommends that a limited set of data elements that are widely available and that represent important characteristics of health effects and occurrence be used as the basis of the screening to select contaminants from the Chemical CCL Universe.**

When thinking about the characteristics related to health effects that would be most important to consider in this first screening step, the Work Group concluded that potency is the most important attribute to consider when selecting contaminants from the Chemical CCL Universe for the PCCL. The Work Group concluded that values for data elements for potency were likely to be available for many agents. The Work Group did not concur with the NRC conclusion that it would be necessary to also consider severity at this stage.

When thinking about the characteristics related to occurrence, the Work Group identified a set of data elements for screening that might be described as representing the potential for exposure. The data elements are intended to reflect two traits: (a) persistence in the environment or in drinking water distribution systems and (b) the potential for contaminants to be present in drinking water. (The latter has also been referred to as a "source screen.") The Work Group recognizes that the selected data elements represent surrogates for the traits of interest and are proposed for use because they are expected to be relatively widely available. Other attributes and characteristics were also discussed. The attribute of "magnitude" was considered but not selected as a focus because it requires estimates of concentrations in water that are not likely to be available.

→ **The Work Group recommends that widely available data elements representing potency be used to reflect health effects. The Work Group recommends that widely available data elements for occurrence that reflect persistence and likelihood that agents will get into drinking water be used to reflect occurrence.**

4.2.3.1 Data Elements for Potency

→ **For potency, the Work Group recommends that data elements reflecting chronic effects, cancer, and acute effects be considered. The data elements that represent the lowest doses at which adverse effects occur are recommended to be used as the basis of the screening for potency. In addition, for carcinogens, data elements that reflect published cancer hazard classification descriptors or cancer slope factors are recommended.**

In general, these would include lowest observed adverse effect levels (LOAELs) for chemicals tested for non-cancer effects. For cancer, equivalent metrics for cancer effects or cancer classifications such as those adopted by the EPA or International Agency for Research on Cancer (IARC) or the National Toxicology Program (NTP) are recommended to be used for screening at this stage. For acute effects, the lowest lethal doses or LD_{50}s in chemicals tested for mortality may be appropriate.

One of the critical issues in developing an approach to screening chemicals from the Chemical CCL Universe for the PCCL is to decide which data elements to use. Many data elements related to potency were identified during the discussions of the Work Group (see July 2003 Attribute Element Crosswalk).

The Work Group sought to select a set of data elements that would be as widely available as possible and that could be estimated using models such as QSARs if values based on experiments were not available. This was in keeping with the overall principle of adopting approaches that allowed for the assessment of as many contaminants as possible. The Work Group selected data elements, in general, that reflect observed values as directly as possible and that do not reflect changes or adjustments for uncertainty and other such factors.

The Work Group has sought to include data elements that will be representative of the major types of health effects of concern: acute effects, non-cancer chronic effects, and cancer. The Work Group has identified possible data elements for each of these three categories. When more than one value is available for these data elements, the Work Group recommends that the lowest dose value be selected.

The LOAEL is a widely reported result for chemical contaminants that are evaluated for non-cancer effects. The LOAEL is the lowest dose at which adverse effects are shown. It may also be appropriate to use an LD_{50} (a measure of acute effects) when LOAELs are not available. Both LOAELs and LD_{50}s can be estimated using QSAR methods, and this is another reason the Work Group recommends the use of these data elements.

For carcinogens, the type of toxicity value typically available is not analogous to a LOAEL. (This is because the toxicity of carcinogens typically is represented in terms of a unit risk or cancer slope value that reflects how quickly risk increases with dose.) EPA will need to consider carefully how to address this. One option is to generate (based on a review of the literature) values comparable to LOAELs for carcinogens. These would be the lowest doses that cause adverse cancer-related effects. Another option is to use the cancer slope factors that are typically used to represent the potency of carcinogens. A third option, which the Work Group specifically recommends, is to use the cancer classifications generated by EPA or other organizations such as IARC or the NTP; those chemicals that are listed using descriptors such as “known,” “probable,” “likely,” or “possible” carcinogen would be considered to have a health effects value of concern.

For acute effects, the lowest dose that causes mortality (LDlo) or the dose that causes mortality in 50% of exposed animals (LD_{50}) would be used for chemicals for which only tests for mortality are available.

The Work Group has not recommended using any data elements drawn from the types of assays that the NRC considered to represent the “potential” for health effects. However, estimates obtained

from QSAR models would represent values that would be considered to represent the potential for health effects.

→ **The Work Group recommends that, for potency, only one data element would be selected for screening each contaminant for health effects. It would be the single data element with the value that is most likely to lead to inclusion of the contaminant on the PCCL. (This might also be called the most health-conservative value.)**

The data elements described measure different things. Consequently, each data element would be assessed separately. The one data element most likely to result in the contaminant being placed on the PCCL (the one that is of the great health concern) would be the one to be used in the screening process.

4.2.3.2 Data Elements for Occurrence

→ **As with the screening for health effects, The Work Group recommends that the EPA develop an approach for assessing potential exposure that uses a limited set of widely available data elements. The data elements that are thought to be most widely available are: 1) those related to the tendency of an agent to persist in the environment or in the water distribution system; and, 2) those that reflect an agent's potential for occurrence in drinking water based on information characterizing its source(s). Information about demonstrated occurrence of contaminants in drinking water should also be used, where available.**

Due to data limitations and the interest in assessing a large number of contaminants, the Work Group concluded that it is important to develop a workable way to screen contaminants with regard to occurrence using data elements that might be viewed as surrogates for potential exposure. The Work Group recommends that EPA develop a screen based primarily on data elements that reflect the potential for agents to reach drinking water and the persistence of agents in the environment, including drinking water systems.

Persistence is included as a key element because a compound that is persistent, if released, could eventually contaminate drinking water. Even though the time scale may be long, the potential to persist should be addressed in this screening process. Conversely, a compound that is not persistent is not likely to remain available in drinking water long enough to pose a concern. The Work Group considers persistence as a characteristic that is best represented at this time as either "persistent" or "not persistent" and not through any kind of continuous metric.

The potential for a chemical to reach drinking water as a result of being produced or released to the environment is important to consider along with persistence, because even a highly persistent compound will not contaminate drinking water if it is never released to the environment. The potential for inadvertent releases should also be considered as should the capacity or propensity of a compound to migrate. A combination of data elements to assess the potential to reach drinking water might be called a "source screen." The Work Group recommends that EPA investigate developing a source screen that could readily be implemented using data sources such as production volumes, amounts released, use in disinfection processes and other such information that would contribute to the potential for an agent to occur in drinking water. If this is fully addressed in the process of assembling

the Chemical CCL Universe, it may need less attention during the screening from the Universe to the PCCL.

The NRC specifically recommended use of solubility in screening agents at this stage, and the Work Group discussed this issue at some length. Solubility is the equilibrium concentration of a compound in water, often expressed in the form of milligrams per liter. Solubility is available for many compounds and can also be estimated using quantitative structure-activity relationship models (QSARs).

The Work Group concluded that it would not be appropriate to use solubility as a screening criterion at this stage because it may screen out contaminants that occur widely, but at low concentrations, or that occur as particulates in suspension. Also, most chemicals occur in solution well below their equilibrium solubility concentrations, so solubility is a poor surrogate for occurrence. However, under certain assumptions, solubility is an indicator of the upper limit of a compound that is likely to occur in solution and may be useful for priority setting at a later stage in the process.

Several other data elements for chemicals also have been considered. Log K_{ow} was rejected because the group felt that it would not add any new information important at this phase of the screening. (K_{ow} is the octanol-water partition coefficient, a measure of the tendency of a dissolved compound to move out of water into a nonpolar material, also used as an index of the tendency to bioaccumulate). Henry's Law Constant was rejected because it may not be applicable for ground water and may be captured in persistence. (Henry's Law Constant is a measure of how much of a substance stays in the water compared to how much evaporates or volatilizes into the air.) In surface waters, which are exposed to the air, compounds that volatilize will not be persistent, simply because they evaporate into the air and are not found in the water. However, in ground water, the same compounds may be persistent because they cannot evaporate. TCE, a common solvent, is a good example.

The Work Group focused on use of contaminant characteristics because they concluded that data about such characteristics is what is likely to be available. However, the Work Group believes it also fully appropriate for EPA to consider data about demonstrated occurrence of contaminants in drinking water or in ambient water bodies in addition to the screening approach discussed. Measurement of contaminants in water is more direct evidence of their occurrence than persistence or source indicators. However, the Work Group does not want lack of such data to create a barrier to full consideration of a contaminant for the PCCL and so has not emphasized such data in this screening approach.

Table 4.4 - Possible Data Elements for Selecting Universe Contaminants for the PCCL

Characteristic	Data Elements	Details
Potency	LOAEL; LDlo; LD_{50}; analogues for carcinogens; carcinogen classifications	Numeric value of potency: mg/kg or mole/kg
Exposure	Persistence	Half-life or other measure; if not available, then the contaminant is assumed to be persistent
	Measured occurrence	Actual measurements of a contaminant in drinking water
	Potential to reach drinking water	Quantities produced or released; on production or release lists

One additional consideration is that there may be contaminants that reach drinking water not by being directly dissolved into water but by being adhered or adsorbed onto particles. Such compounds tend not to be soluble. Such compounds should, however, be included on the PCCL. If there were such compounds that were not identified through the data elements discussed here, it may be appropriate for them to be added to the PCCL.

4.2.4 Screening for Both Health Effects and Occurrence

→ **The Work Group recommends that the contaminants that are screened to the PCCL be those for which values for data elements for both health effects and occurrence reach a level of concern, based on the screening process, for inclusion on the PCCL. Generally, neither alone would be sufficient under this screening process.**

However, in keeping with the recommendations of the NRC that contaminants on the PCCL be those with either demonstrated or potential health effects and occurrence the Work Group recognizes that there will also be cases for which a different approach is appropriate. This approach is not intended to preclude the addition to the PCCL of groups of contaminants of particular concern where expert judgment concludes that they should be included. This approach is intended, instead, to provide a feasible way to screen compounds for which the recommended data elements may not be available. For example, the Work Group recognizes that disinfection by-products are formed and water treatment chemicals are introduced to the drinking water system and may be a concern even if they are not highly persistent.

→ **If an "On or Off" approach is used in the screening, the Work Group recommends that contaminants that have the highest values for data elements related to *either* health effects *or* occurrence but that do not make it onto the PCCL, be subjected to further review to see whether there is cause for concern in drinking water. This supplemental assessment should be done for very high potency values that score too low on exposure, and for very high exposure values that score too low on potency. Expert judgment may conclude that some of these compounds belong on the PCCL even if they fail the criteria for the screening.**

The Work Group recognizes that there are likely to be contaminants that are highly toxic but have low potential for exposure or that have high potential for exposure but do not appear to be highly toxic. Some of these contaminants may pose a concern even if they do not pass the screening process. The Work Group recommends that EPA use a supplemental assessment to identify such agents that should be further investigated and perhaps should be included on the PCCL.

→ **The Work Group recommends that EPA allow expert judgment to be used to correct mistakes or oversights that will arise from this relatively simple process. It will likely be appropriate to add some number of contaminants to the PCCL that pose a concern but that do not fit the process outlined. The Work Group recognizes that unforeseen circumstances will arise, and recommends that EPA allow for supplemental consideration to address them.**

The Work Group concurs with the NRC's view that expert judgment will need to be used in conjunction with the screening. There are likely to be contaminants that do not fit the screening criteria that should be included on the PCCL. EPA should provide for expert review and assessment to allow for the inclusion of such additional compounds when warranted. (See Chapter 2.3.1.)

4.2.5 Tagging Sources of Values for Data Elements and Implications

→ **The Work Group recommends that "tags" be used to retain information about the sources of values used in the screening process and that this be done in such a way as to preserve this information for later steps in the process. The tags should identify values derived from models such as QSARs. The tags should also identify what combination of "demonstrated" and "potential" values for health effects and occurrence were used.**

Under the process proposed by the Work Group, both measured and estimated values may be used for the data elements ultimately selected for the PCCL.

Measurements refer to data obtained from experiments, studies, surveys, or from environmental sampling and analysis. This might include, for example, measurements of the agent in water (occurrence); the results of epidemiological studies relating the presence of an agent in water and the appearance of effects; measurements of a NOAEL or LOAEL (health effects) by the oral route of exposure.

Estimates may be generated by appropriate and credible models (including quantitative structure-activity relationship models, or QSARs, if consistent with the policy for acceptable QSARs addressed in Chapter 2.3.5). Estimates may be derived by analogy, in comparing compounds without data to similar compounds with data, using expert judgment or some other estimation process.

The NRC concluded that the type of data used in the assessment of contaminants for the PCCL should contribute to the priority of the contaminants on the PCCL and that contaminants for which there are demonstrated health effects and demonstrated occurrence would have a higher priority than those where health effects or occurrence (or both) are considered to be potential.

Some members of the NDWAC Work Group believe that measured values have higher quality than estimates. Other members of the Work Group believe that this will not always be true and that

estimates based on good models and robust inputs may be of better quality than measurements based on a small or biased selection of values. Work Group members agreed, however, that tagging data elements used in screening with regard to the source of the data, and recording critical information about the sources of data, will allow appropriate decisions to be made later in the process. It is important to include information about data elements such as whether the values are measured or estimated.

In keeping with the inclusionary principles adopted by the Work Group, the Work Group does not recommend prioritizing contaminants on the PCCL itself. However, the Work Group does see value in retaining information about the types of data used in the screening process. It is important to note that the Work Group's recommendation differs from that of the NRC with respect to prioritizing contaminants on the PCCL. Where the numbered regions on the NRC's Venn diagram of the CCL Universe correspond to priority levels for contaminants included in the PCCL, the Work Group proposes using these tags not to prioritize, but to track origins of the data used.

The challenge of the proposed PCCL process is to develop a means to conduct the screening process as efficiently as possible so that it can be applied to screen the large number of agents in the Chemical CCL Universe using a manageable approach in all or most cases. Screening will be easiest if it is possible to identify the acceptable data elements and apply clear criteria for movement to the PCCL. It is likely that significant work will be needed on existing data sets to understand and standardize these before an efficient search process can be applied.

The Work Group feels that tagging contaminants on the PCCL to indicate underlying categories of data or information is also valuable for the subsequent process of moving from the PCCL to the CCL.

4.2.6 Approaches to Classifying Agents on the Chemical CCL Universe to the PCCL

→ **The Work Group recommends that, as the Agency develops approaches to screen chemical agents from the Universe to the PCCL, it should consider a range of options both for using data element values in the screening process and for establishing appropriate screening criteria to select PCCL contaminants. The screening method developed should be practical and transparent, and should efficiently screen the Universe to the PCCL. The method should also employ a level of precision that appropriately characterizes the nature and type of information used. While the Work Group discussed several options and identified their advantages and disadvantages, it does not recommend a single approach.**

The NDWAC Work Group examined several important technical issues relating to the specific approaches for screening the Chemical CCL Universe to select those agents to be placed on the PCCL. These technical issues fall into three categories:

1) Issues related to assigning specific values for the data elements used in the screening process

2) Issues related to the basis for establishing the appropriate screening criteria / decision rules

3) Issues related to the form of the screening criteria / decision rules to be applied

4.2.6.1 Assigning Specific Values to Data Elements Used in the Screening Process

Sections 4.2.4 and 4.2.5 described the various data elements related to potency and occurrence that the Work Group recommends that EPA consider for screening the Chemical CCL Universe. These are also summarized in Table 4.4. These data elements encompass several different types of specific measurements that are expressed in different ways.

Having data elements expressed in a variety of different ways raises some implementation challenges. The Work Group considered the advantages and disadvantages of using the actual measurements or estimates provided for each data element versus using categorical values based upon the measurements or estimates for the data elements. The Work Group recognized that transforming measurements and estimates into ordered categorical data will, in many cases, result in a substantial "loss of information." Avoiding such "loss of information" is important to avoid a potential decrease in the sensitivity of the screening process to properly differentiate among agents. As discussed further below, EPA will need to develop decision rules for determining which agents on the universe should go to the PCCL and which should not. Transforming the underlying data to categorical values could result in grouping some agents together and treating them the same even though the underlying data for them indicate there are differences among them. Of particular concern would be the categorical values that fall near the decision boundary line that could result in decisions to either include or exclude agents from the PCCL that would be different from decisions that might be made on the basis of the actual data that show there are differences among those similarly categorized agents.

Using the actual measurements or estimates for the data elements in the screening process whenever possible will avoid this "loss of information" problem and the implications as noted. At the same time, the Work Group also recognizes that there will be some circumstances where the use of categories for data elements may be necessary. Ideally, the Agency will develop and implement a screening process that could accommodate both actual data and categorical data. Where it is necessary to use categorical data, it should be approached in a manner that uses a sufficient number of categories to minimize the loss of information, while also not incorporating an excessive number of categories that would inappropriately imply more precision in the underlying data than is appropriate.

4.2.6.2 Basis for Establishing the Screening Criteria / Decision Rules

There are a variety of ways to define the screening criteria or decision rules for determining which agents in the Chemical CCL Universe should be placed on the PCCL. An important set of issues that the Work Group considered are those related to how those decision rules should be established. There are two major options in this regard.

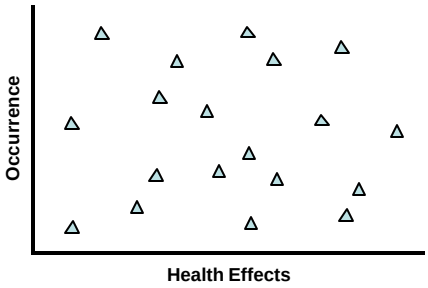
The first major option is to identify levels of concern for data elements used in the screening that have been developed outside the PCCL process, using expert judgment. Such authoritative levels of concern could come from standard references or similar sources.

The second major option is to use the observed values for the data elements in the selection of thresholds for health effects and occurrence that will cause those contaminants most likely to be of concern to move to the PCCL. The thresholds selected could reflect the number of contaminants that are sought for the PCCL as well as the appropriate weighting of the values for data elements for health effects and occurrence.

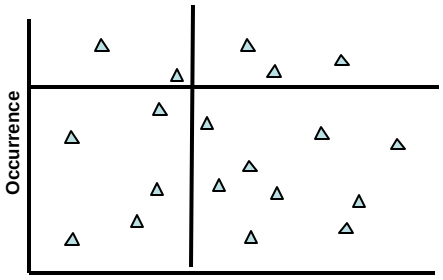
The Work Group did not address how to select threshold values for either method.

There are advantages and limitations to these options. Establishing decision rules in advance is transparent, and can be implemented easily by those with limited technical training once the rules have been established. Moreover, using generally recognized criteria for the levels of concern may add credibility. One potential limitation of this approach is the potential for arriving at either an inappropriately small or large number of contaminants for the PCCL.

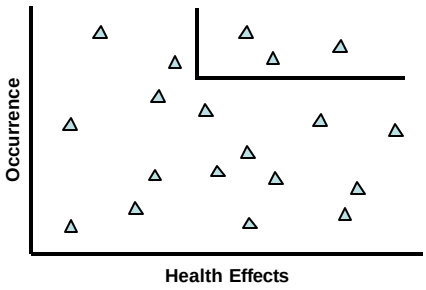
The second alternative also is transparent, in that one can draw a simple graph showing how a proposed approach would divide the data into two groups – one representing those that would be included on the PCCL and the other those that would not. An example is shown in Figure 4.4. One could use a variety of statistical methods to identify groups with similar characteristics, and then define rules for thresholds that would distinguish between them. It would also be relatively easy to conduct a sensitivity analysis of the various thresholds that could be applied to the data to see how they would differ in terms of the chemical compounds to be placed on the PCCL. It also is possible to adjust thresholds to achieve a PCCL of an appropriate size.



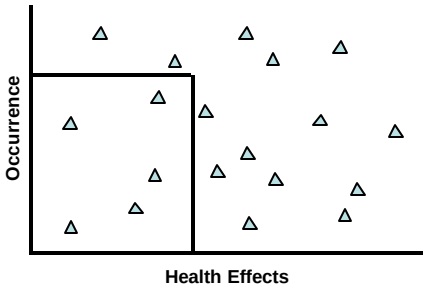
(a) The Data



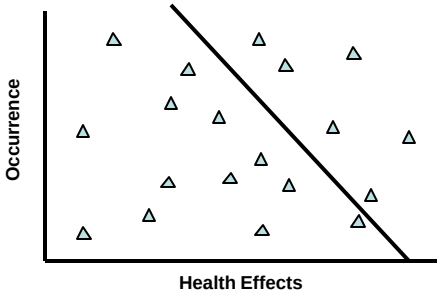
(b) Thresholds



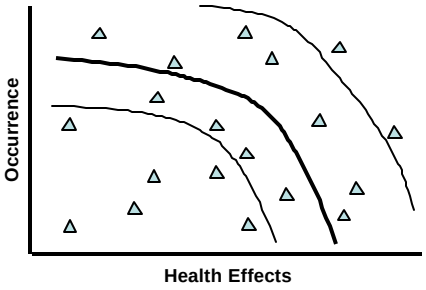
(c) Thresholds: Occurrence and Exposure



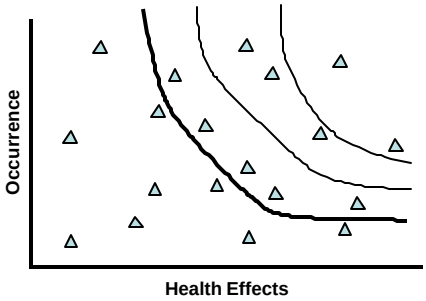
(d) Thresholds: Occurrence or Exposure



(e) Linear Rule



(f) Curvilinear Rule Emphasizing Extremes



(g) Curvilinear Rule Excluding Singular Extremes

Figure 4.4. Examples of Alternative Forms of Screening Criteria / Decision Rules

Figure 4.4 [a] shows a hypothetical data set, with several compounds plotted on two axes, health effects (on the x-axis) and occurrence (on the y-axis).

Figure 4.4 [b-d] shows the application of relatively simple threshold decision criteria. As shown in Figure 4.4 [b], the threshold for occurrence would be reflected by the horizontal line and the threshold for potency by the vertical line. Based on the four different regions defined by the specific levels of concern, one could determine which compounds get classified into the PCCL and which do not. For example, one could establish that only those that have both high health concern and high occurrence are classified onto the PCCL. Those compounds are in the upper right area of the graph defined by the thick lines (see Figure 4.4[c]). Alternatively, one could say that compounds that have *either* high occurrence (without consideration of potency) *or* high potency (without consideration of occurrence), as well as those that are moderate to high in both occurrence and potency, should be classified onto the PCCL (Figure 4.4[d]). For such a rule, those compounds that fall in the lower left quadrant of the graph, defined by the thick lines, are the only ones that would not be classified in the PCCL.

It is also possible to apply rule-based combinations of data elements that reflect more complex sets of interactions between health effects and occurrence than are reflected in the simpler threshold versions shown above. Figure 4.4[e] shows an example of a linear rule-based combination of data elements. This example implies the inclusion of compounds with high health effects and low-to-moderate occurrence, and those with high occurrence and low-to-moderate health effects, as well as those with both moderate-to-high health effects and moderate-to-high occurrence. The slope of the line determines the relative importance of health effects and potency interactions. Figures 4.4[f] and 4.4[g] show the application of still more complex curvilinear rule-based combinations of data elements. These are applied in a similar manner to the linear rule, but are more flexible with respect to what combinations of data elements result in an agent being included or excluded from the PCCL.

4.3 Use of Attributes to Classify Chemical Contaminants

4.3.1 Introduction

Chapter 5 of this report considers the several types of structured decision-making models that could be used by EPA, in conjunction with expert judgment, to determine which chemical contaminants on the PCCL are most appropriately moved forward to the CCL based on their known or potential health effects and on their known or potential occurrence in drinking water. Specific measures related to those health effects and occurrence indicators – that is, the actual values of the data for the various contaminants, or attribute scores based on the actual values of the data – provide the inputs to those decision-making models.

The NRC developed a set of five specific attributes – characteristics of a contaminant that contribute to the likelihood it could occur in drinking water at levels and frequencies that pose a public health risk – that they believed constituted a reasonable starting point for EPA to consider:

- ***Potency*** *and* ***Severity*** *as key predictive attributes for health effects*
- ***Prevalence*** *and* ***Magnitude*** *as key predictive attributes for occurrence*

- *And **Persistence/Mobility**, as characteristics that might predict possible occurrence if direct measures of Prevalence and Magnitude were not available*

As envisioned by the NRC, these five attributes were applicable to both chemical and microbial contaminants, though the NRC recognized that the types of measures and information used to quantify the attributes would differ for these two categories of contaminants. (See Chapter 2, section 2.2.2.3 for a more detailed description and discussion of these attributes.)

