



The Environmental Health Language Collaborative

*Harmonizing Data, Connecting Knowledge, Improving Health*

---

# **2024 Society of Toxicology (SOT) Annual Conference: EHLC Community Presentations**

**Salt Lake City, Utah**

**March 10 - 14, 2024**



## SOT Presentations by EHLC Community Members

This document provides an overview of 2024 SOT presentations from Environmental Health Language Collaborative (EHLC) community members.

Presentation Order	Presentation Title	Presenter, Organization
1	<i>Improving the findability of toxicology studies for decision making in the era of data sharing</i>	Michelle Angrish, PhD, U.S. EPA <a href="mailto:angrish.michelle@epa.gov">angrish.michelle@epa.gov</a>
2	<i>How best to combine data from multiple independent studies?</i>	Jeanette A Stingone, PhD, MPH, Columbia University <a href="mailto:js5406@cumc.columbia.edu">js5406@cumc.columbia.edu</a>
3	<i>Digitizing Relationships between Exposures, Biomarkers, and Clinical Outcomes (In the era of AI and exposomics)</i>	Chirag Patel, PhD, Harvard Medical School <a href="mailto:chirag_patel@hms.harvard.edu">chirag_patel@hms.harvard.edu</a>
4	<i>Challenges and opportunities to improve communication about exposure and risk for collaboration and information exchange</i>	Elke Jensen, PhD, Dow Chemical Company <a href="mailto:elke.jensen@dow.com">elke.jensen@dow.com</a>
5*	<i>Overcoming Barriers to More Scalable Environmental Health Science Research via Harmonized Language*</i>	Andrew Rooney, PhD, NIEHS* <a href="mailto:andrew.rooney@nih.gov">andrew.rooney@nih.gov</a> Steve Edwards, PhD, U.S. EPA <a href="mailto:edwards.stephen@epa.gov">edwards.stephen@epa.gov</a>

*\*presentation materials not included*

## Presentation 1

Presentation Order	Presentation Title	Presenter, Organization
1	<i>Improving the findability of toxicology studies for decision making in the era of data sharing</i>	Michelle Angrish, PhD, U.S. EPA <a href="mailto:angrish.michelle@epa.gov">angrish.michelle@epa.gov</a>

# Improving the findability of toxicology studies for decision making in the era of data sharing

Michelle Angrish  
U.S. EPA



*The author declares no conflict of interest.*

*The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the US EPA.*

# Today's Goals

- Understand the challenges in reusing research.
- Learn how structured data helps to reuse research – and help you!
- Starting practices for making your research findable and therefore, reusable!
- Perspectives from a chemical assessment practitioner with examples:
  - Finding information
  - Bringing structure to unstructured data
  - Standardizing data

# Definitions

- Annotation – labeled text with a tag that indicates the type of thing or concept the text represents
- Interoperable – the ability for information to flow to/from tools
- Controlled vocabulary – non-redundant list of preferred terms
- Standardized data extraction format – template for formatting extracted data
- Template – organization framework for extracted data
- Schema – organization framework for templates and metadata

# Who are we?

## About the Chemical and Pollutant Assessment Division (CPAD)

The Center for Public Health and Environmental Assessment (CPHEA) provides the science needed to understand the complex interrelationship between people and nature in support of assessments and policy to protect human health and ecological integrity. Within CPHEA, sits the Chemical and Pollutant Assessment Division.

### On This Page:

[What We Do](#)  
[Management](#)  
[Branches/Locations](#)

### Related Information

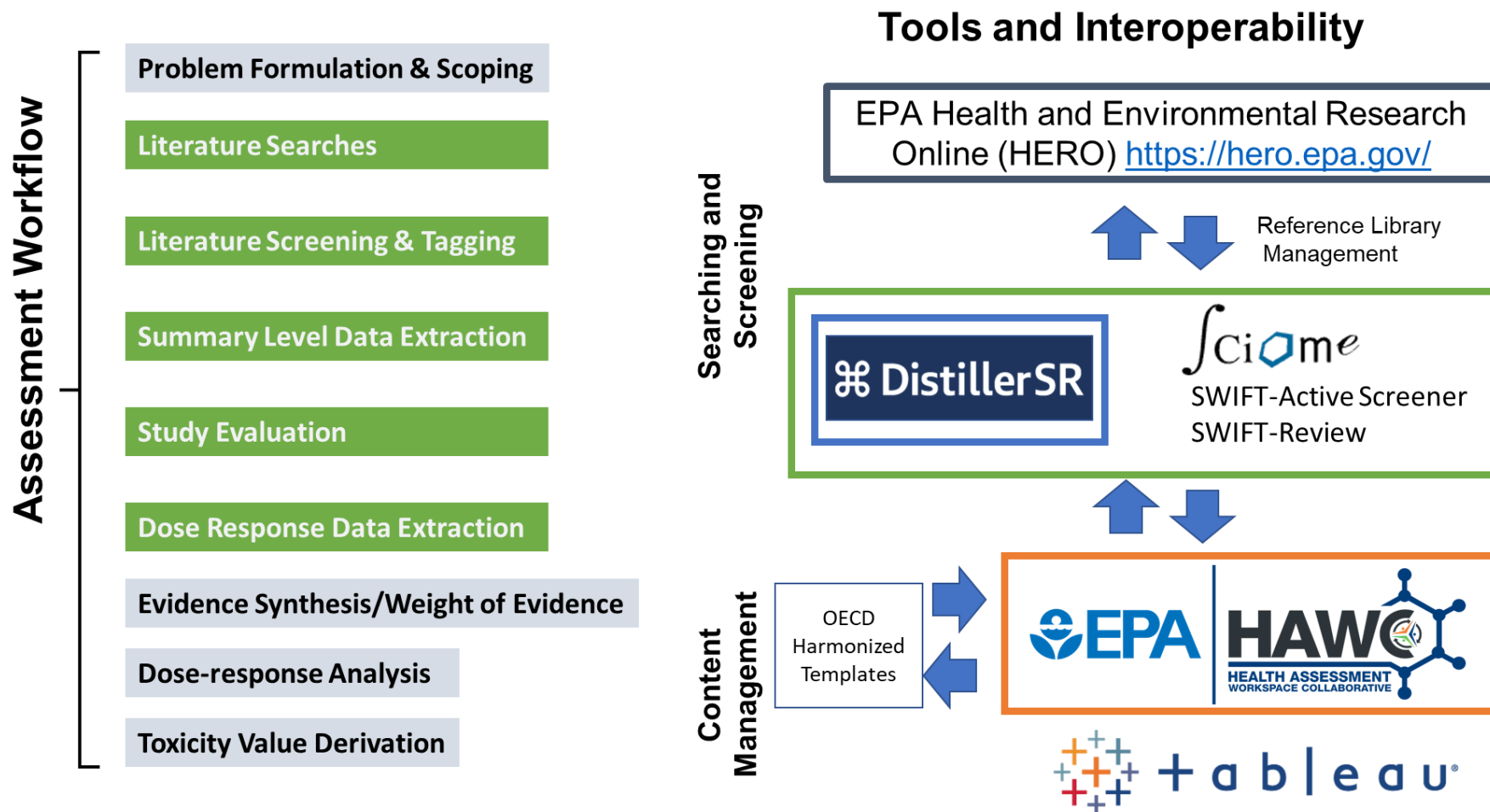
- [About CPHEA](#)
- [Organization Chart for CPHEA](#)
- [About the Office of Research](#)

## EPA's Chemical Pollutant Assessment Division (CPAD)

*We are data consumers.*

CPAD scientists develop a range of fit-for-purpose human health risk assessment products based on the evaluation, synthesis, and analysis of the most up-to-date scientific information. Products include the [Integrated Risk Information System](#) (IRIS) and [Provisionally Peer Reviewed Toxicity Values](#) (PPRTV) assessments. These products are developed through interactions with EPA's program and regional offices, other agencies, the scientific community, industry, policy-makers, and the public. Once finalized, they serve as a major scientific component supporting EPA's regulations, advisories, policies, enforcement, and remedial action decisions. CPAD also conducts cutting-edge research to develop innovative human health risk assessment methods (e.g., systematic review) that facilitate careful evaluation of scientific evidence, as well as tools and models (e.g., [benchmark dose modeling software](#)).

# How do we do this?



We use a workflow that includes:

- interoperable tools
- web accessible applications,
- standardized data reporting frameworks
- machine readable data

to find and use the data that generated by data producers.

# First we have to find your research and we can only search things that are findable

## Data consumers have to know

- what we are looking for and where to find it
- how to search an indexing service
- what services and labels data producers are using

## Data producers have to know

- what information data consumers are looking for
- how to label information so that it can be identified

## Journal

## Title

## Author

## Abstract

> Arch Toxicol. 2016 Jan;90(1):217-27. doi: 10.1007/s00204-014-1391-7. Epub 2014 Nov 5.

## Interaction of perfluoroalkyl acids with human liver fatty acid-binding protein

Nan Sheng <sup>1</sup>, Juan Li <sup>2</sup>, Hui Liu <sup>1</sup>, Aiqian Zhang <sup>3</sup>, Jiayin Dai <sup>4</sup>

Affiliations + expand

PMID: 25370009 DOI: 10.1007/s00204-014-1391-7

## Abstract

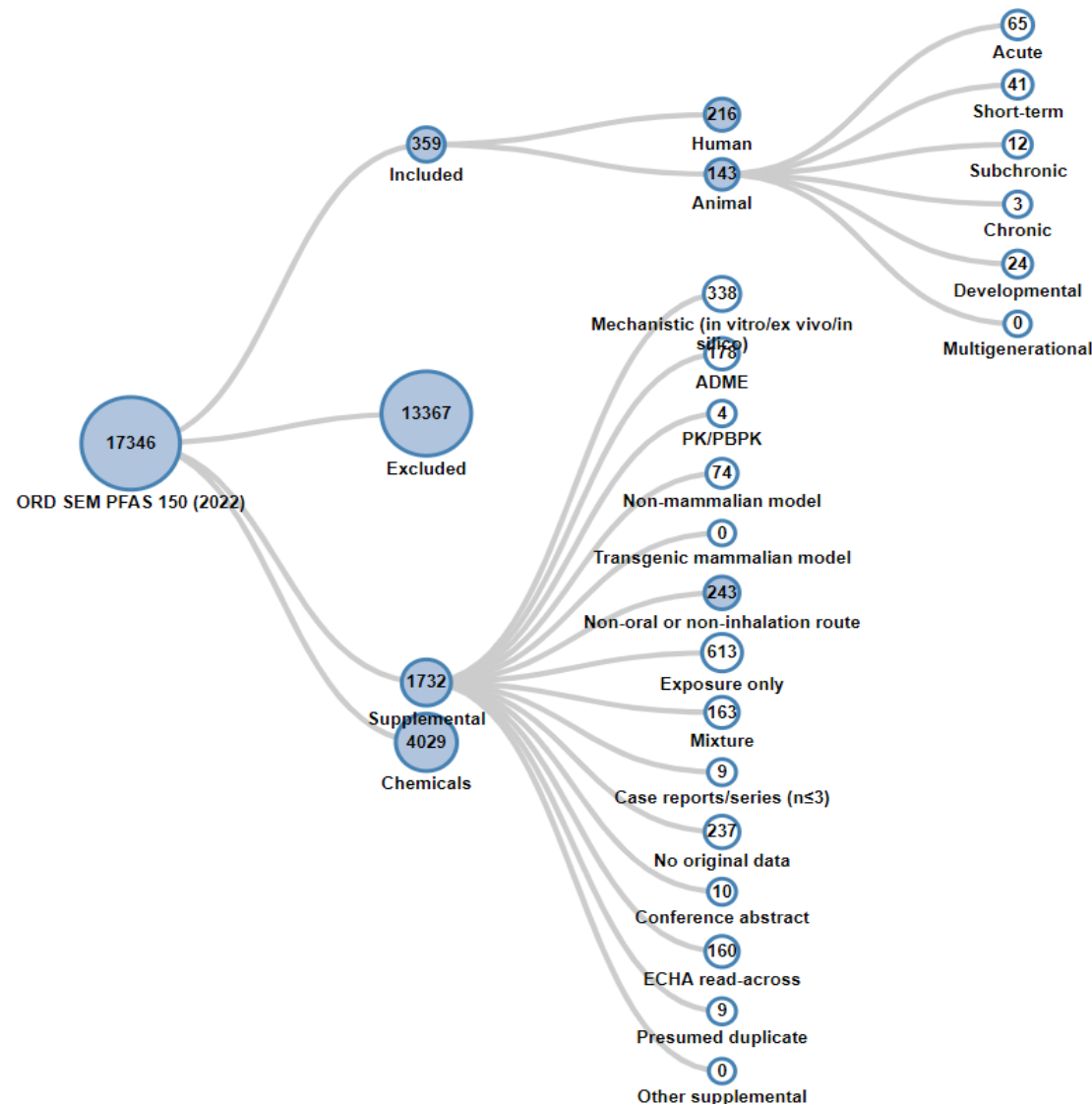
Perfluoroalkyl acids (PFAAs) are highly persistent and bioaccumulative, resulting in their broad distribution in humans and the environment. The liver is an important target for PFAAs, but the mechanisms behind PFAAs interaction with hepatocyte proteins remain poorly understood. We characterized the binding of PFAAs to human liver fatty acid-binding protein (hL-FABP) and identified critical structural features in their interaction. The binding interaction of PFAAs with hL-FABP was determined by fluorescence displacement and isothermal titration calorimetry (ITC) assay. Molecular simulation was conducted to define interactions at the binding sites. ITC measurement revealed that PFOA/PFNA displayed a moderate affinity for hL-FABP at a 1:1 molar ratio, a weak binding affinity for PFHxS and no binding for PFHxA. Moreover, the interaction was mainly mediated by electrostatic attraction and hydrogen bonding. Substitution of Asn111 with Asp caused loss of binding affinity to PFAA, indicating its crucial role for the initial PFAA binding to the outer binding site. Substitution of Arg122 with Gly caused only one molecule of PFAA to bind to hL-FABP. Molecular simulation showed that substitution of Arg122 increased the volume of the outer binding pocket, making it impossible to form intensive hydrophobic stacking and hydrogen bonds with PFOA, and highlighting its crucial role in the binding process. The binding affinity of PFAAs increased significantly with their carbon number. Arg122 and Asn111 played a pivotal role in these interactions. Our findings may help understand the distribution pattern, bioaccumulation, elimination, and toxicity of PFAAs in humans.

**Keywords:** Human liver fatty acid-binding protein; Interaction; Isothermal titration calorimetry; Molecular simulation; Perfluorinated compounds.

## Key words

# We organize your information using tags

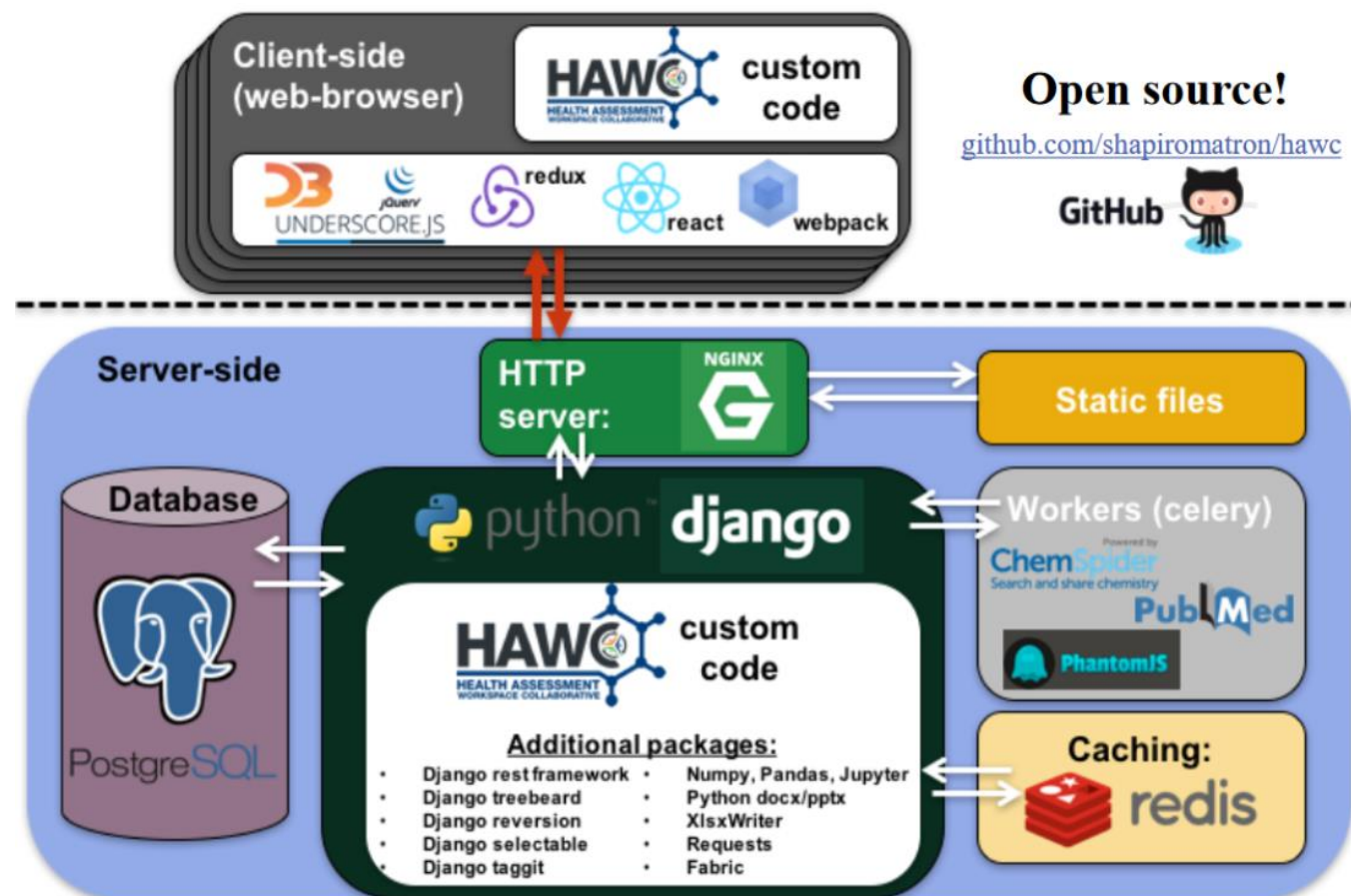
- What are tags and why do we use them?
- Tags or labels are used to **filter** or **flag** records during the review process.
- They are *kind of like a sticky note* and help us to **organize information** into different bins that can be rapidly recalled.
- Tags are **standardized** to picklists and **controlled vocabularies**.
- Tags are applied manually or automatically by computers based upon classifiers (e.g. search strategies that are specified by key words). If you skimp on key word descriptions, we will not find or might filter out your data!



