

Example CCL Universe Data Set: Progress and Recommendations

Presentation to NDWAC CCL Work Group
Washington, DC

July 15, 2003

Review Example CCL Universe Data Set/Analysis

- Purpose/Intent of Example CCL Universe Data Set
- Development of the Data Set
- Findings
 - What insights do we have about this data set?
 - Availability
 - Quality
 - Other
- Will this data set be sufficient to test the proposed “gate” screening?
- What can we say about the use of this data set for attribute/scoring and classification?

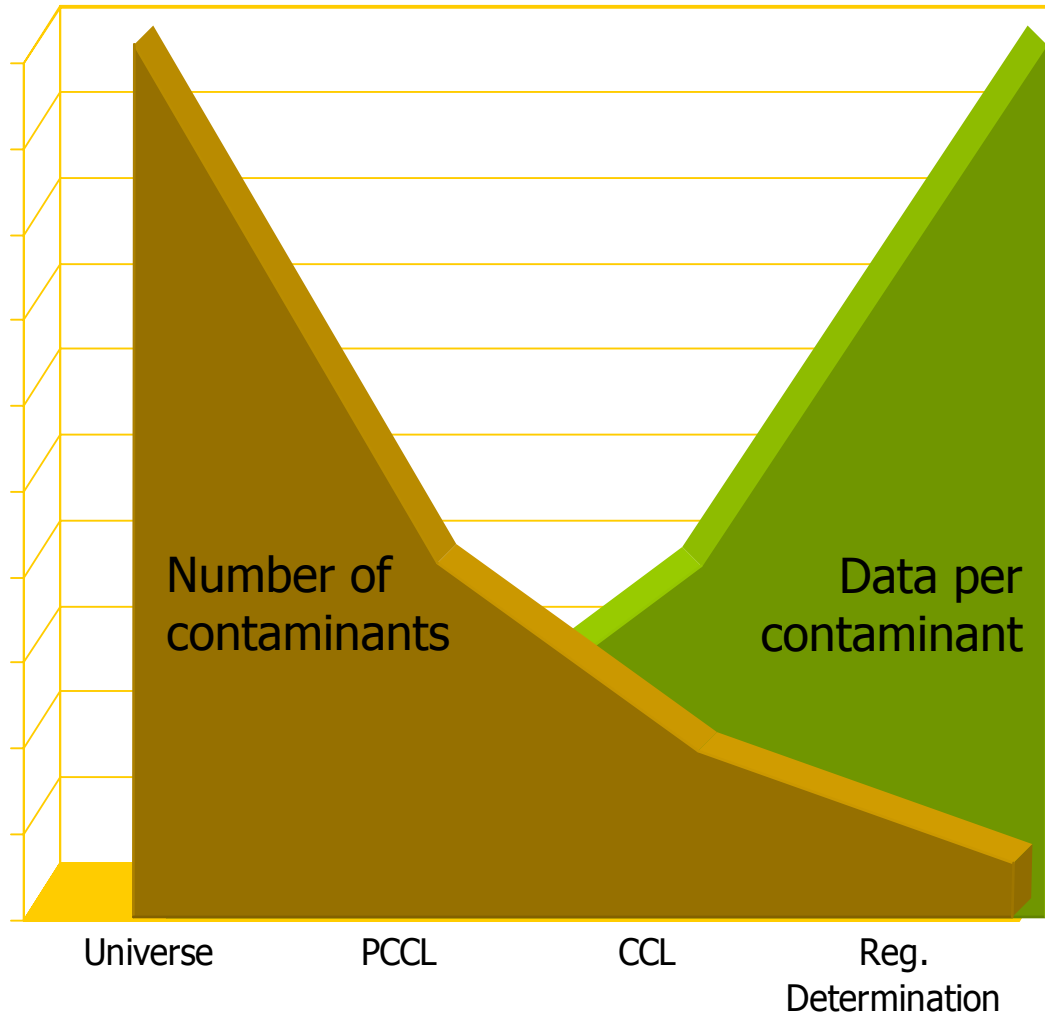
Overview: Purpose

- The NRC has recommended a CCL process requiring data for the following steps:
 - identify the “universe” of potential drinking water contaminants;
 - select a Preliminary CCL from the “universe” by a screening process;
 - develop a training set of compounds to train a prototype algorithm to prioritize drinking water contaminants;
 - apply a prototype algorithm to classify the PCCL into a CCL; and
 - conduct an expert review of the algorithm results.

Overview: Purpose of the Example CCL Universe Data Set

- Gather some data to support NDWAC's process
- Test the proposed approaches:
 - Building the CCL Universe
 - "Gates" screening approach
 - Attribute Scoring
 - Use of a prototype algorithm to classify the PCCL into a CCL
- Gain insights about available data on contaminants, identify data elements, issues and gaps.

Overview: Structure



As the process moves from "universe" to CCL,

the data requirements of each contaminant will grow,

while the pool of contaminants will shrink.