Chapter 5.1 provides a detailed discussion of the Work Group's consideration of the alternatives of using actual values for data elements versus scores based on those values to quantify the attributes so they can be used as inputs to the classification models. This section of the report addresses the attributes for chemicals, focusing on the types of information and data that are expected to be used to quantify them.

4.3.2 Use of Data Elements to Quantify Chemical Attributes

As mentioned in Chapter 2, the characteristic represented by an attribute often can be measured or described by more than one type of data element. For example, for a given chemical compound, the attribute of *potency* may be measured by a Reference Dose, a Cancer Risk Factor, or an LD_{50}. In deliberating on the use of data elements to quantify chemical attributes, therefore, the Work Group considered these questions:

- *What data elements should be used as measures to quantify those attributes?*
- *What should the hierarchy (preferences) among data elements for a given attribute be?*
- *How should quantitative values for a given attribute obtained from using different data elements be normalized for ensuring consistency in their use in the classification models?*

During the Work Group deliberations, EPA staff explored the attribute scoring alternative and developed draft protocols for scoring the five attributes for chemical contaminants, including rules delineating the hierarchy of data elements that should be used for scoring each attribute and the rules (or algorithms) for assigning a specific attribute score based on the specific data element or information item used for scoring. (See Appendix C.) To further help the Work Group gain insights into the practical aspects of using the five attributes and the draft scoring protocols for chemical contaminants in the CCL process, an Attribute Scoring Workshop was held in October 2003. The scoring protocols that were used, the results obtained from applying them in the workshop, and the observations of workshop participants, were presented to and considered by the Work Group in developing the recommendations presented here.

4.3.3 NDWAC Work Group Recommendations

→ **If attribute scoring is conducted, the scoring protocols for chemical contaminants should accommodate multiple data sources and a variety of data elements that may be available to score contaminants on the PCCL.**

→ **If attribute scoring is conducted, the scoring across the different types of data elements for a given attribute should be consistent and allow for a meaningful comparison among scored PCCL contaminants.**

As discussed in Chapter 5, the Work Group did not reach a conclusion regarding whether to recommend the use of actual values of data elements or scores based on those values to serve as inputs to the classification models. If scoring of attributes is carried out, the EPA should develop clear, pragmatic chemical attribute scoring protocols that can accommodate the anticipated variety of data sources and elements to be used in this process for chemical contaminants. The Work Group anticipates that there will be instances where more than one data source provides information on a contaminant. There also will be instances where more than one type of data element is available for a contaminant. These two components of attribute scoring, the data source hierarchy and data element hierarchy, should be evaluated simultaneously for each attribute being scored.

Indeed, whether EPA uses the actual values for data elements or attribute scores based on those values, a data source hierarchy should be developed to provide a descending ranked list of data sources that begins with the more trusted data sources based on pre-determined criteria such as the standard of peer review that is conducted on data prior to submittal. Similarly, the data element hierarchy should provide a descending ranked list of data elements that provides clear instruction about which data element should be used to quantify a contaminant when more than one data element is available. For example, when characterizing the potency of a contaminant the Reference Dose may be preferred to an experimentally derived Lethal Dose. The data element hierarchy for occurrence attributes may descend from measured concentrations or presence to production and release estimates. Following this logic, if the preferred data element according to the hierarchy is unavailable, then the next available data element in the hierarchy should be used to quantify the attribute. There also will be instances when a decision has to be made regarding the use of a preferred data element versus a preferred data source.

When attribute scoring is carried out, the Work Group anticipates that different contaminants' attribute scores will be based on different data elements for the same attribute. The scoring protocols should ensure that the scales for assigning scores could produce an attribute score that would carry the contaminant to the CCL regardless of which data element was used to generate the score for that contaminant. That is, the attribute score should be a function of each data element for the specific contaminant's characteristics that suggests a level of scrutiny or concern, and not a function of each data element's position in the hierarchy.

It is important that the range of scores that a contaminant could receive for a particular attribute is the same regardless of which data element is used to generate that score. For example, a contaminant that is scored using a data element from the low end of the data element hierarchy should be able to receive any score in the range – including the highest possible attribute score – if the specific information provided by that data element warrant it. This is to ensure that even those contaminants lacking data for one of the preferred data elements in the hierarchy still have an opportunity to receive a high attribute score and move forward in the process based on information that is provided by a less preferred data element in the hierarchy.

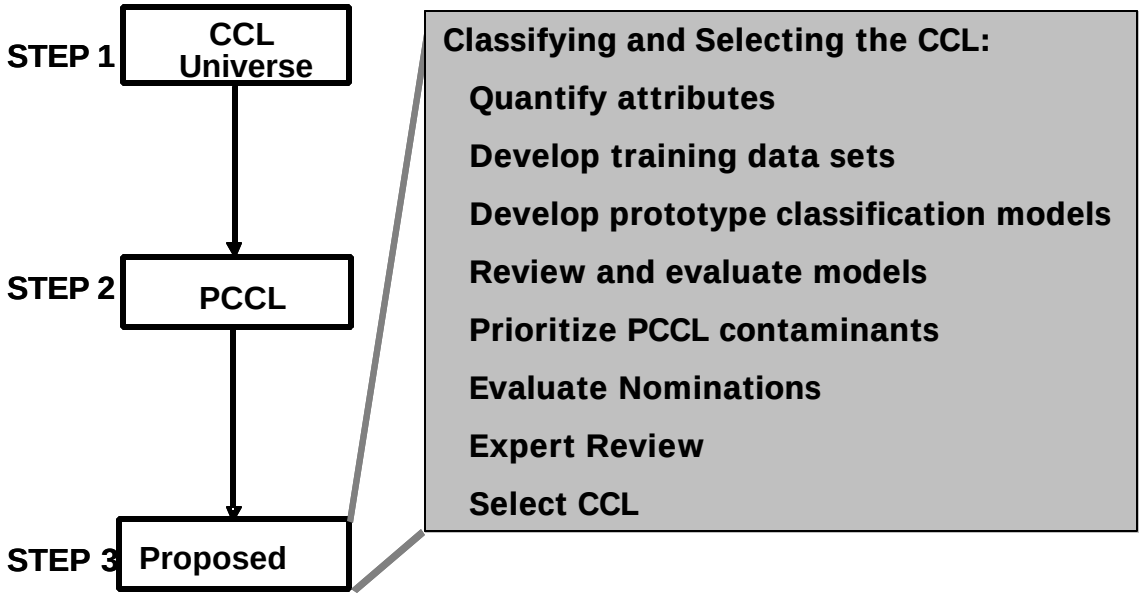
Chapter 5

Moving from the PCCL onto the CCL

The NRC Committee recommended a three-step approach[7] for identifying the list of possible contaminants for future CCLs. This chapter discusses considerations for the third step: quantifying attributes to describe contaminant risk and the use of the selected attributes in a structured decision approach to select the CCL from the PCCL. Section 5.1 discusses options and recommendations for attribute scoring. Section 5.2 presents an overview of various classification approaches and the Work Group's consideration of these approaches. Section 5.3 presents the Work Group's recommended approach to selecting the CCL, using a structured decision-making tool. Section 5.4 discusses issues to consider in the selection of a "training data set" used to inform this decision-making tool, or algorithm. Section 5.5 concludes the report with an overview of documentation required to support key components of the CCL Classification Process.

Figure 5.1 summarizes the recommended actions that EPA will need to further develop to classify the PCCL, conduct expert reviews, and select the CCL.

Figure 5.1 Classifying and Selecting the CCL



[7] The NRC report refers to a "two-step" approach because it does not count the identification of the "Universe" as a first step. The NDWAC Work Group, having elaborated on the processes for identifying both the Chemical and Microbial CCL Universes, considers this a first step; hence the NRC's "two-step" process is referred to in this report as a "three-step" process.

5.1 Quantifying Attributes for Use as Inputs to Classification Models

In its report, the NRC suggested a set of five attributes – two addressing health effects, three addressing occurrence. (General definitions provided by the NRC for each of these five attributes are displayed in the text box in Chapter 2, Exhibit 2.1.) The NRC noted in its report that it was neither explicitly nor implicitly recommending that these specific five attributes be used by EPA, nor that there necessarily be exactly five attributes. These five were offered by NRC as a starting point for developing the appropriate attributes to be used in the CCL process.

Attributes are also discussed in Chapters 3 (for microbes) and Chapter 4 (for chemicals). The remainder of this section discusses alternatives for quantifying the attributes so that they can be used as inputs to the classification models for the PCCL to CCL stage of the process.

5.1.1 The Alternatives: Using Actual Data Values versus Attribute Scoring

The latter portion of this chapter examines the several types of structured decision-making models that could be used by EPA, in conjunction with expert judgment, to determine which chemical and microbiological contaminants on the PCCL are most appropriately placed on the CCL based on their known or potential health effects and on their known or potential occurrence in drinking water. These various models require as inputs some specific measures related to those health effects and occurrence indicators. Attributes – or more specifically, either the actual values or scores generated from the actual values for data elements used to characterize those attributes – can serve as the inputs for these models.

Attributes are defined in the context of the CCL process as characteristics of a contaminant that contribute to the likelihood that it could occur in drinking water at levels and frequencies that pose a public health risk. As noted in Chapter 2, there are various types of measures or descriptors that may be used as a means for quantifying the attributes. These measures and descriptors are referred to as *data elements.*

The NRC developed, implemented and presented results for some attribute scoring metrics for the five attributes that have been discussed previously in this report (see Chapter 2). The NRC indicated that the attribute scoring metrics it explored were to be viewed as illustrative only. In particular, the consideration of attribute scoring in the NRC report involved the use only of categorical scores. That is, attribute scores explored by the NRC were limited to specific integer values with a specified range such as 0 through 10, 1 through 10, or 1 through 3 depending upon the particular attribute.

The NDWAC Work Group considered both using the actual values directly to quantify the attributes and two alternatives to generating attribute scores. The attribute scoring alternatives that NDWAC considered were:

1) Use a set of rules or an algorithm to convert the quantitative value or measurement provided by the data element to a *normalized* numerical score, with a specified range (e.g., 0 – 10) for that attribute, allowing the scores to be continuous values within that range (e.g., 8.26) and not just integer values (e.g., 8).

2) Use a set of rules or an algorithm to convert the quantitative value or measurement provided by the data element to a *categorical* score, with a specified range (e.g., 0 – 10) for that attribute, but limit the resulting scores to specific integer values only (e.g., 8). (This is essentially the approach used in the assessment of attribute scoring conducted by the NRC.)

An important advantage of using actual values to quantify the attributes is that it is, arguably, the most direct reflection of the underlying data with no "loss of information" that could occur with other methods to develop attribute scores. It should be noted that the direct use of the underlying data is what is being recommended for the occurrence and health effects measures in the screening process for going from the Universe to the PCCL.

Some technical challenges can arise, however, with this approach. For example, the data element available for characterizing the attribute may not itself be a specific numerical value. This can occur, for example, in the scoring of severity where the data elements reflect qualitative descriptors of the type of adverse effect(s) caused by the contaminant; in the scoring of magnitude based on production/import volumes expressed in gross terms of "greater than or less than" some poundage; or in the case of persistence where biodegradation rates may be reported only in broad terms such as weeks, months, or as "recalcitrant." Directly using actual values in the classification modeling processes would require that the models be developed to accommodate the various types of values or units for the data elements. For example, an attribute quantified as a concentration value of "10 parts per billion" based on a finished drinking water measurement would need to be treated differently in the classification model than an attribute quantified as "1,500,000 pounds" based on estimates of the amount of the contaminant produced each year.

The first alternative for assigning numerical scores to attributes addresses the potential challenge of having to design the classification model to address a variety of different types of data as input values for the same attribute. This is accomplished by first converting the measurements within each of the data elements that might be used for a given attribute to a unitless attribute score across some specified range of values (for example, 0 through 10). This conversion procedure would involve an algorithm that would reflect a direct relationship between the original data and the resulting attribute score. An appropriate number of significant digits should be maintained in the attribute scores so that information loss in the conversion process is avoided. This implies that, knowing the algorithm and the data element used, one could determine the underlying actual values for the data element used to produce the attribute score. In addition to the conversions within each data element, it would also be necessary that the algorithms employed "normalize" the resulting scores across different data elements used for a given attribute. For example, an attribute score for potency of "3.8" based on a particular LD_{50} value and an attribute score for potency of "3.8" based on a particular Reference Dose should convey the same degree of concern regarding potency even though they are derived from different data elements.

The second approach for assigning attribute scores is conceptually very similar to the first approach with respect to applying algorithms that normalize for the disparate data elements that might be used for a given attribute, except that the algorithm would generate categorical scores to reflect similar levels of concern for that attribute. These category scores would also be in some specified range, for example 0 through 10, but unlike the values for the option above that could have any value (i.e., any number of decimal places, such as 8.26) the scores in this third approach would be limited to

integer values only (such as 8). As a result, this approach would in most instances result in some loss of information upon which the scores are based. While there may be some advantages to this approach in that it groups "like things" together, it does preclude being able to make finer distinctions among those "like things" without going back to the original underlying data. This approach also may allow for other models to be used that accommodate categorical variables.

It should be recognized that there are attributes under consideration (for example, severity) and some data elements as mentioned previously that do not involve quantified measurements or estimates. In those cases, there is likely no option other than to assign a categorical score, probably as an integer value, using a protocol to reflect appropriately the level of concern implied by the underlying data element.

The NRC did not specifically address the alternative of using the actual data values as input to the classification models, but rather addressed only the approach of developing attribute scores (and in this regard only the integer value categorical scores as discussed above). The NDWAC Work Group did not reach a conclusion regarding which approach is preferred and therefore does not make a specific recommendation favoring one over the other. Some of the recommendations that follow refer to aspects of attribute scoring and are therefore relevant where EPA determines that attribute scoring is the preferred approach.

5.1.1.1 Summary of NRC Recommendations on Quantifying Attributes

The NRC (2001) recommended that EPA develop and use a set of attributes to evaluate the likelihood that any particular PCCL chemical or microbial contaminant could occur in drinking water at levels and frequencies that pose a public health risk. The NRC further recommended that attribute scores be the input to a prototype classification approach, used in conjunction with expert judgment, to help identify the highest priority PCCL contaminants for inclusion on a CCL. The NRC suggested that attributes and a scoring process be used because various types of information (i.e. data elements) would need to be used to score the attributes for different contaminants, and because of the widely varying data availability for emerging contaminants. The attribute scoring process would be used to put the different types of data elements for the same attribute on a common scale for evaluation and use in a classification model.

5.1.1.2 NDWAC Work Group Evaluation of Attributes

The Work Group carefully reviewed the information presented in the NRC report on the attributes proposed by the NRC for consideration by EPA. The Work Group also explored a number of important questions regarding attributes.

- *Which attributes appear to be most appropriate for use in the CCL Classification Process and how should those attributes be defined?*
- *What data elements should be used as measures in quantifying those attributes?*
- *What should the hierarchy (preferences) among data elements for a given attribute be?*
- *What are the practical constraints of obtaining and processing the data and information needed to quantify attributes?*

- *How should values for a given attribute obtained from using different data elements be normalized for ensuring consistency in scores?*
- *When scoring is used, do the scales (scoring ranges) need to be consistent across attributes?*
- *How should data quality concerns be considered in the process of quantifying attributes?*

5.1.2 NDWAC Work Group Recommendations

→ **EPA should proceed initially with using the two health effects attributes and three occurrence attributes described by the NRC as input for the PCCL-to-CCL classification modeling for contaminants.**

The Work Group determined that the concept of using attributes as part of the process for selecting those contaminants on the PCCL that are likely to pose the greatest human health risks from drinking water and moving them forward to the CCL is sound. A number of specific questions and concerns, discussed further below, were raised with respect to details about defining attributes and scoring them for use in a classification approach. These concerns generally go to certain specifics of implementing the attribute scoring process, and except as noted in Chapters 3 and 4 for magnitude, do not suggest the need for any major conceptual changes from the NRC recommendations.

→ **EPA should systematically refine and improve upon the details of the attributes as more experience is gained. These should include refinements and improvements in gathering and processing the needed data and information to quantify the attributes and with respect to using the attribute values in the selected classification approach. Further refinements and improvements should include consideration of whether fewer than all five attributes are needed, as well as of the data elements used to quantify the attributes, and, if scoring is used, the scoring protocols, and the actual attribute scoring process itself.**

The Work Group recognizes that there are numerous details concerning how many attributes are needed and how they should be characterized and scored that must be developed in conjunction with the development of the specific classification approach(es) to be used as part of the process of moving contaminants from the PCCL to the CCL. This is in keeping with the NRC observation that the five attributes discussed in its report were illustrative and represented a reasonable starting point for EPA's consideration.

Consistent with the adaptive management approach discussed in Chapter 2, the Work Group recognizes that results obtained by EPA from the initial development and implementation of the classification model, as well as from associated expert judgment processes, will result in additional information about attributes. EPA should consider this information when deciding whether fewer than the five attributes as currently described are needed to adequately prioritize agents on the PCCL for placement on the CCL. In that same vein, it is possible that these initial results will help EPA to improve on the data elements needed and make the information gathering process more focused and efficient. Therefore, it is important for EPA to specifically include as part of its CCL efforts an adaptive process to assess the attributes and make changes to them and the scoring protocols based on experience gained. (This is an example of the adaptive management approach described in Chapter 2.)

→ **Attribute scores, if used, should increase with concern. That is, contaminants that warrant higher scrutiny in the CCL process should receive higher scores for attributes.**

The Work Group recognizes that the classification models that ultimately will make use of the attribute scores if they are used can be structured to allow for a different ordering of the numerical scores (i.e., for some attributes a score of 10 on a scale of 1 to 10 could reflect greatest concern, while for another attribute a score of 1 on that scale could reflect the greatest concern). Nevertheless, to enhance the interpretation of the attribute scores themselves outside of their subsequent use as input to a classification modeling effort, the Work Group recommends that EPA use a consistent order in the scores across all attributes to reflect greater or lesser degrees of concern (i.e., contribution to potential risk) indicated by that particular attribute.

→ **The Work Group recommends that EPA explore the alternative approaches of using actual values to quantify attributes and attribute scoring described above, taking into account the requirements for implementing them – both in terms of the quantification process itself and in terms of their use in the classification models – and the possible implications each approach could have on the outcome of the classification modeling. EPA should also consider using a combination of scoring approaches depending upon the particular attribute rather than selecting one approach only for all attributes.**

→ **If attribute scoring is used, the scoring system selected by EPA for each attribute should enable discrimination among contaminants. If scoring categories are used, there should be a sufficient number of scoring categories so that information loss during characterization of contaminants is limited. At the same time, the scoring categories should not be so numerous that they convey a false sense of precision.**

The major purpose of attribute scoring is to provide relative values for the attributes that can be compared among contaminants to identify those contaminants that merit further consideration. An attribute scoring system set up on two scores may not be sufficient to discriminate contaminants accurately and the process could lose information because it does not provide enough separation of data. Conversely, a scoring system set up on 20 scores may convey a false sense of precision such that a score of 9 versus 10 may not be significant. Therefore, the available data used to score an attribute should be evaluated to determine the number of scoring categories that provides sufficient separation of the contaminants without implying a false precision.

→ **EPA should generate and include, along with the actual values or the attribute scores that are generated, descriptive "tags" that provide additional data quality information that may be used by experts reviewing the data, the attribute scores and/or the PCCL-to-CCL classification modeling results.**

As discussed in Chapter 2, the Work Group recognizes that inherent in the use of the various types of data and information that will be needed to generate attribute scores are issues and concerns about data quality. The Work Group explored some options for how EPA might specifically address data quality concerns as part of the attribute scoring component of the PCCL to CCL effort. Included among the options considered were: 1) integrating a score reflecting data quality into the attribute score itself; and, 2) generating a separate quantitative score for data quality to pair with the actual

attribute score. Integrating data quality into the scoring process increases the complexity of attribute scoring.

Another alternative the Work Group considered is for EPA to include along with each attribute value or score a data quality "tag" to indicate the source and nature of the information used to generate the score. These "tags" are intended to provide some descriptive data quality information that will be suitable for use by experts reviewing the attribute scoring process or the final outcome of the PCCL-to-CCL classification modeling effort. Those reviewers may, then, use the information provided by the tag to determine whether aspects of the underlying data or information used to score the attribute should be taken into account in arriving at the final CCL determinations.

As EPA gains more experience in implementing the procedures for moving contaminants from the PCCL to the CCL, and in particular with quantifying the attributes, it is anticipated that the approach to capturing, characterizing, and using data quality information as part of that process may be refined. It is also expected that the specific approach used by EPA for considering data quality concerns in this step of the process may evolve from the currently recommended "tags" to some other procedure for further consideration of the quality of the information.

→ **If attribute scoring is used, the scoring protocols should be transparent and straightforward.**

Attribute scoring protocols should be clear and easy to follow. A group of users of varying expertise should be able to derive equivalent attribute scores for the same set of contaminants. Evaluation of the scoring protocols should take into consideration varying types of data format and display, data element names, and data units.

5.2 Overview of Classification Approaches

The NRC discussed three general approaches for classifying PCCL contaminants: expert judgment processes, *a priori* rule-based approaches, and *a posteriori* prototype classification algorithms.

Expert judgment processes are consultations with experts on a given subject matter to elicit opinions and possibly consensus for decision-making purposes. Expert judgment processes may occur as workshops, as facilitated discourses, or by eliciting the opinions of individuals, and combining the extracted information in a rigorous framework.

An *a priori* rule is one in which a set of rules for decision-making is constructed through an expert judgment process prior to making the decision. For example, a group of experts might decide that the relevant attributes for CCL listing are potency and magnitude and that potency is twice as important as magnitude. Further, they might decide that any contaminant receiving a total score of greater than 25 should advance to the CCL. Then each contaminant would receive a total score of 2 times the potency score plus the magnitude score. Those contaminants with total scores exceeding 25 would be placed on the CCL.

In contrast, **prototype classification methods** develop *a posteriori* rules for decision-making based on decisions that have already been made. Rather than specifying the relative importance of the attributes *a priori*, an expert process establishes a "training data set." The expert process decides

which contaminants in the training set should be included or not included on the CCL. Based on these decisions a mathematical pattern-recognition algorithm establishes the relative importance of each contaminant attribute in the decisions made by experts. This "trained" algorithm would then serve as a tool to aid experts for future CCL decisions.

These approaches are not necessarily mutually exclusive. For example, a process could first use structured discourse or an *a priori* rule-based approach to create an appropriate training set for the *a posteriori* approach. In such a combined method, the first stage could have participants reflect on and discuss the strengths and weaknesses of different forms of the algorithm and weights of attributes. This discussion could focus on a set of candidate agents that are in some sense "representative" of the range of candidates. The participants could then form judgments regarding which of these candidates should be listed or not in the training set. The training set could then be used in an *a posteriori* approach to develop the final algorithm form and attribute weightings that best explain these judgments.

5.3 Recommended Approach to Selecting the CCL

5.3.1 NRC Recommendations

The NRC recommended prototype classification algorithms be considered over expert processes and rule-based approaches, citing limitations of experts (time, knowledge base, bias) in the former, and the complexity of the expert decision process relative to simple rules in the latter. However, NRC emphasized that with whatever alternative is considered, EPA must continue to rely on expert judgment throughout the classification process, because the data and knowledge for selecting drinking water contaminant candidates – particularly emerging contaminants – are admittedly incomplete.

NRC suggested EPA explore alternative model formulations, conducting sensitivity analysis to validate any findings, and being cognizant of the dangers of over-fitting and loss of generalization in model development. NRC recommended that EPA use a prototype classification approach in conjunction with expert judgment, suggesting that this approach lessens the need for transparency of the algorithm.

5.3.2 NDWAC Work Group Recommendations

The Work Group concurred with the NRC report that EPA should move toward a prototype classification approach. The Work Group concluded that implementation of a classification algorithm could improve the transparency of the decision process and would improve consistency in how the CCL is developed over time – provided that this approach can be developed in accordance with the specific recommendations below. The Work Group recognized that these models are not "objective," as they will try to mimic the decisions that are made for the training data set, but felt that quantifying these decisions made them more explicit and transparent. Thus, careful contaminant selection and attribute scoring for the training data set are imperative.

The rationale for recommending a prototype algorithm is based in large part on the NRC's deliberations. The Work Group agreed with the NRC that a formal approach could improve decision

making over past expert judgment processes, at the very least by increasing the number of contaminants that can be considered in the CCL process.

The NRC trial analysis had compared a linear discriminant model with an artificial neural network (ANN), and found the ANN to outperform the linear discriminant model. The Work Group also conducted an exploratory analysis using four methods: 1) logistic regression, 2) ANN, 3) classification and regression trees (CART), and 4) multivariate adaptive regression splines (MARS). The Work Group did not further consider the linear discriminant model, because its use is based on a set of fairly restrictive assumptions regarding the structure of the input data. All five approaches are briefly discussed in Appendix E.