# Quick note: What is HAWC?

The IRIS Program commonly uses the EPA's version of Health Assessment Workspace Collaborative (HAWC) (<https://hawcprd.epa.gov/portal/>) for structured data extraction and digitization of epidemiological and animal toxicological studies.

- A Python application
  - A web-application data entry in/excel out
  - APIs for automated data in/out
  - Data science stack available for compute
- A relational database
  - Mostly relational data
  - Also binary/nosql data
- An interactive frontend
  - Dynamic visualizations + modern web
- An open-source application
  - We can accept pull-requests from anyone
  - Code freely available on github



# Structured Data Extraction Frameworks

Templates for consistent summary of information included in the HAWC database.

*Promotes consistency, transparency, and efficiency in that a task is done once and uniformly*

Domain/Field Name	Picklist or free text	Help text
Experiment	Domain heading	Domain heading
experiment type	Picklist Short-term (1-30 days) Subchronic (30-90 days) Chronic (>90 days) Mechanistic Reproductive Developmental Acute (<24 hr) Other	Select the study type. If multiple study types are covered by the same data entry form, the specific study type should be selected. If none matches, select 'other', highlight and extract the text, and add a comment into the
Test article	Domain heading	Domain heading
test article name	Free text	Select the chemical name (test material) as reported by authors and the appropriate link to chemical information (if available) from the CompTox Chemicals Dashboard. Link to <a href="https://comptox.epa.gov/dashboard/">https://comptox.epa.gov/dashboard/</a>
CAS number	Free text	Select the appropriate CAS number.
purity	Free text	Description of the chemical purity (%) including information on contaminants, isomers, etc.
test article source	Free text	Description of the chemical source (i.e. manufacturer or supplier) and lot/batch number of test material
vehicle	Free text	Description of the vehicle (use name as described in methods but also add the common name if the vehicle was described in a non-standard way).

- structured fields for consistent data entry
- Picklists for consistent data entry
- Help text to explain the content that should be entered into a field

# How does this work?

## Dosing regime

Route of exposure	Oral gavage
Exposure duration	90 d
Duration observation	91 days

## Female Crl:CD(SD) Rats

Positive control	Name	Female Crl:CD(SD) Rats
Negative control	Species	

Doses	Strain
Description	Sex
	Source
	Lifestage exposed
	Lifestage assessed
	Animal Husbandry
	Diet
	Free text
	Free text
	Free text
	Free text

Test article
test article name
CAS number
purity
test article source
vehicle

## Available endpoints

Endpoint	Organ	Obs. Time	Dose [mg/kg-day]			
			0	10	50	200
N	-	-	10	10	10	10
Alanine Aminotransferase (ALT)	Multi-Organ	Day 90	35 ± 6.5	56 ± 41.3 (60%)	45 ± 19.2 (29%)	36 ± 10.2 (3%) <sup>a</sup>
Albumin (A)	Multi-Organ	Day 90	4.7 ± 0.32	5 ± 0.36 (6%)	5 ± 0.62 (6%)	4.7 ± 0.39 (0%) <sup>a</sup>
Albumin/Globulin (A/G) Ratio	Multi-Organ	Day 90	1.79 ± 0.231	1.89 ± 0.189 (6%)	1.88 ± 0.22 (5%)	2.04 ± 0.237 (14%) <sup>a</sup>
Alkaline Phosphatase (ALP)	Multi-Organ	Day 90	55 ± 13	52 ± 12.7 (-5%)	43 ± 10.9 (-22%)	57 ± 11.6 (4%) <sup>a</sup>
Aspartate Aminotransferase (AST)	Multi-Organ	Day 90	78 ± 16.2	108 ± 54.3 (38%)	92 ± 22.2 (18%)	82 ± 15.3 (5%) <sup>a</sup>
Body Weight, Absolute	Whole Body	Day 90	264 ± 27.5	261 ± 30.2 (-1%)	252 ± 22 (-5%)	257 ± 22.6 (-3%) <sup>a</sup>
Brain Weight, Absolute	Brain	Day 90	1.91 ± 0.095	1.93 ± 0.072 (1%)	1.89 ± 0.068 (-1%)	1.9 ± 0.107 (-1%) <sup>a</sup>
Brain Weight, Relative	Brain	Day 90	0.73 ± 0.057	0.747 ± 0.095 (2%)	0.755 ± 0.055 (3%)	0.74 ± 0.053 (1%) <sup>a</sup>
Calcium (CA)	Multi-Organ	Day 90	11 ± 0.44	11.3 ± 0.53 (3%)	11.2 ± 0.43 (2%)	11 ± 0.51 (0%) <sup>a</sup>
Cholesterol (CHOL), Total	Multi-Organ	Day 90	74 ± 20.1	81 ± 23.5 (9%)	83 ± 23.7 (12%)	71 ± 9.5 (-4%) <sup>a</sup>

# How can the standards contribute to findable, accessible, interoperable, and reusable data?

- Findable: standardized language provides harmonization in the description of environmental health science findings.
- Accessible: The EHV and data normalized to EHV are made available in EPA HAWC.
- Interoperable: Data curation using standardized terminology makes it easier to build connections, map the normalized terms to other databases.
- Reusable: Data are extracted using structured formats and stored as digital assets.



<https://hawc.epa.gov/assessment/public/>

# Data Standardization

## Why do we standardize data?

Assessment teams must standardize the language used to report data so that it can be aggregated. This is done digitally with picklists and controlled vocabularies. Standardization such as the Environmental Health Vocabulary (EHV). Standardization makes information more findable and interoperable within and across assessments.

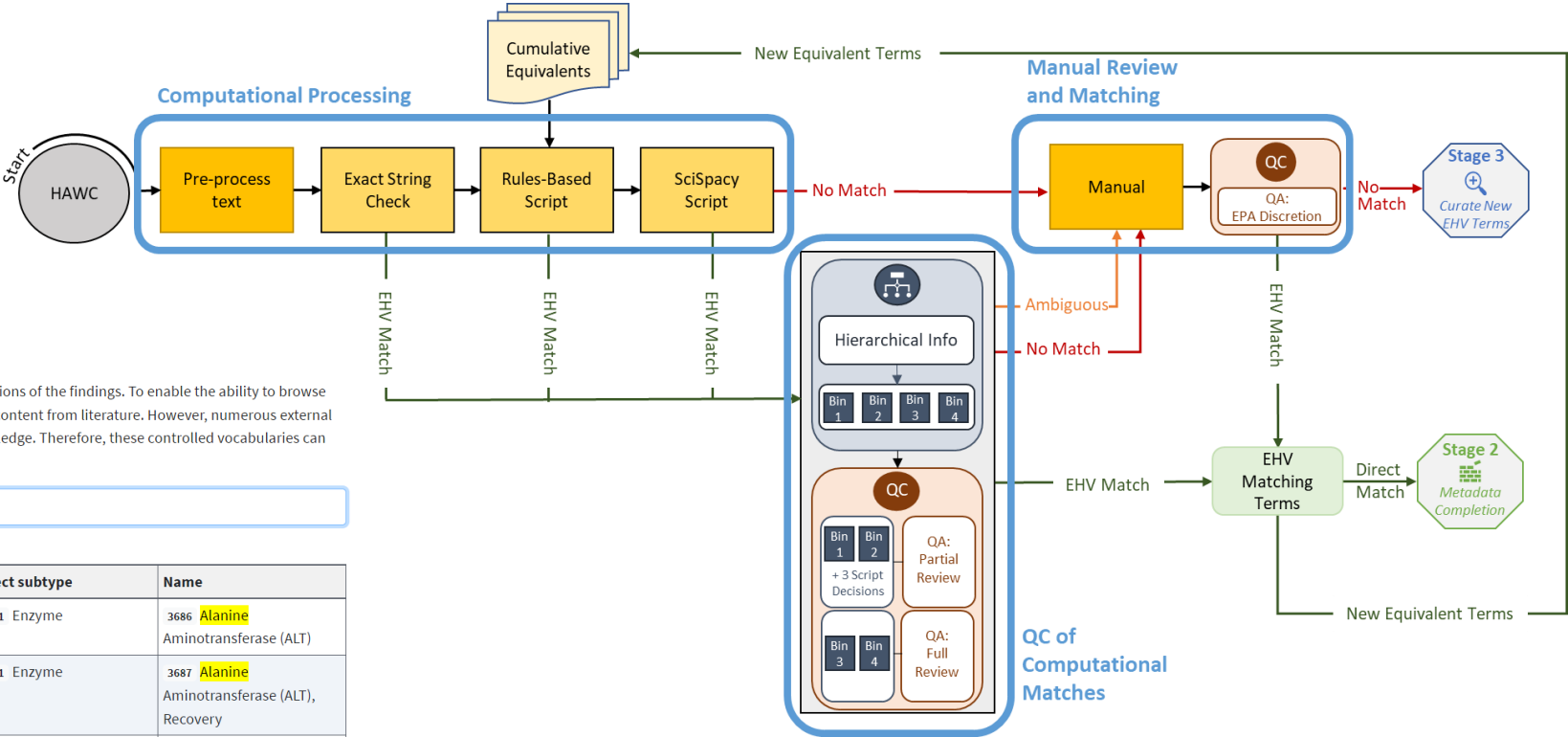
## CPAD’s workflow for curating new EHV terms in HAWC

### Environmental Health Vocabulary (EHV)

Experimental data extracted into HAWC is designed to enable filtering, sorting, and visualizations of the findings. To enable the ability to browse and filter at different levels, a curated controlled vocabulary of terms can be used to extract content from literature. However, numerous external ontologies and thesauri exist in the biological domain that can map health-effects and knowledge. Therefore, these controlled vocabularies can also be mapped to external entities.

7 items found.

System	Organ	Effect	Effect subtype	Name
4430 Metabolic	4471 Multi-Organ	3843 Clinical Chemistry	3981 Enzyme	3686 Alanine Aminotransferase (ALT)
4430 Metabolic	4471 Multi-Organ	3843 Clinical Chemistry	3981 Enzyme	3687 Alanine Aminotransferase (ALT), Recovery



# Examples from the EHV

## Environmental Health Vocabulary (EHV)

Experimental data extracted into HAWC is designed to enable filtering, sorting, and visualizations of the findings. To enable the ability to browse and filter at different levels, a curated controlled vocabulary of terms can be used to extract content from literature. However, numerous external ontologies and thesauri exist in the biological domain that can map health-effects and knowledge. Therefore, these controlled vocabularies can also be mapped to external entities.

7 items found.

System	Organ	Effect	Effect subtype	Name
4430 Metabolic	4471 Multi-Organ	3843 Clinical Chemistry	3981 Enzyme	3686 Alanine Aminotransferase (ALT)
4430 Metabolic	4471 Multi-Organ	3843 Clinical Chemistry	3981 Enzyme	3687 Alanine Aminotransferase (ALT), Recovery

- Environmental Health Vocabulary (EHV) <https://hawc.epa.gov/vocab/ehv/>
- Housed in EPA's Health Assessment Workplace Collaborative (HAWC) <https://hawc.epa.gov/assessment/public/>

# Application of the EHV in an IRIS Assessment

[Home](#) / [PFHxA \(2018\)](#) / [Chengelis, 2009, 2850404](#) / [90-Day Oral](#) / [Female Crl:CD\(SD\) Rats](#) / [Alanine Aminotransferase \(ALT\)](#) / [Update](#)

## Update Alanine Aminotransferase (ALT)

Update an existing endpoint. The [Environmental Health Vocabulary \(EHV\)](#) is enabled for this assessment. Browse to view controlled terms, and whenever possible please use these terms.

Endpoint/Adverse outcome\*

3686

Load ID

Alanine Aminotransferase (ALT)

Selected term: 3686 | Alanine Aminotransferase (ALT) ×

☒ Use controlled vocabulary

Short-text used to describe the data in this form. Please use a controlled vocabulary term if possible and if enabled for your assessment. A separate field, "Endpoint Name in Study", captures the name of endpoint as reported. If no preferred term matches the data extracted, type in the desired description. Do not add units — units are summarized in a separate extraction field. If the endpoint is a repeated measure, indicate the time in parentheses, e.g., running wheel activity (6 wk), using the abbreviated format: seconds = sec, minutes = min, hours = h, days = d, weeks = wk, months = mon, years = y.

System

Metabolic

Selected term: 4430 | Metabolic ×

☒ Use controlled vocabulary

Organ/Tissue/Region

Multi-Organ

Selected term: 4471 | Multi-Organ ×

☒ Use controlled vocabulary

Effect

Clinical Chemistry

Selected term: 3843 | Clinical Chemistry ×

☒ Use controlled vocabulary

Effect subtype

Enzyme

Selected term: 3981 | Enzyme ×

☒ Use controlled vocabulary

# EHV to Facilitate Evidence Assimilation

Ability to  
aggregate  
information  
from various  
studies  
reporting the  
same  
endpoints



<https://hawc.epa.gov/summary/data-pivot/assessment/100500070/pfhexa-animal-toxicology-hepatc-effects-serum-bioma/>

# EHV Facilitates Data Interaction and Use

## Portal

> PFHxA (2018)

» Literature review

» Management dashboard

» Study list

» Study evaluation

» Endpoint list

» Summary tables

» Visualizations

» Executive summary

» Downloads

About HAWC

HAWC Resources

study design vs. system								
System	Cancer							
	Cardiovascular							
	Dermal							
	Developmental							
	Endocrine							
	Female Reproductive							
	Gastrointestinal							
	Hematologic							
	Hepatic	12		13		17	28	70
	Immune							
	Male Reproductive							
	Metabolic							
	Multi-System							
	Musculoskeletal							
	Musculoskeletal							
	Nervous							
	Ocular							
	Reproductive							
	Respiratory							
	Systemic							
	Urinary							
	Whole Body							
	<b>Grand Total</b>	<b>12</b>		<b>13</b>		<b>17</b>	<b>28</b>	<b>70</b>
		1-generation reproductive	Acute (<24 hr)	Cancer	Developmental	Reproductive	Short-term (1-30 days)	Subchronic (30-90 days)
								<b>Grand Total</b>

## Experiment Type

- ☒ 12 1-generation reproductive
- ☒ 13 Cancer
- ☒ 17 Short-term (1-30 days)
- ☒ 28 Subchronic (30-90 days)

## System

- ☐ 0 Female Reproductive
- ☐ 0 Gastrointestinal
- ☐ 0 Hematologic
- ☒ 70 Hepatic
- ☐ 0 Immune

## Endpoint Name

- Concentration
- ☒ 12 Liver Weight, Absolute
- ☒ 2 Liver Weight, Absolute, Recovery
- ☒ 12 Liver Weight, Relative

# EHV Facilitates Data Interaction and Use

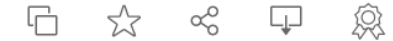
Respiratory			4			25	24	53
Systemic						5		5
Urinary			18			13	28	59
Whole Body	44	2	7	2	16	12	6	89
<b>Grand Total</b>	<b>150</b>	<b>2</b>	<b>196</b>	<b>5</b>	<b>34</b>	<b>323</b>	<b>352</b>	<b>1062</b>
	1-generation reproductive	Acute (<24 hr)	Cancer	Developmental	Reproductive	Short-term (1-30 days)	Subchronic (30-90 days)	<b>Grand Total</b>
Study design								