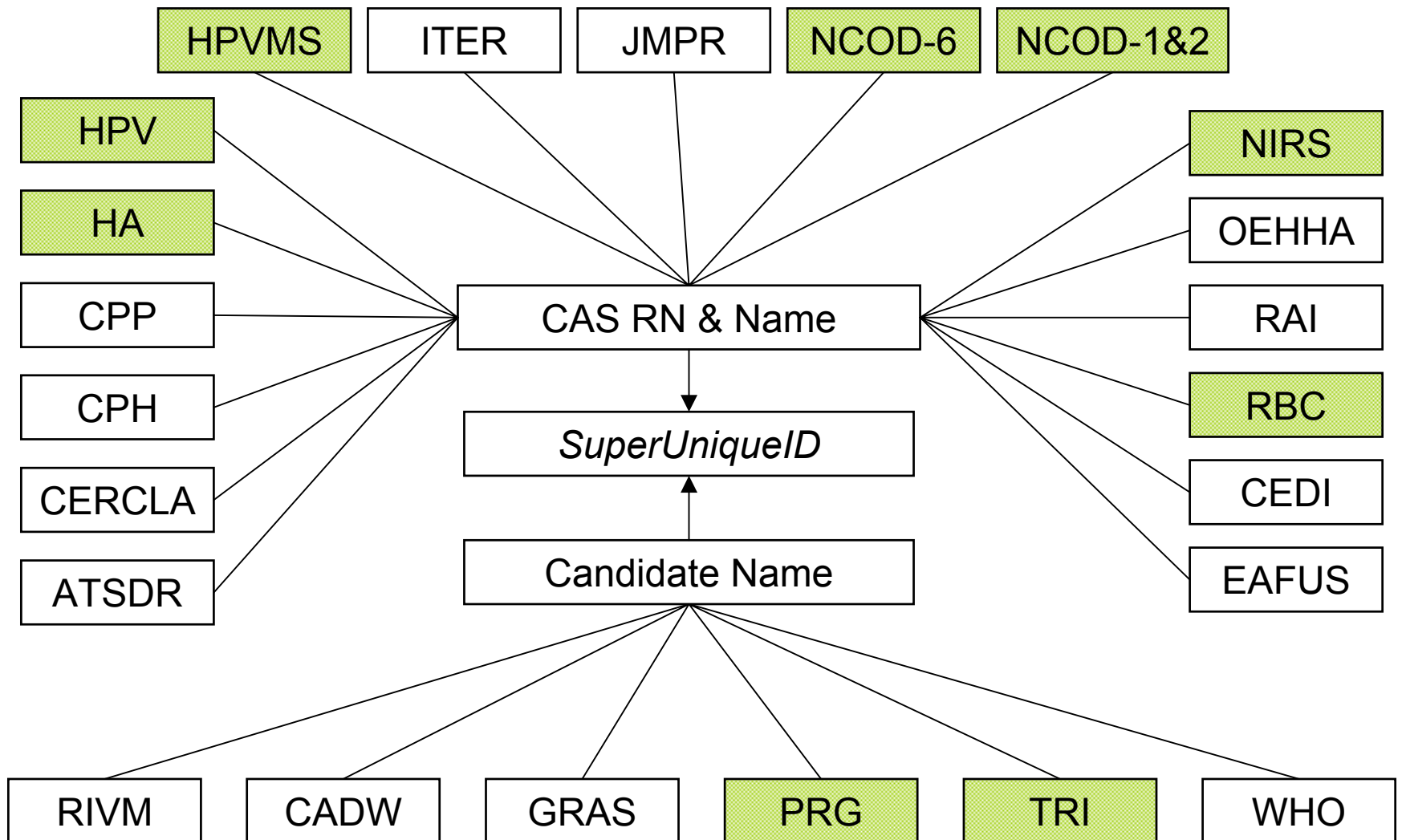
Data Retrieval

- Began with list of more than 200 data sources
- Chose 23 sources based on the following criteria:
 - high-quality data
 - easily accessible electronically
 - contained information relevant to building the CCL Universe (i.e., occurrence data, etc.)
- Downloaded tables from 23 sources
 - a total of 87 tables
- All tables formatted as Microsoft Excel files
- Performed QA/QC review on Excel files to ensure accurate reproduction of data from original source

Linking Data

- ❑ To link data across tables, contaminants need a unique identifier
- ❑ Unique identifier was first assigned to a CAS number
- ❑ If the CAS number was not listed or available, a unique identifying number was assigned to the contaminant by name
- ❑ A table was created to list each contaminant with its corresponding identifier
- ❑ Not a perfect process