The Work Group exploratory analysis used a training data set of 46 contaminants that was based on prior CCL decisions. A "best" model was chosen from each of the four aforementioned methods using appropriate techniques. (See Appendix E for a description of this analysis.) The four "best" models (i.e. one from each method class) were then compared in a cross-validation exercise using randomly chosen subsets of the original training data set. The average misclassification rates were calculated for each model. In this comparison the MARS model had the lowest average misclassification rate, the ANN had the highest misclassification rate, and the logistic and CART models were in between. An additional insight from this analysis was that the "best" model in each category usually incorporated only two or three of the five attributes, with Magnitude and Prevalence consistently included. However, the Work Group does not consider the results to be definitive because the training data set was assembled only for exploratory analysis. To make definitive recommendations, a more extensive and thorough analysis would be needed.

The specific recommendations are as follows.

→ **The Work Group recommends that EPA pursue development of a prototype classification algorithm (*a posteriori* approach) for selecting contaminants for the next CCL. The Work Group recommends moving forward to develop and test one or more prototype models as tools to be used with expert judgment for decisions on classifying contaminants for future CCLs.**

The Work Group did not have time to evaluate the alternatives and recommend a particular prototype model. Several features of the CART models, including their graphical depiction, which could aid transparency, and their ability to partially accommodate missing data, make them particularly attractive. However, more definitive testing and validation of the candidate approaches is required to make a definitive recommendation. The Work Group recognizes that it may be useful to have several models that are used in concert to corroborate results. Additionally, it may be necessary to develop separate models for chemical and microbiological contaminants, or models that differentiate chemicals and microbes within the model structure. *The development of any model should be an adaptive process, and should be reviewed by experts, with consideration given to updating the training data set, with each successive CCL cycle.*

Implementation of this recommendation depends on a well-constructed, reasonably reliable, training data set. Section 5.4 provides further recommendations for constructing a training set.

These models are tools to help prioritize contaminants for CCL, not the final decision of whether a contaminant should be listed. Experts should make the final decisions by review of the available data, including information regarding the quality and uncertainty of the data used in attribute scoring.

The rationale for this recommendation is to ensure that EPA conducts adequate evaluation of models before deciding whether or which models to use. Time constraints prohibited the Work Group from conducting detailed evaluations of the models to make more specific recommendations regarding their use.

→ **The Work Group recommends that the entire model development process be as transparent as possible. The development process should be viewed as iterative, and EPA should involve experts and allow opportunities for meaningful public comment on the evaluation.**

The details of and justification for the decisions that are made should be carefully documented and publicly available. Some issues to consider in comparing algorithms could include:

- How well algorithms predict CCL classification (misclassification rates)
- How algorithm output is affected by changes to individual training set decisions
- How algorithm output is affected by changes to the attribute scoring rules
- The importance of including all five attributes (reducing their number can reduce the labor of gathering data and scoring attributes)
- The relative performance of the different competing algorithms
- The relative performance with different training data sets
- How much of the input information is used in the evaluation
- How well algorithms work with missing or incomplete input data
- How well the results can be communicated to a non-technical audience

The purpose of these recommendations is to assure transparency as well as provide guidance and direction to EPA during the development of the model(s) and to provide a systematic framework for the experts reviewing the models and their performance for a diverse range of chemical and microbiological contaminants.

→ **EPA should use another approach for selecting CCL contaminants in the near term (i.e., for CCL3) if there are difficulties in the model development process that cannot be overcome.**

This approach may include expert processes and/or *a priori* approaches. The Work Group does not recommend alternatives be developed in parallel; however, the Work Group wants to ensure that EPA's schedule for algorithm testing, development, and review allow adequate time for implementation of an alternative approach for the next CCL, including appropriate public involvement.

→ **The Work Group recommends that experts should be involved throughout the process of narrowing a PCCL to a CCL, specifically as advisors in the design of an approach,**

development of a training set, scoring of contaminant attributes, evaluation of algorithm results, and ultimate selection of CCL contaminants.

The rationale for recommending expert involvement and review stems from a concern that EPA's use of prototype classification algorithms or other models be as tools in conjunction with broad participation by experts and others in their development and evaluation.

5.4 Training Data Set

The discussion below applies to both microbial and chemical training data sets. The Work Group recommends against combining these two kinds of contaminants into a single training set, at least for the next round of CCL. Ultimately, EPA should work toward the construction of a unified system of attribute scoring, training and classification, but for the moment, separate treatments will be required for microbial and chemical contaminants.

Training data will play an important role in methods used to classify PCCL contaminants (as being on or off the CCL). The goal is for the training set to be the template for subsequent CCL decisions. Developing the training data set will be a complex activity, requiring significant data synthesis, attribute scoring, and decision-making components.

A training data set consists of numerous contaminants, together with their health effects data/information, occurrence data/information, scored attributes (if scoring is performed), and listing status ("on" or "off" the CCL). Training contaminants and their supporting data/information must be carefully considered because the listing status of the training set contaminants will inform which PCCL contaminants should be listed. Historically listed contaminants may warrant additional evaluation as new information becomes available. In the process of selecting a classification algorithm it will be necessary to compare the performance of the candidate algorithms and validate their respective performances. It is likely that this will be accomplished, in part, by choosing random subsets of the training data set, training the algorithms with the randomly chosen subset, then evaluating the performance of the algorithms on the contaminants that were excluded from the randomly chosen subset. Thus, the training data set should contain enough contaminants to facilitate an informative validation and comparison procedure.

5.4.1 NRC Recommendations on Training Data Set

The NRC suggested that it would be a relatively straightforward exercise to construct a training set suitable for "training" a prototype classification algorithm capable of differentiating between contaminants that should be included on the CCL and those that should not. Specifically, the NRC recommended using a training data set consisting of chemicals, microorganisms, and other types of (potential) drinking water contaminants that clearly belong on the CCL (such as currently regulated contaminants), and those that clearly do not (such as food additives generally recognized as safe by the US Food and Drug Administration).

The NRC also recommended that EPA include in the training data set contaminants for which values of some of the attributes are unknown, and that EPA investigate the importance of different attributes by leaving out certain attributes in the training data set and examining the effect on

classification of the training data. Finally, the NRC also recommended that EPA withhold some contaminants from inclusion in the training set for use in validation testing to assess the predictive accuracy of any classification algorithm developed. If similar results were achieved using different training data sets, this would help ensure a robust classification process.

The NRC noted that a classification tool could be perceived as lacking transparency, in that there is no obvious indication of how it is working. Though the model might be difficult for the public to understand, the NRC indicated that the judgments embodied in the training data set would be things the public would be able to understand and that this would be one reason that this process could be more transparent than a rule-based approach. The NRC also indicated that such an approach can be made relatively transparent by clear communication of the basis for the attribute scoring scheme, the basis for the training data set, and the basis for evaluating the accuracy of the algorithms' predictions. If these aspects of the process are perceived to be sound, the derived algorithms will be easier to justify and defend.

5.4.2 NDWAC Work Group Recommendations

The Work Group considered the NRC recommendations and further investigated some of the technical issues that they raise. In particular, the Work Group disagreed with the NRC's assessment that training data set construction could be "relatively straightforward." The NRC indicated that a usable training data set could be constructed using only contaminants that clearly belong or clearly don't belong on the CCL and about which consensus would be readily reached. The investigations of the Work Group suggest that this may not be the case and that a more extensive training set with a number of contaminants not readily classified would be needed. This raises such questions and concerns as how to develop an appropriate training set, and, given the need for a relatively large number of contaminants of diverse status, how this can be made transparent to the public.

The Work Group recommends the following principles to guide training set development.

→ **The training data set should consist of contaminants (and corresponding decisions to "list" or "not list" each contaminant) that reflect technically sound, consistent judgments about what should and should not be included on the CCL.**

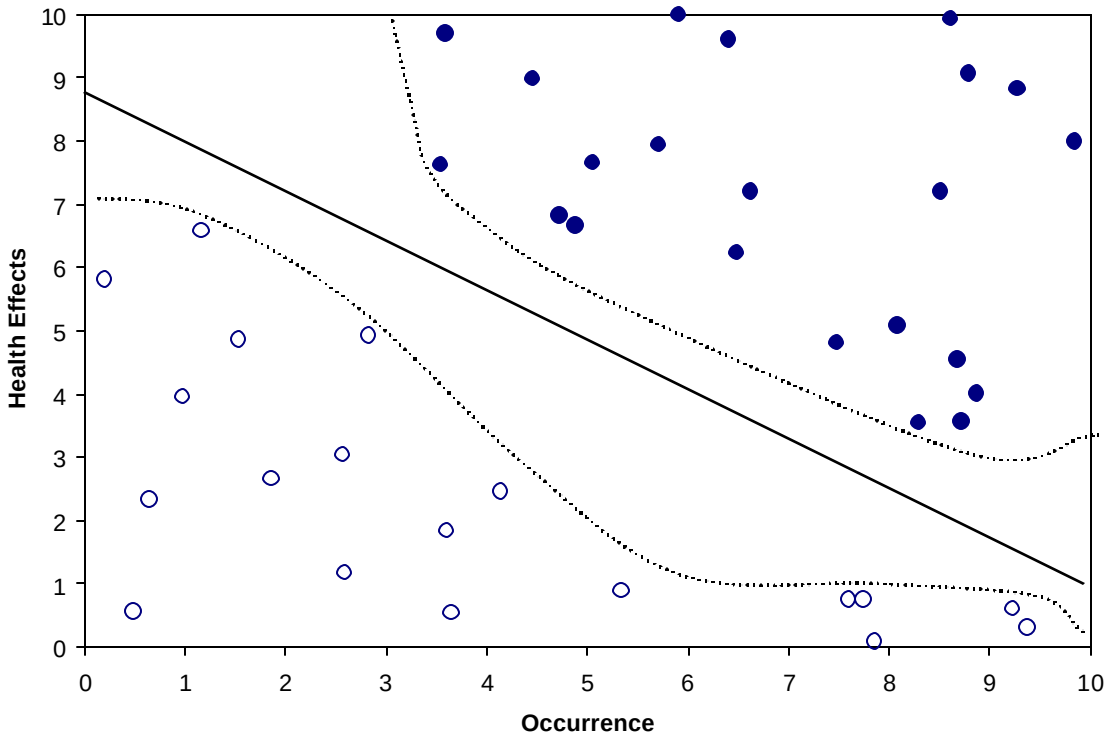
Some of the decisions will be obvious, but others will be more complex or less easily differentiated.

→ **The training set should include contaminant attribute data that are distributed throughout the attribute space, and the training set should be selected to define the discriminant surface (the function that defines "include" and "exclude" decisions) as precisely as possible.**

Training an algorithm to make difficult decisions may be important and requires that "difficult" contaminants (i.e., contaminants for which the correct decision as to whether to list or not list is not obvious) be included in the training set. By failing to include such difficult contaminants, the decision-maker may be left with too wide a range of possible algorithms, resulting in a poorly specified algorithm. The result could be classification decisions that would change if another of the possible algorithms had been used. This is explained through the two figures below.

Figure 5.2 shows a set of contaminants, each characterized by two attributes (Occurrence and Health Effects). The solid dots are contaminants clearly judged to be listed, and the open circles are contaminants clearly judged to be non-listed. Because the analysis considered only clearly listed and clearly non-listed contaminants, there are no contaminants found in a broad band between these two groups. The solid line shows one possible "discriminant" or line separating "listed" and "non-listed" contaminants. Either of the dashed lines, however, would also explain the decisions underlying this training set. The result is an inability to specify precisely where the discriminant should be placed to separate the two groups. The result will be substantial uncertainty in classifying contaminants that eventually are found to lie between the two dashed lines. Understanding this, and having little indication of which PCCL contaminants were assigned with confidence, EPA would need to review many listing decisions indicated by the algorithm, and this would reduce the benefit of utilizing the prototype classification algorithm.

Figure 5.2 – "Separated" Contaminants Poorly Define the Discriminant



By contrast, Figure 5.3 shows a case where the training set includes contaminants in this "border" region. The resulting discriminant is now better specified, lying somewhere between the two dashed lines in that figure. Note the trade-off between the two figures. The discriminant in Figure 5.2 is successful at classifying 100% of the contaminants in the training set, but is located imprecisely. The discriminant in Figure 5.3 does not classify 100% of the contaminants correctly (look at the dots near the discriminant), but places the discriminant much more precisely.

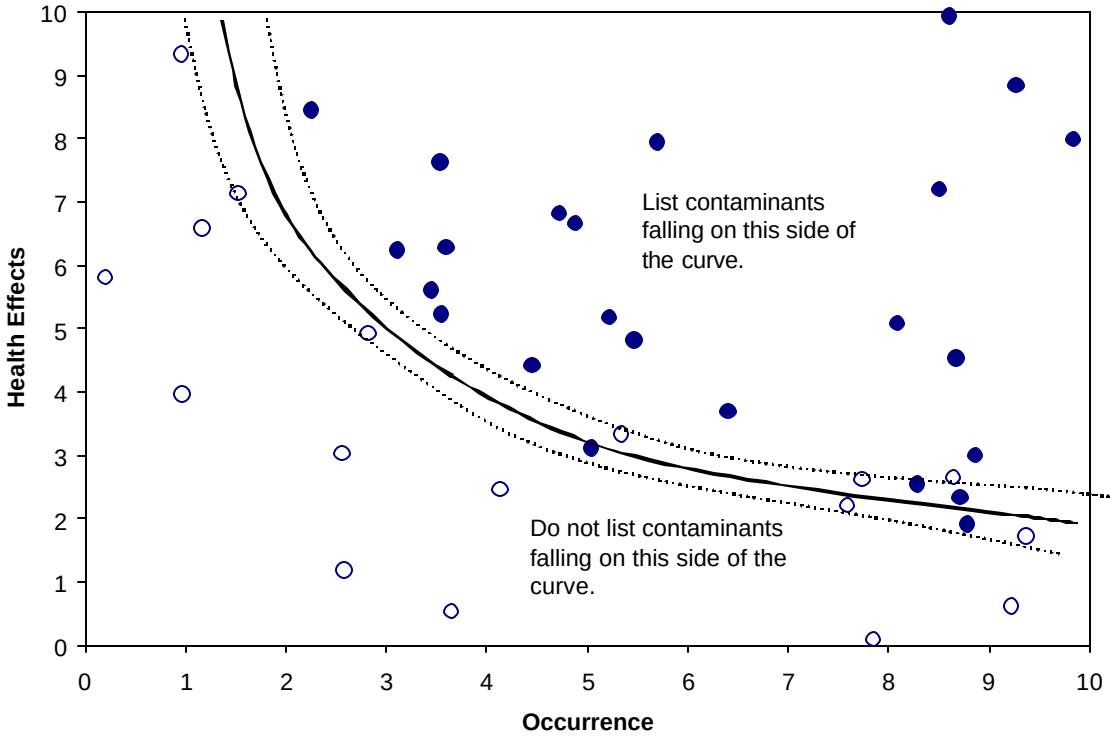


Figure 5.3 – A Discriminant Function on the Basis of Two Attributes

→ **The Work Group recommends that EPA maintain transparency and clarity when developing the training data set. To the extent feasible, EPA should document training data set development and communicate its rationale for assigning decisions to training set contaminants.**

→ **The rationale for the number and distribution of training set contaminants should be described. Quantitative rationale should be expressed for the prototype classification approach.**

These are important considerations for determining if the training set and models have been adequately developed to begin processing PCCL contaminants. The rationale should include a description of the methods used for calibration and validation, and measures used to assess goodness of fit such as misclassification rates.

Glossary of Terms

The purpose of this glossary is to define terms that may be used in the discussion of the 2001 National Research Council report, *Classifying Drinking Water Contaminants for Regulatory Consideration*, National Academy Press, and terms used by Work Group members that may be subject to interpretation. These are suggested definitions, in some cases summarized from the NRC report, presented in alphabetical order and referenced.

Adaptive Management: A continuing process of action-based planning, monitoring, researching, evaluating, and adjusting with the objective of improving implementation and achieving the goals of the selected alternative (41).

NDWAC defines an adaptive management approach as a process that involves the following steps: 1) identify an approach, 2) define evaluative criteria (factors to evaluate), 3) iteratively implement the approach, 4) transparently assess based on obscured results and evaluative criteria and 5) make changes to improve performance of the approach (this report).

Adherence: The ability of microbes to stick (adhere) to surfaces (49, 41).

Adhesin: Microbial surface antigens that frequently exist in the form of filamentous projections (pili or fimbriae) and bind to specific receptors on epithelial cell membranes; usually classified according to their ability to induce agglutination of erythrocytes from various species, their differential attachment to epithelial cells of various origins, or their susceptibility to reversal of such binding activities in the presence of mannose (49, 41).

Aflatoxin: A fungal toxin that is a powerful liver carcinogen. A group of closely related toxic metabolites are designated as mycotoxins. They are produced by *Aspergillus flavus* and *Aspergillus parasiticus*. Members of the group include aflatoxin b1, aflatoxin b2, aflatoxin g1, aflatoxin g2, aflatoxin m1, and aflatoxin m2 (41).

Agent: Any physical, chemical, or biological substance (this report).

Algorithm: A procedure for obtaining a result. It can be applied to solving a mathematical problem in a finite number of steps that frequently involves repetition of an operation; *broadly* **:** a step-by-step procedure for solving a problem or accomplishing some end especially by a computer (10).

a priori: A Latin phrase that refers to something formed or conceived before data or events are reviewed (10).

a posteriori: A Latin phrase that refers to derived by reasoning from observed facts (10).

Assessment: combination of analysis of facts and inference of possible consequences concerning a particular object (2).

Attribute: Characteristics of contaminants or potential contaminants that contribute to the likelihood that a particular contaminant or related group of contaminants could occur in drinking water at levels and frequencies that pose a public health risk (this report).

NRC identified five attributes to characterize PCCL contaminants for classification: severity, potency, prevalence, magnitude, and persistence-mobility. ***Severity*** and ***potency*** describe health effects and ***prevalence***, ***magnitude***, and ***persistence-mobility*** refer to occurrence (1).

Bayesian: Probabilistic inference that combines prior knowledge and newly acquired information via Bayes Theorem (10).

Binning: an approach for sorting agents, by classifying them into two or more "bins" or groups. NDWAC has discussed the use of a two bin (on or off) approach for selecting PCCL contaminants from the universe (this report).

Biofilm: a community of microorganisms growing on a surface in a matrix of polysaccharides and glycoproteins (41).

Bioinformatics: An interdisciplinary approach to biology that combines elements of mathematics, statistics, computer science, and information theory, with genetics, medicine epidemiology, pharmacology, molecular biology, physiology, biochemistry and microbiology (5).

Capsule: Thick gel like material attached to the wall of gram-positive or gram negative bacteria, giving colonies a smooth appearance. May contribute to pathogenicity by inhibiting phagocytosis. Mostly composed of very hydrophilic acidic polysaccharide, but considerable diversity exists (49, 41).

CART: ***Classification and Regression Tree*** analysis is a statistical technique yielding a class of models called tree-based models. It is an exploratory technique for uncovering structure in data, and produces a graphic of a branched tree indicating splits in the data. The points at which the data are split are called nodes, and the splitting of the data into groups occurs such that the homogeneity of each group is maximized. Data are split by binary recursive partitioning into groups of increasing homogeneity (18).

CAS (Chemical Abstracts Service): CAS is a team of scientists who create a digital information environment for scientific research. CAS provides pathways to published research in the scientific literature back to the beginning of the 20th century (13).

CASRN: Chemical Abstract Services Registry Number. Unique substance identification number defined by Chemical Abstract Services. Also represented as CAS Reg. No. (13).

CCL: Drinking Water Contaminant Candidate List (1). The Safe Drinking Water Act (SDWA) Amendments of 1996 require that the Environmental Protection Agency (EPA) publish a list of unregulated chemical and microbial contaminants and contaminant groups every five years that are known or anticipated to occur in drinking water systems and that may pose a public health risk in drinking water and may require regulation (1).

CCL Universe: A list of identified, new, and emerging potential drinking water contaminants used to develop the CCL. The CCL Universe includes contaminants that have demonstrated or potential occurrence in drinking water or that have demonstrated or potential health effects. This is the NRC definition of the CCL Universe, and does not apply to the microbial CCL Universe as it is being proposed in this report (23).

Classification system: A system for the sorting of data into discrete groups (1). Three broad types of systems have been considered for sorting potential contaminants including: expert judgment, rule based systems, and prototype classification algorithms (1).

Colonization (factors): Formation of compact population groups of the same type of microorganism, as the colonies that develop when a bacterial cell begins reproducing (41).

Comparative risk assessment: A process to attempt to evaluate the relative magnitude of risks and set priorities among a wide range of environmental problems (4).

Conservation: Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue (48).

Contaminant: The Safe Drinking Water Act (SDWA) defines "contaminant" as any physical, chemical, biological, or radiological substance or matter in water (41 USC Sec. 300f).

NDWAC defines contaminant as any physical, chemical, or biological substance in water. For this report, the Work Group used contaminant to indicate any agent for which data exist that suggests that the agent belongs on the PCCL (this report).

Continuous: A property of data. A variable is *continuous* if between any two possible values of the variable, there exists another possible value for the variable. This is in contrast to categorized data.

Criteria: standards on which a judgment or decision may be based; or characterizing marks or traits (10).

Cyanobacteria: A division of photosynthetic bacteria, formerly known as blue-green algae, that can produce strong toxins (51).

Cyanotoxin: Toxin produced by cyanobacteria (51).

Cytotoxins: Substances elaborated by microorganisms, plants, or animals that are specifically toxic to individual cells; they may be involved in immunity or may be contained in venoms (41).

Data: factual information (as measurements) used as a basis for reasoning, discussion, or calculation (10).

Database: a collection of data organized especially for search and retrieval (as by a computer) (10). A key feature of a database is the relation of one data element to the next by a unique identifier for each entry.

Data element: one of the necessary data or values on which calculations or conclusions are based (10). In this case: A readily identifiable descriptor that characterizes information about a contaminant (e.g., its identity, form, properties, test conditions, and study endpoints).

Data-poor: a qualitative description of the relative lack of availability of information regarding data elements for contaminant or group of contaminants.

Data-rich: a qualitative description indicating the availability of information regarding data elements for contaminant or group of contaminants.

Data source: Generally refers to a database or other source of information. To date, 242 data sources have been identified, and the list continues to expand. Because large numbers of contaminants are anticipated to be used to produce the Contaminant Candidate List, electronic databases and data sources encompass most of the list (23).

Disease: Any change from a state of good health or interruption in the normal functioning of the body, an organ, or tissue (5).

Ecology: A branch of science concerned with the interrelationship of organisms and their environments. The totality or pattern of relations between organisms and their environment (10).

Emerging agents: a subset of known physical, chemical, or biological substances previously evaluated as not requiring inclusion in the CCL Universe, for which new information becomes available which heightens concern and triggers reevaluation (this report).

Endemic Disease: continued prevalence of a disease in a specific population or area (22).

Estimates: estimates are any evidence of potency or exposure (or both) in drinking water-which may have been derived through a process of inference and/or judgment based on data that are clearly relevant to, but not necessarily directly concerned with drinking water. Estimates may be generated by appropriate and credible models (including quantitative structure-activity relationship models, or QSARs, if consistent with the policy for acceptable QSARs addressed elsewhere). Estimates also may be derived from arguments by analogy, from measures in media other than water, through expert judgment or some other estimation process. (this report)

Expedited process: As new agents are identified, or as new information becomes available, there may be justification to accelerate their passage to the CCL Universe, from the universe to the PCCL, or from the PCCL to the CCL. A re-evaluation process based on key criteria may be considered to allow contaminants of immediate concern to be expedited or "fast-tracked." (this report).

***Expert*:** one who has special skill or knowledge derived from training or experience relevant to the particular subject matter or technical analysis at hand (this report).

Expert Judgment: opinion of an expert(s) on a particular subject based upon relevant technical analysis or garnered as a technical consensus based on available information (this report).

Expert Review: critical or deliberate examination of a decision or process by an expert(s). As used in this report, expert reviews may involve various types of expert consultation and collaboration,

up to and including formal peer reviews. Expert reviewers are qualified individuals (or organizations) who are independent of those who performed the work, but who are collectively equivalent in technical expertise to those who performed the original work ("peers"). EPA uses such review for enhancing a scientific or technical work product so that the decision or position taken by EPA, based on that product, has a sound, credible basis (this report).

Exposure: Amount of a particular agent that reaches a target system. It is usually expressed in numerical terms of a concentration (2).

Genbank: Database of all published DNA, RNA and protein sequences maintained by the National Center for Biotechnology Information (NCBI) (5).

Gene: The functional unit of inherited information, often expressed as a single trait. Genes are located on the chromosomes. Each gene is encoded by a specific sequence of nucleotides in the nucleic acid of the organism (5).