Endpoint Name		✓	✕
<input checked="" type="checkbox"/>	2 Albumin (A), Recovery		
<input checked="" type="checkbox"/>	8 Albumin/Globulin (A/G) Ratio		
<input checked="" type="checkbox"/>	2 Albumin/Globulin (A/G) Ratio, Recovery		
<input checked="" type="checkbox"/>	10 Alkaline Phosphatase (ALP)		
<input checked="" type="checkbox"/>	2 Alkaline Phosphatase (ALP), Recovery		
<input checked="" type="checkbox"/>	2 Alveolar Macrophages		

Study Citation	Experiment Name	Animal Group Name	System	Organ	Effect	Endpoint Name	Doses	Dose Units Name	NOEL	LOEL	BMD	BMDL
Klaunig, 2015, 2850075	2-Year Cancer Bioassay	Male Crl:CD(SD) Rats	Whole Body	Whole Body	Clinical Observation	Survival	0, 2.5, 15, 100	mg/kg-day				
Klaunig, 2015, 2850075	2-Year Cancer Bioassay	Male Crl:CD(SD) Rats	Whole Body	Whole Body	Clinical Observation	Body Weight, Absolute	0, 2.5, 15, 100	mg/kg-day				
Klaunig, 2015, 2850075	2-Year Cancer Bioassay	Male Crl:CD(SD) Rats	Hematologic	Multi-Organ	Hematology	White Blood Cell (WBC)	0, 2.5, 15, 100	mg/kg-day	100			
Klaunig, 2015, 2850075	2-Year Cancer Bioassay	Male Crl:CD(SD) Rats	Hematologic	Multi-Organ	Hematology	White Blood Cell (WBC)	0, 2.5, 15, 100	mg/kg-day	100			

# Application of EHV in a Systematic Evidence Map

PFAS-150 Evidence Map Visualizations by [literature inventory](#)



ReadMe Animal Studies Human Studies

## Toxicological Studies Examining Exposure to PFAS by Study Design and Health System



### Heat Map

### References

	acute							short-term			subchronic		chronic		developmental, F1				Grand Total
	mouse	rat	guinea pig	hamster	rabbit	dog	not reported	mouse	rat	not reported	mouse	rat	mouse	rat	mouse	rat	rabbit	not reported	
Cancer													2						2
Cardiovascular		3				4		1	10		1	6	2			5			30
Dermal		1							2			2				2			7
Developmental															1	21	3	1	24
Reproductive		4						1	12		1	9	2		1	20	3		49
Endocrine									9			7	2			6			24
Exocrine		1																	1
Gastrointestinal		7							6		1	5	1			4			24
Hematologic									11		1	10	2		1	7			31
Hepatic	1	8	1			1		9	17		2	9	2		1	10	1		59
Immune		4						3	10			9	2		1	5			34

<https://public.tableau.com/app/profile/literature.inventory/viz/PFAS-150EvidenceMapVisualizations/HumanStudies>

## Systematic Evidence Map for 150+ Per- and Polyfluoroalkyl Substances (PFAS)

# Take Homes

- Be nice to future you!
  - Make your research findable
    - If key information are not in the title, abstract, key words, author lists we probably are not going to find it.
  - Use standards (if they exist) before creating new ones
  - Use a structured process for documenting (extracting) and reporting data
  - Have fun and make data sharing common place and unexceptional!

# Useful Resources

- <https://force11.org/info/the-fair-data-principles/>
- U.S. EPA. ORD Staff Handbook for Developing IRIS Assessments (2022). U.S. EPA Office of Research and Development, Washington, DC, EPA/600/R-22/268, 2022.
- [Health Assessment Workspace Collaborative \(HAWC\) \(epa.gov\)](https://www.epa.gov/health-assessment-workspace-collaborative)

# Thank you for listening!

- Questions?

Contact:

Michelle Angrish

- [angrish.michelle@epa.gov](mailto:angrish.michelle@epa.gov)

## Acknowledgements

Andy Shapiro

Sean Watford

Kris Thayer

Paul Whaley

Charles Schmitt

Kaitlyn Hair

David Mellor





# The Environmental Health Language Collaborative

*Harmonizing Data, Connecting Knowledge, Improving Health*

---

## Presentation 2

Presentation Order	Presentation Title	Presenter, Organization
2	<i>How best to combine data from multiple independent studies?</i>	Jeanette A Stingone, PhD, MPH, Columbia University <a href="mailto:js5406@cumc.columbia.edu">js5406@cumc.columbia.edu</a>

# HOW BEST TO COMBINE DATA FROM MULTIPLE INDEPENDENT STUDIES?

Jeanette A Stingone PhD MPH  
Assistant Professor, Department of Epidemiology  
Columbia University  
New York, NY USA  
[j.stingone@columbia.edu](mailto:j.stingone@columbia.edu)

# Conflict of Interest Disclosure Slide

I have no conflicts of interest to disclose.

# Acknowledgements

## EHLC Data Harmonization Use Case Members

Charles Schmitt  
Maria Shatz  
Mireya Diaz  
Hina Narayan  
Elaine Faustman  
Kara Fecho  
Ram Gouripeddi  
Philip Holmes  
David Kaeli  
Oswaldo Lozoya  
Andrew Rooney  
Kelly Shipkowski  
And others....

## ICF

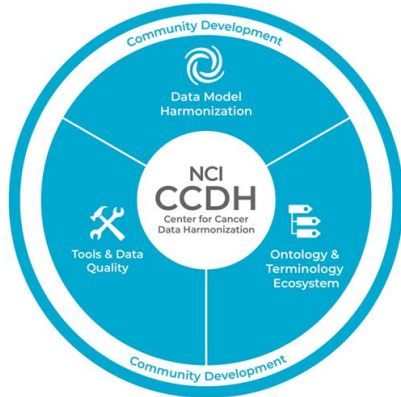
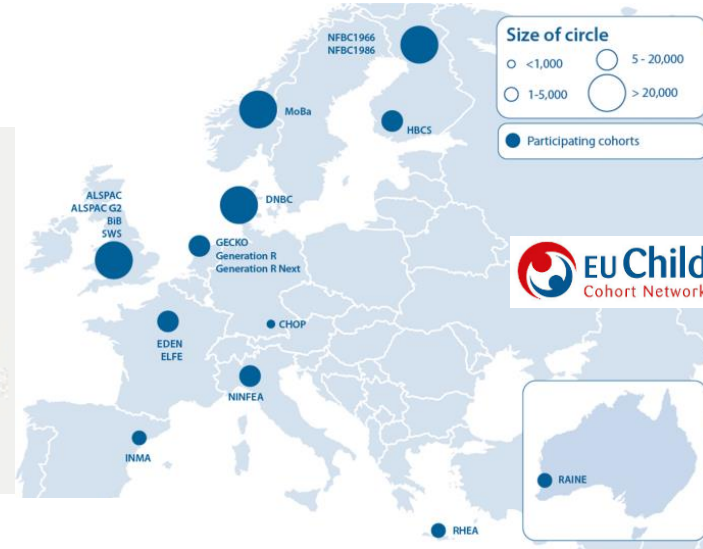
HC Bledsoe  
Jennifer Freed  
Pearl Kaplan  
Jess Wignall

Environmental Health Language  
Collaborative

*Harmonizing Data. Connecting Knowledge. Improving Health.*



# Growing Interest in Data Harmonization



But what does language have to do with it???

# Background and Purpose of Data Harmonization Use Case within the Environmental Health Language Collaborative (EHLC)



- Increased sharing and interoperability of environmental health data has the potential to foster innovation and enhance data-driven discoveries.
- The **purpose** of our use case is to address the feasibility of and to identify the barriers to using harmonized language for combining data across independent research studies.
- Our **goal** is to develop tools and strategies to facilitate data sharing and harmonization through use of data and metadata standards and annotation of datasets.

# Specifics of Our Use-Case: Harmonizing Data Across Two Epidemiologic Research Studies

Two studies from the Human Health Exposure Analysis Resource (HHEAR) Data Repository focused on measures of air pollution exposure and childhood asthma

Can we harmonize data across the two studies with the goal of conducting a pooled data analysis? What resources exist? What do we still need?

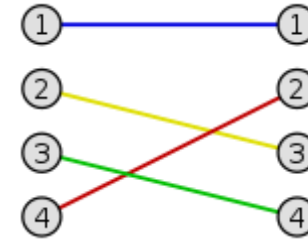


# Retrospective vs Prospective Harmonization

## Tools Developed for Retrospective Harmonization

Human-centered protocols/ “brute force”

Software to facilitate mapping between terms



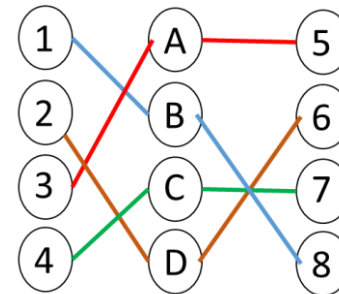
Importance of identifying  
Commonality across language

## Prospective Data Collection/Generation: With what do we align? How do we prepare?

Importance of Community-Agreed Upon Standards

Consideration of Interoperability

Enables Greater Flexibility with Harmonization

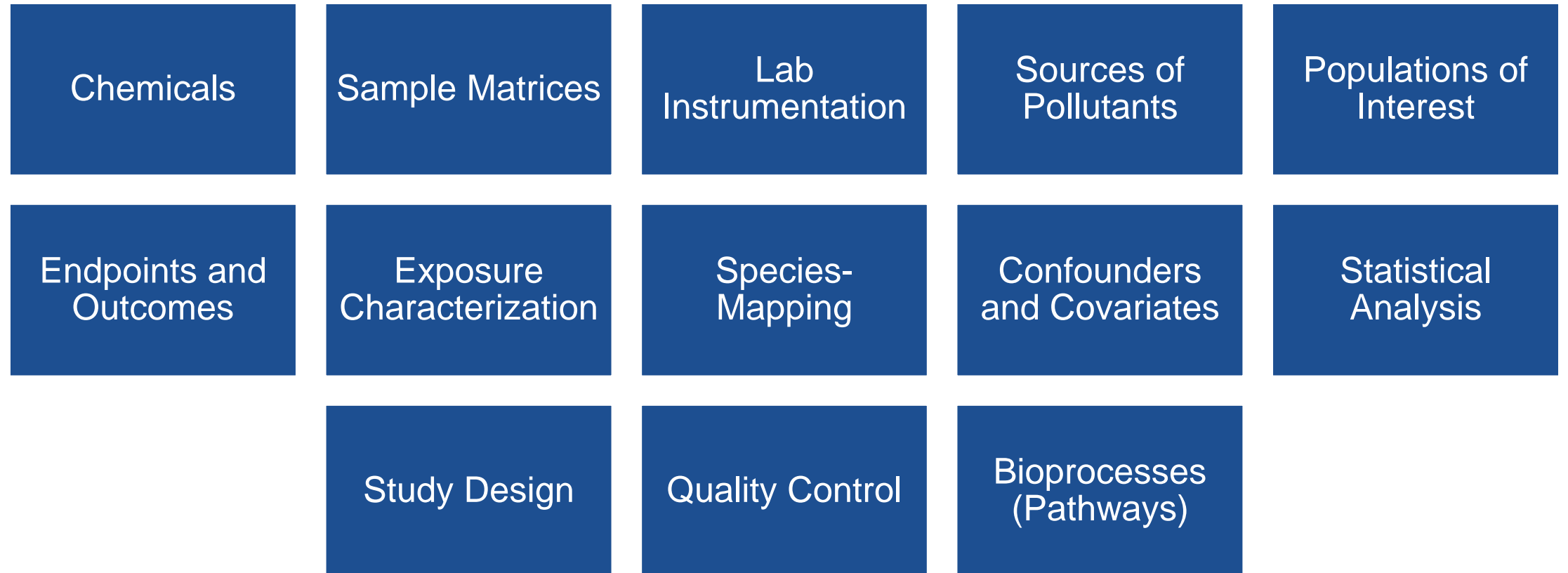


Importance of having standard  
language that can be mapped  
to diverse sources

What resources exist to identify common language to enable prospective approaches to harmonization?



# Domains within Human Epidemiology Studies



# Identifying Resources within Domains

## CHEMICALS

ChEBI: chemical entities of biological interest

Pubchem

CompTox Chemicals Dashboard

Substance Registry System

ChemSpider

## SOURCES OF POLLUTANTS

ENVO: The Environment Ontology

EPA/TSCA has a fair bit here. Follow-up with exposure considerations

ECTO: Environmental Conditions, Treatments and Exposures Ontology

## HUMAN ENDPOINTS/OUTCOMES

DOID: Human Disease Ontology

HPO: Human Phenotype Ontology

CMO: Clinical Measurement Ontology

COGAT: Cognitive Atlas Ontology

OECD Harmonized Templates/IUCLID

UMLS

**Gap:** Positive outcomes, wellness, etc.

Sequence Ontology (SO)

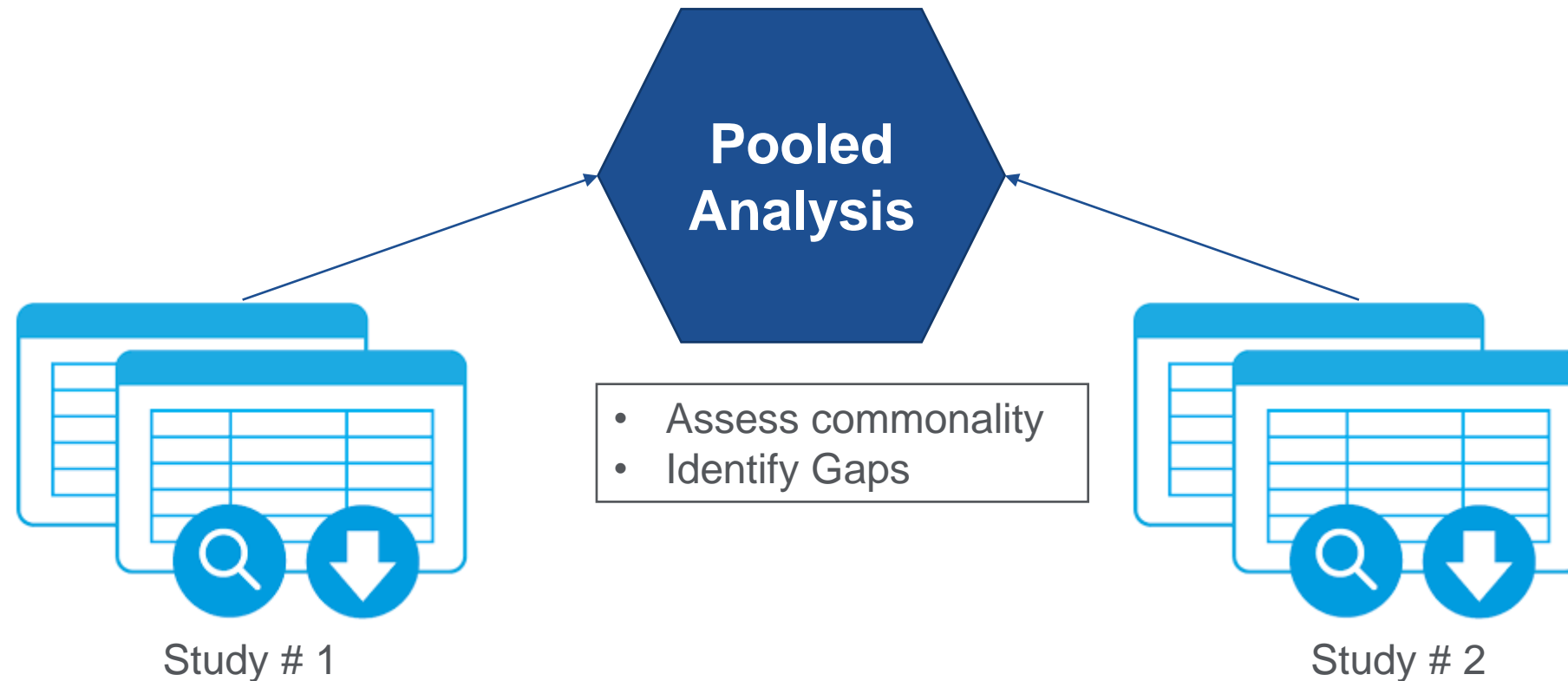
PROMIS®: Patient-Reported Outcomes Measurement Information Systems

PhenX

CompTox Chemicals Dashboard

# Back to Use-Case:

Can we harmonize two studies on similar research question together?