Data Sources



Classifying Data Sources

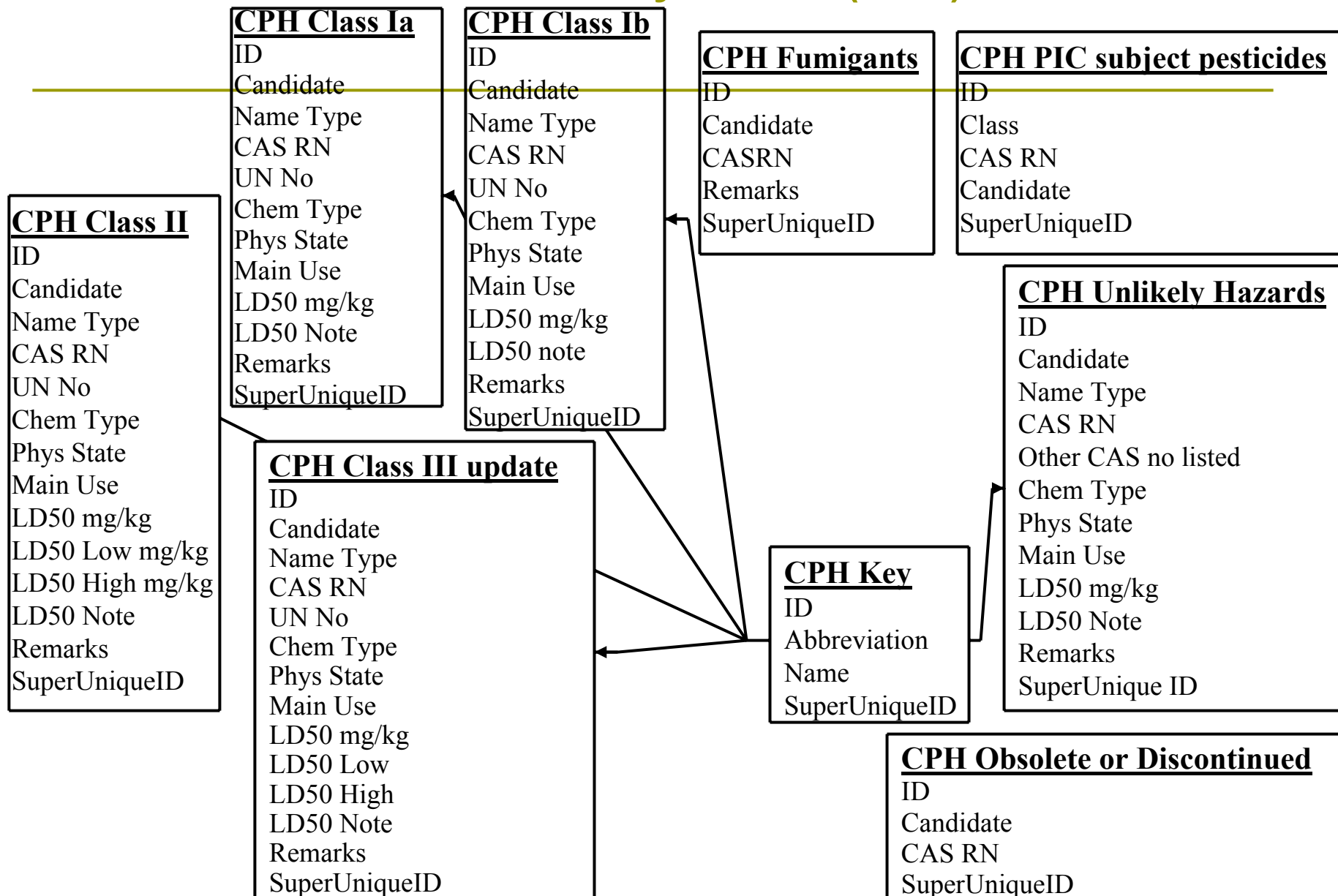
- The 23 data sources were classified according to the type(s) of data and/or information available in each
 - First it was determined what data elements were in each data sources
 - Then the data elements were classified, and the data sources were classified according to what kinds of data elements were available in the source
 - Data sources could be classified into one or more categories: HE Data, HE Info, Occ Data, Occ Info, or No Data or Info

Types of Data and/or Information in the Data Sources

Example Data Set Sources With

HE Data	HE Info	Occ Data	Occ Info	No Data or Info
ATSDR MRLs	ATSDR MRLs	NIRS	CADW	ATSDR CERCLA
CADW	CADW	NCOD Round 1 and 2	PRGs	HPV
CPH	CEDI	NCOD Six Year Review	RAIS	HPV Master Summary Table
CPP	HA		RBCs	GRAS
HA	ITER		TRI	EAFUS
ITER	OEHHA			
WHO	PRGs			
OEHHA	RBCs			
PRGs	RVM			
RBCs				
RVM				
WHO				
CEDI				
JMPR				

The World Health Organization (WHO): Classification of Pesticides by Hazard (CPH)



Classifying Data Elements

Type of Data	NRC 2001 definition	NDWAC CCL CP Workgroups draft definition
measurements in natural waters	demonstrated occurrence	occurrence data
other information about contaminant occurrence	potential occurrence	occurrence information
health effects measurements from toxicological studies that relate to exposures via drinking water	demonstrated health effects	health effects data
other information about adverse health effects	potential health effects	health effects information

Data Mapping

- All data elements (848 total) were mapped to 5 categories:
 - Health Effects Data (108)
 - Occurrence Data (49)
 - Health Effects Information (67)
 - Occurrence Information (69)
 - Other (555)
- Classification of data elements to these categories allows summarizing the candidates by each of the screening 'gates'

Data Set Findings: Overview

□ Overview

- 1,500 records (out of 30,000, or 5 percent) had a chemical name but no CAS RN
 - Assigned CAS RN to 721 (~50 percent) of these
- 11,000 records had multiple chemical names assigned a single CAS RN
- 108 records had multiple CAS RNs assigned to one chemical name

Example CCL Universe Data Set Summary Statistics

Database Subset	Data Sources	Chemicals	Data Elements	
		Contaminants	Available Data or Info Only ¹	Total (data, info, other)
Entire Example Data Set	23	10,360	293	848
All Chemicals with Health Effects Data or Information AND Occurrence Data or Information	18	774	293	848
Chemicals with Health Effects Data AND Occurrence Data	16	62	157	<848

Data Set Summary Statistics Compared With CCL

Database Subset	Data Sources	Contaminants	Chemicals	
			Out of 262 on Draft CCL	Out of 50 on Final CCL
Entire Example Data Set	23	10,360	244	45
All Chemicals with Health Effects Data or Information AND Occurrence Data or Information	18	774	188	40
Chemicals with Health Effects data AND Occurrence data	16	62	35	18