Genome: The complete set of genes carried by an individual or organism. The genome serves as a master blueprint for all cellular structures and activities for the lifetime of the cell or organism (5).

Genomics: The study of genes and their function (34).

Hazard: inherent property of an agent or situation capable of having adverse effects on something. Hence, the substance, agent, source of energy or situation having that property (2).

Health Advisory: An estimate of acceptable drinking water levels for a chemical substance based on health effects information; a Health Advisory is not a legally enforceable Federal standard, but serves as technical guidance to assist Federal, State, and Local officials (7).

Health effects, demonstrated*:* NRC defines contaminants with demonstrated health effects as those that are associated with (1):

1) Human health data showing health effects; or

2) Toxicological studies on whole animals.

Health effects, potential: NRC defines contaminants with potential health effects as those that are associated with any toxicologic data or data from experimental models that predict biological activity (other than human health data, and toxicologic data on whole animals which indicates demonstrated health effects) (1).

Host: A human or other living animal, including birds and arthropods, that affords subsistence or lodgement to an infectious agent under natural conditions. Some protozoa and helminths pass successive stages in alternate hosts of different species (5).

Immune response: Alteration in the reactivity of an organism's immune system in response to an antigen, in vertebrates this may involve antibody production, induction of cell-mediated immunity, complement activation or development of immunological tolerance (47).

Immunocompromised persons: Immunocompromised persons have reduced immune responsiveness due to infection, disease, malnutrition, immnosuppressive drug therapy, or other factors (5).

Incubation period: The time from the moment of inoculation (exposure) to the development of clinical manifestations of a particular infectious disease (47).

Infection: The entry and survival or multiplication of an infectious agent in the body of humans and animals. The process by which a pathogen establishes itself in a host includes transmission, invasion, and multiplication. The target organ (e.g., intestinal tract) must come in contact with sufficient numbers of an agent, the agent must possess specific virulence factors, the virulence factors must be expressed, and an immune response may be elicited. Infection may be asymptomatic or result in disease (5).

Infective dose: The number of organisms required to produce an infection in humans or animals (5).

Insertion sequence: Small (< 2.5 kb) variable genetic elements with simple genetic organization that are capable of inserting at multiple sites in a target DNA molecule (49, 46).

Invasion: The attack or onset of a disease; the entrance of bacteria into the body or deposition in the tissues, as distinguished from infection; the infiltration and active destruction of surrounding tissue, a characteristic of malignant tumors (49, 50).

Known agent: physical, chemical or biological substances that have been identified in the technical literature and adequately characterized to enable a judgment regarding their inclusion in the CCL Universe (this report).

LD50 (50% Lethal Dose): a dose that causes mortality in 50% of exposed animals for chemical contaminants (26).

Lowest Lethal Concentration/Dose (LC/LDlo): The lowest concentration/dose to cause death in test animals (28).

Lowest-Observed-Adverse-Effect Level (LOAEL): The lowest exposure level at which there are biologically significant increases in frequency or severity of adverse effects between the exposed population and its appropriate control group (44).

Lowest-Observed Effect Level (LOEL or LEL): In a study, the lowest dose or exposure level at which a statistically or biologically significant effect is observed in the exposed population compared with an appropriate unexposed control group (44).

Magnitude: An attribute, defined by NRC as, the concentration or expected concentration of a contaminant relative to a level that causes a perceived health effect (1). see ***Attribute.***

MCL (Maximum Contaminant Level): The highest level of a contaminant that is allowed in drinking water under federal regulations, which is set as close to the MCLG as feasible using the best available treatment (20).

MCLG: Maximum Contamination Level Goal. The maximum level of a contaminant in drinking water at which no known or anticipated adverse effect on the health of persons would occur, and which allows an adequate margin of safety. Maximum contaminant level goals are non-enforceable health goals (12).

Measurements: Measurements refer to data showing directly an agent of interest occurs in water, or that produces health effects via drinking water exposure or both. This might include, for example, measurements of an agent in water; measurements of a health effect by the oral route of exposure(this report).

Microarray: A large number of nucleic acid probes (100s - >10,000) immobilized on small glass or nylon supports (5).

Mobility: An attribute defined by NRC to identify whether a contaminant is likely to be found in water, suggested by NRC to be considered with persistence as an attribute, particularly when there are no available data indicating demonstrated occurrence in water. Mobility refers to a biological or chemical contaminant=s ability to move in water, defined for chemicals as properties such as aqueous solubility, octanol water partition coefficient, Henry=s Law constant, recalcitrance, and for microbes as properties that affect transportability in water, such as sedimentation velocity, size and adsorption capability (1).

Morbidity rate: Sickness rate, the number of people who become sick compared with the number who are well in a defined group over a defined time period (47).

Mortality rate: The proportion of deaths in a population or a specific subpopulation (47).

Neural network: A prototype classification system (i.e. one that uses prototypes rather than fixed features), a neural network is a mathematical representation of a network of biological neurons. Input data are fed into the network and the output from the network is computed based on the architecture of the network and the operative mathematical functions (1).

New agent: physical, chemical, or biological substances that are or may be newly-discovered or synthesized, for which little is known about their potential occurrence or adverse health effects (this report).

No-Observed-Adverse-Effect Level (NOAEL): The highest exposure level at which there are no biologically significant increases in the frequency or severity of adverse effect between the exposed population and its appropriate control; some effects may be produced at this level, but they are not considered adverse or precursors of adverse effects (44).

No-Observed-Effect Level (NOEL): An exposure level at which there are no statistically or biologically significant increases in the frequency or severity of any effect between the exposed population and its appropriate control (44).

Nuclease: A general term for enzymes of the hydrolase class that catalyze the cleavage of phosphodiester linkages in nucleic acids to form nucleotides or oligonucleotides. The nucleases are classified in subgroups on the basis of their substrate specificity; they may be endonucleases or

exonucleases, each of which may be specific for the ribonucleic acids (ribonucleases) or deoxyribonucleic acids (deoxyribonucleases) (49, 50).

Nucleic Acid: Any of the numerous large acidic biological polymers that are found concentrated in the nuclei of all living cells. Nucleic acids contain phosphoric acid, sugar, and purine and pyrimidine bases. Two types are DNA and RNA (5).

Occurrence: The presence or prevalence of a contaminant in the environment.

Occurrence, demonstrated: demonstrated occurrence of a contaminant in drinking water is indicated by (in the NRC-recommended hierarchical order of importance) (1):

1) observations in tap water;

2) observations in distribution systems;

3) observations in finished water of water treatment plants; and

4) observations in source water.

Occurrence, potential: potential occurrence of a contaminant in drinking water is defined by NRC as indicated by (1):

1) observations in watersheds and aquifers;

2) historical contaminant release data; and

3) chemical production data.

Outbreak: The occurrence of two or more of cases of a disease in a short period of time associated with a common exposure (47).

Pathogens: Microorganisms that can cause disease in humans, animals or plants. They may be bacteria, viruses, protozoa, or parasitic worms and are found in sewage in runoff from animal farms or rural areas populated with domestic and/or wild animals, and in water used for swimming (12).

Pathogenicity islands: Fitness islands that confer pathogenicity or virulence in the organism in which they are found (49, 43).

PCCL: Preliminary CCL (1). NRC suggests a broadly defined universe of potential drinking water contaminants is identified, assessed, and culled to a preliminary CCL (PCCL) using simple screening criteria and expert judgment. NDWAC recommends the screening criteria for selecting PCCL contaminants be based on health effects and occurrence (this report).

PCR (Polymerase chain reaction): The in vitro exponential replication of a specific DNA sequence. The resulting amplification to detectable levels facilitates qualitative or quantitative genetic analysis (1).

Persistence: The ability of a biological or chemical contaminant to remain in the environment over time (6). For microbes, the ability of an organism (e.g., pathogen) to survive or a compound to exist; that is, to remain in the environment (e.g., water) or in the host for extended periods of time (5). See ***Attribute.***

Persistence-Mobility: An NRC attribute defined as the likelihood that a contaminant would be found in the aquatic environment based solely on its physical properties (1). See ***Attribute***, ***Persistence***, and ***Mobility.***

Pili: Hair-like projection from surface of some bacteria. Involved in adhesion to surfaces (may be important in virulence) and specialized sex-pili are involved in conjugation with other bacteria. Major constituent is a protein, pilin (49, 45).

Plasmid: A small, independently-replicating, piece of cytoplasmic DNA that can be transferred from one organism to another. Circular DNA molecules capable of autonomous replication found both in eukaryotes and prokaryotes. Widely used in genetic engineering as vectors of genes (cloning vectors) (45).

Potency: An NRC attribute defined as the amount of a contaminant required to cause an adverse health effect (1). Potency of a pathogen may refer to the number of organisms required to cause disease, while potency of a chemical refers to the dose required to cause disease. For example, some pathogens require relatively few (*Shigella*); others require a large number of organisms (*Salmonella typimurium*). (5). See ***Attribute.***

Potential Exposure: NDWAC defines potential exposure as any information that suggests exposure to an agent could occur via drinking water (this report).

Prevalence: NRC defines prevalence as how commonly a contaminant is found in drinking water (1). For microbes, prevalence is one of the two most common broad measures of frequency used in epidemiology (i.e. incidence and prevalence). The proportion of individuals in a population who have the disease at a specific instant; prevalence provides an estimate of the probability (risk) that an individual will be ill at a point in time (34). See ***Attribute.***

Proteomics: Proteomics a discipline within functional genomics, is the study of proteomes, protein sets expressed when the genomic blueprint of an organism is translated into functional molecules (5).

Prototype classification algorithm: One based on prior classification of examples or prototypes. Prototype classification methods develop relationships among contaminant attributes based on past decisions (*a posteriori*). Rather than specifying the relative importance of the attributes a priori, an expert process establishes a "training data set" that is used to develop the algorithm (this report).

QSAR: Quantitative structure-activity relationship (1). Quantitative structure-activity relationships comprise one class of techniques used to predict behavior of novel chemicals based upon similarities to chemicals for which specific behaviors have been empirically determined (11). QSAR models are used to estimate properties when empirical data are not available.

Radionuclides: An unstable isotope of an element that decays or disintegrates spontaneously, emitting radiation (17).

Recombination events: Chromosomal recombination during reduction division in the formation of sex cells, a major mechanism of eukaryotic genetic variability (43).

Reference Concentration / Dose (RfC/D): Term used for an estimate of air exposure concentration / daily oral exposure dose to the human population (including sensitive subgroups) that is likely to be without appreciable risk of deleterious effects during lifetime. RfC/Ds have been derived for acute, subchronic, and chronic exposure scenarios (25, 31).

Regression analysis: A statistical procedure for estimating unknown model parameter values, and their uncertainty, based on available data (5).

Risk: The probability of realization of adverse consequences or events (12).

Rule-based system: *A priori* classification models that use various features or parameters [attributes] of a contaminant and weigh and combine these features according to an algorithm that is decided upon in advance-usually as a result of some expert judgment (1).

Sensitive populations: Groups of individuals who respond biologically at lower levels of exposure to a contaminant in drinking water or who have more serious health consequences than the general population. These groups may include infants, children, pregnant women, the elderly, or individuals with a history of chronic illness (52).

Sequelae: Conditions following as a consequence of a disease (47).

Severity: An NRC attribute defined as the degree to which a potential contaminant can cause an adverse health effect. (1). see ***Attribute.***

Scaling: Changing the units of measurement, usually for the numerical stability of an algorithm. (37).

Slime layer (polysaccharide): A diffused layer of polysaccharide exterior to the bacterial cell wall (49, 41).

Slope Factor (SF): Value, in inverse concentration or dose units, derived from the slope of a inhalation dose-response curve; in practice, limited to carcinogenic effects with the curve assumed to be linear at low concentrations or doses. The product of the slope factor and the exposure is taken to reflect the probability of producing the related effect (25).

Spatial prevalence: The proportion of locales in which the contaminant can be found (1).

Temporal prevalence: The average fraction of time that a contaminant is found in a given locale (1).

Toxicogenomics: The collection, interpretation, and storage of information about gene and protein activity in order to identify toxic substances in the environment, and to help treat people at the greatest risk of diseases caused by environmental pollutants or toxicants and to set policies that will protect sensitive populations (40).

Toxicological study endpoint: A data element representing a summary statistic or observation from an empirical health effects study. Population response endpoints express chemical-induced effect concentrations in relation to a specified level of response among the test population (e.g., and LD49).

Toxicity threshold endpoints express concentrations at the onset of an observed adverse effect, irrespective of the level of response (Health Effects Commonalities, June 2003).

Toxin: For microorganisms, a noxious or poisonous substance that is either (1) an integral part of the cell or tissue, (2) an extracellular product (e.g. exotoxin), or (3) represents a combination of the two situations formed or elaborated during the metabolism, death, or growth of certain microorganisms (e.g. endotoxin). Toxin producers include *Clostridium botulinum, E. coli* serovar such as 0157:H7, *Shigella*, *Vibrio cholerae*, and some cyanobacteria (49, 5).

Transposon: Genetic element that can transpose (move) to a different position in a genome or to another genome. Transposons can be divided into two classes based on their structure. Elements of one class, known as compound or composite transposons, have copies of insertion elements (IS elements) at each end, transposition of composite transposons requires transposases coded by one of their terminal IS elements. Transponsons of the second class have terminal inverted repeats of about 30 base pairs and do not contain sequences from IS elements (49, 42).

Training Data Set: A prototype algorithm is developed using a training data set. A training set comprises information about the problem to be solved as input stimuli. The training sets are used in an iterative process to allow the prototype to ›learn‹ how to weight inputs until classification by the algorithm is adequate. (19). A training data set consists of numerous contaminants, together with their health effects data/information, occurrence data / information, scored attributes, and listing status (on or off the CCL) (this report).

Transparency: Transparency provides explicitness in the risk assessment process. It ensures that any reader understands all the steps, logic, key assumptions, limitations, and decisions in the risk assessment, and comprehends the supporting rationale that lead to the outcome. Transparencyachieves full disclosure in terms of:

a) the assessment approach employed

b) the use of assumptions and their impact on the assessment

c) the use of extrapolations and their impact on the assessment

d) the use of models vs. measurements and their impact on the assessment

e) plausible alternatives and the choices made among those alternatives

f) the impacts of one choice vs. another on the assessment

g) significant data gaps and their implications for the assessment

h) the scientific conclusions identified separately from default assumptions and policy calls

i) the major risk conclusions and the assessor‹s confidence and uncertainties in them

j) the relative strength of each risk assessment component and its impact on the overall assessment (e.g., the case for the agent posing a hazard is strong, but the overall assessment of risk is weak because the case for exposure is weak) (3).

Treatment Technique, TT: A required process intended to reduce the level of a contaminant in drinking water (21).

Validation: process of assessing whether the predictions or conclusions reached are correct (2).

Virulence: The degree of pathogenicity; the degree of intensity or severity of disease produced by a pathogen. Severity of disease does not necessarily reflect severity of infection. See also virulence factor activity relationships (VFARs) (5).

Virulence factors: The variability in the virulence of a pathogen may be characterized by one or more its biological characteristics. These characteristics, sometimes referred to as virulence factors, may include genetic elements, proteins, toxins, attachment and invasion mechanisms, metabolic pathways, and/or other architectural and biological characteristics of the pathogen. See also virulence factor activity relationships (VFARs) (5).

Virulence factor activity relationships (VFARs): a novel approach proposed for investigation to identify emerging waterborne microorganisms for the CCL. The terminology VFAR was coined to refer to a presumed or demonstrated linkage or relationship between the presence of identified genomic sequences of a microorganism and the ability of the microorganism to cause harm in humans. When the linkage or correlation of virulence factors with potency, pathogenicity, and/or the intensity/severity of disease yields a consistent statistical relationship, the relationship (i.e., the VFAR) for known pathogens may then be used as a predictive model for assessing the potency, pathogenicity, and/or virulence properties of related microbes (5).

Zoonotic*:* transmissible from animals to humans under natural conditions; pertaining to or constituting a zoonosis (50).

References

Discussion Draft for NDWAC CCL Workgroup. Dimensioning the Microbial Universe. February 25, 2003.

Discussion Draft for NDWAC CCL Workgroup. Defining the Microbial Universe. July 7, 2003.

Taylor, Latham and Woolhouse. 2001. Risk factors for human disease emergence (Appendix A). Phil. Trans. R. Soc. Lond. B 256:983-98.

Wang, D., L. Coscoy, M. Zylberberg, P. C. Avila, H. A. Boushey, D. Ganem, and J. L. DeRisi. 2002. Microarray-based detection and genotyping of viral pathogens. PNAS 99(24):15687-15692.

Glossary References

1) National Academy of Sciences, National Research Council. 2001. Classifying Drinking Water Contaminants for Regulatory Consideration. National Academy Press. Washington, DC.

2) Joint OECD/IPCS Project on the Harmonization of Hazard/Risk assessment Terminology, available at: http://www.who.int/terminology/ter/PDF_documents/tsh.pdf.

3) EPA, December 2000. Risk Characterization Handbook, EPA 100-B-00-002, available at: http://www.epa.gov/osp/spc/2riskchr.htm

4) USEPA. August 1996. Proposed Guidelines for Ecological Risk Assessment. Available on the web at: http://www.epa.gov/ORD/WebPubs/ecorisk/guide.pdf

5) EPA 2002. Glossary.(Craun, Stine, et al)

6) EPA 1999. *Class V: Underground Injection Control Study.* Appendix E: *Contaminant Persistence and Mobility factors.* Available on the web at: http://www.epa.gov/ogwdw000/uic/cl5study.html

7) Glossary of Terms Used in ITER. Available at:http://iter.ctcnet.net/publicurl/glossary.htm

8) Disaster Advise Glossary, Phoenix International Consultancy Ltd., available on the web at: http://www.disasteradvice.co.uk/glossary_search.asp

9) New England Foundation for Research, Science and Technology Glossary, available at: http://contamsites.landcareresearch.co.nz/glossary.htm

10) Merriam-Webster Dictionary Online, Available at: http://www.m-w.com/home.htm

11) Fundamentals of Aquatic Toxicology, Effects, Environmental Fate, and Risk Assessment 2nd ed, Rand G.M., 1995, Taylor & Francis Publishers since 1798, 1101 Vermont Ave., N.W., Suite 200, Washington, D.C. 20005-3511.

12) U.S. EPA, Office of Ground Water and Drinking Water. 2002. A Dictionary of Technical and Legal Terms Related to Drinking Water. Available at: http://www.epa.gov/safewater/pubs/gloss2.html

13) Chemical Abstract Service (CAS) website, available at: www.cas.org/EO/regsys.html.

14) NSF (National Sanitation Foundation) International, Drinking Water Standards Set, Document number: NSF/ANSI DWA Set NSF International , available for sale at: http://www.techstreet.com/cgi-bin/detail?product_id=100428

15) Threshold of Regulation Policy- Deciding Whether A Pesticide with a Food Use Pattern Needs a Tolerance. EPA 1999. October 18, 1999. Available at: http://www.epa.gov/fedrgstr/EPA-PEST/1999/October/Day-27/6041.pdf

16) Michael Dourson, personal communication.

17) U.S. Nuclear Regulatory Commission Glossary. Available at: http://www.nrc.gov/reading-rm/basic-ref/glossary.html#R

18) Qian, S.; and Anderson C., Exploring Factors Controlling the Variability of Pesticide Concentrations in the Willamette River Basin Using Tree-Based Models. *Environmental Science and Technology.* Vol. 33, No. 19, 1999.

19) Envistat Data Products, Neural Network Glossary. Available at: http://envisat.esa.int/dataproducts/meris/CNTR5-2-5.htm

20) National Primary Drinking Water Regulations: Consumer Confidence Reports. Federal Register Documents. Available at: http://www.epa.gov/OGWDW/ccr/ccr-frne.html

21) 2002 Edition of the Health Advisories and Drinking Water Standards. Available at: http://www.epa.gov/waterscience/drinking/standards/dwstandards.pdf

22) Stedman's Online Medical Dictionary. 27th Edition. 2003. Available at: http://www.stedmans.com/section.cfm/44

23) Building the CCL Universe and Associated Issues, Discussion Draft, May 22, 2003, Available on USEPA National Drinking Water Advisory Council, Contaminant Candidate List Classification Process Work Group Web page, under: Data Workgroup, activity materials.

24) Draft Proposed Universe to PCCL Process, Discussion Draft, May 13, 2003. Available on USEPA National Drinking Water Advisory Council, Contaminant Candidate List Classification Process Work Group Web page, under: Methods Workgroup, activity materials.

25) Glossary for Chemists of Terms Used in Toxicology, *Pure and Applied Chemistry*, Vol. 65, No.9, pp. 2003-2122, 1993. Available at: www.sis.nlm.nih.gov/Glossary/main.html

26) Glossary of Terms Used in ATSDR Chemical Profiles. Available within each chemical profile at: www.atsdr.cdc.gov/toxpro2.html

27) Epidemiology in Medicine, Little, Brown and Company, Boston/Toronto. 1987.

28) The Pesticide Management Education Program at Cornell University, Extoxnet Glossary. 2002. Available at: pmep.cce.cornell.edu/profiles/extoxnet/TIB/extoxnetglossary.html

29) Agency for Toxic Substances and Disease Registry (ATSDR), Minimal Risk Levels (MRL=s) for Hazardous Substances. Available at: www.atsdr.cdc.gov/mrls.html.

30) U.S. Food and Drug Administration=s Food Contact Substance Notification Program=s CEDI/ADI Database, available at: http://www.cfsan.fda.gov/~dms/opa-edi.html

31) Joint Meeting on Pesticide Residues - Inventory of Pesticide Evaluations; available at: http://www.inchem.org/documents/jmpr/jmpeval/jmpr2001.htm

32) EPA 2000. Benchmark Dose Technical Guidance Document, EPA/630/R-00/001, External Review Draft, October. Available at http://www.epa.gov/ncea/bnchmrk/bmds_peer.htm

33) Dictionary of Epidemiology. University of Cambridge. Last updated: 1/1/2003. Available on the Internet at: http://www.albany.net/~tjc/gloss96.html#O

34) Human Genome Project Information. Genome Glossary. DOE Human Genome Program. Available on the Internet at: http://www.ornl.gov/TechResources/Human_Genome/glossary/

35) INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY, Clinical Chemistry DivisionCommission on Toxicology. GLOSSARY FOR CHEMISTS OF TERMS USED IN TOXICOLOGY. 1993. Available at: http://sis.nlm.nih.gov/Glossary/main.html

36) California Department of Health Services. Review of MCLs in Response to PHGs. Last Update: June 9, 2003. Available at: http://www.dhs.cahwnet.gov/ps/ddwem/chemicals/PHGs/

37) Mathematical Programming Glossary. 1996-2004. Available at: http://carbon.cudenver.edu/~hgreenbe/glossary/index.php)

38). MathWorld: Wolfram Research. 1999-2004. Available at: http://mathworld.wolfram.com/)

(39) US EPA. 2000 Peer Review Handbook 2nd Edition. EPA 100-B-00-001.

40) MedicineNet. Available on the internet at: http://www.medterms.com/script/main/art.asp?articlekey=30715

41) Department of Medical Oncology, University of Newcastle upon Tyne. The On-line Medical Dictionary. 1997-2002. Available on the Internet at: http://cancerweb.ncl.ac.uk/omd

42) The Encyclopedia of Molecular Biology, 1995, Blackwell Science, Inc. 237 Main Street, Cambridge, MA 02141. Editor in chief: Sir John Kendrew

43) Hacker, J; Carniel, E; Ecological fitness, genomic islands and bacterial pathogenicity: A Darwinian view of the evolution of microbes, EMBO, 2 (5). 2001. pp this report6-371.

44) Integrated Risk Information System (IRIS). USEPA. 2004. Available on the web at: http://www.epa.gov/iris/gloss8.htm

45) Lackie, J.M. and J.A.T. Dow. The Dictionary of Cell and Molecular Biology. 1999. London: Harcourt Brace and Company. Internet edition maintained by Julian Dow and may be accessed at: http://www.mblab.gla.ac.uk/dictionary/

46) Mahillon, J; Chandler, M. Insertion Sequences. Microbiology and Molecular Biology Reviews, Sept. 1988, p. 725-774.

47) Medine Plus Health Information, Cancerweb Dictionary, Available at: http://www.nlm.nih.gov/medlineplus/dictionaries.html, last updated 2002.

48) National Center for Biotechnology Information, BLAST info More Information Glossary, revised June 7, 2000, available at: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html

49) NDWAC VFAR Subgroup Findings; Attachment 8. Glossary of GenBank Keywords and Terms used in the December Findings Document; Dec 6, 2002

50) Saunders, W.B.: Harcourt Health Sciences. Dorland's Illustrated Medical Dictionary. 2002. Available on the Internet at: http://www.mercksource.com/pp/us/cns/cns_health_library_frame.jsp?pg=/pp/us/cns/cns_hl_dorlands.jsp?pg=/pp/us/common/dorlands/dorland/dmd_a-b_00.htm&cd=3d

51) Sydney Catchment Authority (SCA) Annual Water Quality Monitoring Report 2000-2001: Technical Terms. Available on the internet at: http://www.sca.nsw.gov.au/awqr/glossary-technical-p135.html

52) US EPA. 2000. Report to Congress: EPA Studies on Sensitive Subpopulations and Drinking Water Contaminants EPA 815-R-00-015. Office of Water and Office of Research and Development. Washington DC 20460.