# ILLUSTRATION OF Harmonization EXERCISE

## DataSet 1

<b>Study 2016_1450</b>
demo_hisp_latino
demo_racialafam
demo_racialwhite
demo_racialasian
demo_racialaian
demo_racialafrican
demo_racialhawaiian
demo_racialno
demo_racialno
demo_income_code
actest_asthma_affect_visit1
controller_treatment_b_visit1
controller_treatment_b_visit1
daps_spiro_age_visit1
daps_spiro_gender_visit1

## DataSet 2

<b>Study 2016_1407</b>
ethnicity_form03
race_form03_black
race_form03_white
race_form03_asian
race_form03_am_indian
race_form03_hawaii
race_form03_unknown
race_form03_refused
household_income
stop_play_symp_14days
prescription_control
prescription_control_now
age
gender_form03
symptoms_14days
maxsx
symptoms_14days
maxsx
wake_up_14days
wake_up_14days
fef_25_75_per_predict
fev1_per_predict
fvc_per_predict
pft_data_accepted
fvc_best
fev1_best
composite_score_form10
bmi_pct
bmi
height_average
weight_average
date_form03
fev1_best/fvc_best

## Mapping Criteria

- Substring match
- Data match
- Heuristic match
- Language model match
- Ontology match

## Mapping Options

DAPS\_SPIRO\_AGE\_VISIT1 – AGE AGREEMENT  
 Variable name substring match score: 100%/14%  
 Language model similarity: 64%  
 Heuristic match: age heuristic, time heuristic  
 Data type match: numeric/numeric  
 Data distribution match: 95%  
 Ontology match: Age category

DAPS\_SPIRO\_AGE\_VISIT1 – SYMPTOMS\_14days AGREEMENT  
 Variable name substring match score: 0%/0%  
 Language model similarity: 31%  
 Heuristic match: time heuristic  
 Data type match: numeric/numeric  
 Data distribution match: 27%  
 Ontology match: Time category

DAPS\_SPIRO\_AGE\_VISIT1 – HEIGHT\_AVERAGE AGREEMENT  
 Variable name substring match score: 9%/9%  
 Language model similarity: 13%  
 Heuristic match: no match  
 Data type match: numeric/numeric  
 Data distribution match: 5%  
 Ontology match: No common category

## Criteria Explanation

- Substring Match: score based on the syntactic similarity.
- Heuristic Match: various heuristic matching criteria, such as having dates in the variable values or preset keyword lists (e.g., BMI).
- Ontology Match: if the variables have been mapped to an ontology, number of steps to a common ancestor. May have to incorporate multiple ontologies.
- Data match: whether the data types (ordinal, numeric) are the same, and if yes, whether the distribution of values is comparable.
- Language model match: similarity of embedding scores for the variables and their descriptions.

Tool Main Mapping Page

Slide Courtesy of C. Schmitt

# Lessons Learned from Harmonization Exercise around Importance of Common Language



- Language used for variable names and data dictionaries often requires human assessment for mapping
- Combination of lack of standard language AND lack of metadata
- Reminder: Our goal is to develop tools and strategies to facilitate data sharing and harmonization through use of data and metadata standards and annotation of datasets.



# Recommendations for the Broader Scientific Community

## Tool Development

- Reliance on human annotation is not practical for large-scale, timely and consistent harmonization

## Community Data Standards

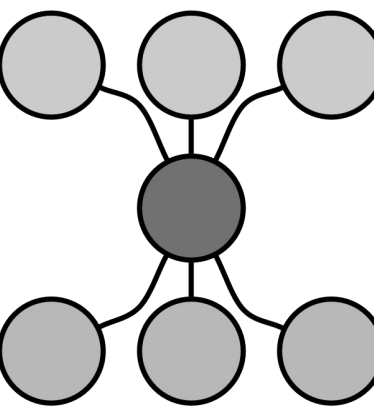
- Gap: Need for common language around context and perspective
  - Models and paradigms used for research; Biases; Evidence-Forms, quality and weight; Evaluation of Evidence

## Promotion of Data Harmonization Efforts as part of Standard Scientific Pipelines

- Thinking about FAIR at study design and data collection phases of research

## Presentation 3

Presentation Order	Presentation Title	Presenter, Organization
3	<i>Digitizing Relationships between Exposures, Biomarkers, and Clinical Outcomes (In the era of AI and exposomics)</i>	Chirag Patel, PhD, Harvard Medical School <a href="mailto:chirag_patel@hms.harvard.edu">chirag_patel@hms.harvard.edu</a>



# Digitizing Relationships between Exposures, Biomarkers, and Clinical Outcomes (In the era of *AI* and *exposomics*)

Chirag Patel

Society of Toxicology 2024, Salt Lake City

3/12/2024



**HARVARD**  
MEDICAL SCHOOL

DEPARTMENT OF  
Biomedical Informatics

[chirag@hms.harvard.edu](mailto:chirag@hms.harvard.edu)  
@chiragjp  
[www.chiragjpgroup.org](http://www.chiragjpgroup.org)

# Disclosures

- No financial conflicts
- Research funding from National Institutes of Health
  - National Institute on Aging (NIA)
  - National Institute of Environmental Health Sciences (NIEHS)

# What are the *biological processes and biomarkers* associated with *exposure and how do they relate to the potential for an adverse outcome*?

## Purpose

---

The purpose of this use case is to explore how harmonized language can help answer the question “What are the biological processes and biomarkers associated with exposure and how do they relate to the potential for an adverse outcome associated with a given exposure?” We are doing this by building upon the other use cases by utilizing their interim results and providing feedback on the general utility of their outputs. Our goal is to connect measured biomarkers to exposure-response relationships by:

- Extending the semantic description of the exposure event to explicitly include measurements as previously done for adverse outcome pathways
- Semantically linking the exposure event to adverse outcomes by connecting the perturbed biological processes with toxicity mechanisms
- Supporting the integration of existing data and resources (e.g., ‘omics measurements, adverse outcome pathways)

<https://www.niehs.nih.gov/research/programs/ehlc/use-cases/bio>

# Participants of the working group

Albert Donnay (JHU/Donnay Detoxicology LLC)

Andrew Rooney (NIEHS)

Anna Maria Masci (NIEHS)

Annie Jarabek (US EPA)

Bren Ames (Aye Open Outcomes)

Carmen Marsit (Emory University)

Carol Hamilton (RTI International)

Charles Schmitt (NIEHS)

Chirag Patel (Harvard University)

David Hines (RTI International)

David Reif (NIEHS)

Elaine Faustman (University of Washington)

Ginger Chew (CDC)

Grace Cooney (ICF)

Hina Narayan (University of Otago)

Joseph Romano (University of Pennsylvania)

Karamarie Fecho (Copperline Professional Solutions)

Ken Wilkins (NIH)

Maria Shatz (NIEHS)

Megan Meinel (ICF)

Michelle Heacock (NIEHS)

Mireya Diaz Insua (Western Michigan Univ.)

Oswaldo Lozoya (RTI International)

Phillip Holmes (NCIT)

Rebecca Boyles (RTI)

Rong-Lin Wang (US EPA)

Sam Hall (ICF)

Shannon Bell (RTI)

Stephanie Holmgren (NIEHS)

Steve Edwards (EPA)

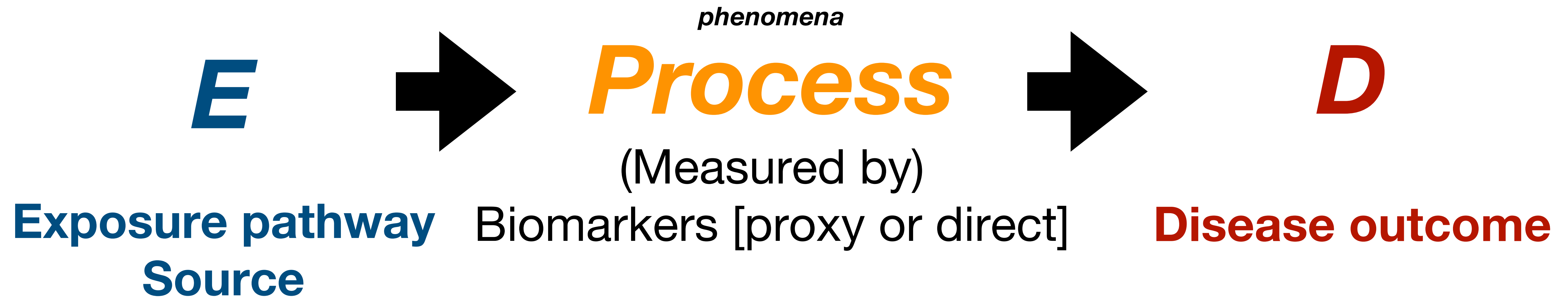
Thomas Hartung (Johns Hopkins University)

Vasu Kilaru (US EPA)

# EHLC biomarkers working group process

- Led by Chirag Patel, Stephen Edwards; facilitated by Charles Schmitt, Samantha Hall (ICF), Stephanie Holmgren, NIEHS
- Met virtually ~bimonthly-quarterly from 2021-23
- Used the “Integrated Science Assessment” from the EPA as a practical example to map exposures, processes, biomarkers, and disease
- ***PM2.5 and lung related outcomes*** (asthma, COPD, decreased lung function)

What are the biological processes and biomarkers associated with exposure and how do they relate to the potential for an adverse outcome?  
*probabilistic*



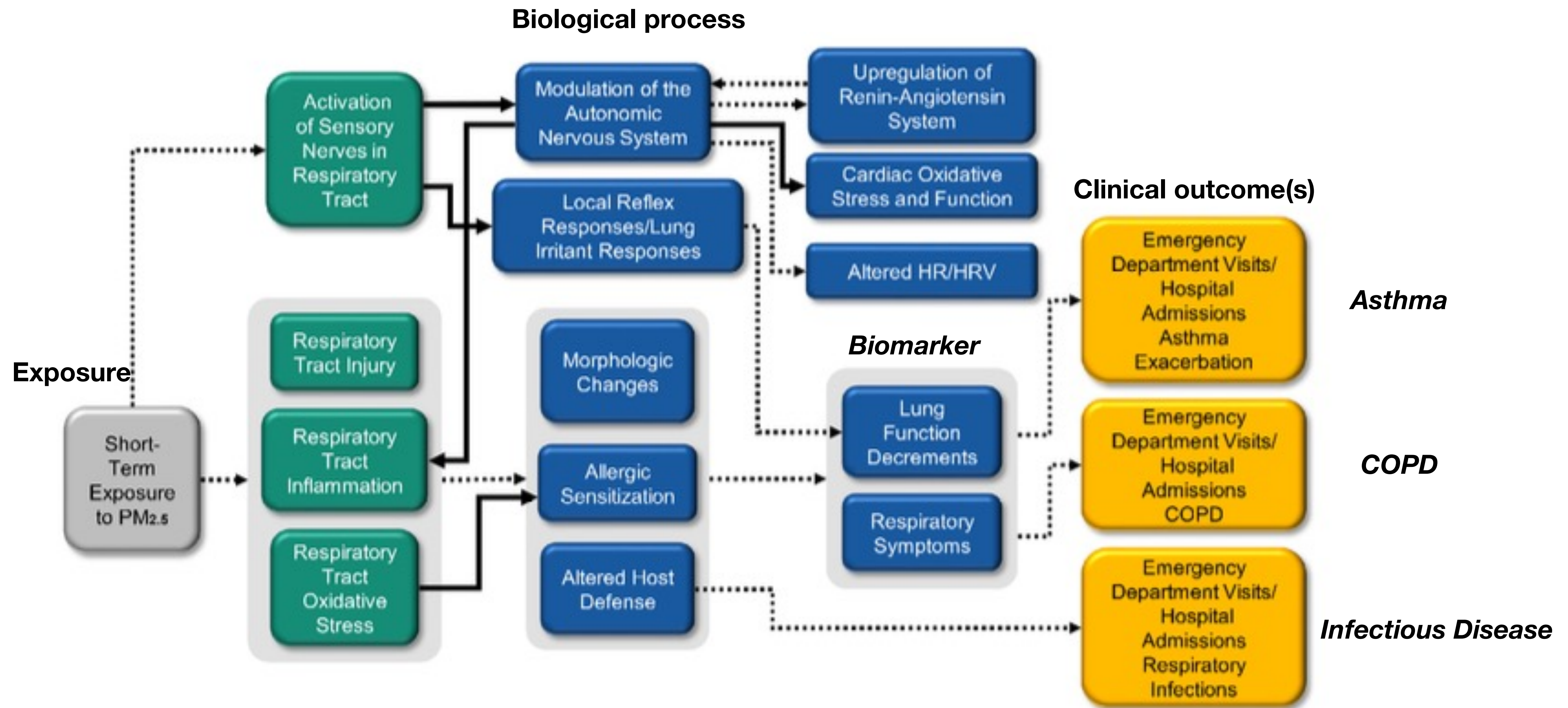
# Integrated Science Assessment for Particulate Matter



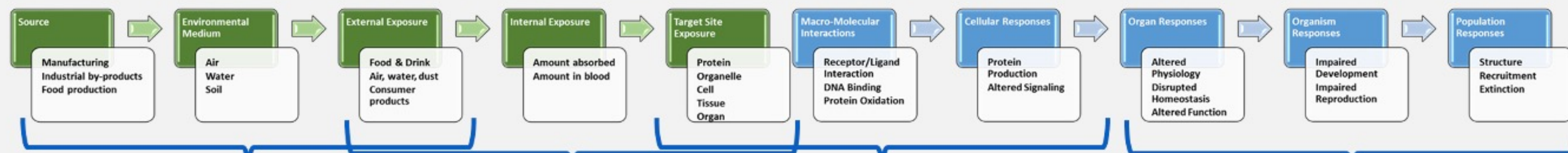
### ***Summary of Causality Determinations for Short- and Long-Term Particulate Matter (PM) Exposure and Respiratory Effects***

This chapter characterizes the scientific evidence that supports causality determinations for short- and long-term PM exposure and respiratory effects. The types of studies evaluated within this chapter are consistent with the overall scope of the ISA as detailed in the [Preface](#) (see [Section P.3.1](#)). In assessing the overall evidence, strengths and limitations of individual studies were evaluated based on scientific considerations detailed in the [Appendix](#). The evidence presented throughout this chapter support the following causality determinations. More details on the causal framework used to reach these conclusions are included in the Preamble to the ISA ([U.S. EPA, 2015](#)).

<b>Size Fraction</b>	<b>Causality Determinations</b>
<i>Short-Term Exposure</i>	
PM <sub>2.5</sub>	Likely to be causal
PM <sub>10-2.5</sub>	Suggestive of, but not sufficient to infer
UFP	Suggestive of, but not sufficient to infer
<i>Long-Term Exposure</i>	
PM <sub>2.5</sub>	Likely to be causal
PM <sub>10-2.5</sub>	Inadequate
UFP	Inadequate