Screening Gates

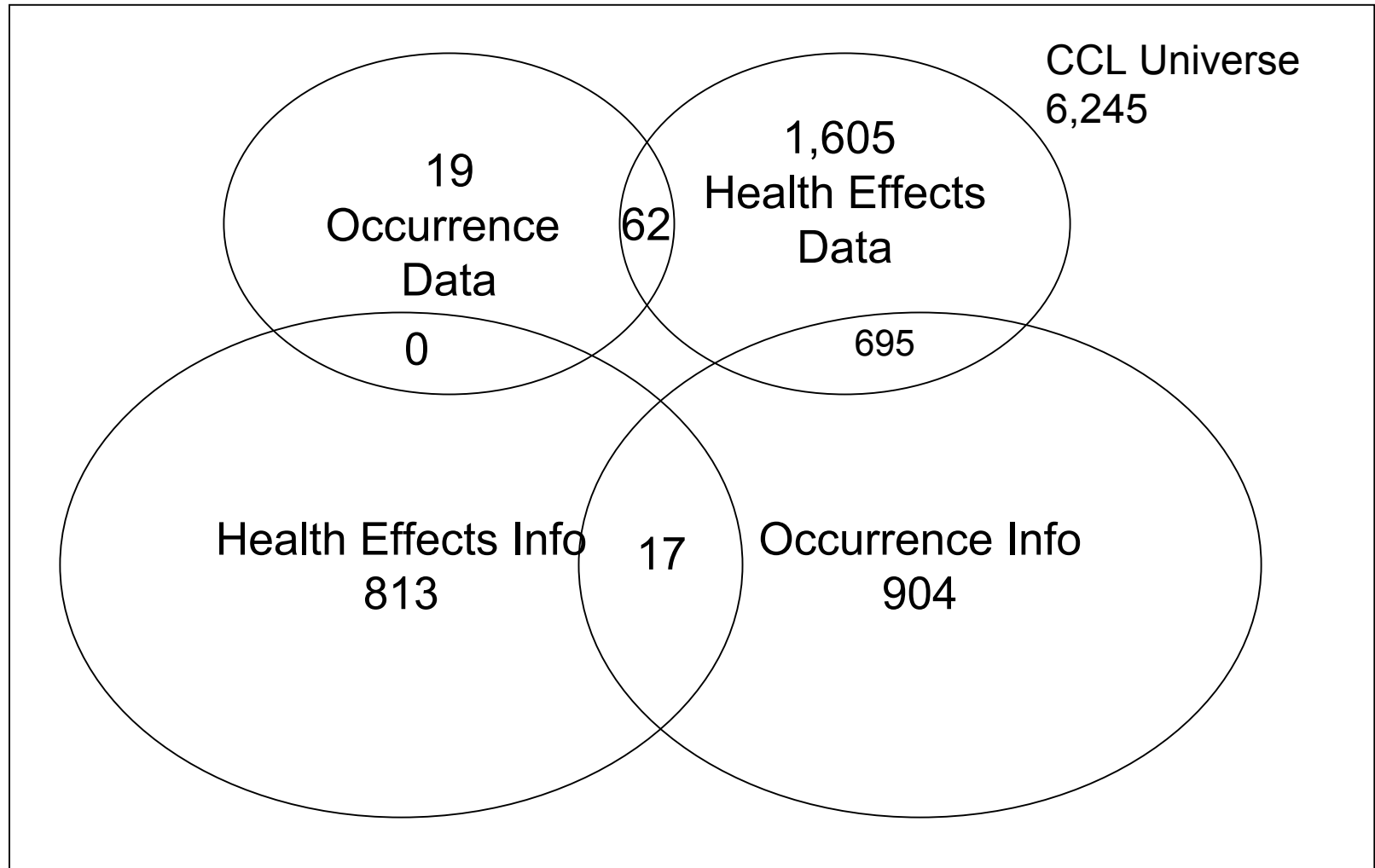
- “Gates” will be used to screen contaminants from the Universe to the PCCL based on varying criteria:
 - Gate 1: Contaminants have Health Effects Data and Occurrence Data
 - Gate 2: Contaminants have Health Effects Information and Occurrence Data
 - Gate 3: Contaminants have Health Effects Data and Occurrence Information
 - Gate 4: Contaminants have Health Effects Information and Occurrence Information
 - Gate 5: Expert Judgement

Number of Chemicals with Data/Information

- Number of chemicals in various categories related to data/information elements in the Example CCL Universe Data Set

Category or Criteria of Data/Information Available; Gate Screening Criteria	Number of Chemicals in Example Data Set Meeting the Criteria	Percentage of Chemicals in Example Data Set Meeting the Criteria
Health Effects DATA AND Occurrence DATA; Gate 1	62	0.6%
Health Effects Information AND Occurrence DATA; Gate 2	0	0.0%
Health Effects DATA AND Occurrence Information; Gate 3	695	6.7%
Health Effects Information AND Occurrence Information; Gate 4	17	0.2%
Subtotal of Candidates Through Gates 1-4:	774	7.5%