Appendix A

Summary of NAS-NRC Recommendations from Classifying Drinking Water Contaminants for Regulatory Consideration

Recommendations of the NRC Committee on Drinking Water Contaminants (NRC, 2001)	**Page Number**
Executive Summary	
The committee recommends that EPA develop and use a two-step process for creating future CCLs. In summary, a broadly defined universe of potential drinking water contaminants is first identified, assessed, and culled to a preliminary CCL (PCCL) using simple screening criteria and expert judgment. All PCCL contaminants are then individually assessed using a "prototype" classification tool in conjunction with expert judgement to evaluate the likelihood that they could occur in drinking water at levels and frequencies that pose a public health risk to create the corresponding CCL.	4
The committee recommends that this two-step process be repeated for each CCL development cycle to account for new data and potential contaminants that inevitably arise over time.	4
All contaminants that have not been regulated or removed from the existing CCL should be automatically retained on each subsequent CCL.	4
The committee recommends that the process for selecting contaminants for future CCLs be systematic, scientifically sound, and transparent. The development and implementation of the process should involve sufficiently broad public participation.	6
The committee recommends that the definition of vulnerable subpopulations should not only comply with the amended language of the SDWA, it should also be sufficiently broad to protect public health.	6
EPA should begin by considering a broad universe of chemical, microbial, and other types of potential drinking water contaminants and contaminant groups.	7
EPA should rely on databases and lists that are currently available and under development along with other readily available information to begin the identification of the universe of potential contaminants that may be candidates for inclusion on the PCCL.	7
As an integral part of the development process for future PCCLs and CCLs, all information used from existing or created databases or lists should be compiled in a consolidated database to provide a consistent mechanism for recording and retrieving information on the contaminants under consideration.	8
To generally assist in the identification of the universe of potential contaminants and a PCCL, the committee recommends EPA consider substances based on their commercial use, environmental location, or physical characteristics.	8
The committee recommends the use of a Venn diagram approach to conceptually distinguish a PCCL from the broader universe of potential drinking water contaminants.	8

Executive Summary (continued)	
Regarding the development of screening criteria for health effects, the committee recommends that human data and data on whole animals be used as indicators of demonstrated health effects and that other toxicological data and data from experimental models that predict biological activity be used as indicators of potential health effects.	8
A variety of metrics could be used to develop screening criteria for occurrence of contaminants in drinking water. These are identified in a hierarchical framework in the committee's first report and include (1) observations in tap water, (2) observations in distribution systems, (3) observations in finished water of water treatment plants, (4) observations in source water, (5) observations in watersheds and aquifers, (6) historical contaminant release data, and (7) chemical production data. The committee recommends that the first four of these should be used as indicators of demonstrated occurrence, and information that comes from items 5-7 should be used to determine potential occurrence.	10
Each PCCL should be published and thereby serve as a useful record of past PCCL and CCL development and serve as a starting point for the development of future PCCLs.	10
Development of the first PCCL should begin as soon as possible to support the development of the next (2003) CCL; each PCCL should be available for public and other stakeholder input (especially through the Internet) and should undergo scientific review.	10
The committee recommends EPA develop and use a set of attributes to evaluate the likelihood that any particular PCCL contaminant or group of related contaminants could occur in drinking water at levels and frequencies that pose a public health risk.	11
These contaminant attributes should be used in a prototype classification approach such as described in Chapter 5 and in conjunction with expert judgment to help identify the highest priority PCCL contaminants for inclusion on a CCL.	11
Should EPA choose to adopt a prototype classification approach for the development of future CCLs, the committee recommends that options for developing and scoring contaminant attributes should be made available for public and other stakeholder input and undergo scientific review.	12
The assessment of severity should be based, when feasible, on plausible exposures via drinking water. The committee also recommends that EPA give consideration to different severity metrics such as ranking through use of either quality adjusted or disability adjusted life years lost from exposure to a contaminant.	12

Executive Summary (continued)	
Regarding the assessment of contaminant prevalence, in some cases (particularly where contaminants have been included on a PCCL on the basis of potential occurrence rather than demonstrated occurrence), insufficient information will be available to directly assess temporal or spatial prevalence (or both). Thus, EPA should consider the possibility of including information on temporal and regional occurrence to help determine (score PCCL) contaminant prevalence. The issue of changing (or incorporating) "thresholds" for contaminant detection, rather than relying on continually decreasing detection limits, is one that needs explicit attention and discussion by EPA and stakeholders.	12
As existing and readily available databases may not be sufficient to rapidly and consistently score health effect and occurrence attributes for individual PCCL contaminants, all information from existing or created databases or lists used in the development of a CCL and PCCL should be compiled in a consolidated database (as previously recommended).	12
Contaminant databases used in support of the development of future CCLs should report summary statistics on all data collected, not only the quantifiable observations. In this regard, EPA should formalize a process for reporting means and/or medians from data with large numbers of non-detect observations. In addition, EPA may want to consider providing other measures of concentration in water supplies such as the 95^{th} percentile of contaminant concentration.	12
The committee presents two alternative models for use in the prototype classification scheme—a linear model and a neural network. Although the neural network performed better than the linear model, the committee cannot at this time make a firm recommendation as to which model EPA should use as a prototype classification scheme due to uncertainties in the training data set used by the committee. Thus, the committee recommends that EPA explore alternative model formulations and be cognizant of the dangers of overfitting and loss of generalization.	13
The committee strongly recommends that EPA greatly increase the size of the training data set that was used illustratively in this report to improve predictive capacity.	14
EPA should accurately and consistently assign attribute scores for all contaminants under consideration, i.e., contaminants in the training data set as well as contaminants to which the prototype classification algorithm will be applied for a classification determination. To do this, EPA will need to collect and organize available data and research for each PCCL contaminant and document the attribute scoring scheme used to help ensure a transparent and defensible process.	14
EPA will need to withhold contaminants from inclusion in the training data set to serve as validation test cases that can assess the predictive accuracy of any classification algorithm developed for use in the creation of future CCLs.	14

Executive Summary (continued)	
The committee recommends that EPA should use several training data sets to gauge the sensitivity of the method as part of its analysis and documentation if a classification approach is ultimately adopted and used to help create future CCLs.	15
The committee recommends the establishment of a scientific Virulence Factor Activity Relationships (VFAR) Working Group on bioinformatics, genomics, and proteomics, with a charge to study these disciplines on an ongoing basis, and to periodically inform the Agency as to how these disciplines can affect the identification and selection of drinking water contaminants for future regulatory, monitoring, and research activities.	19
The committee recommends that the findings of this report, and especially that of the Biotechnology Research Group (*the Interagency Report on the Federal Investment in Microbial Genomics*), should be made available to a VFAR Working Group at its inception.	19
The VFAR Working Group should be charged with the task of delineating specific steps and related issues and timelines needed to take VFARs beyond the conceptual framework of this report to actual development and implementation by EPA.	19
With the assistance of the Working Group, EPA should identify and fund pilot bioinformatic projects that use genomics and proteomics to gain practical experience that can be applied to the development of VFARs while simultaneously dispatching its charges outlined in the two previous recommendations.	19
EPA should employ and work with scientific personnel trained in the fields of bioinformatics, genomics, and proteomics to assist the Agency in focusing efforts on identifying and addressing emerging waterborne microorganisms.	19
EPA should participate fully in all ongoing and planned U.S. federal government efforts in bioinformatics, genomics, and proteomics as potentially related to the identification and selection of waterborne pathogens for regulatory consideration.	19

Chapter 1: Drinking Water Contaminant Candidate List: Past, Present, and Future	
An ideal CCL development process would include the following features: **$** Meet all statutory requirements of the SDWA Amendments of 1996, such as requirements for consultation with the scientific community and opportunities for public comment. **$** Begin with identification of the entire universe of potential drinking water contaminants prior to any attempts to rank or sort them. **$** Address risks from all potential routes of exposure to water supplies, including dermal contact and inhalation as well as ingestion. **$** Use the same identification and selection process for chemical, microbial, and all other types of potential drinking water contaminants. **$** Use mechanisms for identifying similarities among contaminants and contaminant classes to assess potential risks of individual contaminants– especially emerging contaminants. **$** Result in CCLs that contain only contaminants that when regulated would reduce disease, disability, and death, and excludes contaminants that have few or no adverse effects on human health (e.g., contaminants removed or detoxified through conventional drinking water treatment methods).	42
As recommended in its second report titled *Identifying Future Drinking Water Contaminants*, **the committee continues to recommend that EPA develop and use a two-step process for creating future CCLs**. In summary, a broadly defined universe of potential drinking water contaminants is first identified, assessed, and narrowed to a preliminary CCL (PCCL) using simple screening criteria and expert judgment. All PCCL contaminants are then individually assessed using a "prototype" classification tool in conjunction with expert judgment to evaluate the likelihood that they could occur in drinking water at levels and frequencies that pose a public health risk to create the corresponding (and much smaller) CCL. The "universe" of potential drinking water contaminants includes: (1) naturally occurring substances, (2) water-associated microbial agents, (3) chemical agents, (4) products of environmental transformation of chemical agents, (5) reaction byproducts, (6) metabolites in the environment, (7) radionuclides, (8) biological toxins, and (9) fibers. PCCL includes: (1) contaminants that are demonstrated to occur in drinking water and demonstrated to cause adverse health effects, (2) contaminants that are demonstrated to occur in drinking water and have the potential to cause adverse health effects, (3) contaminants that are demonstrated to cause adverse health effects and have the potential to occur in drinking water, and (4) contaminants that have the potential to occur in drinking water and the potential to cause adverse health effects	43
The committee also continues to recommend that this two-step process be repeated for each CCL development cycle to account for new data and potential contaminants that inevitably arise over time. In addition, all contaminants that have not been regulated or removed from existing CCL should be automatically retained on each subsequent CCL.	44

Chapter 2: Sociopolitical Considerations for Developing Future CCLs	
The committee believes that public participation procedures should satisfy the criteria of equity, fairness, and justice. General recommendations to facilitate public participation in environmental programs are provided in *The Model Plan for Public Participation*, developed for the EPA by the Public Participation and Accountability Subcommittee of the National Environmental Justice Advisory Council. In addition, Hampton (1999) provides the following recommendations: **$** The public should be involved in defining the process of participation. **$** Public involvements should be early in the process (e.g., at the time of agenda setting or when value judgements become important to the process). **$** Participants should have access to appropriate resources such as the information that is necessary in order to participate fully in the process, access to scientists, technical assistance, and sufficient time to prepare for the deliberations. **$** Prior agreement should be reached with the participants as to how the output of the procedure (e.g., recommendations, decisions) will be used and how it will affect agency policy decisions.	68
The committee recognizes that the development of a PCCL from the universe of potential drinking water contaminants, as well as contaminant movement from a PCCL to the corresponding CCL, is a complex task requiring numerous difficult classification judgements in a context where data are often uncertain or missing. In order to be scientifically sound as well as publicly acceptable, the process for developing future CCLs must depart considerably from the process used to develop the first (1998) CCL. **The committee recommends that the process for selecting contaminants for future CCLs be systematic, scientifically sound, and transparent. The development and implementation of the process should involve sufficiently broad public participation.**	69
The ultimate goal of the contaminant selection process is the protection of public health through the provision of safe drinking water to all consumers. To meet this goal, the selection process must place high priority on the protection of vulnerable subpopulations.	69
The committee recommends that the list of vulnerable subpopulations described in the amended SDWA should not be seen as a minimum list, but rather as several examples of possible vulnerable subpopulations. A minimum list must go much further than this. The definition of vulnerable subpopulations should not only comply with the amended language of the SDWA, it should also be sufficiently broad to protect public health and, in particular, EPA should consider including (in addition to those subgroups mentioned as examples in the amended SDWA) all women of childbearing age, fetuses, the immuno-compromised, people with acquired or inherited genetic disposition that makes them more vulnerable to drinking water contaminants, people who are exceptionally sensitive to an array of chemical contaminants, people with specific medical conditions that make them more susceptible, people with poor nutrition, and people experiencing socioeconomic hardships and racial/ethnic discrimination.	69

Chapter 2: Sociopolitical Considerations for Developing Future CCLs (continued)	
Transparency should be incorporated into the design and development of the classification and decision-making process for future CCLs in addition to being an integral component in communicating the details of the process to the public. Otherwise, the public may perceive that the process is subject to manipulation to achieve or support desired results. Therefore, sufficient information should be provided such that citizens can place themselves in a similar position as decision-makers and arrive at their own reasonable and informed judgements. This may require making available to the public the software and databases used in the process.	70
The central tenet that the public is, in principle, capable of making wise and prudent decisions should be recognized and reflected in the choice of public participation procedures used to help create future CCLs. A "decide-announce-defend" strategy that involves the public only after the deliberation process is over is not acceptable. Substantive public involvement should occur throughout the design and implementation of the process. EPA should strive to "get the right participation" (i.e., sufficiently broad participation that includes the range of interested and affected parties) as well as "get the participation right" (e.g., incorporating public values, viewpoints, and preferences into the process).	70
Chapter 3: The Universe of Potential Contaminants to the Preliminary CCL	
In general, greater consideration should be given to including substances on the PCCL that cause serious, irreversible effects as opposed to those that cause less serious effects. The committee is not suggesting that less serious health effects such as cholinesterase inhibition should be ignored, however, it recognizes that health effects such as cancer or birth defects may be given greater weight.	85
The committee believes that generally contaminant concentration alone should not be used as a relevant metric for culling from the universe to the PCCL, although the committee recognizes that some consideration of concentration may be needed as analytical procedures continue to reduce detection limits. EPA may want to consider binary data, such as found or not found in public water systems, for selecting chemicals for the PCCL from the universe. Also, the committee believes that frequency over time should not be used as the sole relevant metric for this step as this may place undue emphasis on contaminants that are repeatedly found and eliminate those that may have a significant impact but occur infrequently. The committee believes that prevalence at a large number of public water systems or prevalence at systems that serve large numbers of people is an important metric to determine inclusion into the demonstrated occurrence category.	86
Of the metrics that serve as indicators of potential occurrence, the committee recommends that EPA use production or release data, combined with physical properties, to serve as useful indicators of the potential for chemical occurrence in watersheds and aquifers.	86

Chapter 3: The Universe of Potential Contaminants to the Preliminary CCL (continued)	
For chemicals, a binary approach would serve to categorize the universe of chemicals being produced commercially (i.e., would not include byproducts or chemicals formed in the environment) into four bins for potential occurrence. The committee recommends that if such an approach were used for commercial chemicals, all chemicals except those with those with both low production volume and low water solubility should be considered for inclusion on a PCCL.	87
EPA should review contaminants already included in the potential occurrence category ("ring") to determine if they have any important environmental degradation products, production or reaction by-products or metabolites in the environment that should also be considered for inclusion on the potential occurrence list.	88
EPA should also review naturally occurring substances and fibers to determine if any of them need to be included on the potential occurrence list. EPA may also want to review data for specific watersheds and aquifers to determine if any other contaminants should be included on the potential occurrence list.	88
In keeping with its inclusive nature, the PCCL should not be expected to maintain a more-or-less fixed number of potential drinking water contaminants.	88
EPA should begin by considering a broad universe of chemical, microbial, and other types of potential drinking water contaminants and contaminant groups. The total number of contaminants in this universe is likely to be on the order of tens of thousands of substances and microorganisms, given that the Toxic Substances Control Act inventory of commercial chemicals alone includes about 72,000 substances (NRC, 1999b). This represents a dramatically larger set of substances to be initially considered in terms of types and numbers of contaminants than used for the creations of the 1998 CCL.	89
EPA should rely on databases and lists that are currently available and under development along with other readily available information to begin identifying the universe of potential contaminants that my be candidates for inclusion on the PCCL. For example, EPA should consider using the Endocrine Disruptor Priority-Setting Database (EDSPD) database to help develop future PCCLs (and perhaps CCLs). While relevant databases and lists exist for many "universe categories" of potential drinking water contaminants, others have no lists or databases (e.g., products of environmental degradation). **Thus, EPA should initiate work on a strategy for filling the gaps and updating the existing databases and lists of contaminants (e.g., through involvement of the National Drinking Water Advisory Council or panels of experts) for future CCLs.** This strategy should be developed with public, stakeholder, and scientific community input.	90

Chapter 3: The Universe of Potential Contaminants to the Preliminary CCL (continued)	
As an integral part of the development process for future PCCLs and CCLs, all information from existing or created databases or lists used should be compiled in a consolidated database that would provide a consistent mechanism for recording and retrieving information on the contaminants under consideration. Such a database could function as a "master list" that contains a detailed record of how the universe of potential contaminants was identified and how a particular PCCL and its corresponding CCL were subsequently created. It would also serve as a powerful analytical tool for the development of future PCCLs and CCLs. As a starting point, **the committee recommends that EPA review its developing EDSPD database to determine if it can be expanded and used as this consolidation database or whether it can serve as a model for the subsequent development of such a database**. Regardless, the (re)design, creation, and implementation of such a database should be made in open cooperation with the public, stakeholders, and the scientific community.	90
To generally assist in the identification of the universe of potential contaminants and a PCCL, the committee recommends EPA consider substances based on their commercial use, environmental location, or physical characteristics. EPA should be as inclusive as possible in narrowing the universe of potential drinking water contaminants down to a PCCL. The committee envisions that a PCCL would contain on the order of a few thousand individual substances and groups of related substances, including microorganisms, for evaluation and prioritization to form a CCL. However, preparation of a PCCL should not involve extensive analysis of data, nor should it directly drive EPA's research or monitoring activities.	90
The committee recommends the use of a Venn diagram approach to conceptually distinguish a PCCL from the universe of potential drinking water contaminants. However, due to the extremely large size of the universe of potential drinking water contaminants, well-conceived screening criteria remain to be developed that can be rapidly and routinely applied by EPA in conjunction with expert judgement to cull the universe to a much smaller PCCL. Thus, the PCCL should include those contaminants that have a combination of characteristics indicating that they are likely to pose a public health risk through their occurrence in drinking water. These characteristics are demonstrated or potential occurrence in drinking water and demonstrated or potential ability to cause adverse health effects.	91
Regarding the development of screening criteria for health effects, the committee recommends that human data and data on whole animals be used as indicators of demonstrated health effects and that other toxicological data and data from experimental models that predict biological activity be used as indicators of potential health effects.	91

Chapter 3: The Universe of Potential Contaminants to the Preliminary CCL (continued)	
A variety of metrics could be used to develop screening criteria for occurrence of contaminants in drinking water. These are identified in a hierarchical framework in the committee's first report (NRC, 1999a) and include (1) observations in tap water, (2) observations in distribution systems, (3) observations in finished water of water treatment plants, (4) observations in source water, (5) observations in watersheds and aquifers, (6) historical contaminant release data, and (7) chemical production data. **The committee recommends that the first four of these should be used as indicators of demonstrated occurrence, and information that comes from items 5-7 should be used to determine potential occurrence.** For commercial chemicals, their potential for occurrence in drinking water may be estimated using a combination of production volume information and water solubility. Most likely occurrence would involve high production volume chemicals with high water solubility.	91
A new PCCL should be generated for each CCL development cycle to account for new data and emerging contaminants.	91
Each PCCL should be published and thereby serve as a useful record of past PCCL and CCL development and serve as a starting point for development of future PCCLs.	91
Development of the first PCCL should begin as soon as possible to support the development of the next (2003) CCL; each PCCL should be available for public and other stakeholder input (especially through the Internet) and should undergo scientific review.	91
Chapter 4: PCCL to CCL: Attributes of Contaminants	
To overcome the limitations of current chemical fate and persistence models, the committee recommends consideration of three general characteristics of contaminants that would foster their persistence and/or mobility in water systems: **$** High potential for amplification by growth under ambient conditions (applies to microbial contaminants and not to chemicals). **$** High solubility in water (applies primarily to chemicals); though transportability of microorganisms may be assessed through sedimentation velocities and size and adsorption capabilities. **$** Stability in water, i.e., resistence to degradation via mechanisms such as hydrolysis, photolysis, or biodegradation in the case of chemicals; death or the ability to produce non-culturable states or resistant states (e.g., spores and cysts) in the case of microorganisms.	100
Expanding upon the Chapter 3 recommendation for EPA to review the EDSPD database to determine if it can be used to help develop a PCCL and perhaps help select PCCL contaminants for inclusion on a CCL, **the committee also recommends that EPA consider the possibility of including information on temporal and regional occurrence**.	103

Chapter 4: PCCL to CCL: Attributes of Contaminants (continued)	
The committee recommends EPA develop and use a set of attributes to evaluate the likelihood that any particular PCCL contaminant or group of related contaminants could occur in drinking water at levels and frequencies that pose a public health risk. More specifically, these contaminant attributes should be used in a prototype classification algorithm approach such as described in Chapter 5 and in conjunction with expert judgement to help identify the highest priority PCCL contaminants for inclusion on a CCL.	105
Should EPA choose to adopt a prototype classification approach for the development of future CCLs, the committee recommends that options for developing and scoring contaminant attributes should be made available for public and other stakeholder input and undergo scientific review.	105
The assessment of severity should be based, when feasible, on plausible exposures via drinking water. **The committee also recommends that EPA give consideration to different severity metrics such as ranking through use of either quality adjusted or disability adjusted life years lost from exposure to a contaminant.**	106
Regarding the assessment of contaminant prevalence, in some cases (particularly where contaminants have been included on a PCCL on the basis of potential occurrence rather than demonstrated occurrence), information will often be insufficient to directly assess temporal or spatial prevalence (or both). **Thus, EPA should consider the possibility of including information on temporal and regional occurrence to help determine (score PCCL) contaminant prevalence.** When prevalence cannot be assessed, this attribute must then go unscored and the attribute persistence/mobility used in its stead. The issue of changing (or incorporating) "thresholds" for contaminant detection, rather than relying on continually decreasing detection limits, is one that needs explicit attention and discussion by EPA and stakeholders.	106
Existing and readily available databases may not be sufficient to rapidly and consistently score health effect and occurrence attributes for individual PCCL contaminants for promotion to a CCL. **As recommended in Chapter 3, all information from existing or created databases or lists used in the development of a CCL and PCCL, should be compiled in a consolidated database that would provide a consistent mechanism for recording and retrieving information on the PCCL contaminants under consideration. As a starting point and as recommended in Chapter 3, EPA should review its developing EDSPD database to determing if it can be expanded and used (or served as a model for the development of) such a consolidated database and to help develop future PCCLs and CCLs.**	106
Contaminant databases used in support of the development of future CCLs should report summary statistics on all data collected, not only the quantifiable observations. In this regard, **EPA should formalize a process for reporting means and/or medians from data with large numbers of non-detect observations**. In addition, EPA may want to consider providing other measures of concentration in water supplies such as the 95th percentile of contaminant concentration.	106

Chapter 5: PCCL to CCL: Classification Algorithm	
A ranking process that attempts to sort contaminants in a specific order is not appropriate for the selection of drinking water contaminants already on a CCL. In the absence of complete information, the output of the prioritization schemes was found to be uncertain. **A linear model and a neural network were discussed and demonstrated for potential use in a prototype classification scheme**. It is recommended that EPA give careful consideration and experiment with developing a prototype classification approach using a neural network or similar methods. Furthermore, EPA should use several training sets to gauge the sensitivity or the adopted model. Neural networks provide the flexibility to capture linear as well as nonlinear dependencies. While the neural network performed better than the linear model (with respect to minimizing the number of misclassified contaminants), at this time the committee cannot make a firm recommendation as to which model EPA should use due to the aforementioned uncertainties in the training data set. Thus, the committee recommends that EPA explore alternative model formulations and be cognizant of the dangers of overfitting and loss of generalization.	140
To adopt and implement the recommended approach for the creation of future CCLs, EPA will need to employ or work with persons knowledgeable of prototype classification methods and devote appreciable time and resources to develop and maintain a comprehensive training data set. In this regard, **the committee strongly recommends that EPA greatly increase the size of the training data set that was used illustratively in this chapter to improve predictive capacity**. One way that EPA can expand the training data set and classification algorithm is to allow for the expected case of missing data. That is, purposefully include in the training data set contaminants for which values of some of the attributes are unknown and develop a scheme that allows prediction for contaminants for which some of the attributes are unknown.	140
EPA will also have to accurately and consistently assign attribute scores for all contaminants under consideration. To do this, EPA will need to collect and organize available data and research for each PCCL contaminant and document the attribute scoring scheme used to help ensure a transparent and defensible process, the importance of which was discussed in Chapter 2. To implement this scheme, EPA must purposefully include in the training data set contaminants for which values of some of the attributes are unknown and develop a scheme that allows prediction for these contaminants. As recommended in Chapter 3, the creation of a consolidated database that would provide a consistent mechanism for recording and retrieving information on the contaminants under consideration would be of benefit.	140