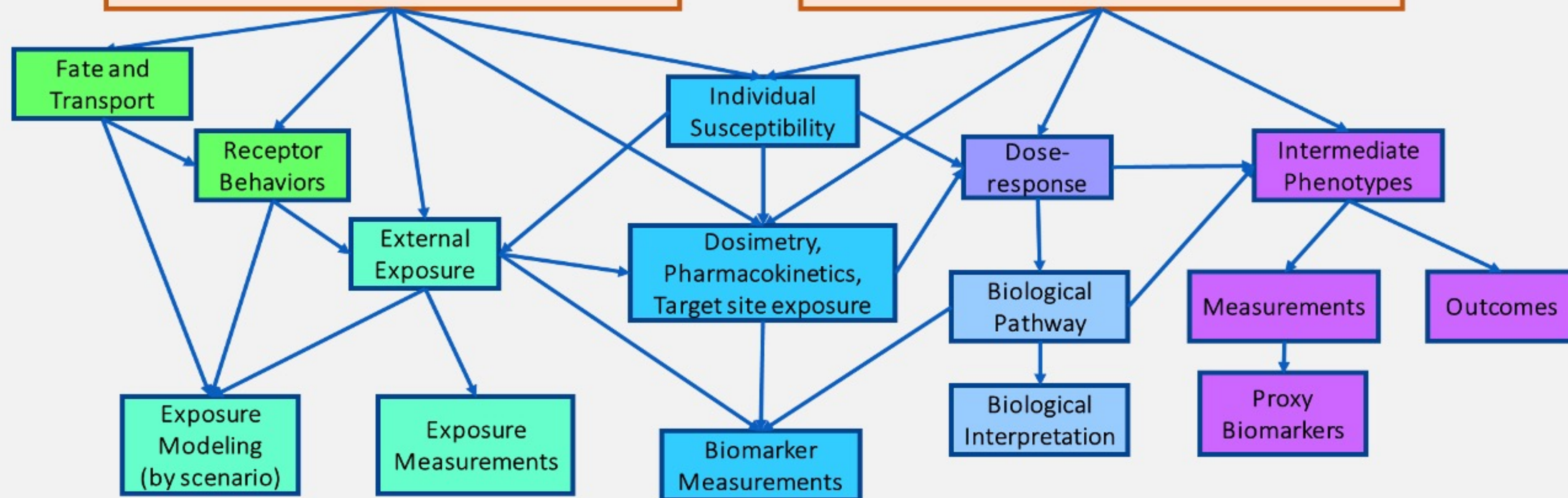


**Figure 5-1. Short term effects of exposure to PM<sub>2.5</sub> in Lung Disease**



## Exposure Pathways and Exposure Event

## Biological Processes and Outcomes



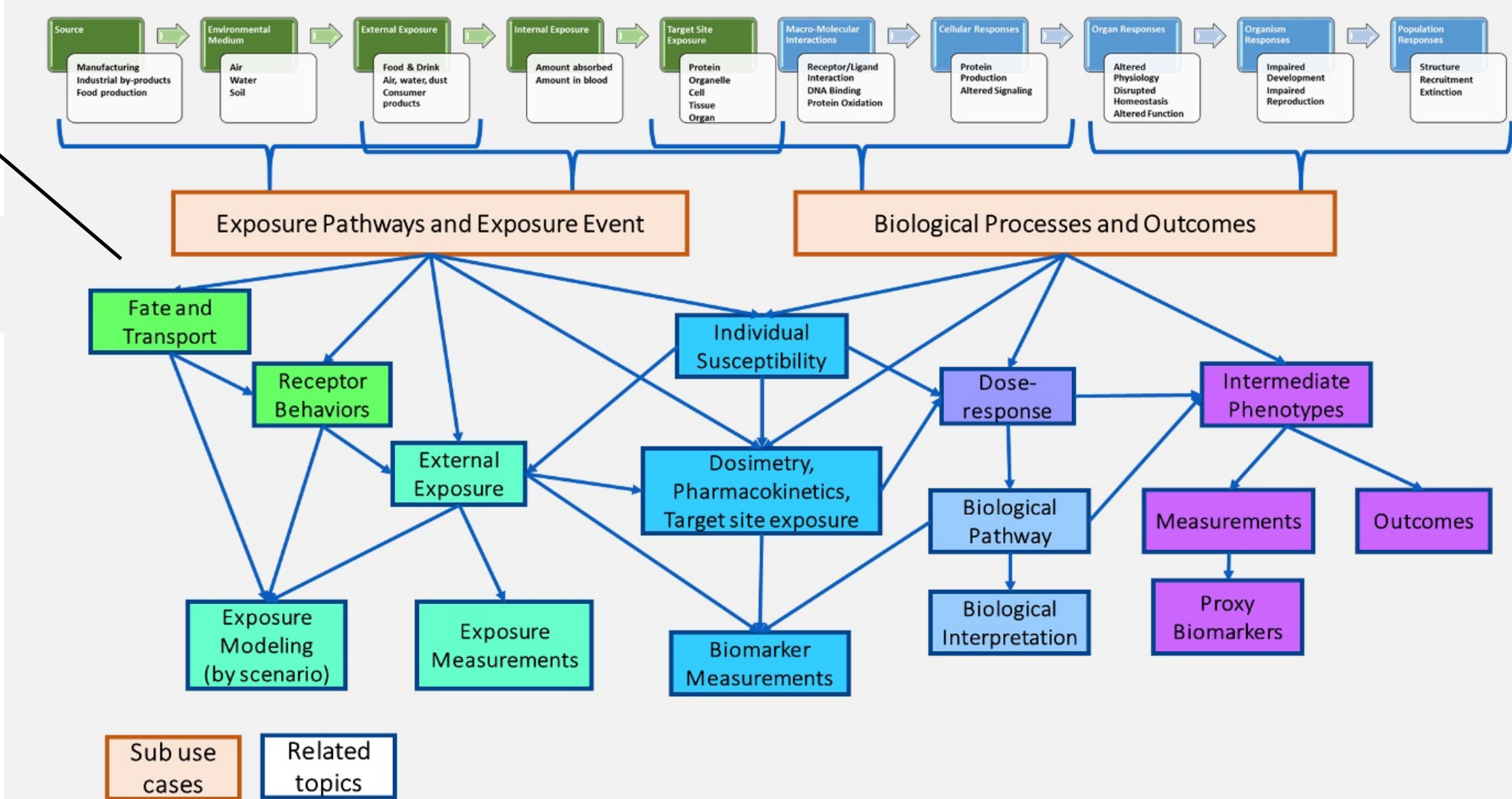
Sub use cases

Related topics

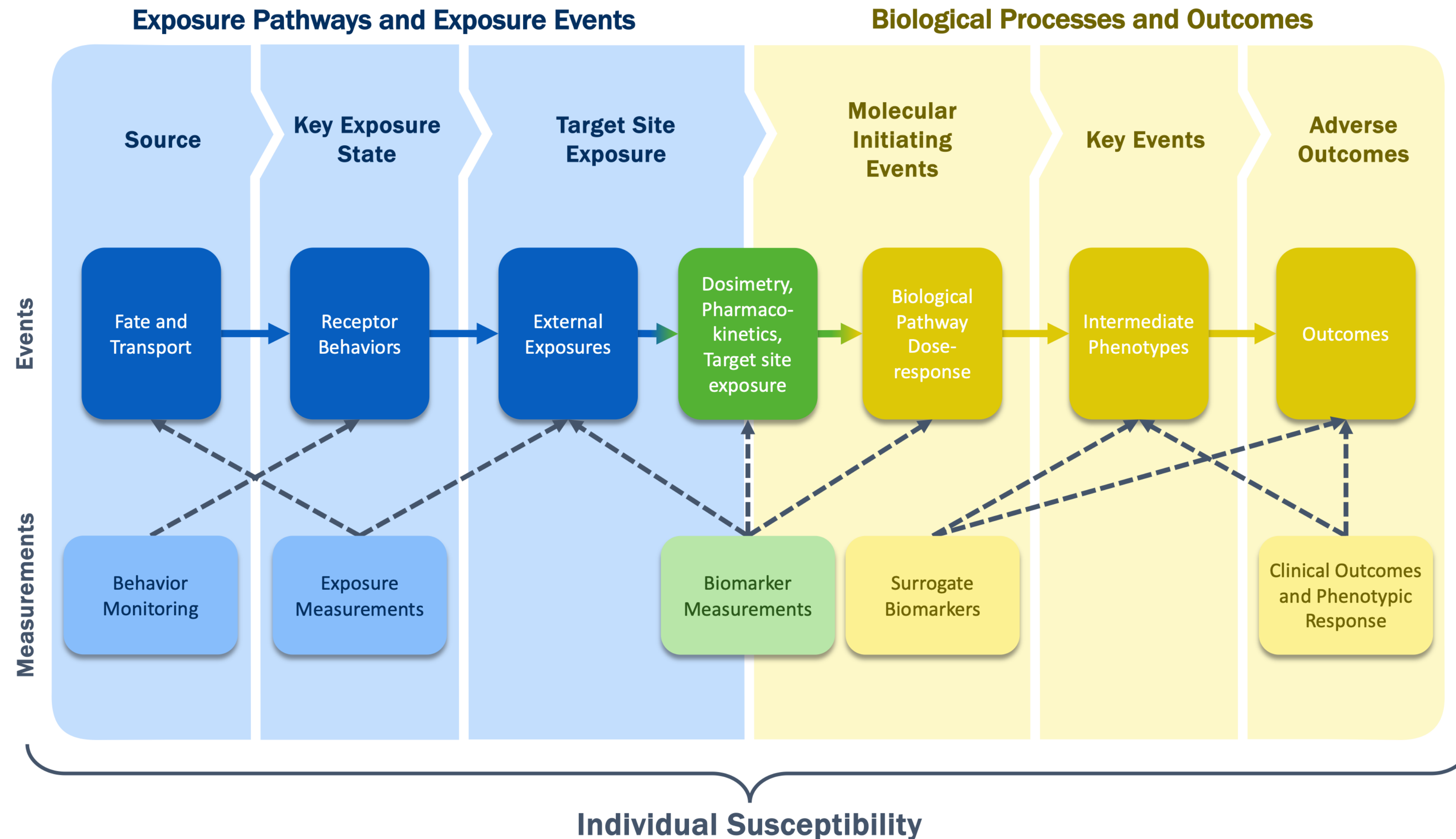
**Table 5-14 Study-specific details from animal toxicological studies of short-term PM<sub>2.5</sub> exposure and respiratory effects in healthy animals.**

Study/Study Population	Pollutant	Exposure	Endpoints
<a href="#">Amatullah et al. (2012)</a> Species: mouse Sex: female Strain: BALB/c Age/weight: 6–8 weeks, 18 g	PM <sub>2.5</sub> CAPs Toronto Particle size: PM <sub>0.15–2.5</sub> Control: HEPA-filtered air	Route: nose-only inhalation Dose/concentration: PM <sub>0.15–2.5</sub> 254 µg/m <sup>3</sup> Duration: 4 h Time to analysis: at end of exposure Modifier: baseline ECG	Pulmonary function BALF cells
<a href="#">Aztatzi-Aguilar et al. (2015)</a> Species: rat Sex: male Strain: Sprague-Dawley	PM <sub>2.5</sub> CAPs Mexico City Particle size: PM <sub>2.5</sub> Control: filtered air	Route: inhalation Dose/concentration: PM <sub>2.5</sub> 178 µg/m <sup>3</sup> Duration: acute 5 h/day, 3 days Subchronic 5 h/day, 4 days/week, 8 weeks Time to analysis: 24 h	Gene expression and protein levels—lung tissue IL-6, components of the RAS and kallikrein-kinin endocrine system-heme oxygenase-1
<a href="#">Budinger et al. (2011)</a> Species: mouse Sex: male Strain: C57BL/6 wild type and IL-6 knockouts Age/weight: 8–12 weeks	PM <sub>2.5</sub> CAPs Chicago, IL Particle size: PM <sub>2.5</sub> Control: filtered ambient air	Route: whole-body inhalation Dose/concentration: 88.5 ± 13.4 µg/m <sup>3</sup> Duration: 8 h/day for 3 days	BALF and lung tissue-protein level and gene expression of inflammatory mediators Plasma—biomarkers of coagulation
<a href="#">Chiarella et al. (2014)</a> Species: mouse Sex: male Strain: C57BL/6 wild type and Adrβ knockouts Age/weight: 8–12 weeks	PM <sub>2.5</sub> CAPs Chicago, IL Particle size: PM <sub>2.5</sub> Control: filtered ambient air	Route: whole-body inhalation Dose/concentration: 109.1 ± 6.1 µg/m <sup>3</sup> Duration: 8 h/day for 3 days	BALF and lung tissue—IL-6, norepinephrine Brown adipose tissue—norepinephrine
<a href="#">Clougherty et al. (2010)</a> Species: rat Sex: male Age/weight: 12 weeks	PM <sub>2.5</sub> CAPs Boston, MA Particle size: PM ≤ 2.5 µm Control: filtered air	Route: whole-body inhalation Dose/concentration: 374 µg/m <sup>3</sup> With large variance Duration: 10 days, 5 h/day Time to analysis: respiratory data was collected during exposure at 10 min. intervals using Buxco Coexposure: stress	Pulmonary function <ul style="list-style-type: none"> <li>• Peak inspiratory flow</li> <li>• Minute volume</li> <li>• Breathing frequency</li> <li>• Inspiratory time</li> <li>• Expiratory time</li> <li>• Expiratory flows</li> <li>• Tidal volume</li> </ul>


				nd Exposure Event	Exposure Pathways and Exposure Event		Exposure Pathways and Exposure Event		Exposure Pathways, Exposure Event, Biological Processes and Outcomes		Exposure Pathways, Exposure Event, Biological Processes and Outcomes		Exposure Pathways, Exposure Event, Biological Processes and Outcomes		Biological Processes and Outcomes	
Citation	ISA Figure/Table	Assigned To	Due Date	Exposure Modeling - by Scenario (Ontology/Mapping)	External Exposure (Example from Paper)	External Exposure (Ontology/Mapping)	Exposure Measurements (Example from Paper)	Exposure Measurements (Ontology/Mapping)	Individual Susceptibility (Example from Paper)	Individual Susceptibility (Ontology/Mapping)	Dosimetry, Pharmacokinetics, Target site exposure (Example from Paper)	Dosimetry, Pharmacokinetics, Target site exposure (Ontology/Mapping)	Biomarker Measurements (Example from Paper)	Biomarker Measurements (Ontology/Mapping)	Dose-Response (Example from Paper)	Dose-Response (Ontology/Mapping)
<a href="#">Samat et al., 2012</a>	<a href="#">ISA Figure 5-5</a>	Chirag	5/12/22				10, PM2.5 at school; during 2 48 hour sampling sessions per week. Measurements at school and city wide		age 6-12							
<a href="#">Silverman et al., 2010</a>	<a href="#">ISA Figure 5-2</a>	Chirag	5/13/22		24 hour average PM 2.5 and ozone		EPA Air Quality System monitors; averaged over 24 hours; 20 monitors within 20 miles of NYC		Susceptible groups by age (<6 y, 6-18, 19-49, 50+)						Relative Risk per IQR range	
<a href="#">Zhao et al., 2016</a>	<a href="#">ISA Figure 5-2</a>	Chirag	5/13/22		24 hour average PM 2.5, PMc, SO2, ozone, NO2, Temp		Dongguan Air Monitoring system; averaged over 24 hours								Relative Risk per IQR range	
<a href="#">Stieb et al., 2009</a>	<a href="#">ISA Figure 5-3</a>	Charles Schmitt	6/3/22		Hourly max concentration of CO, NO2, O3, SO2, PM10, PM2.5	Pollutants: chemical IDs, ECTO	National Air Pollution Surveillance system; Environment Canada's weather archive									
<a href="#">Hebbern and Cakmak, 2015</a>	<a href="#">ISA Table 5-1</a>	Charles Schmitt	6/3/22		Hourly max concentration of	Pollutants: chemical IDs, ECTO; Clinical	National Air Pollution Surveillance system; Aerobiology Research Laboratory / Rotational Impact									



# Conceptual diagram: Mapping the trajectory between exposure, pathways, events, and biological outcomes



# Existing knowledge-base-related resources: simple as integrating them together? (A non-exhaustive list)

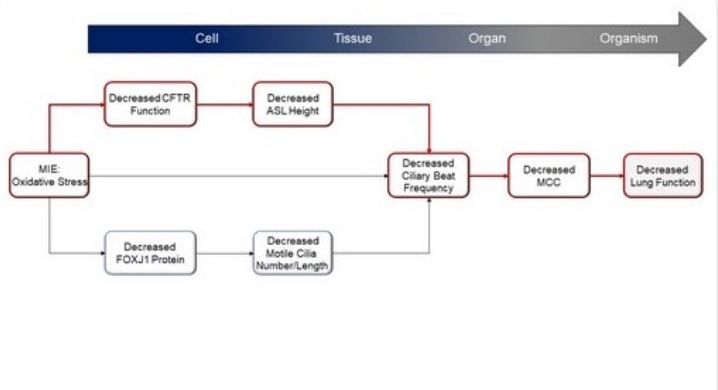
 This AOP is licensed under the [BY-SA license](#). This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.