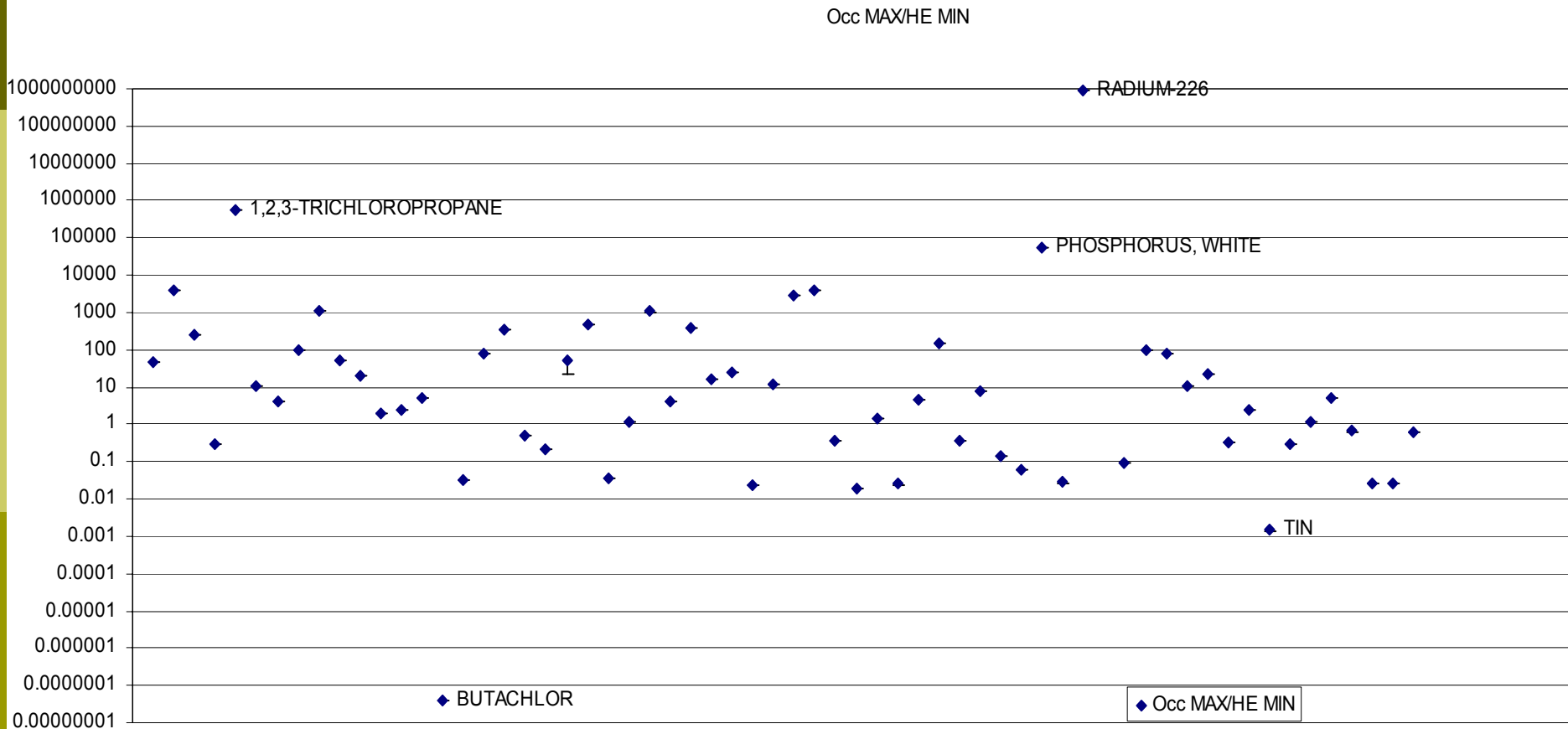
The Example CCL Data Set and the NRC'S Venn Diagram