Chapter 5: PCCL to CCL: Classification Algorithm (continued)	
EPA will also need to withhold contaminants from inclusion in the training data set to serve as validation test cases that can assess the predictive accuracy of any classification algorithm developed While the committee was able to withhold 5 contaminants presumed worthy of regulatory consideration (T = 1) for this purpose, it had insufficient numbers of contaminants presumed not worthy of regulatory consideration (T = O) to similarly withhold. All withheld validation contaminants were correctly classified as belonging in the T = 1 category and such results provide (albeit limited) additional supporting evidence of the validity of the classification algorithm approach. EPA should make every effort to increase the number of both types of validation test cases (especially for T = O contaminants) to more thoroughly assess the predictive accuracy of any classification algorithm developed for use in the creation of future CCLs.	141
If neural networks are used for prototype classification, the transparency in understanding which contaminant attributes determine the category of a contaminant will be less than that of a linear model or more traditional rule-based scheme. However, if one acknowledges that the underlying process that maps attributes into categorical outcomes is very complex, then there is little hope that an accurate rule-based classification scheme can be constructed. The fact that the nonlinear neural network performed better than the linear classifier is a strong indicator that the underlying mapping process is complex and it would be a difficult task for a panel of experts to accurately specify the rules and conditions of this mapping. Furthermore, the loss in transparency in using a neural network is not inherent, but rather derives from the difficulty in elucidating the mapping.	141
The underlying mapping in a neural network classifier can be examined just as one would conduct experiments to probe a physical system in a laboratory. Through numerical experimentation, one can probe a neural network to determine the sensitivity of the output to various changes in input data. While a sensitivity analysis was not conducted due to time constraints, the **committee recommends that EPA should use several training data sets to gauge the sensitivity of the method as part of its analysis and documentation if a classification approach is ultimately adopted and used to help create future CCLs**.	141
Finally, EPA should realize that the committee is recommending a prototype classification scheme to be used in conjunction with expert judgment for the future selection of PCCL contaminants for inclusion screening on a CCL. Thus, transparency is less crucial (though no less desired) at this juncture than when selecting contaminants from the-CCL for regulatory activities as discussed in the committee's first report.	142
Chapter 6: Virulence-Factor Activity Relationships	
The committee believes that "virulence-factor activity relationships" or VFARs can be a powerful approach for examining emerging waterborne pathogens, opportunistic microorganisms, and other newly identified microorganisms, and for predicting the virulence of these pathogens.	145

Chapter 6: Virulence-Factor Activity Relationships (continued)	
The committee defines VFAR as the known or presumed linkage between the biological characteristics of a microorganism, and its real or potential ability to cause harm. VFARs are conceived as being the relationship that ties specific descriptors (genetic elements, surface proteins, toxins, attachment factors, metabolic pathways, invasion factors, and other possible virulence attributes) with outcomes of concern (virulence, potency, and persistence).	148
The committee has concluded that methods *other* than culture must be used to fully evaluate microbial contamination of drinking water (e.g., PCR-based methods).	161
The committee anticipates that, in a very short period of time, microarrays could be developed that are labeled with all of the genes for a variety of virulence factors identified within enteric bacteria, pathogenic viruses, opportunistic protozoa, and other (waterborne) microorganisms. These gene chips could be used to assay environmental and drinking water samples for the presence of genetic virulence factors of concern.	163
It is one of the committee's central assertions that the assessment of persistence (survival) in the environment using molecular techniques may be superior to some of the older methods.	172
The same prototype classification method developed to distill the PCCL to the CCL can also be applied to VFARs. Training sets of descriptor and response variables could be developed and used in conjunction with the prototype classification methods to help derive VFARs.	180
Establish a scientific VFAR Working Group on bioinformatics, genomics, and proteomics, with a charge to study these disciplines on an ongoing basis, and to periodically inform the Agency as to how these disciplines can affect the identification and selection of drinking water contaminants for future regulatory, monitoring, and research activities. The committee acknowledges the importance of several practical considerations related to the formation of such a working group within EPA, including how it should be administered, supported (e.g., logistically and financially), or where it could be located. However, the committee did not have sufficient time in its meetings to address these issues or make any related recommendations.	184
The findings of this report, and especially that of the Biotechnology Research Group (the *Interagency Report on the Federal Investment in Microbial Genomics*), should be made available to such a working group at its inception. The committee views the activities of a VFAR Working Group as a continuing process in which developments in the fields of bioinformatics, genomics, and proteomics can be rapidly assessed and adopted for use in EPA's drinking water program.	185
This Working Group should be charged with the task of delineating specific steps and related issues and timelines needed to take VFARS beyond the conceptual framework of this report to actual development and implementation by EPA. All such efforts should be made in open cooperation with the public, stakeholders, and the scientific community.	185

Chapter 6: Virulence-Factor Activity Relationships (continued)	
With the assistance of the Working Group, EPA should identify and fund pilot bioinformatic projects that use genomics and proteomics to gain practical experience that can be applied to the development of VFARs while simultaneously dispatching its charges outlined in the two previous recommendations.	185
EPA should employ and work with scientific personnel trained in the fields of bioinformatics, genomics, and proteomics to assist the Agency in focusing efforts on identifying and addressing emerging waterborne microorganisms.	185
EPA should fully participate in all ongoing and planned U.S. federal government efforts in bioinformatics, genomics, and proteomics as potentially related to the identification and selection of waterborne pathogens for regulatory consideration.	185

Identifying Future Drinking Water Contaminants (NRC, 1999)

1999 Recommendations of the NRC Committee on Drinking Water Contaminants	
Committee Report: A Conceptual Approach for the Development of Future Drinking Water Contaminant Candidate Lists	**Page Number**
An ideal CCL development process would include the following features: ▪ It would meet the statutory requirements of the SDWA Amendments of 1996, including requirements for consultation with the scientific community and opportunities for public comment. ▪ It would start by identifying the entire universe of potential drinking water contaminants prior to any attempt to rank or sort them. ▪ It would consider risks from all potential routes of exposure to water supplies, including dermal contact and inhalation as well as ingestion. ▪ It would use the same identification and selection process for microbial, chemical, and all other types of potential drinking water contaminants. ▪ It would have mechanisms for identifying similarities among contaminants and contaminant classes that can be used to assess potential risks of individual contaminants. ▪ It would result in CCLs containing only contaminants that, when regulated, would reduce disease, disability, and death, and it would exclude contaminants that have few or no adverse effects on human health (e.g.,contaminants entirely removed or detoxified through conventional drinking water treatment methods). However, EPA's resources are constrained, ...the committee believes that EPA can and should develop and use a process that: ▪ starts broadly, using existing lists of potential drinking water contaminants, supplemented by readily available information; ▪ considers microbiological, chemical and other types of potential contaminants in a common selection process; ▪ takes advantage of structure-activity relationships to help overcome deficiencies in health effects and occurrence data; ▪ expands the knowledge base over time; ▪ uses simple criteria, supplemented by expert judgement, to initially cull the candidates to a preliminary list; and ▪ employs a prioritization scheme, again supplemented by expert judgement, to identify final candidates for inclusion on a CCL.	3
EPA should develop a two-step process for creating future CCLs. In this process, a broad universe of potential drinking water contaminants is examined and then narrowed to a preliminary drinking water contaminant candidate list (PCCL) using simple screening criteria and expert judgment. Then, the PCCL is narrowed to a CCL using a quantitative screening tool in conjunction with expert judgment.	18
EPA should be as inclusive as possible in narrowing the universe of contaminants (perhaps on the order of 100,000 substances) down to a PCCL. The committee envisions that a PCCL would contain on the order of thousands of potential drinking water contaminants of all types for subsequent evaluation, prioritization, and culling to a CCL.	18

As a start, a PCCL should contain all substances and microbes that are known to cause significant adverse health effects (regardless of exposure route) and have the potential to occur in drinking water and those demonstrated to occur in drinking water supplies (unless they are known not to pose a significant health risk). A PCCL should also include all substances that may pose a drinking water risk based on their potential for occurrence and health effects.	18
Preparation of a PCCL should not involve extensive collection or analysis of data, nor should it drive research or monitoring activities. However, the committee recognizes that it will be necessary to develop and use screening criteria (e.g., production values of commercial chemicals) to shorten the list of contaminants for a PCCL.	18
Development of a PCCL should begin as soon as possible to support the development of the next CCL; the PCCL should be available for public and other stakeholder input (especially through the Internet) and should undergo scientific review.	18
A new PCCL should be generated for each CCL development cycle to account for new data and potential contaminants.	19
As an integral part of the CCL development process, the committee recommends the use of a comprehensive database that provides a consistent mechanism for recording and retrieving information on all the contaminants under consideration. A well-designed relational database can function as a "master list" that contains a detailed record of how the PCCL and CCL were developed, as well as providing a powerful analytical tool for the development of future CCLs.	19
To help identify commercial chemicals that might pose risks in drinking water, EPA should consider exercising its authority under the Toxic Substances Control Act (TSCA) to collect production and import data on both organic and inorganic chemicals by use category.	19
To assist in the evaluation of microbial pathogens, it also may be useful to identify common mechanisms of pathogenicity among contaminants in order to include them on future CCLs. An approach analogous to chemical structure-activity relationships (SARs) for microorganisms does not currently exist, but EPA should develop such a prioritization tool for microbial contaminants through use of gene data banks and with the cooperation and support of other federal and state health organizations.	19
Preparation of a CCL from a PCCL will require collection and evaluation of all available health effects and occurrence data for each substance on the PCCL. To cull a list of thousands of potential drinking water contaminants of all types to approximately a hundred for inclusion on the CCL, EPA must combine expert judgment equally with a single prioritization tool that can be used to evaluate any type of PCCL contaminant.	19

EPA should develop a prioritization tool to help narrow the PCCL to a CCL. The tool should be kept as simple as possible and be developed with regular public and other stakeholder input. Ensuring transparency throughout its development and avoiding "black-box" decision-making are critical steps. The tool should be validated using contaminants with extensive health effects and occurrence data and well-established risks. The tool must be able to identify and effectively address data gaps for each contaminant. Following a re-examination of 10 existing chemical hazard ranking schemes, the committee concluded that none was directly suitable for developing a CCL from a PCCL. However, 3 of the schemes (Cadmus, ITC, and WMPT) contain features that would be suitable for CCL development and could conceivably be adapted for this purpose.	19
The committee strongly recommends that no factors or components (e.g. measures of occurrence of health effects) of the tool should be weighted in any way (even through expert judgement).	13
The committee further recommends that any prioritization tool should be subjected to validation and scientific review prior to use.	13
EPA should reserve a number of contaminants or a percentage of the CCL for contaminants that are listed based solely or primarily due to expert judgment. EPA must also retain the ability to remove contaminants from inclusion on a CCL based on expert judgment.	19
The CCL should consist of roughly equal numbers of contaminants ready for regulatory decisions and those requiring further research to drive such efforts equally. This recommendation is consistent with EPA's partitioning of the first CCL into equivalent future action categories.	19
Regardless of what process is adopted by EPA to develop future CCLs, the committee strongly recommends that all contaminants that have not been regulated or removed from the existing CCL (and future CCLs) should be automatically retained on each subsequent CCL for reevaluation.	20
As in the previous report, the committee recognizes that the need for policy judgments by EPA cannot and should not be removed from any CCL development process. In making these decisions, EPA should use common sense as a guide and err on the side of public health protection.	20

Appendix B

Summary of the NDWAC CCL Work Group Investigation of QSAR Models as Sources of Data / Information for the CCL Development Process

The NRC (2001) suggested that QSAR analytical methodology be used to identify chemicals with potential occurrence in drinking water and potential adverse human health effects. This summary discusses work conducted toward evaluating the feasibility of applying specific QSAR models for future CCLs.

Background

A variety of QSAR models have been developed for human health endpoints and "packaged" into user-friendly commercial or public-use programs. Human health effect endpoints predicted by QSAR models include: mutagenicity, carcinogenicity, teratogenicity, neurotoxicity, reproductive and developmental toxicity, skin/eye sensitization and irritation, and systemic toxicity. QSAR development for other endpoints is underway by a number of EPA Office of Research and Development (ORD) organizations, other regulatory institutions and the private sector (Benigni and Richard, 1998).

The more popular commercial QSAR packages for human health include The Open Practical Knowledge Acquisition Toolkit (TOPKAT) (http://www.accelrys.com/products/topkat/), MultiCase (MCASE) (http://www.multicase.com/), and the Deductive Estimation of Risk from Existing Knowledge (DEREK) (http://www.chem.leeds.ac.uk/LUK/derek/index.html). Two general types of models can be distinguished: statistically-based models such as TOPKAT, and rule-based models such as MCASE. Two issue papers developed by the technical team for the NDWAC Work Group reviewed these models [*Screening Models and Algorithms (1/28/03)* and *Status and Feasibility of Using (Quantitative) Structure-Activity Relationship ((Q)SAR) Models for CCL Development* (8/7/03)]. Concise characterizations of these and other QSAR packages appear on OECD's (Organisation for Economic Cooperation and Development) web-site at http://webdomino1.oecd.org/comnet/env/models.nsf.

A broader array of QSAR packages are available for endpoints related to chemical occurrence than for health effects, and have much larger domains (i.e., are applicable to more chemicals). For example, EPA's Assessment Tools for the Evaluation of Risk (ASTER) (http://www.epa.gov/med/databases/aster.htm) includes a database of more than 56,000 chemicals, and batch searches of the entire European Inventory of Existing Commercial Chemical Substances (EINECS) directory of 166,000 chemicals have been conducted by the Danish EPA. Notably, for homeland security reasons public access to ASTER has been suspended; the status of other packages in this regard has not been determined.

The OECD web-site lists predictive models for environmental fate and exposure pathways, including human health routes of exposure. One with possible application in identifying chemicals with potential occurrence in drinking water is the physical-chemical predictive package, the Estimation Programs Interface for Windows (EPIWIN) (as part of the EPI Suite) (http://www.epa.gov/oppt/newchems/denver). Also, the Persistent, Bioaccumulative and Toxic

(PBT) Profiler is a versatile new QSAR package developed through Office of Prevention, Pesticides, and Toxic Substances (OPPTS) Sustainable Futures Program (www.pbtprofiler.net). It was recently released by OPPTS as a resource for industries to voluntarily use for screening chemicals.

Persistence and biodegradation estimates from packages such as CATABOL (http://btu6.btu.bg/main/software/catabol/) require mechanistic understanding of enzyme-mediated processes, similar to those available for predicting toxicity. Hence, QSAR models for persistence and biodegradation are more limited in scope and accuracy than packages that predict physical-chemical parameters.

Model Assessments

The technical support team for the NDWAC CCL Work Group conducted a limited assessment of the performance of two QSAR models, TOPKAT for predicting LOAELs, and EPI Suite for predicting chemical properties. The purpose of the evaluation was to inform NDWAC Work Group members regarding the potential for using QSAR derived data for future CCLs.

TOPKAT is a commercial computational toxicology package that uses chemical structural information (2-D descriptors of structural fragments) and QSAR models to estimate a range of human health toxicological and non-human ecological endpoints. Predictions are made for untested chemicals by comparison with structural fragments contained in the model's training set.

TOPKAT was selected for evaluation for several reasons. It includes the capability to predict rat chronic oral LOAELs for a variety of chemicals. It is currently being used by EPA ORD scientists in the National Center for Environmental Assessment (NCEA), who maintain a current license for its use for EPA research. NCEA has compiled a substantial historical database of LOAEL predictions for diverse chemicals, and made the database available for this exercise. Through the cooperation and technical assistance of NCEA in Cincinnati, OH, 525 compounds from each of three test groups of chemicals were run by technical support staff on NCEA's computers in order to expand the existing data set.

Developers of the statistically-based TOPKAT model estimated that predictions of rat oral LD50's fell within a factor of 5 from test results for 86-100 percent of the 4,000 chemicals in the model (Health Designs, Inc., 1997). An evaluation of TOPKAT by the Danish EPA, using 1,840 chemicals not contained in the training data set, gave somewhat poorer results: an $R^2 = 0.31$; they concluded that 86 percent of QSAR estimates fell within a factor of 10 from test results (Wedebye and Niemela, 2000). An evaluation by an EPA QSAR researcher in ORD estimates that for approximately one-third of compounds, the model indicates that no prediction is possible due to nonconformance with the training set domain (parameters are outside the "Optimum Prediction Space"). This means that the training sets used in the NCEA TOPKAT model are limited; a more expanded training set would reduce nonconformance. For 60 percent of the remaining two-thirds of substances, TOPKAT estimates of Lowest Observable Adverse Effects Levels (LOAELs) may roughly be within a factor of two. Since the model was developed, however, there have been additional chemicals assayed and not yet included in the training set database.

The EPI (Estimation Program Interface) Suite™ is a Windows-based suite of physical/chemical property and environmental fate estimation models[1] developed by the EPA's Office of Prevention, Pollution, and Toxics and Syracuse Research Corporation (SRC). EPI Suite™ uses a single chemical identifier called SMILES notation (Simplified Molecular Input Line Entry System) as input for 14 regression models to predict a suite of chemical parameters. EPI Suite™ includes algorithms for calculating aerobic biodegradability and water solubility, two of the properties in this limited analysis. BIOWIN™ was used to estimate aerobic biodegradability of organic chemicals and WSKOWWIN™ was used to estimate chemical water solubility. EPI Suite™ is routinely used by OPPT's Chemical Control Division for evaluation of new chemicals and chemical uses, as required under the Toxics Substances Control Act (Premanufacturing Notice, PMN).

EPI Suite™ is publicly available, user-friendly, and considered reasonably well-validated for the modules used in this assessment. EPI Suite™ enables the user to simultaneously run 10 estimation programs for a selected chemical.

The QSAR model evaluations were for three groups of chemicals, initially:

- 262 Draft CCL1 chemicals plus 22 Deferred Potential Endocrine Disruptors
- 262 Non-CCL1 chemicals with selected toxicity and environmental fate parameters
- 262 Non-CCL1 chemicals with no available data for selected toxicity and environmental fate endpoints.

The results for the first two groups of chemicals indicate how well the QSAR model predictions of chemical properties and toxicity endpoints for human health compare with published empirical data. The third data set provides some indication of the portion of chemicals lacking empirical data for which the models were able to predict the needed chemical properties and toxicity endpoints. EPA's National Center for Environmental Assessment (NCEA) provided some LOAELs already generated with TOPKAT, as well as measured values for comparison. To this list, CCL1 chemicals were added, as were chemicals with available data from the CCL Example Universe Data Set. Regulated contaminants and those identified in model training data sets were excluded. The initial data set was culled to a test data set of 695 chemicals.

Data Gathering

Input data required to run both TOPKAT and EPI Suite™ are chemical structural formulae represented by the SMILES notation. A database of over 103,000 SMILES notations are included in EPI Suite™, retrieved using chemical CAS Registry Numbers. Notations not identified in EPI Suite™ were developed manually.

To compare model predictions with available data, LOAELs were gathered from the Registry of Toxic Effects of Chemical Substances (RTECS). The lowest LOAEL from a rat or mouse oral

---

[1] See http://www.epa.gov/opptintr/exposure/docs/episuite.htm for more information on EPI Suite™. The model, including underlying algorithms and training sets, are available for free download from the U.S. Environmental Protection Agency at: http://www.epa.gov/oppt/exposure/docs/episuitedl.htm.

LOAELs from studies of 28 days or longer (TDlo value) was used, converted to a daily dose from the reported cumulative dose by dividing dose by study length (mg/kg-day). LOAELs for two hundred and twenty seven (227) chemicals were in RTECS.

Five data sources were used in a hierarchy to identify the chemical properties of solubility and biodegradation: SRC Chemfate Database, Physical –Chemical Properties and Environmental Fate Handbook, the Hazardous Substances Data Bank (HSDB), the International Program on Chemical Safety, and the National Toxicology Program. Biodegradation data was found for 100% of chemicals, and solubility measurements for 206 chemicals.

Estimated LOAELs were generated from TOPKAT, run on computers at NCEA in Cincinnati, with oversight by EPA staff. Solubility and biodegradation estimates were made with EPISuite™ models downloaded from EPA's web site. Two programs were downloaded and run for this evaluation: WSKOWWIN for predicting water solubility, and BIOWIN for predicting biodegradability. WSKOWWIN estimates water solubility (mg/L) of an organic compound by regression of the octanol-water partition coefficient (Kow). BIOWIN estimates the time required for a compound to biodegrade under aerobic conditions with mixed cultures of microorganisms. The half-life for biodegradation of a chemical in water is determined using the ultimate biodegradation expert survey module of BIOWIN. This estimation program provides an indication of a chemical's environmental biodegradation rate in relative terms such as hours, hours to days, days-weeks, and so on. The rate is estimated from the chemical's "half-life," i.e., the time required for one-half of the chemical to "completely" degrade (i.e., mineralization to H_2O and CO_2).

Findings

TOPKAT was able to predict LOAELs for 45% of 525 chemicals tested (QSAR predicted LOAELs were provided by NCEA for 170 chemicals). TOPKAT identified 55% of the chemicals as outside of the predictive domain. This evaluation of TOPKAT is comparable with preliminary results from NCEA, given study design differences. The comparison of TOPKAT LOAEL predictions with empirical results was difficult because of variability in empirical measurements. Using the lowest reported LOAEL resulted in 20 percent of model results within a factor of two of the empirical data, 53% within a factor of 5, and about 65% of predictions within a factor of 10 of empirical values. These are slightly lower than a comparative study by NCEA, using LOAELs from EPA's IRIS database, as compared to the use of RTECS LOAELs by this study.

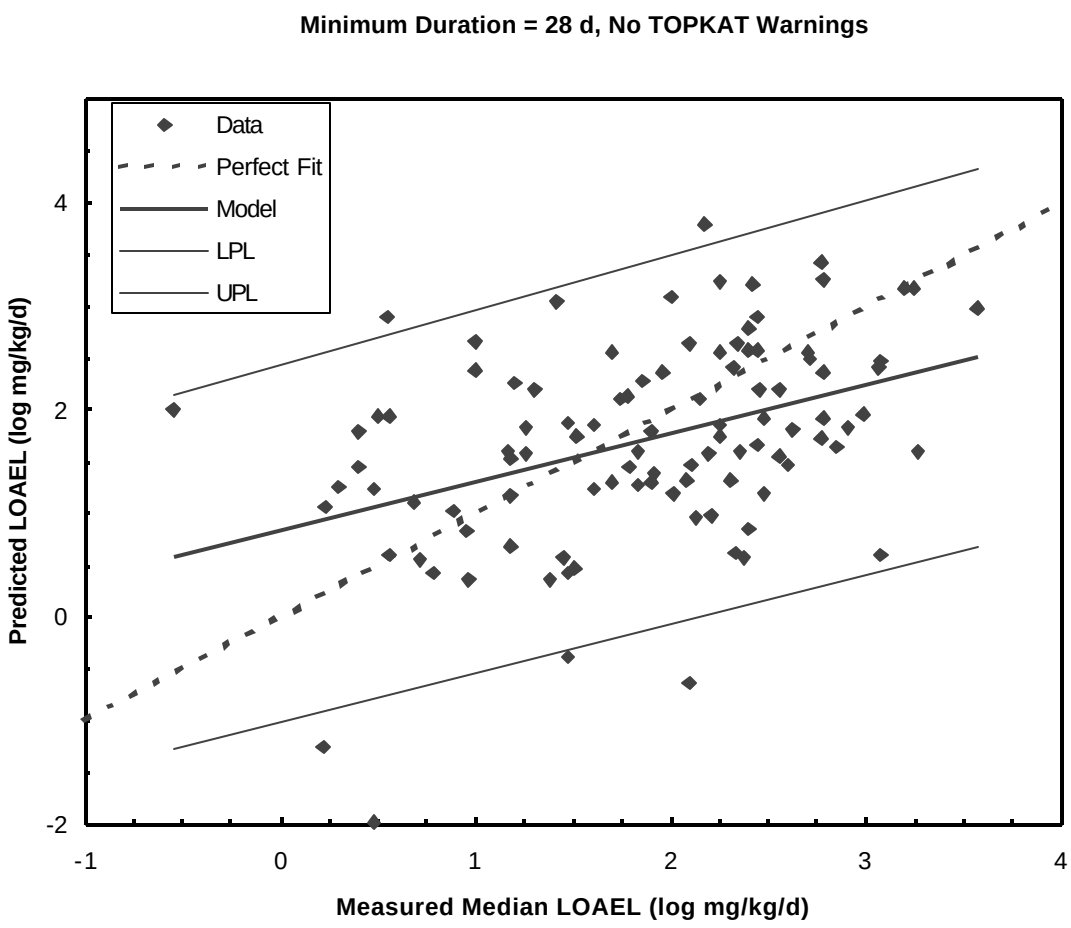
WSKOWWIN predicted water solubility for organic chemicals within a factor of five for 54% of chemicals evaluated. Relatively high variability was observed among 46 chemicals with multiple empirical values, due to methods diversity and variability. WSKOWWIN performance was user-friendly, and appears to be transparent and applicable to broad range of chemicals.