AOP: 424

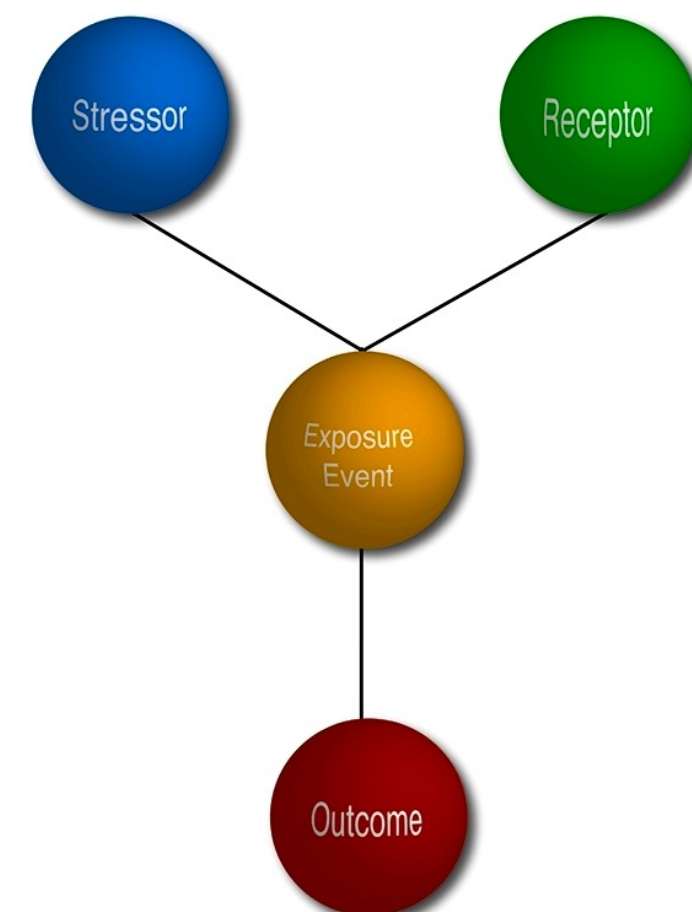
**Title**  
[Oxidative stress Leading to Decreased Lung Function via CFTR dysfunction](#)

**Short name**  
Ox stress-mediated CFTR/ASL/CBF/MCC impairment

**Graphical Representation**  
[Click to download graphical representation template](#) [Explore AOP in a Third Party Tool](#)



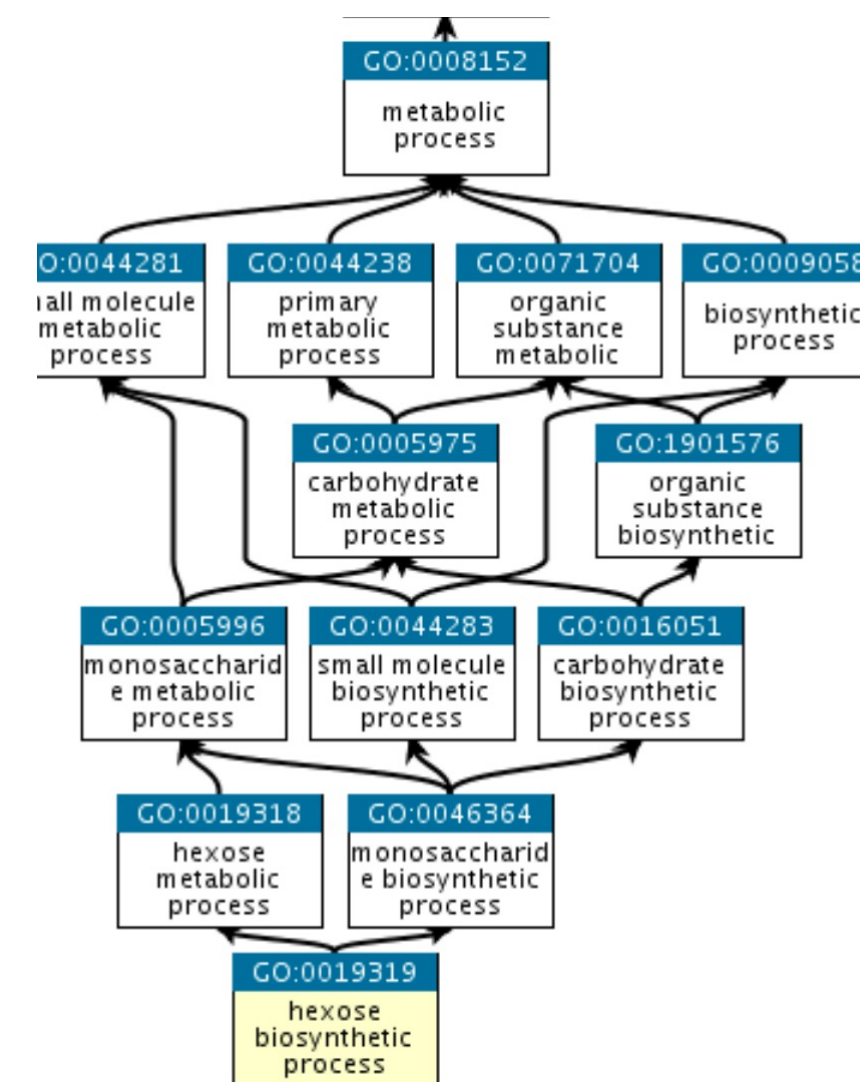
**AOP Framework**  
**AOPkb**  
**AOPwiki**



**Exposure  
Ontology (ExO)**



**Comparative Toxicogenomics  
Database**



**Gene Ontology**

Metadata		<a href="#">Submit Comment</a>	<a href="#">Visualize</a>
ID	D0ID:3083		
Name	chronic obstructive pulmonary disease		
Definition	An obstructive lung disease that is a chronic and progressive disorder of small airways in the lungs and that is characterized by irreversible airflow obstruction, typically identified by reductions in quantitative spirometric indices, induced forced expiratory volume at 1 second (FEV1) and the ratio of FEV1 to forced vital capacity (less than 0.7 is diagnostic of COPD). Lung volume is increased and pulmonary hypertension may occur. The pathologic changes result in the disruption of the airflow in the bronchial airways. Signs and symptoms include shortness of breath, wheezing, productive cough and chest tightness. COPD is a consequence (an end result) of chronic bronchitis, emphysema or both. <a href="https://pubmed.ncbi.nlm.nih.gov/28513453/">https://pubmed.ncbi.nlm.nih.gov/28513453/</a> , <a href="https://pubmed.ncbi.nlm.nih.gov/32745458/">https://pubmed.ncbi.nlm.nih.gov/32745458/</a> , <a href="https://pubmed.ncbi.nlm.nih.gov/32800196/">https://pubmed.ncbi.nlm.nih.gov/32800196/</a> , <a href="https://www.nlm.nih.gov/health-topics/copd">https://www.nlm.nih.gov/health-topics/copd</a>		
Xrefs	EFO:0000341 <a href="#">ICD10CM:I44.9</a> <a href="#">MESH:D029424</a> <a href="#">NCIC3199</a> <a href="#">OMIM:606963</a> SNOMEDCT_US_2023_03_01:13645005 <a href="#">UMLS_CUI:C0024117</a>		
Alternateids	D0ID:11500 D0ID:6144		
Subsets	DO_RAD_slim NCItthesaurus		
	chronic obstructive airway disease [EXACT] chronic obstructive lung disease [EXACT]		

**Disease  
Ontology**

<https://aopwiki.org/>

<https://ctdbase.org/>

<https://geneontology.org>

<https://disease-ontology.org/>

**Table 5-2    Epidemiologic studies of short-term exposure to PM<sub>2.5</sub> and respiratory symptoms and medication use in children with asthma.**

Study	Study Population	Exposure Assessment	Concentration (µg/m <sup>3</sup> )	PM <sub>2.5</sub> Copollutant Model Results and Correlations
<a href="#">†Spira-Cohen et al. (2011)</a> Bronx, NY 2002–2005	n = 40, ages 10–12 yr 86% with rescue inhaler use Daily diary for 1 mo No information on participation rate 89% time spent indoors	School outdoor and total personal 24-h avg <i>r</i> = 0.17 school and personal children walk to school	Mean School: 14.3 Total personal: 24.1	Correlation ( <i>r</i> ): NA Copollutant models with: NA
<a href="#">†Zora et al. (2013)</a> El Paso, TX March–June 2010	n = 36, ages 6–11 yr 33% ICS use, 47% atopy Weekly measures for 13 weeks 95% follow-up participation	School outdoor 96-h avg Two schools: High and low traffic area <i>r</i> = 0.89 between schools, 0.91 between monitors, 0.73–0.86 school and monitor	Mean, max School 1: 13.8, 24.9 School 2: 9.9, 18.5	Correlation ( <i>r</i> ): (School 1, School 2) –0.33, –0.19 NO <sub>2</sub> ; –0.02, 0.25 benzene; 0.10, 0.33 toluene; 0.47, 0.28 O <sub>3</sub> Copollutant models with: NA
<a href="#">†Rabinovitch et al. (2011); Rabinovitch et al. (2006)</a> Denver, CO 2002–2005	n = 82 (3-yr study), 73 (2-yr study) 65–86% moderate/severe asthma, 82–90% ICS use Daily measures for 4–7 mo No information on participation rate	One monitor 24-h avg, 10-h avg (12–11 a.m.), 1-h max (12–11 a.m.) 4.3 km from school <i>r</i> = 0.92 monitor and school	Mean, max for yr 1–3 24-h avg: 6.5–8.2, 20.5–23.7 10-h avg: 7.4–9.1, 22.7–30.2 1-h max: 16.8–22.9, 39–52 (95th)	Correlation ( <i>r</i> ): NA Copollutant models with: NA
<a href="#">†Escamilla-Nuñez et al. (2008)</a> Mexico City, Mexico 2003–2005	n = 147, ages 9–14 yr 43% persistent asthma, 89% atopy Daily diary for mean 22 weeks 94% follow-up participation	One monitor 24-h avg Within 5 km of school or home <i>r</i> = 0.77 monitor and school	Mean: 27.8	Correlation ( <i>r</i> ): 0.62 NO <sub>2</sub> , 0.54 O <sub>3</sub> Copollutant models with: NA

**Study design characteristics captured:**

- Study population (inclusion criteria)
- Pollutant
- Exposure and assessment
- Endpoints and outcomes

**Some characteristics difficult to extract:**

- Risk estimates and standard error
- Outcome definition and phenotyping heterogeneity
- Covariates and modeling approach
- Linkages to external data resources
- “Quality” of a study

# Study design plays a large role in making statements about risk: checklists and guidelines for evidence

## GRADE Handbook

### Introduction to GRADE Handbook

Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. Updated October 2013.

**Editors:** Holger Schünemann ([schuneh@mcmaster.ca](mailto:schuneh@mcmaster.ca)), Jan Brožek ([brozekj@mcmaster.ca](mailto:brozekj@mcmaster.ca)), Gordon Guyatt ([guyatt@mcmaster.ca](mailto:guyatt@mcmaster.ca)), and Andrew Oxman ([oxman@online.no](mailto:oxman@online.no))

### About the Handbook

The GRADE handbook describes the process of rating the quality of the best available evidence and developing health care recommendations following the approach proposed by the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) Working Group ([www.gradeworkinggroup.org](http://www.gradeworkinggroup.org)). The Working Group is a collaboration of health care methodologists, guideline developers, clinicians, health services researchers, health economists, public health officers and other interested members. Beginning in the year 2000, the working group developed, evaluated and implemented a common, transparent and sensible approach to grading the quality of evidence and strength of recommendations in health care. The group interacts through meetings by producing methodological guidance, developing evidence syntheses and guidelines. Members collaborate on research projects, such as the DECIDE project ([www.decide-collaboration.eu](http://www.decide-collaboration.eu)) with other members and other scientists or organizations (e.g. [www.rarebestpractices.eu](http://www.rarebestpractices.eu)). Membership is open and free. See [www.gradeworkinggroup.org](http://www.gradeworkinggroup.org) and Chapter [The GRADE working group](#) in this handbook for more information about the Working Group and a list of the organizations that have endorsed and adopted the GRADE approach.

The handbook is intended to be used as a guide by those responsible for using the GRADE approach to produce GRADE's output, which includes evidence summaries and graded recommendations. Target users of the handbook are systematic review and health technology assessment (HTA) authors, guideline panelists and methodologists who provide support for guideline panels. While many of the examples offered in the handbook are clinical examples, we also aimed to include a broader range of examples from public health and health policy. Finally, specific sections refer to interpreting recommendations for users of recommendations.

### 4.2 GRADE Evidence Profile

See online tutorials at: [cebgrade.mcmaster.ca](http://cebgrade.mcmaster.ca)

The **GRADE evidence profile** contains detailed information about the quality of evidence assessment and the summary of findings for each of the included outcomes. It is intended for review authors, those preparing SoF tables and anyone who questions a quality assessment. It helps those preparing SoF tables to ensure that the judgments they make are systematic and transparent and it allows others to inspect those judgments. Guideline panels should use evidence profiles to ensure that they agree about the judgments underlying the quality assessments.

A GRADE evidence profile allows presentation of key information about all relevant outcomes for a given health care question. It presents **information about the body of evidence** (e.g. number of studies), the **judgments about the underlying quality of evidence**, key **statistical results**, and the **quality of evidence rating for each outcome**.

A GRADE evidence profile is particularly useful for presentation of evidence supporting a recommendation in clinical practice guidelines but also as summary of evidence for other purposes where users need or want to understand the judgments about the quality of evidence in more detail.


The standard format for the evidence profile includes:

- A list of the **outcomes**
- The **number of studies** and **study design(s)**
- Judgements about each of the **quality of evidence factors** assessed; risk of bias, inconsistency, indirectness, imprecision, other considerations (including publication bias and factors that increase the quality of evidence)
- The **assumed risk**; a measure of the typical burden of the outcomes, i.e. illustrative risk or also called baseline risk, baseline score, or control group risk
- The **corresponding risk**; a measure of the burden of the outcomes after the intervention is applied, i.e. the risk of an outcome in treated/exposed people based on the relative magnitude of an effect and assumed (baseline) risk
- The **relative effect**; for dichotomous outcomes the table will usually provide risk ratio, odds ratio, or hazard ratio
- The **absolute effect**; for dichotomous outcomes the number of fewer or more events in treated/exposed group as compared to the control group
- Rating of the **overall quality of evidence** for each outcome (which may vary by outcome)
- Classification of the **importance** of each outcome
- **Footnotes**, if needed, to provide explanations about information in the table such as elaboration on judgements about the quality of evidence

Example 1: GRADE Evidence Profile

<https://gdt.grade.pro.org/app/handbook/handbook.html#h.9rdbelsnu4iy>

*Probabilistic* statements (e.g., epidemiological risk) are a challenge to estimate, but required for evidence synthesis



HAWC Home

Public Assessments

Traffic-related air pollution and hypertensive disorders during pregnancy (2019)

Literature review

Study list

Risk of bias

Endpoint list

Visualizations

Downloads

About HAWC

HAWC Resources

Contact UsPublic AssessmentsLogin

Public Assessments / Traffic-related air pollution and hypertensive disorders during pregnancy (2019)

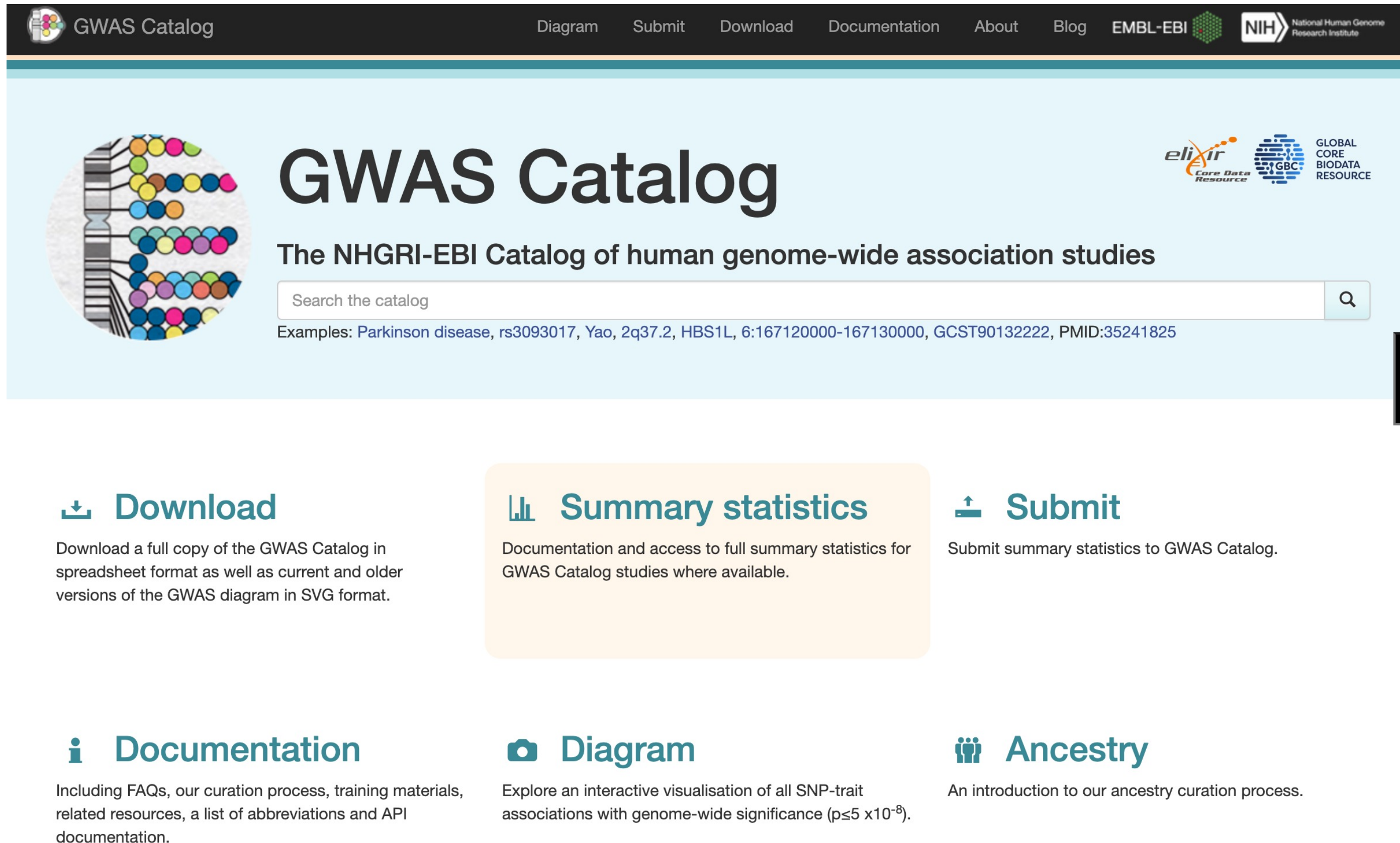
Traffic-related air pollution and hypertensive disorders during pregnancy (2019)

Actions

Assessment name	Traffic-related air pollution and hypertensive disorders during pregnancy
Year	2019
Version	Draft
Objective	<p><i>This evaluation, including the DRAFT NTP Monograph, and content of the HAWC project space is distributed solely for the purpose of pre-dissemination peer review under the applicable information quality guidelines. It has not been formally disseminated by NTP. It does not represent and should not be construed to represent any NTP determination or policy</i></p> <p>The overall objective of this systematic review is to develop NTP hazard conclusions on the association between exposure to traffic-related air pollution (TRAP) and pregnancy-associated hypertensive disorders by integrating levels of evidence from human and experimental animal studies along with relevant mechanistic data.</p> <p>Additional information on this evaluation including the DRAFT NTP Monograph and review protocol can be found on NTP's Office of Health Assessment and Translation project webpage.</p> <ul style="list-style-type: none"><li><a href="https://ntp.niehs.nih.gov/go/trap">https://ntp.niehs.nih.gov/go/trap</a></li></ul>
Authors	NIEHS/NTP
Conflicts of interest	The study assessment team had no financial conflicts of interest.
Funding source	This work was supported by the National Toxicology Program at the National Institute of Environmental Health Sciences, National Institutes of Health with portions of this work performed by ICF under contract to NTP.