Quantitative “GATE 1” Screening Results

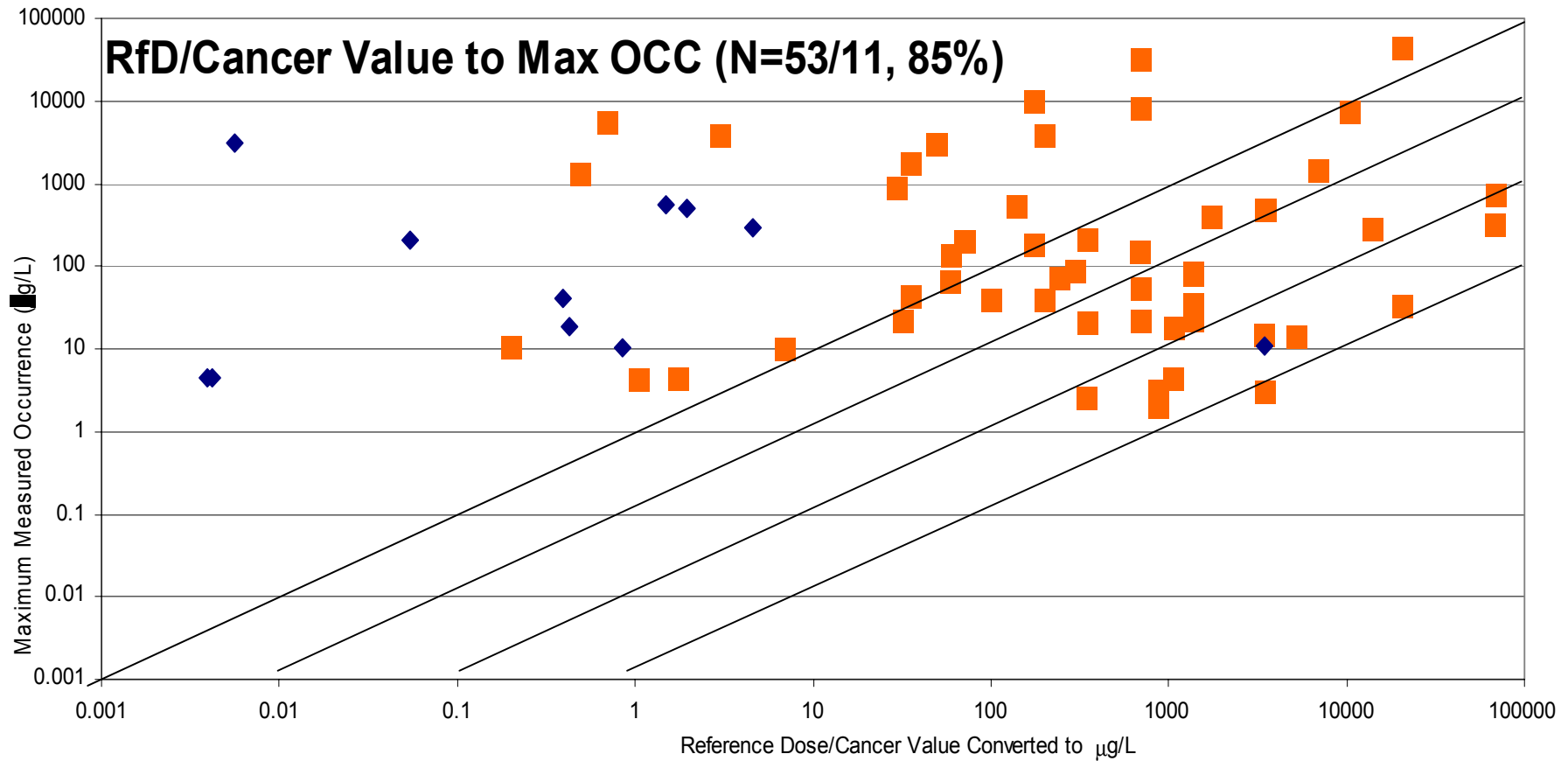


Ratio of Maximum Occurrence to Minimum Health Effect Level

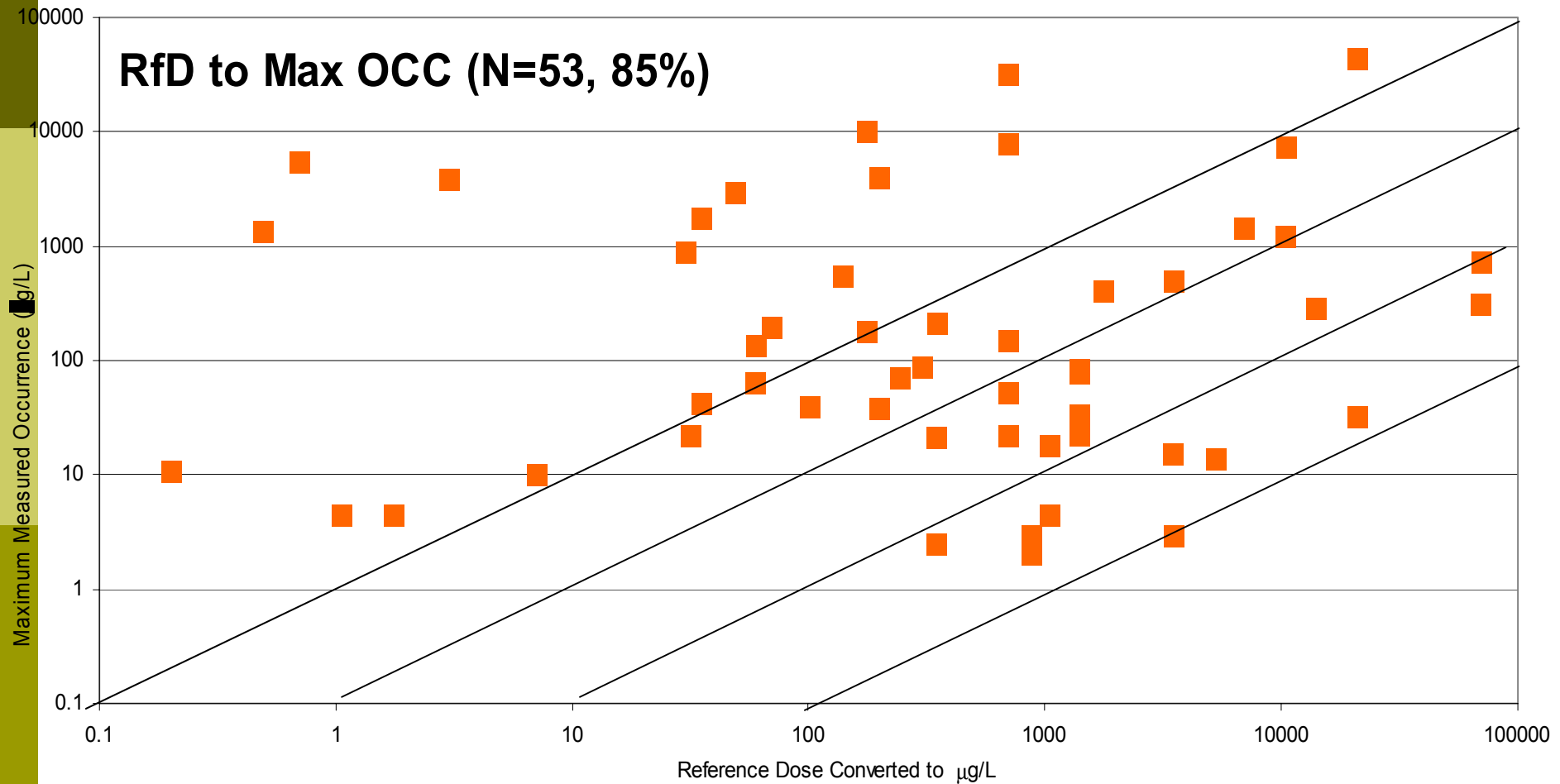


- **Most GATE 1 chemical occurrence is within two orders of magnitude (1 to 100) of lowest health effect level with exceptions noted.**