BIOWIN predictions broadly distinguished chemicals that degrade rapidly in the environment from those that degrade slowly. Empirical data from certain test procedures limited comparisons of predictions of degradation rates (e.g., weeks, months) from BIOWIN.

Conclusions

QSAR modeling using TOPKAT for health effects may require greater selectivity in chemicals (coverage of training sets) and health effects modeled. Comparison of QSAR model results to empirical data was limited by missing and highly variable measurements reported. This may generally limit the ability to fully evaluate QSAR model predictions for chemicals outside their respective training sets. Based on this validation exercise, TOPKAT appears to have limited utility for estimating rat chronic oral LOAELs. First, many compounds evaluated for future CCLs are likely to be outside the model domain because of the limiting training sets; model performance for these compounds was, therefore, less than optimal. Second, for compounds within the TOPKAT model domain, model performance was modest at best, though performance was similar to prior model evaluations. Of greatest concern was the apparent underestimation of toxicity for those compounds that are the most toxic, and therefore likely to be of greatest potential concern (see Figure 1). In addition to these two concerns, TOPKAT requires additional modules to be run in batch modes , and the training data set is proprietary.

Figure 1. Results of TOPKAT compared with empirical LOAEL data.



The frequency with which error codes were reported significantly limited the breadth and applicability of TOPKAT. Of 525 chemical queries conducted, 288 (55 percent) were accompanied by error codes indicating that they were outside the predictive domain. Most commonly, SI (poor fit training set data with queried chemical), TK (queried chemical contains an element not included in the training set) and OPS (outside the prediction space) codes were encountered. This is useful information to judge the utility of TOPKAT, a key goal of this exercise, though it will limit the ability to assemble quantitative data. TOPKAT predicted LOAELs for half of the chemicals it was able to evaluate within a factor of 5 of empirical data. However, the empirical data used for comparison may be a contributor to the poor correlation.

Part of the difficulty is that a LOAEL, even when specified as a rat chronic oral LOAEL as in TOPKAT, is not a specific health endpoint. That is, there are numerous toxicological mechanisms and specific adverse effects that can result from chemical exposure that are difficult to predict specifically based on chemical structure. It is generally recognized that limitations in understanding the mechanism of toxicological action for chemicals also limits the ability for developing QSAR models to predict these mechanisms and outcomes, and the exposure levels that cause them. EPA scientists and others have suggested that QSAR models for health effects with less biologically complicated outcomes, e.g., mortality as measured by the LD50.

Based on this limited evaluation, QSAR modeling appears to be a feasible approach for estimating solubility of organic compounds that lack empirical data. The WSKOWWIN module in the EPI Suite™ was able to provide reliable estimates of water solubility of organic compounds based on this limited analysis of experimental and predicted values. Further, this package is easily accessible, user friendly and the training data set is publicly available. WSKOWWIN does not address ionizing chemicals in its water solubility calculations.

The BIOWIN model of the EPI Suite™ package also appears to be useful for predicting chemical biodegradability. BIOWIN provides semi-quantitative estimates of time to complete mineralization, and these predictions have been shown to be reasonably reliable in terms of identifying chemicals that biodegrade fast or slow. Another model, the OASIS_Catabol software package, estimates percent biodegradation of chemicals in a 28 day inherent test, a different endpoint that could be a quantitative data element for evaluating chemical persistence.

Model selection for this exercise was limited by several factors, including time and staffing resources. Other QSAR models that predict other endpoints should be considered and tested to evaluate use of QSAR for CCL data. These may include QSAR models that predict: other health effects endpoints, other potential exposure data, degradation rates due to hydrolysis, photolysis or degradation mechanisms other than aerobic biodegradation.

Appendix C

Draft Scoring Protocols Developed and Used for Trial Attribute Scoring Exercise

This appendix provides five draft attribute scoring protocols used in an Attribute Scoring Exercise conducted as part of the NDWAC Work Group process. Protocols for Magnitude, Persistence-Mobility, Potency/Severity, and Prevalence follow.

Magnitude Scoring

This document describes how to assign a numerical score for the attribute magnitude, one of the five evaluated in the October 21, 2003 attribute scoring workshop.

NRC Definition of Magnitude

The National Research Council (NRC) defines magnitude as the concentration or expected concentration of a contaminant relative to a level that causes a perceived health effect.[1] NRC recommended that magnitude be scored on the basis of data on concentration and potency.

Approach for Magnitude Scoring

In this document, we describe an approach to attribute scoring that relies on concentration data alone.

Protocol for Magnitude Scoring based on Concentration Only

Step One: Identify highest-ranked data element

When more than one data element is available for a particular contaminant, use the hierarchy below to select the preferred element. Exhibit 1 presents the hierarchy of data elements to be used in the magnitude scoring process. Note that the Magnitude element should be correlated with the value used to score the attribute Prevalence.

[1] NRC 2001. *Classifying Future Drinking Water Contaminants for Regulatory Consideration*. Washington, D.C.: National Academy Press.

Exhibit 1. Hierarchy of Magnitude Data Elements

Rank	Magnitude Data Element	Type of Data
M1	Finished Drinking Water – Median of detected concentrations from Public Water Systems with detections	National scale finished drinking water occurrence data (NCOD, NIRS)
M2	Median of detected concentrations from ambient / raw source monitoring sites with detections	National scale ambient monitoring data (NAWQA)
M3	Median of detected concentrations from ambient / raw / source water samples with detections (Note: use combined surface / ground water if available and higher of SW/GW if not	National scale / representative data (NREC)
M4	Finished Drinking Water – Median of detected concentrations from Public Water Systems with detections	Individual state / small regional finished drinking water data
M5	Median of detections from ambient / raw / source water samples with detections	Individual state / small regional data
M6a	Environmental release data, total pounds or tons reported as released (TRI)	From Toxics Release Inventory
M6b	Environmental release data (ATSDR HazDat)	From ATSDR HazDat
M7	Production or Use data	Various. From EPA list (e.g. HPV) or actual production amount if available, or other information about use (e.g. consumer use insecticide), NCFAP.
M8	Manufactured Chemicals	Various. From lists (e.g. TSCA, HPV, IUR)

Step Two: Use look-up table to find attribute score for value identified in Step One.

For each ranked data element, there is a corresponding look-up table, which contains a range of data values assigned to a numerical magnitude score. Locate the look-up table associated with the highest-ranking data element identified in step one. Use the look-up table to determine the numerical score associated with the data value for the chemical being scored.

Place the magnitude score in the scoring worksheet. To generate a Magnitude Score as defined by NRC, multiply the Magnitude score by the Potency Score, take the square root, and record the resulting score.

LOOK-UP TABLES

Look-Up Table M1:
Finished Drinking Water – Median Detections from Public Water Systems with Detections

Median Detection Concentration (mg/L)			Magnitude Attribute Score
Combined (SW and GW)	SW	GW	
<0.00020	<0.00022	<0.00020	1
0.00020 - <0.00050	0.00022 - <0.00044	0.00020 - <0.00050	2
0.00050 - <0.00051	0.00044 - <0.00055	0.00050 - <0.00051	3
0.00051 - <0.00075	0.00055 - <0.00072	0.00051 - <0.00080	4
0.00075 - <0.00100	0.00072 - <0.00100	0.00080 - <0.00100	5
0.00100 - <0.00101	0.00100 - <0.00101	0.00100 - <0.00110	6
0.00101 - <0.00125	0.00101 - <0.00135	0.00110 - <0.00130	7
0.00125 - <0.00160	0.00135 - <0.00200	0.00130 - <0.00160	8
0.00160 - <0.00250	0.00200 - <0.00400	0.00160 - <0.00220	9
> 0.00250	≥0.00400	≥0.00220	10

If both surface water and ground water data are available, score the simple total of SW and GW values. If only surface water or ground water data are available, score the value using the corresponding surface water or ground water column.

Look-Up Table M2: Median Concentration Values (µg/L) from Drinking Water Sites with Detected Concentrations

Magnitude Range (µg/L)	Magnitude Attribute Score
0.0 - 0.049	1
0.05 - 0.099	2
0.1 - 0.5	3
0.51 - 0.99	4
1.0 - 1.5	5
1.51 - 3.0	6
3.01 - 5.0	7
5.01 - 9.99	8
10.0 - 50.0	9
50.01 +	10

Look-Up Table M3: Median of Source Water Samples with Detections (National Data)

Magnitude Range (µg/L)	Magnitude Attribute Score
0.0 - 0.049	1
0.05 - 0.099	2
0.1 - 0.5	3
0.51 - 0.99	4
1.0 - 1.5	5
1.51 - 3.0	6
3.01 - 5.0	7
5.01 - 9.99	8
10.0 - 50.0	9
50.01 +	10

Look-Up Table M4: Public Water Systems with Detections (Regional Data)

Median Detection Concentration (mg/L)			Magnitude Attribute Score
Combined SW + GW	SW	GW	
<0.00020	<0.00022	<0.00020	1
0.00020 - <0.0005	0.00022 - <0.00044	0.00020 - <0.00050	2
0.00050 - <0.00051	0.00044 - <0.00055	0.00050 - <0.00051	3
0.00051 - <0.00075	0.00055 - <0.00072	0.00051 - <0.00080	4
0.00075 - <0.00100	0.00072 - <0.00100	0.00080 - <0.00100	5
0.00100 - <0.00101	0.00100 - <0.00101	0.00100 - <0.00110	6
0.00101 - <0.00125	0.00101 - <0.00135	0.00110 - <0.00130	7
0.00125 - <0.00160	0.00135 - <0.00200	0.00130 - <0.00160	8
0.00160 - <0.00250	0.00200 - <0.00400	0.00160 - <0.00220	9
> 0.00250	≥0.00400	≥0.00220	10

If both surface water and ground water data are available, score the simple total of SW and GW values. If only surface water or ground water data are available, score the value using the corresponding surface water or ground water column.

Currently, no data were located for this table (from the 41 data sources sought). It is anticipated that these data will become available in the future, and when they do, will fall here in the hierarchy.

Look-Up Table M5: Median of Source Water Samples with Detections (Regional Data)

Magnitude Range (µg/L)	Magnitude Attribute Score
0.0 - 0.049	1
0.05 - 0.099	2
0.1 - 0.5	3
0.51 - 0.99	4
1.0 - 1.5	5
1.51 - 3.0	6
3.01 - 5.0	7
5.01 - 9.99	8
10.0 - 50.0	9
50.01 +	10

Look-Up Table M6a: Environmental Release Data (TRI)

Toxic Release Inventory Total Reported Release in 2001 (Quantity in Pounds)	Magnitude Attribute Score
< 10 pounds	1
11 - 300	2
301 - 1,000	3
1,001 - 10,000	4
10,001 - 50,000	5
50,001 - 300,000	6
300,001 - 1,000,000	7
1,000,001 - 8,000,000	8
8,000,001 - 40,000,000	9
> 40 million	10

Look-Up Table M6b: Other Environmental Release Data (ATSDR HazDat)

Maximum Concentration (mg/L) Con	**Magnitude attribute score**
≤0.001	1
≤0.01	2
≤0.03	3
≤0.08	4
≤0.17	5
≤0.34	6
≤0.80	7
≤2.30	8
≤10	9
≤100	10

Look-Up Table M7: Data for Pesticides

Mass of Pesticides Applied or Used	**Magnitude Attribute Score**
Default for any pesticide for non-environmental use (restricted hospital or indoor use)	3
Default for any pesticide in environmental use without data	5
≥ 100,000 lbs	6
≥ 1,000,000 lbs	7
≥ 2,000,000 lbs	8
≥ 20,000,000 lbs	9
≥ 50,000,000 lbs active ingredient applied	10

Look-Up Table M8: Mass Produced/Imported Annually

Mass Produced/Imported Annually (TSCA, HPV)	Corresponding Score
no data available on production from any source	1
data available from other sources and < 10,000 lbs	3
≥10,000 lbs (CUS/IUR)	5
≥10,000 lbs and other factors warrant a higher score: e.g., consistently reported at this level since CUS/IUR reporting began; or in routine/wide commercial use; or other sources indicate production $ 100,000 lbs	6
≥ 1,000,000 lbs (CUS/IUR, HPV)	7
≥ 1,000,000 lbs and other factors warrant a higher score: e.g., consistently reported at this level since CUS/IUR reporting began; or in routine/wide commercial use	8
≥ 1,000,000,000 lbs (CUS/IUR)	9
≥ 1,000,000,000 lbs and other factors warrant a higher score: e.g., consistently reported at this level since CUS/IUR reporting began; or in routine/wide commercial use	10

Persistence - Mobility Scoring

This document describes the process for assigning a numerical score for the attribute persistence-mobility, one of the five attributes to be tested in the October 21, 2003 attribute scoring workshop. In the protocol, Persistence - Mobility may be scored as a fifth attribute, or as a surrogate measure for the attribute Prevalence, as the lowest element in the hierarchy.

NRC Definition of Magnitude

The National Research Council (NRC) defines persistence-mobility as a surrogate measure when prevalence is unavailable, describing the likelihood that a contaminant would be found in the aquatic environment based solely on it physical properties.[2] NRC recommended that persistence-mobility be scored on the basis of data on physical chemical properties such as solubility and half-life.

Approach for Persistence-mobility Scoring

The approach for scoring includes assigning two scores, one for persistence and one for mobility, on a numeric scale of 1 through 3, representing low, medium, and high. Using a hierarchy of physical property data elements, each contaminant is scored for both persistence and mobility. The average of these two scores is multiplied by 10/3 to obtain the persistence-mobility score. Below are two tables that include a hierarchy of available properties for each data element representing either persistence or mobility.

Protocol for Persistence-mobility Scoring

Step One: Identify and score highest-ranked data value for Persistence

When more than one data element value is available for a particular contaminant candidate, use the hierarchy below to select the preferred element. Exhibit 1, below, describes the hierarchy of data elements to be used in the Persistence scoring process. When several values for a physical property are available, the highest scoring value should be used for scoring, unless that value is not representative of environmental conditions in drinking water. Enter the element type, source, and attribute score in the Persistence Mobility Worksheet. Also record any available supporting information, e.g. test conditions, in the Notes column.

[2] NRC 2001. *Classifying Future Drinking Water Contaminants for Regulatory Consideration*. Washington, D.C.: National Academy Press.

Exhibit 1. Hierarchy of Persistence Data Elements

Hierarchy	Element	1 (low)	2 (medium)	3 (high)
P1	Half Life (T1/2)	< 1 week	>1week - < 4 weeks	> 4 weeks
P2	Stability (abiotic and biotic degradation)	measured or calculated biotic or abiotic half-life in environmental or laboratory waters (excluding abnormal conditions like activated sludge, extreme pH) is less than one week	measured or calculated biotic or abiotic half-life in environmental or laboratory waters (excluding abnormal conditions like activated sludge, extreme pH) is less than one month.	measured or calculated biotic or abiotic half-life in environmental or laboratory waters (excluding abnormal conditions like activated sludge, extreme pH) is one month or longer.
P3	Biodeg rate (measured)	days days-weeks	weeks weeks - months	months recalcitrant
P4	Biodeg rate (estimated)	days days-weeks	weeks weeks - months	months recalcitrant

Step Two: Identify and Score Highest Ranking Value for Mobility

The hierarchy of physical properties for scoring mobility is in Exhibit 2. Select the left-most data element available for scoring. When several values for a physical property are available, the highest scoring value should be used for scoring, unless that value is not representative of environmental conditions in drinking water.

Exhibit 2. Mobility Scoring Hierarchy

Hierarchy	Element	1 (Low)	2 (Medium)	3 (High)
M1	Organic Carbon Partition Coefficient (Koc)	>300	100-300	<100
M2	Log of Octanol-Water Partition Coefficient (Log Kow)	>4	1-4	<1
M3	Dissociation Constant (Kd) (cm^3/g)	<5	1-5	>5
M4	Henry's Law Constant (HLC) (atm m^3/ mol)	$>10^{-3}$	10^{-7} - 10^{-3}	$<10^{-7}$
M5	Solubility (mg/L)	<1	1-1,000	>1,000

Step Three: Multiply the average of the persistence and mobility scores by 10/3 for the persistence-mobility score.

Step 3B. Alternately, use one of the two elements (multiply score by 10/3) if only one is available.

If using persistence-mobility as surrogate measures for prevalence, use the Persistence-Mobility Score for Prevalence. This will be used in conjunction with the use of production data for scoring Magnitude (see Prevalence and Magnitude Attribute Scoring Protocols for details).

Potency/Severity Scoring

This document describes the process for assigning a numerical score for potency and severity, during the October 21, 2003 attribute scoring workshop.

Protocol for Potency Scoring

Step One: Open the spreadsheet for Potency and Severity Scoring

Step Two: Enter the name of the chemical in the column labeled contaminant

Step Three: Identify and score highest-ranked data element for potency using the following hierarchy of values.

$ RfD or equivalent>NOAEL>LOAEL>LD50
$ Measured > Modeled
$ EPA RfD> ATSDR MRL (Chronic> Intermediate >Acute)**> Cal EPA PHG >WHO/EU/Health Canada**
$ OPP> IRIS for Pesticides

Step Four: Enter the selected measure of potency into the appropriate column of the spread sheet. Make sure that the units are in mg/kg/day.

Step Five: Select a measure for cancer potency if one is available. The preferable measure will be the E-4 risk concentration in drinking water in mg/L. If the risk is expressed at levels other than E-4, convert the value to the target risk (E-4). If the cancer potency measure is the slope factor, calculate the E-4 risk concentration using the following equation:

$$\text{E-4 Risk concentration} = \frac{10{,}000 \times 35 \text{ kg/day/L}}{\text{Slope Factor (mg/kg/day)}^{-1}}$$

Step Six: Choose the higher of the non-cancer or cancer potency score as the measure of potency.

Protocol for Severity Scoring

Step 1: Enter the critical effect that goes with the potency score selected through the Potency Protocol in the appropriate column of the Potency-Severity Spreadsheet.

- **$** If the potency is based on a tumorigenic response enter cancer as the critical effect. If information on tumor type is provided include that information.
- **$** If the potency score is derived from a non-cancer parameter enter the critical effect(s) that go with the RfD, LOAEL, or NOAEL (no observed effect).
- **$** If the potency score is from an LD50 study (measured or modeled) enter death as the critical effect.
- **$** If the potency score is a modeled LOAEL examine the Health Effects Information to determine if an appropriate critical effect can be determined.

Step Two: Use the Severity Scoring Sheets (A and B) to give the Critical Effect a Score and enter that score on the Potency/ Severity Scoring Sheet.

- **$** Severity Score A should be selected from the nine point scale and Severity Score B from the five point scale.

Prevalence Scoring

This document describes how to assign a numerical score for prevalence, one of the attributes evaluated for the October 21, 2003 Attribute Scoring Workshop.

Definition of Prevalence

The National Research Council (NRC) defines prevalence as how commonly a contaminant occurs, or would occur, in drinking water[3]. Prevalence ideally involves both spatial and temporal occurrence, and should be scored based on seven measurements (in order of preference): tap water, distribution systems, finished water of treatment plants, and source water used for supplying drinking water. If no information is available to demonstrate occurrence in drinking water, use: observations in watersheds/aquifers, historical contaminant release data, or chemical production data.

Approach for Prevalence Scoring

We have followed the approach recommended by NRC when scoring candidates for prevalence. A wide variety of data sources exist which could be used for this exercise. Some of these sources are better than others, and every effort was made to use the best and most complete data available.

Protocol for Prevalence Scoring

Step One: Identify highest-ranked data value

When more than one data value is available for a particular contaminant candidate, a pre-established hierarchy ensures that scoring decisions are made consistently. Exhibit 1, below, describes the hierarchy of data elements to be used in the prevalence scoring process. Each data source has been given a rank, corresponding with Exhibit 1, with 1 being the top of the hierarchy. Note that the data element used to score Prevalence should be correlated with the value used to score the attribute Magnitude. That is, the attribute Magnitude should be score with the element in the corresponding M rank, in the accompanying Magnitude Scoring Protocol.

[3] NRC 2001. *Classifying Future Drinking Water Contaminants for Regulatory Consideration*. Washington, D.C.: National Academy Press.

Exhibit 1. Hierarchy of Prevalence Data Elements

Rank	Prevalence Data Element	Type of Data
P1	Finished Drinking Water - Percentage of Public Water Systems (PWSs) with Detections	National scale / representative data (NCOD,NIRS)
P2	Percentage of Ambient/Raw/Source Monitoring Sites with Detections	National scale / representative data (NAWQA)
P3	Percentage of Ambient/Raw/Source Monitoring Samples with Detections	National scale / representative data (NREC)
P4	Finished Drinking Water - Percentage of PWSs with Detects	Individual state / small regional data
P5	Percentage of Ambient/Raw/Source Monitoring Sites with Detects	Individual state / small regional data
P6a	Environmental release data, number of states reporting releases	From Toxics Release Inventory
P6b	Hazardous substance release data, number of states	HazDat
P7	Production or Use data for Pesticides	Various. From EPA (e.g. HPV) list or actual production amount if available, or other information about use (e.g. consumer use insecticide) - NCFAP
P8	Persistence / Mobility data	Physical chemical properties

Data elements corresponding to rank P1 should be looked for first. If this element exists, this is the one to use, no other elements need to be considered. If there are no data for rank P1, data for rank P2 should be sought, and so on down the list until the highest ranked element is located.

Step Two: Use look-up table to find attribute score for value identified in Step One.

For each rank there is a corresponding "look up table" which contains a range of data values assigned to a numeric prevalence score between 1 and 10. Once a data value has been found for a particular element, that value can be looked up on these tables to determine the prevalence score. The lookup tables are listed below.

Look-Up Table P1: Finished Drinking Water - Percentage of PWSs with Detections (national data)

Category Score	Total (SW and GW)	SW Only	GW Only
	% PWSs with detections	% PWSs with detections	% PWSs with detections
1	≤ 0.10	≤ 0.22	≤ 0.07
2	0.11 - 0.16	0.23 - 0.37	0.08 - 0.12
3	0.17 - 0.25	0.38 - 0.63	0.13 - 0.18
4	0.26 - 0.44	0.64 - 1.00	0.19 - 0.29
5	0.45 - 0.61	1.01 - 1.50	0.30 - 0.45
6	0.62 - 1.00	1.51 - 2.00	0.46 - 0.71
7	1.01 - 1.30	2.01 - 3.10	0.72 - 1.20
8	1.31 - 2.50	3.11 - 6.00	1.21 - 2.50
9	2.51 - 10.00	6.01 - 15.00	2.51 - 10.00
10	> 10.00	> 15.00	> 10.00

Record the attribute scores of all three measures. Use the Total for the Prevalence Score to look up the Magnitude Score. In addition, if both surface water and ground water data are available, score both SW and GW values and identify which provides a higher category score for prevalence. If there is no distinction between SW or GW values score with the value provided. If only surface water or ground water data are available, score the value using the corresponding surface water or ground water column. Data for use with Table P1 can be found in the data sources NCOD and NIRS.

Look-Up Table P2: Percentage of Ambient/Raw/Source Monitoring Sites (national data) with Detections

Prevalence Range (% sites w/ Detects)	**Corresponding Score**
0.0 - 0.05	1
>0.05 - 0.1	2
>0.1 - 0.5	3
>0.5 - 1.0	4
>1.0 - 2.0	5
>2.0 - 5.0	6
>5.0 - 10	7
>10 - 20	8
>20 - 40	9
>40 - 100	10

Data for use with Table P2 may be found in NAWQA.

Look-Up Table P3: Percentage of Ambient/Raw/Source Monitoring Samples (national data) with Detections

Prevalence Range (% samples w/ Detects)	Corresponding Score
0.0 - 0.05	1
>0.05 - 0.1	2
>0.1 - 0.5	3
>0.5 - 1.0	4
>1.0 - 2.0	5
>2.0 - 5.0	6
>5.0 - 10	7
>10 - 20	8
>20 - 40	9
>40 - 100	10

Data for use with table P3 are found in NREC.