Environmental Health Vocabulary (EHV; available at <https://hawc.epa.gov/vocab/ehv/>), which is implemented in [Health Assessment Workspace Collaborative \(HAWC\)](#).

# Finding inspiration in genome-wide association studies GWAS (G-P): standardized genetic variant, analytic approaches, and study designs



The screenshot shows the GWAS Catalog website. At the top is a navigation bar with links: Diagram, Submit, Download, Documentation, About, Blog, EMBL-EBI, and NIH National Human Genome Research Institute. The main header features the GWAS Catalog logo, a search bar with the text "Search the catalog", and a list of example search terms: "Parkinson disease, rs3093017, Yao, 2q37.2, HBS1L, 6:167120000-167130000, GCST90132222, PMID:35241825". Below the header are six main sections: Download, Summary statistics, Submit, Documentation, Diagram, and Ancestry, each with a brief description of its function.

**Download**  
Download a full copy of the GWAS Catalog in spreadsheet format as well as current and older versions of the GWAS diagram in SVG format.

**Summary statistics**  
Documentation and access to full summary statistics for GWAS Catalog studies where available.

**Submit**  
Submit summary statistics to GWAS Catalog.

**Documentation**  
Including FAQs, our curation process, training materials, related resources, a list of abbreviations and API documentation.

**Diagram**  
Explore an interactive visualisation of all SNP-trait associations with genome-wide significance ( $p \leq 5 \times 10^{-8}$ ).

**Ancestry**  
An introduction to our ancestry curation process.

<https://www.ebi.ac.uk/gwas/>

**3,567** publications (as of 9/18/18)

**71,673 *G-P*** associations

**3,955** publications (as of 4/21/19)

**136,287 *G-P*** associations

**4,493** publications (as of 3/10/20)

**179,364 *G-P*** associations

**5,690** publications (as of 5/11/22)

**372,752 *G-P*** associations

**6,245** publications (as of 1/31/23)

**471,482 *G-P*** associations

**6,715** publications (as of 1/30/24)

**571,148 *G-P*** associations

# GWAS catalog: mapping variants, genes, and disease to enhance identification of gene function and disease etiology

Trait: chronic obstructive pulmonary disease

GWAS / Traits / EFO\_0000341

Trait information

Trait label ⓘ

EFO ID ⓘ

Synonyms

Mapped terms ⓘ

Description

Reported Traits ⓘ

Child traits ⓘ

chronic obstructive pulmonary disease

EFO\_0000341

58 synonyms +

11 mapped terms +

A chronic and progressive lung disorder characterized by the loss of elasticity of the bronchial tree and the air sacs, destruction of the air sacs wall, thickening of the bronchial wall, and mucous accumulation in the bronchial tree. The pathologic changes result in the disruption of the air flow in the bronchial airways. Signs and symptoms include shortness of breath, wheezing, productive cough, and chest tightness. The two main types of chronic obstructive pulmonary disease are chronic obstructive bronchitis and emphysema. +

53 reported traits +

3 child traits +

Trait in OLS ↗

Trait in OXO ↗

Trait in Open Targets ↗

Trait in PGS Catalog ↗

Available data:

Associations 1150

Studies 121

Full summary statistics 81

LocusZoom

Download Associations ⬇

☐ Include background traits data ⓘ

☒ Include child trait data

Associations 1150

⬆

# GWAS catalog: mapping variants, genes, and disease to enhance identification of gene function and disease etiology

Trait: chronic obstructive pulmonary disease

GWAS / Traits / EFO\_0000341

Trait information

Trait label ⓘ

EFO ID ⓘ

Synonyms

Mapped terms ⓘ

Description

Reported Traits ⓘ

Child traits ⓘ

chronic obstructive pulmonary disease

EFO\_0000341

58 synonyms +

11 mapped terms +

A chronic and progressive lung disorder characterized by the loss of elasticity of the bronchial tree and the air sacs, destruction of the air sacs wall, thickening of the bronchial wall, and mucous accumulation in the bronchial tree. The pathologic changes result in the disruption of the air flow in the bronchial airways. Signs and symptoms include shortness of breath, wheezing, productive cough, and chest tightness. The two main types of chronic obstructive pulmonary disease are chronic obstructive bronchitis and emphysema. +

53 reported traits +

3 child traits +

Trait in OLS ↗

Trait in OXO ↗

Trait in Open Targets ↗

Trait in PGS Catalog ↗

Available data:

Associations 1150

Studies 121

Full summary statistics 81

LocusZoom

Download Associations ⬇

- ☐ Include background traits data ⓘ
- ☒ Include child trait data

Associations 1150

⬆

***GWAS catalog contains underlying risk estimates (e.g., odds ratios) - can we do the same for the “exposome” ?***

Associations 1150

Show 5 entries

Column visibility

Export

Clear search

Variant and risk allele	P-value	P-value annotation	RAF	OR	Beta	CI	Mapped gene	Reported trait	Trait(s)	Background trait(s)	Study accession	Location
rs2869967-C	6 x 10 <sup>-10</sup>	(EA)	0.41	1.38	-	[1.25-1.53]	FAM13A	Chronic bronchitis and chronic obstructive pulmonary disease	chronic obstructive pulmonary disease, chronic bronchitis	-	GCST002625	4:88948181
rs34391416-A	5 x 10 <sup>-8</sup>	(EA)	0.05	1.93	-	[1.53-2.45]	CRACR2B	Chronic bronchitis and chronic obstructive pulmonary disease	chronic obstructive pulmonary disease, chronic bronchitis	-	GCST002625	11:831818
rs139257032-T	3 x 10 <sup>-7</sup>	(EA)	0.02	3.35	-	[2.12-5.30]	CFAP221	Chronic bronchitis and chronic obstructive pulmonary disease	chronic obstructive pulmonary disease, chronic bronchitis	-	GCST002625	2:119571288
rs12910412-G	5 x 10 <sup>-7</sup>	(EA)	0.46	1.3	-	[1.17-1.44]	LINC01581, H3P40	Chronic bronchitis and chronic obstructive pulmonary disease	chronic obstructive pulmonary disease, chronic bronchitis	-	GCST002625	15:94163844
rs13141641-T	3 x 10 <sup>-6</sup>	(EA)	0.58	1.27	-	[NR]	KRT18P51, HHIP-AS1	Chronic bronchitis and chronic obstructive pulmonary disease	chronic obstructive pulmonary disease, chronic bronchitis	-	GCST002625	4:144585304

Showing 1 to 5 of 1,150 entries

«

1

2

3

4

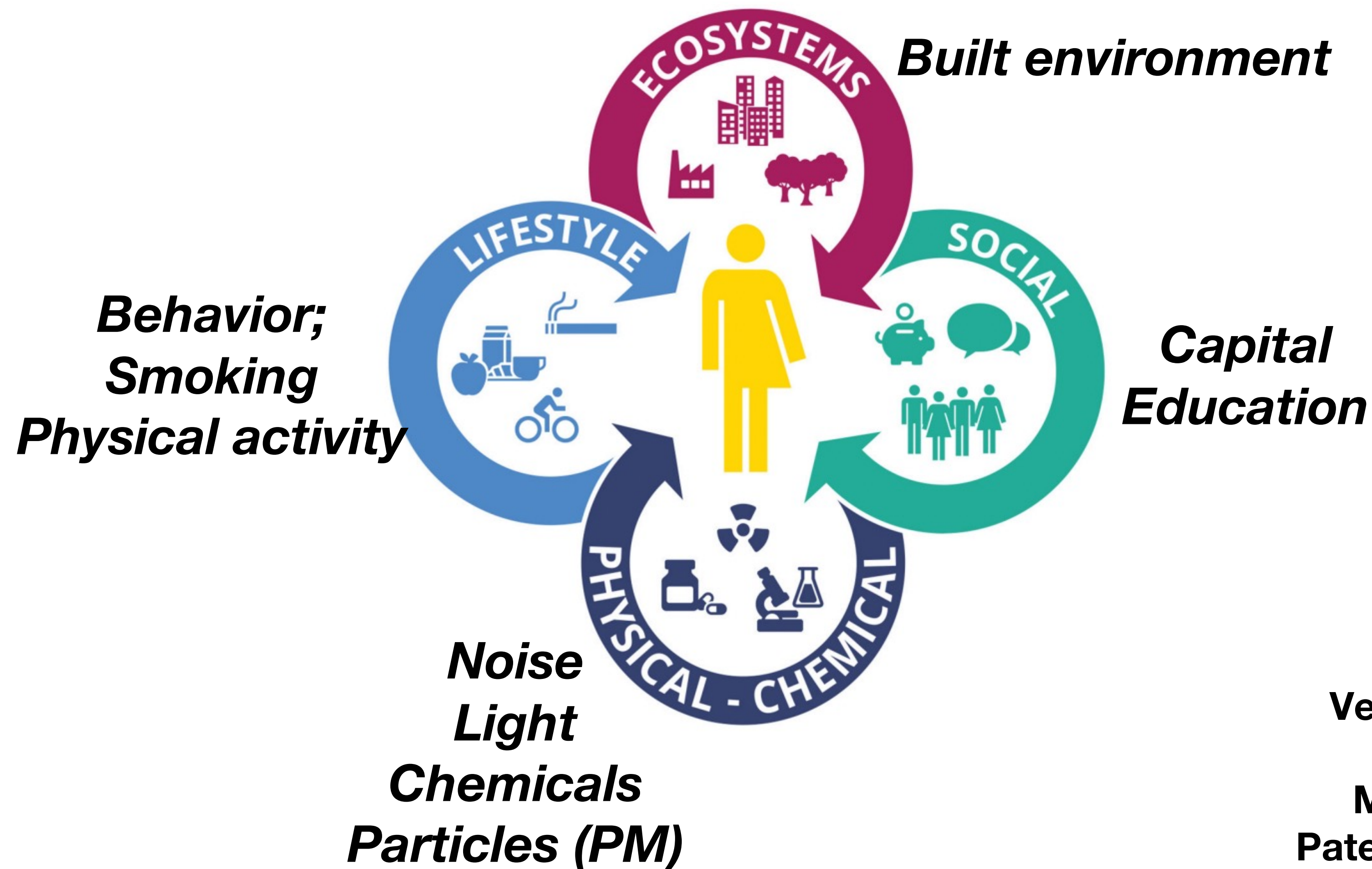
5

...

230

»

# The *exposome*: toward a taxonomization of systematic exposures across domains & modalities



Vermeulen R, Science 2020  
Wild, Int J Epi 2012  
Manrai et al., ARPH 2017  
Patel and Ioannidis JAMA 2014  
Ioannidis et al. STM 2009








# Many modalities of the *exposome* to taxonomize

<u>Modality</u>	<u>Type</u>	<u>Examples</u>
Targeted mass spec	Tabular; spectra	Lead; Cadmium; PFAS
Geospatial markers	Area-level; 2D spectra	Zipcode-level PM 2.5
Self-report questionnaire	Tabular; hierarchical	Nutritional recall
Untargeted mass spec	Tabular; spectra	Mass-charge ratio
Sensor-based behaviors	Tabular; spectra	Accelerometers

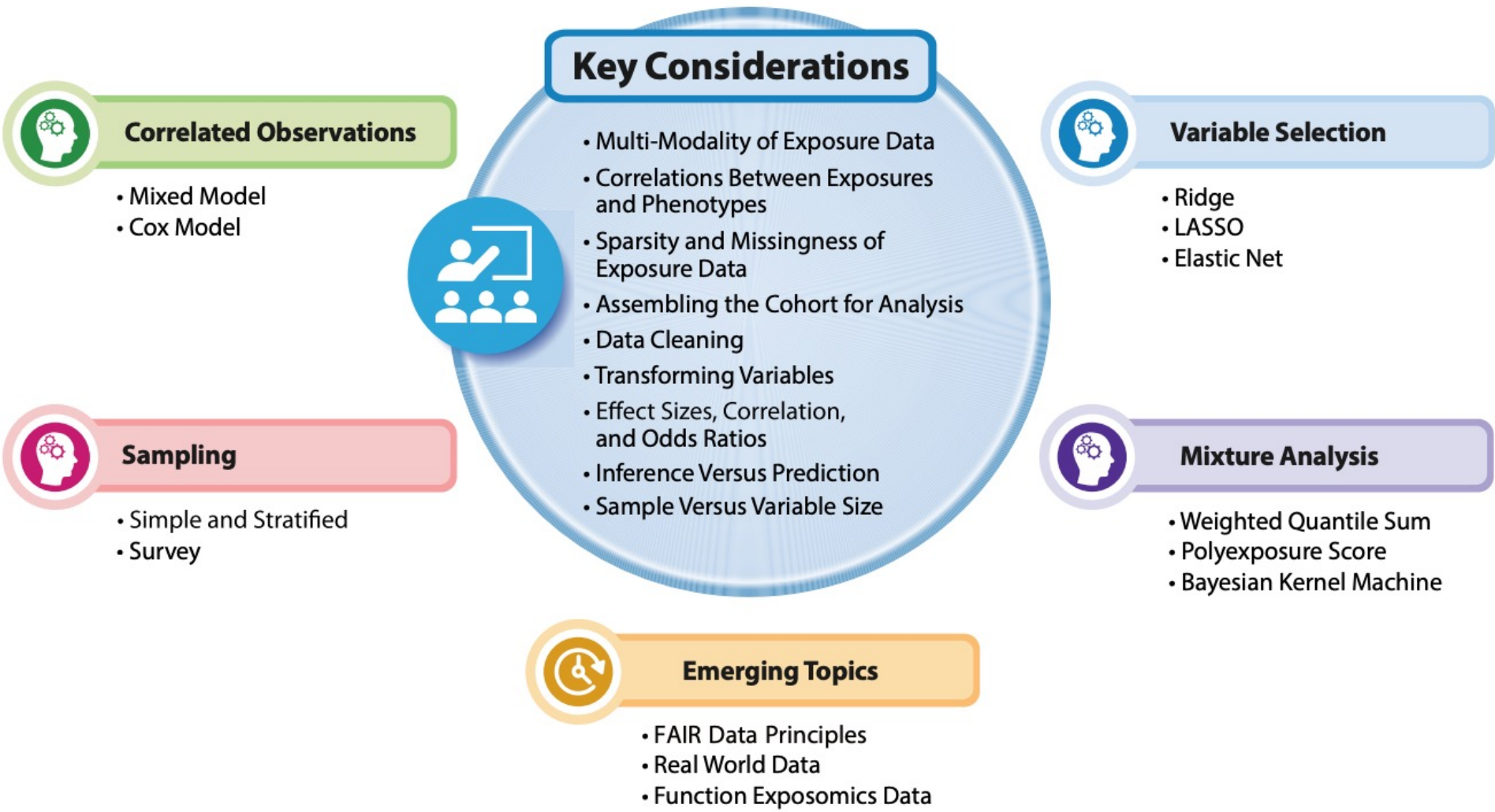
Patel et al, CEBP 2017  
Manrai et al, ARPH 2017  
Vermeulen et al, Science 2020