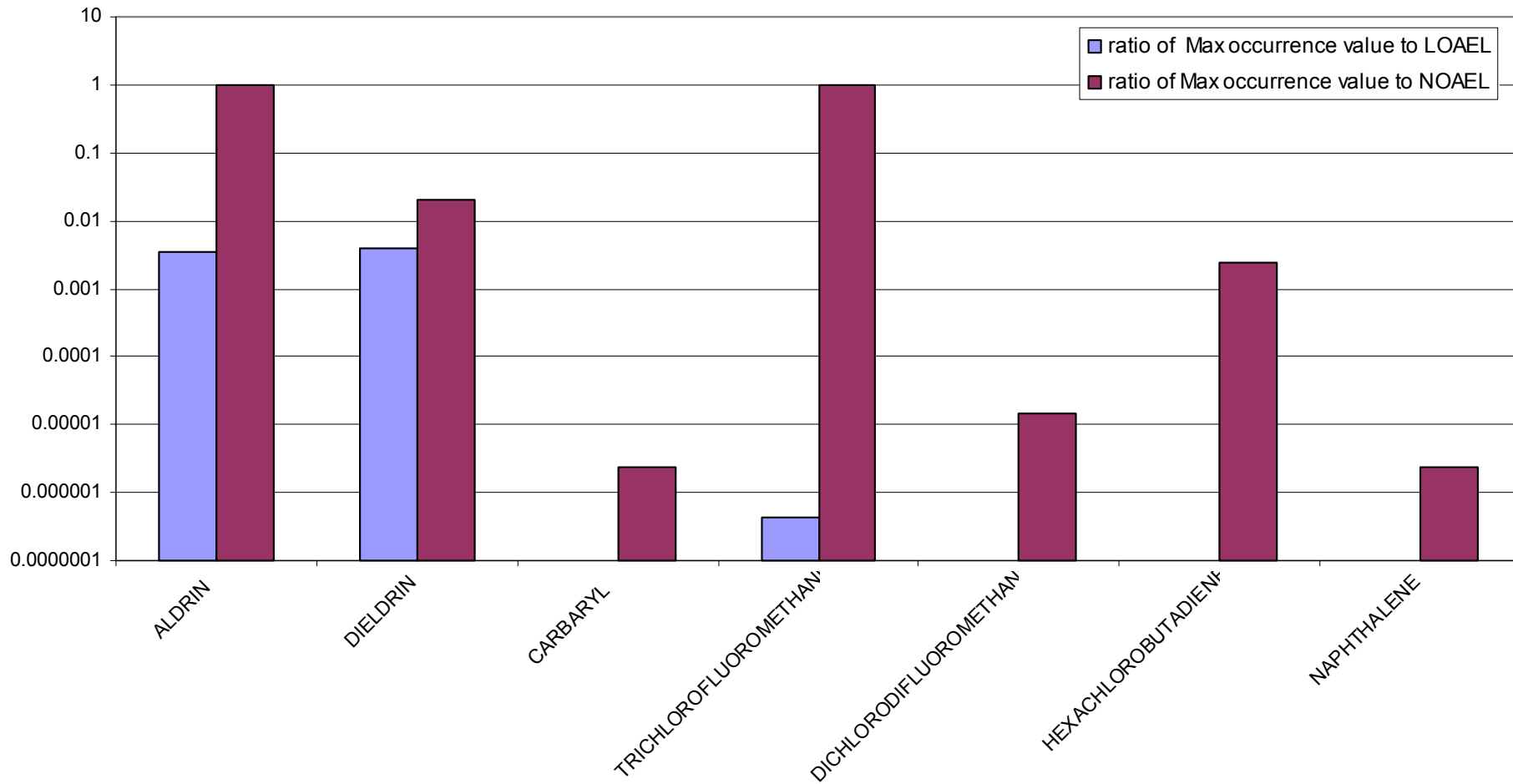
Ratio of Max Occurrence to Reference Dose: screen out ~ 60% of Gate 1 if 1:1