Look-Up Table P4: Finished Drinking Water - Percentage of PWSs (local/state data) with Detects

	All PWSs	SW Only	GW Only
Prevalence Score	Weighted Average % PWSs with detections	% PWSs with detections	% PWSs with detections
1	≤ 0.10	≤ 0.22	≤ 0.07
2	0.11 - 0.16	0.23 - 0.37	0.08 - 0.12
3	0.17 - 0.25	0.38 - 0.63	0.13 - 0.18
4	0.26 - 0.44	0.64 - 1.00	0.19 - 0.29
5	0.45 - 0.61	1.01 - 1.50	0.30 - 0.45
6	0.62 - 1.00	1.51 - 2.00	0.46 - 0.71
7	1.01 - 1.30	2.01 - 3.10	0.72 - 1.20
8	1.31 - 2.50	3.11 - 6.00	1.21 - 2.50
9	2.51 - 10.00	6.01 - 15.00	2.51 - 10.00
10	> 10.00	> 15.00	> 10.00

For the workshop, no data were located for this table (from the 41 data sources). It is anticipated that these data will become available in the future. When these data are available they will fit into this level of the hierarchy.

Look-Up Table P5: Percentage of Ambient/Raw/Source Water Sites (local/state data) with Detects

Prevalence Range (% sites w/ Detects)	Corresponding Score
0.0 - 0.05	1
>0.05 - 0.1	2
>0.1 - 0.5	3
>0.5 - 1.0	4
>1.0 - 2.0	5
>2.0 - 5.0	6
>5.0 - 10	7
>10 - 20	8
>20 - 40	9
>40 - 100	10

For the workshop, no data were located for this table (from the 41 data sources). It is anticipated that these data will become available in the future. When these data are available they will fit into this level of the hierarchy.

Look-Up Table P6a: Number of States Reporting TRI releases

Number of States reporting discharges	Corresponding Score
1	1
2	2
3	3
4	4
5	5
6	6
7-10	7
11-15	8
16-25	9
> 25	10

Look-Up Table 6b: Number of States Reporting Contaminant in ATSDR HazDat (database of hazardous waste sites)

Number of States reporting discharges to surface water	Corresponding Score
1	1
2	2
3	3
4	4
5	5
6	6
7-10	7
11-15	8
16-25	9
> 25	10

Look-Up Table P7: Surrogate Data for Pesticides

Number of States in which Pesticide was used in 1997	Corresponding Score
Default for any pesticide for non-environmental use	3
Default for any pesticide in environmental use without use data	5
<6 in database	6
6-10	7
11-15	8
16-25	9
>25	10

The available data for Table P7 are from the source NCFAP.

Look-Up Table P8: Persistence Mobility

Please refer to the Persistence-Mobility Scoring Protocol for using P8.

Appendix D
Microbial Protocols / Attribute Scoring

Exhibit D1 lists the data elements associated with health effect and pathogen occurrence that constitute the basis for the development of the attribute scoring system proposed in this appendix.

Exhibit D1 - Data Elements for Pathogen Scoring

Attribute	Elements
Health Effect	Severity of disease (manifestations, duration, sequelae, etc.)
	Susceptible populations (number, immune status, etc.)
	Incidence of disease (all sources vs. water-borne)
	Availability and efficacy of available treatment
	Infectious dose
Pathogen Occurrence	Presence in source water
	Persistence-mobility (stability, growth potential in water)
	Documented water-borne outbreaks reported
	Route of transmission (ingestion, inhalation or dermal contact)
	Size of exposed population

Potency

Potency is defined as the amount of a contaminant that is needed to cause illness. For microbes the infective dose is the most useful marker of potency, however the infective dose is not known for many pathogens. Microbiologists frequently speak in terms of the minimum infective dose, but the terms LD_{50} and lethal dose apply only to animal studies or *in vitro* cell culture assays. Some pathogens cannot be grown in the laboratory and their infective dose can only be estimated. In the future, quantitative virulence-factor activity relationships may become available for determining the relative potency of a pathogen.

The data elements for scoring potency include knowledge of water-related disease, the class of pathogen, i.e. bacteria, viruses, protozoa, the burden of disease in the population, the infectious does of the pathogen, the likelihood of fecal or urinary shedding in humans and animals, and the presence of genomic sequences conferring virulence.

A proposed system for scoring potency is shown in Exhibit D2. Data elements providing answers to Category 1 questions are readily available from reference sources, however data elements for Category 2 and Category 3 questions are not available for many pathogens on the PCCL. For this reason, the questions are constructed in a manner to allow for uncertainty or unavailability of data, while admitting the use of available information.

Exhibit D2 - Potency Scoring Protocol

Category 1[1,2]	Category 2[3]	Category 3		Score[4]
Causes water-related disease in otherwise healthy individuals	Morbidity rate high			11
	Morbidity rate low or uncertain	Viruses or protozoa		10
		Bacteria or fungi	Enteric	9
			Non-enteric	8
No water-related disease, but organism is a primary or opportunistic pathogen	Published human ID_{50} value available	$ID_{50} < 10^2$		7
		$ID_{50} > 10^2$		5
	Published humanID_{50} value not available	Viruses or protozoa		7
		Bacteria or fungi	Enteric	6
			Non-enteric	5
Animal pathogen shed in feces or urine				4
Genetic sequences are available in searchable databases	Organism has known virulence genes or gene products	Known pathogenicity islands		3
		No known pathogenicity islands		2
	Virulence genes not documented			1

[1]Assumes all microbes on PCCL are pathogens, or potential pathogens, and occur or may occur in water.

[2]For scoring potency of toxins (e.g., various cyanotoxins, aflatoxins), assume that a toxin is a pathogen.

[3]Infective dose based on single dose exposure to healthy individuals.

[4]Use single highest score for each microbe.

The most obvious data point for potency scoring is infective dose, however infective dose data are rarely available, and extremely variable due to strain and host variability. Biological properties of pathogens may be used to estimate potency where infective dose data are unavailable. Assumptions built into the algorithm presume that viruses and protozoa have a lower infective dose than bacteria, hence they score higher. Preliminary scoring exercises on CCL organisms revealed little separation among scores. A more inclusive test data set would probably provide a range of scores useful in ranking potency of pathogens. The position of VFARs in the algorithm is controversial since genomic sequences are available for few pathogens on the CCL, however the Work Group determined that manifestation of disease (genomic expression) suggested higher potency than genomic potential (presence of virulence genes or pathogenicity islands), or stated another way, functional data carry more significance than structural data.

Severity

NRC defines severity as the seriousness of the health effect, and suggests severity be based on "the most sensitive health endpoint for a particular contaminant, and considering vulnerable subpopulations; ... [and] should be based, when feasible, on plausible exposures via drinking water."

The risk assessment terminology applicable to chemicals becomes problematic in the microbiological context of the host-pathogen relationship. For microbial agents, severity may be defined in terms of colonization, infection, immune response, disease, sequella, or death. The host-pathogen relationship is variable and dynamic. This continuum may be unrecognizable at

various stages. The most sensitive endpoint indicative of host-pathogen interaction is an immune response, however this is not a practical end point for assessment of health effects, since immunodeficient populations may be infected without eliciting an immune response. While chemical health effects may be immediate or cumulative, microbiological health effects may be unapparent for an extended time, depending upon the incubation period of the pathogen, and the manifestation of disease.

The data elements for scoring severity include recognition of significant morbidity and mortality, the location and intensity of infectious processes, the extent of contagion, the amount of time lost to illness, the extent to which medical intervention is required for recovery, and chronic manifestations or disabilities associated with the disease.

A central issue with severity scoring is whether to score on acute manifestations of disease in normal populations, or to score the worst possible outcome in the most sensitive population. Because most frank pathogens are capable of killing some segment of the population, using worst possible outcome in the most sensitive host inflates and clusters scores. The initial severity scoring tables were constructed to use median outcome in normal populations, with case fatality rate and patient population classification and percentage of patients in the population classifications as weighting factors. This approach was criticized as overly complex, and potentially contentious, and the Work Group sought alternative scoring approaches.

One such approach applied the attribute characteristics to the population for which the most data and information were available, then recalculating scores to acknowledge special circumstances and to apply additional stringency. This proposed system applied worst case scoring criteria for healthy and sensitive sub-populations, thereby driving many pathogens to maximal scores. In an effort to overcome the complexities and limitations of a scoring system using case fatality rates and population-based weighting factors, the Work Group proposed a series of questions carefully constructed so that a ‘yes’ answer signified significance while a ‘no’ answer did not.

Twelve questions were constructed to capture progressively severe outcomes of disease, and the sum of ‘yes’ answers constitutes the numerical severity score for a particular pathogen (Exhibit D3). This binary scoring process was conducted for typical and worst case disease outcomes in normal and sensitive sub-populations for the microbial contaminants on the current CCL. The results provided reasonable spread for both patient populations, although scoring ‘worst case’ tended to cluster pathogens toward the upper end of the scale. While ‘worst case’ scoring is believed to provide the highest level of public health protection, it fails to consider existence of other reservoirs and transmission routes for pathogens besides drinking water, and places undue responsibility for prevention of infectious diseases on the EPA regulatory process. By limiting the manifestations of disease to those related to infections acquired by ingestion, inhalation, and dermal contact with drinking water, the binary scoring system produces plausible results for severity of illness.

Exhibit D3 - Severity Scoring Protocol

Question	Yes [1]	Question
1		Does the organism cause a CDC notifiable disease?
2		Does the organism cause significant morbidity (> 1,000/year) in the U.S.?
3		Is diarrhea a symptom of illness?
4		Does the illness require medical intervention for resolution?
5		Does the organism disseminate from the gastrointestinal tract to other organs?
6		Does the organism cause mild disease in normal populations, but severe disease in individuals with predisposing conditions?
7		Is illness associated with 3 or more days lost from school or work?
8		Is person-to-person spread a typical component of the disease syndrome?
9		Does illness usually require hospitalization?
10		Does the organism cause pneumonia, meningitis, hepatitits, encephalitis, endocarditis, or other severe manifestations of illness?
11		Does illness result in long-term disability or sequella?
12		Does the organism cause significant mortality (> 1/1,000 cases)?
Total Score[2]		

[1]Enter 1 for each yes answer.

[2]Add the numbers in the column to arrive at a final score.

Prevalence

For the occurrence attributes, NRC defines prevalence as, "How commonly does or would a contaminant occur in drinking water?" Prevalence may be determined via the seven measures proposed by NRC in the PCCL screening criteria for demonstrated or potential occurrence (in order of preference): (1) tap water, (2) distribution systems, (3) finished water of water treatment plants, and (4) source water used for supplying drinking water. If no information is available to demonstrate occurrence in water, NRC recommended evaluating the potential for occurrence in water through: (5) observations in watersheds/aquifers, or (6) historical contaminant release data. It should be emphasized that prevalence involves the consideration of both geographical (spatial) and temporal ranges of occurrence.

Most pathogen occurrence data are based upon indicator monitoring, hence they become surrogate information, not pathogen occurrence data. True pathogen occurrence data come from epidemiological investigations following outbreaks, research studies on pathogen distribution, and detection method evaluations. There is little pathogen information and less pathogen data regarding environmental and drinking water occurrence.

The Work Group developed a conceptual framework for prevalence, based upon actual detection in drinking water, actual detection in source water, potential for zoonotic transmission through water contamination, and potential for zoonotic agents to infect humans (host range).

These factors are the basis of Exhibit D4. Prevalence scoring using these criteria proved to be more straight forward that other attributes, primarily because occurrence data are either available or not available, limiting the number of criteria in the scoring system.

Exhibit D4 - Prevalence Scoring Protocol

Category 1[1]	Category 2	Score[2]
Detected in drinking water		7
Not detected in drinking water, but detected in source water	Documented WBD in swimmers in the U.S.	6
	Common in source water	5
	Monitored but infrequently detected	4
	Rarely monitored or never detected	3
Not detected in drinking water or source water	Broad host range for animals and humans	2
	Narrow host range limited primarily to humans	1

[1]Based upon worldwide occurrence data.

[2]Select the single highest score for each organism.

Persistence-Mobility

NRC used a persistence/mobility attribute as a surrogate for potential occurrence when information is unavailable for a contaminant regarding its demonstrated occurrence in water. For microorganisms, the following three characteristics pertain to their persistence and/or mobility: high potential for amplification under ambient conditions, sedimentation velocities and absorption capabilities, and death or the ability to produce non-culturable or resistant states (e.g. spores and cysts). When a contaminant already has data on demonstrated occurrence in water, and thus information for the prevalence and magnitude attributes, those attributes will take precedence over persistence/mobility.

Pathogenic microorganisms are genetically adapted to their hosts, and they do not typically survive the rigors of the ambient environment. Some fastidious pathogens such as *Treponema pallidum* and HIV are inactivated within seconds of exposure to the ambient environment. The factors determining the persistence of microorganisms in aquatic systems include:

- ability to withstand ambient conditions of temperature, pH, ionic strength, radiation and oxygenation
- ability to compete with other microorganisms for substrate
- ability to produce or sequester themselves in biofilms or adsorbed to particles
- ability to produce resting forms, e.g. cysts or spores
- ability to persist in viable but non culturable state
- ability to enter into commensal or symbiotic relationships with other microorganisms
- ability to resist disinfectants
- presence of predators, e.g. amoeba, cilliates, etc.

Persistence implies steady state occurrence or amplification of microorganisms in water. This occurs in surface water by production of resistant forms such as spores, cysts, oocysts, by colonization of other life forms serving as a reservoir, through symbiotic relationships with amoebae, by adsorption to particles, or production of quiescent forms such as viable but non-culturable bacteria. In water treatment plants and distribution systems, persistence is associated with colonization of infrastructure, e.g. production of biofilm. Organisms that amplify are given higher scores than organisms that produce resistant forms but do not amplify in water. This

scoring scale may overemphasize relatively innocuous organisms that produce biofilms but rarely or never cause disease in humans.

Data elements for scoring persistence-mobility include survival time in water under ambient conditions, ability to amplify, ability to produce resistant forms, relationship to particles, and potential for symbiotic relationships enhancing survival.

The persistence-mobility scoring table (Exhibit D5) emphasizes non-turbid waters, i.e. groundwater and treated drinking water, but the Work Group believes that all source water should be included in scoring. The rationale for excluding turbid waters was that organisms adsorbed to particles persist considerably longer than organism in non-turbid waters. Amplification frequently occurs in surface source water due to large amounts of available nutrient, whereas the assimilable organic carbon is limited in groundwater and treated water, slowing or restricting amplification. For example, *Aeromonas hydrophila* grows to population densities in excess of 8 logs per mL in sewage, from 4-6 logs per mL in surface water, while maximal levels in distribution water rarely exceed 2 logs per mL (typical levels range from 10^{-2} to 10 CFU/mL), and groundwater is typically less than 1 CFU/mL. Persistence of bacteria, which amplify under environmental conditions is highly variable, and the extent to which they persist and move is largely a function of their population density. It may be inappropriate to equate persistence-mobility of organisms in surface waters with persistence-mobility in non-turbid waters.

Mobility is not limited to chemicals, since microorganisms move though the aqueous environment and in distribution system water actively (motility) and passively (adsorbed to particulates, in symbiotic relationship with amoebae, and by hydrostatic flow). Organisms percolate through soil layers to contaminate groundwater. Viruses are particularly mobile because of their extremely small size and their relatively long survival times in the environment. Because mobility is associated with the hydrodynamics of distribution systems, presence of biofilms, presence of particulates, and opportunity for symbiotic relationships, it is considered together with persistence for scoring purposes.

Exhibit D5 - Persistence-Mobility Scoring Protocol

Stability in Non-Turbid Water[1]	**Score[2]**
Usually dies rapidly in water (days)	2
Stability uncertain, no amplification	3
Stable for weeks to months, no amplification[3]	4
Stable for weeks, months, or years, with amplification or protection from symbiotic relationships[4]	5

[1]Non-turbid water is defined as ground water or filtered surface water.

[2]Select the single highest score for each organism.

[3]Development of endospores, cysts or oocysts

[4]Capsule or slime production, protection by amoebae, autotrophic metabolism

Magnitude

NRC defines magnitude as "the concentration or expected concentration of a contaminant relative to a level that causes a perceived health effect" (NRC 2001). For characterizing the attribute of magnitude, ideally two data elements are needed: the concentration of a contaminant in water, and the concentration associated with an adverse health effect. NRC recommended the use of a median water concentration in combination with a measure of potency, if available.

Magnitude, in a microbiological context, implies delivery (persistence-mobility) of an infective dose (potency) to the customer's tap with resulting illness. The Work Group proposes to score magnitude according to the number and frequency of waterborne disease outbreaks reported in the U. S. and around the world, pathogen distribution, and biological properties determining pathogen distribution. A scoring table is shown in Exhibit D6. This algorithm has not been adequately evaluated using a test data set.

Exhibit D6 - Magnitude Scoring Protocol

Category[1,2]	Score[3]
Has caused numerous recently documented WBDOs in the U.S. or other developed country	6
Rarely causes documented WBDOs in the U.S. or other developed country	5
Has not caused documented WBDO in the U.S. or other developing countries, but has caused documented foodborne outbreaks	4
Has caused numerous recent documented WBDO in developing countries, but its biological properties would mitigate against causing WBDO in the U.S.	3
Rarely causes documented WBDOs in developing countries, but its biological properties would mitigate against causing WBDO in the U.S.	2
Has never caused WBDOs in any country, or its biological properties mitigate against causing WBDOs in the U.S.	0

[1]U.S. is defined as the 50 states and territories.

[2]Waterborne disease outbreaks (WBDOs) associated with drinking water, fresh water used for recreation, and outbreaks associated with hot tubs, swimming pools, etc. where makeup water is drawn from potable water sources.

[3]Select the single highest score for each organism.

Work Group discussion on organisms known to cause waterborne disease outbreaks concluded that such pathogens could be placed directly on the CCL (known pathogens causing WBDO are already on the CCL, with exclusions based upon treatment efficacy), and that attribute scoring would not play a significant role in moving them from the CCL Universe to the CCL. While priority would be given to domestic outbreaks, organisms causing WBDO in other countries would be evaluated for their public health significance in the U.S.

Microbial Data Elements for Attribute Scoring

By obtaining information on the attributes of each of these elements for each known or prospective pathogen, it is possible to assess the relative risk and prioritize pathogens according to their occurrence and health effects.

A. Elements Considered in Pathogen Occurrence

Spatial distribution (clumping, particle-association, clustering)

Concentrations in environmental vehicles and foods
Seasonality and climatic effects
Temporal distribution, duration, and frequency
Niche (potential to multiply or survive in specific media)
Amplification, die-off, persistence
Indicators/surrogates predictive of pathogens

B. Elements Considered in Exposure Analysis
Identification of water and other media
Unit of exposure
Temporal nature of exposure (single or multiple; intervals)
Route of exposure and transmission potential
Demographic of exposed population
Size of exposed population
Behavior of exposed population

C. Elements Considered in Pathogen Characterization
Virulence and pathogenicity of the microorganism
Pathological characteristics and diseases caused
Survival and multiplication of the microorganism
Resistance to environmental control measures
Host specificity
Infection mechanism and route; portal of entry
Potential for secondary spread
Taxonomy and strain variation

D. Elements Considered in Host Characterization
Demographics of the exposed population (age, density, etc.)
Immune status
Pregnancy
Concurrent illness or infirmity
Nutritional status
Genetic background
Behavioral and social factors

E. Elements Considered in Health Effects
Morbidity, mortality, sequelae of illness
Severity of illness
Duration of illness
Chronic or recurrent
Potential for secondary spread

Appendix E

Prototype Classification Methods/ Results of Pilot Demonstration

Following is a brief description of different model classes discussed by the NDWAC Work Group that might be used for a prototype based classification approach. Four of these were part of a demonstration exercise.

Linear Discriminant Analysis

Linear discriminant analysis can be thought of as a special case of linear regression analysis. In linear regression analysis a response variable is described as a linear function of one or more predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

where Y is the response variable, β_0 is an intercept term, X_1 is a predictor variable, β_1 is the slope parameter and ε is an error term which acknowledges that Y may not be perfectly predictable by knowing X_1. Linear regression analysis refers to the procedure in which optimal values for β_0 and β_1 are estimated, given a data set that consists of observations for Y and X_1. A linear regression model predicts the value of the response variable, given the values of a set of predictor variables, while the linear discriminant model predicts which category the response variable is likely to belong to, given the values of the predictor variables.

Logistic Regression

Generalized linear models are a class of models that have a linear model as their basis, but the response variable is now a function of the linear model:

$$Y = f(\beta_0 + \beta_1 X_1) + \varepsilon$$

where f represents some mathematical function. Logistic regression is a special case of the generalized linear model, in which the response variable is categorical, with two categories.

Artificial Neural Networks (ANN)

Artificial Neural Networks consist of a base node and several "hidden" nodes. The base node is typically a logistic model and the hidden nodes further refine the results of the logistic model. The number of hidden nodes included in the model is determined by evaluating the improvement in model prediction with each additional node. Generally, adding hidden nodes requires a large data set to ensure that the additional model structure is not just capturing the idiosyncrasies of a small sample.

Classification and Regression Trees (CART)

A CART Model is analogous to a dichotomous key, used in biology to determine an animal or plant species based on observable characteristics of the organism. The value or category of the response variable is determined by evaluating the way in which the response variable is related to the predictor variables as a set of "if – then" statements (i.e. if X_1 is greater than a value, and X_2 is less than some value then Y belongs in a particular category). This model can be visually

depicted in a diagram that looks like a branching tree. Recent implementations of CART include the ability to accommodate some missing data not only in the training data set, but also among new observations.

Multivariate Adaptive Regression Splines (MARS)

Generalized Additive Models are a class of nonlinear models in which the relationship between the response and predictor variables is not pre-specified by a particular mathematical function. Instead the relationship is developed from the observed data. The procedure divides the data into regions, similar to a moving window, and estimates a smooth nonlinear relationship among the data within each region. Multivariate Adaptive Regression Splines are a form of generalized additive model that allow the inclusion of interactions among the predictor variables similar to interactions that might be included in various forms of linear regression or analysis of variance models.

Model Pilot Demonstration

An exercise by the technical support team for the Work Group demonstrated how these models work. First, 46 contaminants were selected to comprise a training set. Each contaminant was scored for five attributes (severity, potency, prevalence, magnitude, and persistence/mobility) according to a draft scoring protocol. Next, each contaminant was assigned a decision: either "list" or "do not list." Finally, the data were used to inform the Logistic Regression, ANN, CART, and MARS models. This was done in two steps. First, the structure of a "best-fit" model was determined for each of the four model classes. In the second step, the models from each class were compared to show how each class of model performed with an example data set.

Model Selection within each of the four model classes

"Over-fitting" is a concern when selecting a best-fit model. Any of these four model types could be made to fit a particular data set very well by making the model more complex (this usually means estimating more model parameters). However, the addition of model complexity can come at the cost of a loss of generality; the added complexity may capture the idiosyncrasies of the specific training data set, and may not be representative of the broader processes that generate the data. Several methods were used as guidance to avoid over-fitting, depending on the specific model being fit. Cross-validation is a technique in which the data set is repeatedly, randomly sub-divided and a model is fit to a subset of the data, then used to predict the complementary subset that was "left-out" of the fitting process. A second method is the Bayesian Information Criterion a combined measure of a model's predictive capability and complexity. For the logistic regression, standard classical methods of assessing "statistical significance" were used as guidance for the number of predictor variables that should be included in the model.

Comparison of the pilot example models

Each of the four model classes was assessed using a "ten-fold" cross validation procedure. The data set was randomly divided into 10 roughly equal sized groups. The four models were fit 10 times using the selected models, each time setting aside one group of the data. The fitted models

were than applied to the left-out data set and the predictive misclassification rates for each model were recorded. The exercise showed that the four model classes could be compared by misclassification rates of a training data set.

Lessons Learned

Lessons learned were limited by the use of a small set of contaminants scored by a draft (not final) attribute scoring protocols. Specific findings, such as estimated classification error rates, could be inaccurate predictors of performance in the future, when the models are informed by a complete training set, with well-justified "list"/"do not list" decisions, and final scoring protocols.

The major lessons learned in the exercise were:

- The major cost of running any model is development of the training data set (i.e., developing attribute scores, selecting the training data set, and scoring the training data set contaminants). Once the training data set is available, computer processing (training, cross-validation, diagnostics) is relatively quick . Additional time and resources may be required to modify the training data set based on the results of the diagnostic exercises.
- All four models can classify contaminants based on complete training data sets.
    - All models could deal with integer attribute scores.
    - All models could deal with raw attribute data
    - The CART model could deal with missing data/scores. Other models could not, and therefore had smaller training data sets. This may also require consideration in developing the attribute scoring protocols and selecting training set contaminants.
- All four models provided diagnostic information:
    - Estimated classification error rates can indicate whether the training set is of adequate size.
    - Estimated classification error rates may allow rejection of one or more models.
- Comparing results across models provides information on the training set contaminants:
    - All four models correctly classify most contaminants.
    - Four contaminants were misclassified by all four models, suggesting that the five attribute protocols scores may need to be refined to account for characteristics of these contaminants.