2022 NIEHS Catalytic Workshop Series on the Exposome

# Decoding the exposome: data science methodologies and implications in exposome-wide association studies (ExWASs)

Ming Kei Chung <sup>1,2,3</sup>, PhD, John S. House <sup>4</sup>, PhD, Farida S. Akhtari<sup>4</sup>, PhD, Konstantinos C. Makris <sup>5</sup>, PhD, Michael A. Langston<sup>6</sup>, PhD, Khandaker Talat Islam<sup>7</sup>, PhD, Philip Holmes<sup>8</sup>, PhD, Marc Chadeau-Hyam <sup>9</sup>, PhD, Alex I. Smirnov<sup>10</sup>, PhD, Xiuxia Du<sup>11</sup>, PhD, Anne E. Thessen <sup>12</sup>, PhD, Yuxia Cui<sup>13</sup>, PhD, Kai Zhang<sup>14</sup>, PhD, Arjun K. Manrai<sup>1</sup>, PhD, Alison Motsinger-Reif <sup>4,\*</sup>, PhD, Chirag J. Patel <sup>1,†,\*</sup>, PhD and Members of the Exposomics Consortium

<sup>1</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA



## Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health

Arjun K. Manrai,<sup>1</sup> Yuxia Cui,<sup>2</sup> Pierre R. Bushel,<sup>2</sup> Molly Hall,<sup>3</sup> Spyros Karakitsios,<sup>4</sup> Carolyn J. Mattingly,<sup>5</sup> Marylyn Ritchie,<sup>3,6</sup> Charles Schmitt,<sup>7</sup> Denis A. Sarigiannis,<sup>4</sup> Duncan C. Thomas,<sup>8</sup> David Wishart,<sup>9</sup> David M. Balshaw,<sup>2</sup> and Chirag J. Patel<sup>1,10</sup>

Chung et al, *Exposome* 2024  
Manrai et al., *ARPH* 2017

Table 1. Data-related recommendations

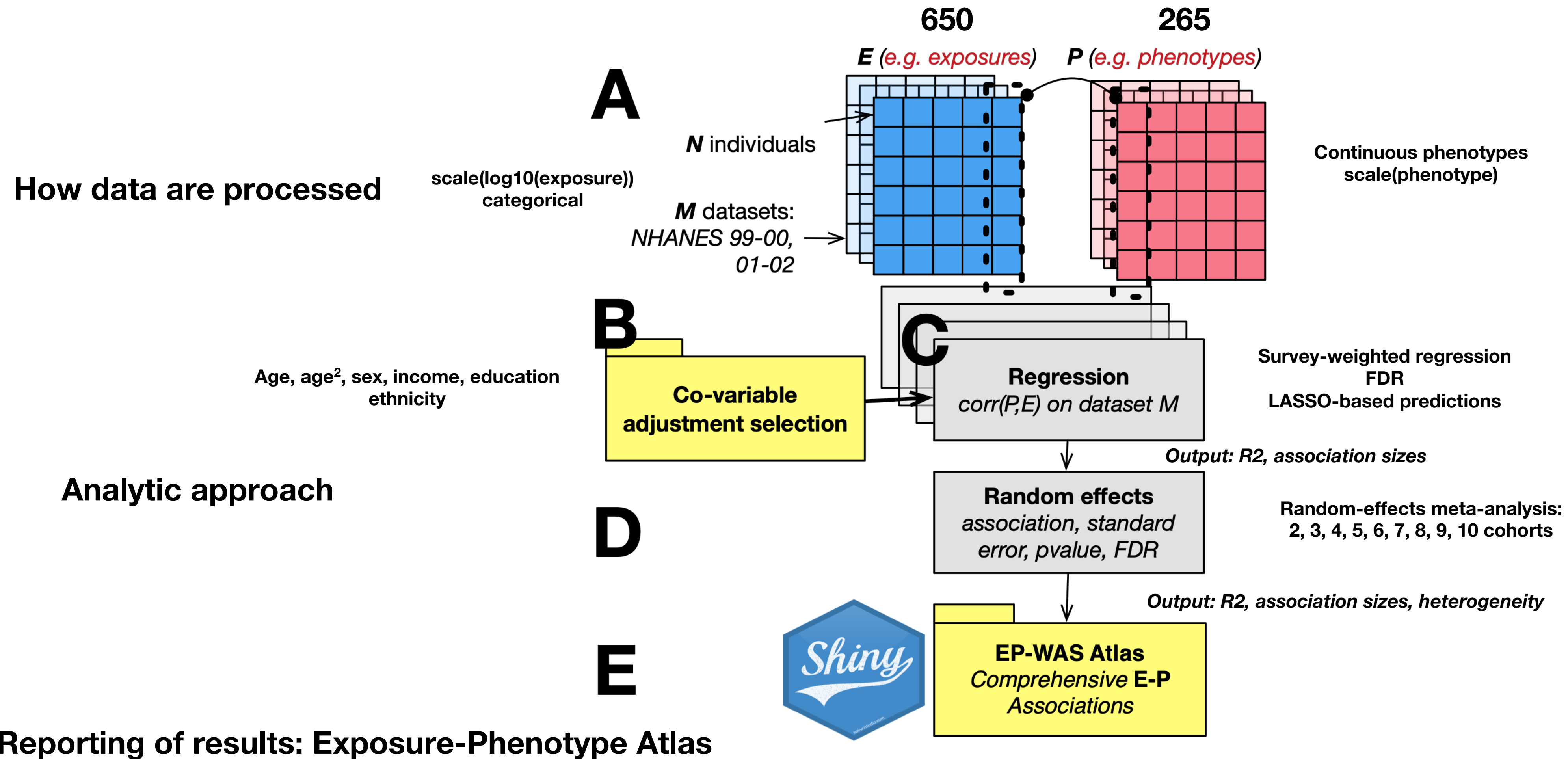
Recommendation		Examples
1	Catalog contributions of environmental exposures to disease risk (e.g., susceptibility, variance explained) to strengthen the case for exposome research.	Develop requirements for an exposome-disease association catalog.
2	Identify high-throughput (e.g., ‘omics, sensor-based) technologies and gaps to allow agnostic assessment of the exposome.	Develop infrastructure to characterize the variability of the exposome in various populations, akin to the NHANES.
3	Incentivize other parties (e.g., ‘omics investigators in other disciplines, funding institutions, industrial entities) to integrate the exposome in their programs and develop high-throughput analytics methods to analyze exposome data.	Develop big data analytics and visualization tools to accelerate exposome-related research (e.g., exposome–phenome association studies). Identify how existing ‘omics statistical methods can be extended for the exposome research and identify gaps for new method development. Encourage a shift in focus from “one exposure–one phenotype” to multiple exposures, genes, and phenotypes. Develop methods to link the internal and the external exposome. Develop methods to support varieties of study designs (e.g., longitudinal studies) to strengthen inference and credibility.

Figure 1. Key considerations for Exposome-Wide Association Studies.

... many analytic approaches to map *E-P* associations

# Benchmarking exposome-phenome relationships: *ExWAS* between 650 *E* & 265 *P* in US NHANES

## Grand total of ~400k E-P associations



Toward an “exposome atlas”: cataloging between exposures, processes, and clinical outcomes (e.g., “abstracting” Table 1 & 2 of published studies)

<u>Study Type</u>	<u>Exposure Factor</u>	<u>Method of Association</u>	<u>Phenotype</u>
Cross sectional	PM 2.5	Linear Regression	Forced expiratory volume
Case-control	PFAS	Logistic Regression	Body Mass Index
Sample size			C-Reactive Protein
<u>Inclusion criteria</u>	<u>Exposure Media</u>	<u>Association Type</u>	<u>Clinical Outcome</u>
Demographics	Geocode	Odds ratio	COPD
Location of study	Blood biomarker	Hazard Ratio	
	<u>Exposure Dose</u>	<u>Association Size and Error</u>	
	Per 10ug/m3	1.1 (0.001)	
	Mg/dL	10 (1.5)	

# Conclusions: digitizing the biological pathways phenomena between exposures and clinical outcomes

- Possible to put together existing resources to map between exposures and clinical outcomes
- However, to enhance triangulation of evidence, risk estimates are required
- A prerequisite for assimilating evidence includes documenting parameters around the study design and the association
- The ***exposome*** provides an opportunity to produce a “catalog” of benchmarks between exposures and biomarkers across experimental study design (e.g., tox and epi)
- Multi-modal AI approaches can introduce new ways of using text to refine knowledge between exposures and disease outcomes but need to be evaluated at scale



# The Environmental Health Language Collaborative

*Harmonizing Data, Connecting Knowledge, Improving Health*

---

## Presentation 4

Presentation Order	Presentation Title	Presenter, Organization
4	<i>Challenges and opportunities to improve communication about exposure and risk for collaboration and information exchange</i>	Elke Jensen, PhD, Dow Chemical Company <a href="mailto:elke.jensen@dow.com">elke.jensen@dow.com</a>



SYMPOSIUM: OVERCOMING BARRIERS TO MORE SCALABLE  
ENVIRONMENTAL HEALTH SCIENCE RESEARCH VIA HARMONIZED LANGUAGE

Challenges and opportunities to improve  
communication about exposure and risk for  
collaboration and information exchange

Elke Jensen, PhD, Dow Chemical Company  
SOT 2024 Salt Lake City, Utah

# DISCLAIMER AND COI

- *The content of this presentation is for information and discussion purposes only. This material is presented with the understanding that neither Dow nor the presenter are rendering legal, business or professional advice or opinion, and accordingly, Dow assumes no liability whatsoever in connection with use of the information presented herein. This presentation may not be reproduced without the express permission of the author.*
- *Dr. Jensen is employed by the Dow chemical Company, a manufacturer of chemicals and chemical products. No external compensation or financial interest was involved in the development of this presentation. Dr. Jensen has no conflicts of interest to declare.*

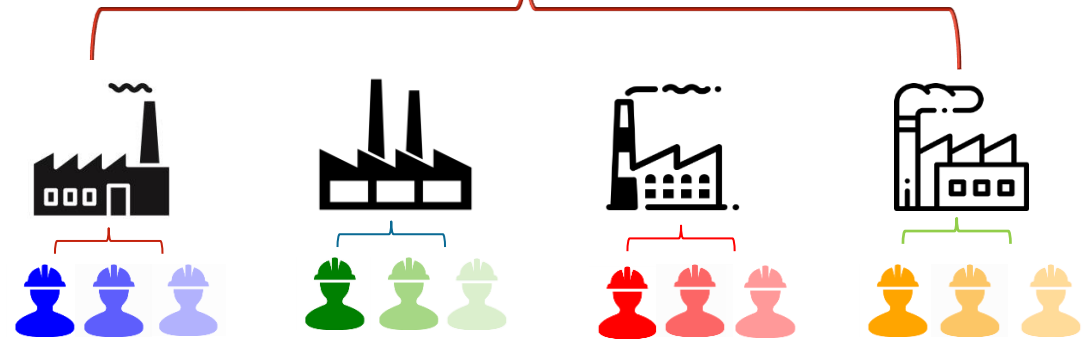
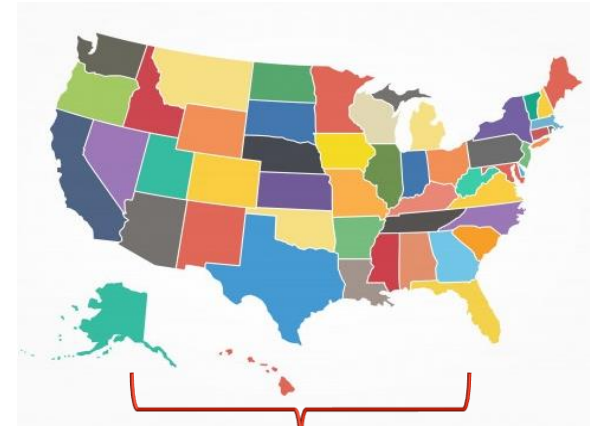
# ONE CHALLENGE FOR TSCA<sup>3</sup> RISK EVALUATION

RE must be general and  
broad and cover all COU.



IH is highly specific and  
difficult to generalize.

COU = conditions of use  
RE = risk evaluation  
IH = industrial hygiene



# EPA's Data Needs: Elements of Occupational Exposure Assessment

## Use Information

- ☐ End-Uses of Chemical Substance
- ☐ Life Cycle of Chemical Substance
  - Industries involving the chemical substance that are parts of the supply chains for the end-uses
  - Recycling operations
  - Disposal operations
- ☐ Production Volume Associated with Each Life Cycle Step

## Facility Information

- ☐ Process Description (including concentration)
- ☐ Operations Information
  - Days of operation per year
  - Worker activities
  - Number of sites
- ☐ Industrial Hygiene Information
  - Existing OELs
  - Physical form
  - Potential exposure routes, durations and frequencies
  - Engineering controls
  - Administrative controls
  - PPE
  - Number of potentially exposed workers

## Monitoring / Testing Information

- ☐ Inhalation Exposure Mass Concentration
  - Worker and ONU
  - Personal and area concentrations
  - TWA, short-term and peak values
  - Central tendency and high-end values
  - OES-specific or surrogate data
  - Exposure duration & frequency
- ☐ Dermal Applied Dose & Exposure Frequency
- ☐ Dermal Percent Absorption

## Modeling Information

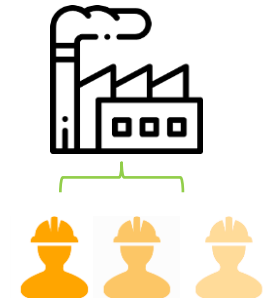
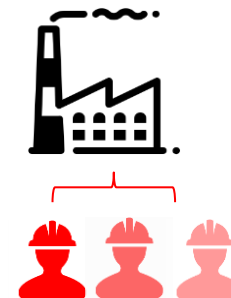
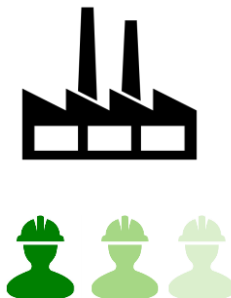
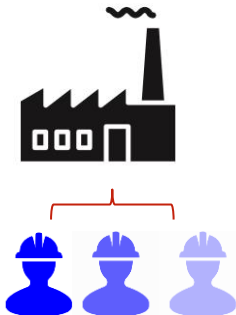
- ☐ Throughput of the Chemical
- ☐ Use Rate of the Chemical
- ☐ Emissions Rate
- ☐ Duration of Operation or Worker Activity
- ☐ Ventilation Rate
  - Exchange rate
  - Workspace volume
- ☐ Dermal Applied Dose and Percent Absorption

# WHY LANGUAGE MATTERS...

- ONU – Occupational Non-Users
  - New term introduced under TSCA
  - This term does not exist under OSHA
  - By-standers defined for plant protection (i.e., pesticides) but does not apply in industrial settings (either you're a worker or not)
- Who do we monitor?

# HYPOTHETICAL IH META DATA

	Company A	Company B	Company C	Company D
Employee	Engineer	Process engineer	Technician	Process engineer
Activity	Collect 4 oz samples	Sampling, 50 ml	Sampling, 1 L	Sampling, volume not specified
Sampling	Task monitoring	Task monitoring	Full shift monitoring	Full shift monitoring
Exposure modifiers	Not specified	10 minutes	2x per shift	Specified PPE, 5 minutes, 1/week
Engineering controls	Outdoors	Closed loop	Indoors Needle/septum	Outdoors, open jar



# HARMONIZATION ↔ COMMUNICATION

- Descriptors need to be well defined, mutually understood
- Meta-data need to be harmonized – especially for combining data sets, understanding aggregate and co-exposures
- [Industrial Hygiene Data Standardization \(aiha.org\)](http://aiha.org)

# LEVERAGING EXISTING EXPOSURE/MONITORING DATA

- Merging exposure data from different sources
  - Data collected for different purposes
  - Some existing sources but are organized NOT as centralized database platform rather but a distributed infrastructure (links to external holders of exposure data)
  - [IPCHeM Portal \(europa.eu\)](http://europa.eu)
  - [ECETOC heatDB](#)
  - ...

# MOVING FORWARD...

We need to speak the same language – have the same understanding of scenarios, activities, and other exposure descriptors

- Permit stakeholders to provide, generate data that is fit-for-purpose

More dialog between stakeholders

- Manufacturers

- Customers

- Regulatory agencies

Consistent approach to exposure assessment → better risk assessment and risk management

# WHAT MIGHT A TSCA PLAYBOOK LOOK LIKE?

Start collecting and generating information ASAP

Communication

- Define conditions of use
- Collect data and information for each COU
  - Products, concentrations, downstream uses / supply chain
  - IH monitoring data
  - Other reporting data: CDR, TRI, etc...
  - Emission controls
- What are best practices? For an enterprise? For an industry?

Communication

Communication

# SUMMARY

- To characterize risk properly, must understand exposure
- That means risk managers and risk assessors must understand each other
- *Mutual understanding of the exposure scenario details*
- *Common language and terminology*
- *Harmonized meta data*
- *Broader sharing of data **in context***

# THANK YOU

- Co-panelists
- SOT
- Dow colleagues
- YOU



The Environmental Health Language Collaborative

*Harmonizing Data, Connecting Knowledge, Improving Health*

---

Questions related to these presentations?

Reach out to: **[EHLC@icf.com](mailto:EHLC@icf.com)**