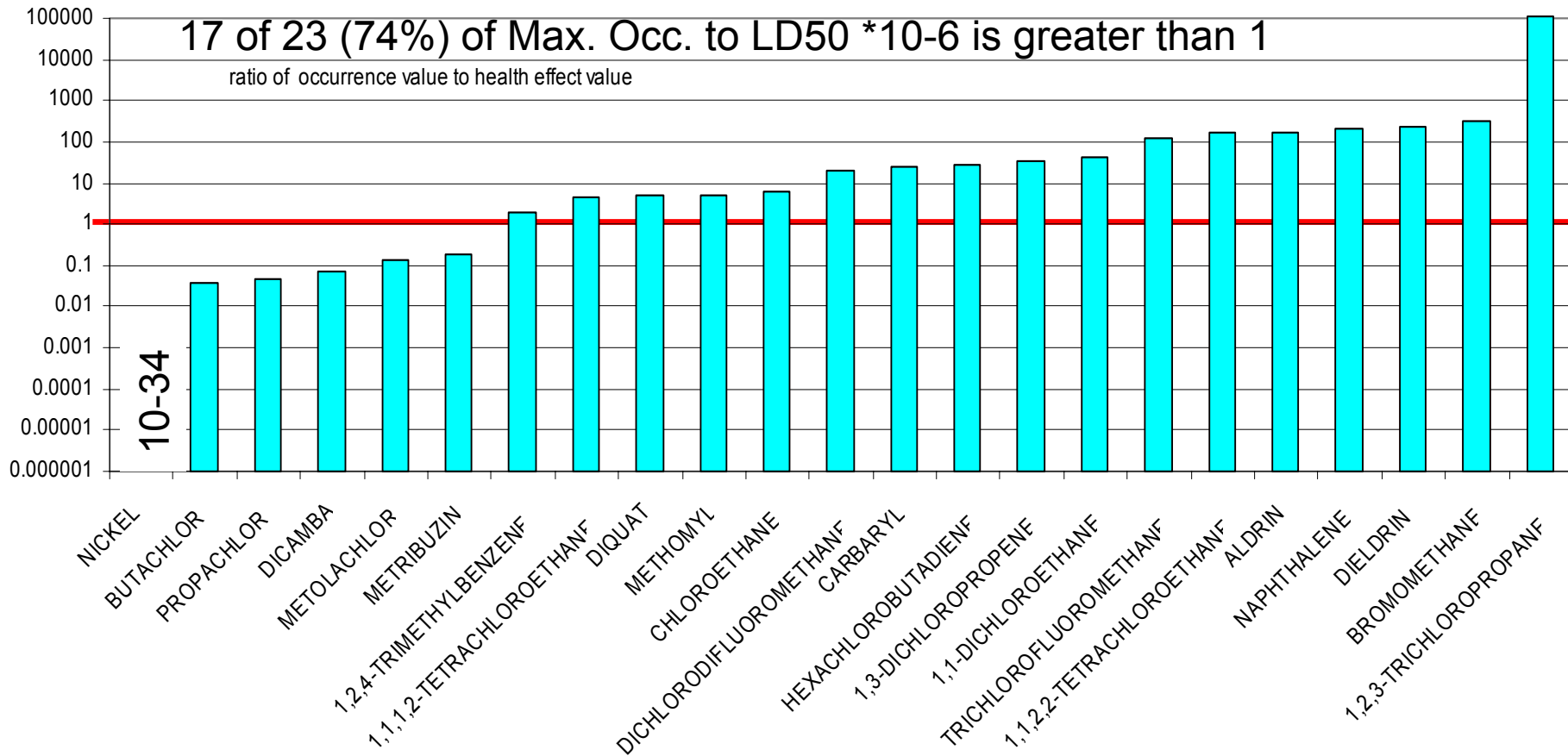
Ratio of Max Occurrence to Reference Dose : screen out ~ 60% of Gate 1 if 1:1



Max. Occurrence to NOAEL/LOAEL



Ratio of Max Occ to LD50(*10-6)



Summary of Gate 1 Quantitative Screening

Gate 1 Screening Criteria Ratio (Occ:HE)	Max Occurrence/ Lowest HE	Max Occurrence/ Lowest RfD	Max Occurrence/High est NOAEL/ Lowest LOAEL	Max Occurrence/ Lowest LD50
1:1 or greater	65%	42%	0%	74%
1:0.1	81%	64%	0%	83%
1:0.01	97%	83%	14%	96%
1:0.001	98%	98%	43%	96%
1:0.0001	98%	100%	86%	96%
>1:0.00001	100%	100%	100%	100%
N	62	53	7	23

Insights

- Data Availability is better for health effects than for water occurrence. Lack of contaminants for Gate 2 an indicator that water occurrence data may be a limiting factor
- Used high quality sources – few additional sources will be of similar quality – need data quality measures for additional sources

Relationships among Data Sources:

- Use of prioritized lists informative for screening, but may want background data for attribute scoring (iterative process)
- E.g. ATSDR data sources
 - HazDat Data
 - CERCLA Priority List
 - Toxicity Profiles
 - Minimal Risk Levels
- Another Example: ITER
 - Derived endpoints e.g. Reference Dose
 - Can get background data from RAIS, EPA, etc.
 - ITER useful for screening, may want additional data for attribute scoring

How Representative are the 23?

- 10% of reviewed data sources included
- 10,360 contaminants (~10% of anticipated CCL Universe)
- 750 (about 7.5%) get through “qualitative” “gate” screening
- 90% of CCL 1998 chemicals included
- 80% of CCL 1998 through “qualitative” “gates” (40 of 50)

Is the Example Data Set Adequate for Screening?

- ❑ Sufficient data are available for screening
- ❑ Gate 1 screening demonstrates adequacy for decision making
 - Qualitative screening is simple, and effective for this example
 - With larger numbers (e.g. Actual CCL Universe), quantitative screening may be needed to limit size of PCCL
 - Other gates probably require a quantitative screening
- ❑ Would increase representativeness to include additional occurrence data (e.g. NAWQA, NREC) in Example CCL Data Set

- ❑ **RECOMMENDATION:**
 - **Use Qualitative Screening for NDWAC Example PCCL**
 - **Add more Occurrence Data sources**

Findings and Recommendations

- ❑ Using tabular sources maximizes the number of elements for screening
- ❑ Have many derived health effect endpoints and redundancies
- ❑ Have included national drinking water surveys
- ❑ Could identify additional Gate 1 contaminants with other natural water surveys (e.g. NAWQA)
- ❑ Additional data sources would be helpful for attributes scoring

What can we say about the use of this data set for attribute/scoring and classification?

- ❑ Additional sources may provide needed elements and perhaps should be added to data set

- ❑ Additional elements may be helpful for attribute scoring.
 - For example, critical effect information to score severity from derived “potency” endpoints may not be available.

- ❑ Additional sources may require manual manipulation and judgment for use in PCCL to CCL
 - E.g., additional occurrence data is in raw data format, so statistics must be calculated