# Project 1: Predicting the Emergence of 'Chemicals of Interest' via the Study of Scientific Publications

*CEER Technical Report*

Jason M. Whyte[1,2]

[1]The Centre for Environmental and Economic Research (CEER), School of BioSciences, University of Melbourne
[2]Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), School of Mathematics and Statistics, University of Melbourne

January 29, 2021

# Contents

# List of Figures

# List of Tables

# Acknowledgements

We gratefully acknowledge Clarivate for providing access to the Web of Science Lite API. This research was supported by Melbourne Research Cloud and by use of the Nectar Research Cloud. The Nectar Research Cloud is a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy (NCRIS).

# Notices

This report is a product of the Centre for Environmental and Economic Research (CEER) at the University of Melbourne. The author(s) and/or the University assumes no responsibility or liability for any errors or omissions in the content of this publication. In no event will the author(s) and/or the University be liable to you or anyone else for any decision made or action taken in reliance on the information contained in this publication or for any consequential, special or similar damages, even if advised of the possibility of such damages.

# 1. Executive summary

Modern economies employ a large volume and variety of chemicals. Yet, regulators typically have limited resources for conducting risk assessments. Further, the data available for chemical assessments is often insufficient. Consequently, chemicals in various jurisdictions (such as the per- and poly-fluoroalkyl substances, PFAS) have entered into widespread use, at which point Adverse Effects (AEs, including environmental impacts and health risks) are recognised. A regulator must then ′reactively′ manage chemical risks and impacts.

The difficulties of reactive management may encourage progress towards a ′proactive′ approach, seeking to anticipate those chemicals most likely to cause AEs. This should allow a regulator to direct its efforts towards assessing the "Chemicals of Interest" (CoI) recognised. Ideally, a regulator′s targeted actions would lead to avoidance of future AEs, improving environmental and health outcomes.

The New South Wales Environment Protection Authority (NSW EPA) has an ongoing interest in improving its capacity to proactively manage AEs due to chemicals. NSW EPA invited the Centre for Environmental Research (CEER) to consider how to anticipate which chemicals are likely to require future regulation. NSW EPA supplied a list of chemicals (or subfamilies of chemicals, e.g. Octachlorinated furans). NSW EPA assigned each of these chemicals to a particular group: Group 1, 2, or 3. Briefly, if a chemical has a history of concern (and possibly regulation) for at least one regulatory body (from a list supplied by NSW EPA, including regulators outside of Australia), it is labelled as Group 1. A Group 3 chemical is not a CoI under current use conditions.

CEER approached the task via a consideration of research interest (in scientific publications) associated with supplied chemicals. We sought to investigate a hypothesis:

**Hypothesis 1** *Research interest for chemicals previously assigned to Groups 1 and 3 may contain characteristic features. By recognising these features, we may use their presence in the research interest of ungrouped chemicals in deciding which are likely to become future CoI.*

This pilot project has two main parts. The first relates to data collection. We obtained data by applying two types of queries to Web of Science (WoS) over the publication range from 1980–2019 by modifying CEER′s R code previously developed for a similar task. We were able to consider the content of titles, abstracts, and keywords of scientific publications to determine matching documents for queries (′hits′). The first query type sought documents containing any mention of a chemical (or its alternative names) in the publication range (′total-hits′). The second query type (for ′near-hits′) was more selective. This returned any hits where a name variant was within three words of at least one ′modifier′ (from a list supplied by NSW EPA) relating to AEs. We found near-hits appropriate for our purposes.

The second part of the project relates to the classification of chemicals based on their research interest. We sought a classification system that could use a chemical′s near-hits time series in classifying a chemical as CoI, or not. The system has two parts. The first is a collection of tests, each associated with an outcome: the year in which the test is satisfied, or a non-result otherwise. Outcomes allow us to judge the ability of tests to provide a regulator with timely advice.

Tests are designed to be sequentially applied to subsets of the data starting from 1980; the entire data set from 1980–2019 is not required initially. (To illustrate this, if we were applying a test to the near-hits series available for 1980–2000, we do not act as if we can "see the future" and access data up to 2019.) This feature mimics the situation where a regulator would assess the literature available up to some point in time, and would repeat this assessment in subsequent years in order to utilise whatever data has become available since the last assessment.

We describe the question asked of the data by each test (broadly for those tests which have quite technical conditions) below.

| Test | Summary |
|---|---|
| Test 1 | **Sustained initial interest:** Starting from 1980, is there any three-year interval for which the near-hits meets or exceeds five hits in each year? |
| Test 2 | **Volume of interest:** Starting from 1980, does the cumulative sum of near-hits reach at least 150? |
| Test 3 | **Year of substantial interest:** Starting from 1980, is there a year in which the near-hits reaches at least 10? |
| Test 4 | **Abrupt increase in interest:** Starting from 1981, is there some year $t$ in which the increase in near-hits counts from years $t-1$ to $t$ is at least 5? |
| Test 5 | **Sustained increase in interest:** disregard any part of the near-hits data which occurs before the first non-zero value. Transform the retained data. Does the slope of a line of best fit applied to the transformed data from an interval of four consecutive years exceed some minimum (positive) threshold? |
| Test 6 | **Well-separated times of interest growth:** it is necessary to transform the near-hits series so as trends are more readily apparent. Does the transformed series show at least some specified level of growth within an interval of years, and then greater growth in some subsequent interval? |

The second part of the system is a classification rule. This uses a chemical's test outcomes in classifying that chemical.

We applied our system to the ("training set" of) near-hits time series of approximately half of the chemicals belonging to each of Group 1 and Group 3. In "system training", we adjusted system features to produce an acceptable fraction of correct classifications. We made this judgement by comparing classifications against chemicals' regulatory history, supplied by NSW EPA.

We simulated the use of a mature version of our system by applying our trained system to the research interest of the remaining chemicals (the "validation set"). System performance is reasonable given the small data sets available.

These results provide some support for Hypothesis 1. Our investigations have suggested aspects of our methods that will benefit from further refinement. For example, we may further investigate types of queries so as to improve our ability to find publications which relate chemicals to adverse effects. This may improve our ability to distinguish between groups. Similarly, exploration of a larger data set is likely to illuminate features of research interest that we may use to improve the performance of our classification system.

# 2. Introduction

We begin this chapter with an overview of the problem we seek to address in Section 2.1, which also includes a summary of our work on similar problems to date. In Section 2.2 we present a summary of our approach to the current problem, including an introduction to key features of necessary inputs. In Section 2.3 we present the chemicals (or subfamilies) we shall consider, and provide a summary of the remainder of this report.

## 2.1. Overview

Owing to modern industrial and agricultural practices, many developed countries now employ a large volume and variety of chemicals. This is demonstrated by a public inventory of over 40 000 chemicals available for Australian industrial use (National Industrial Chemicals Notification and Assessment Scheme, n.d.). The need to monitor a large number of chemicals can pose challenges to a regulatory body. Due to limited resources, a regulator may not be able to conduct the necessary risk assessments for all new chemicals. Further, the task of monitoring a chemical does not necessarily conclude with the conduct of a risk assessment. New information on a chemical may become available, which may oblige a regulator to revise an earlier risk assessment, further adding to the regulator's workload. As a result of the varied and ongoing demands placed on regulators, chemicals with harmful properties may enter into use. This may have effects including harm to human or environmental health, loss of environmental amenity, expensive environmental remediation, and a reduction of public confidence in the ability of regulators to perform their function.

One prominent example of unanticipated risk that caused health concerns is shown by the family of per- and poly-fluoroalkyl substances (PFAS). Of this group, PFOS, and later PFAS, to take just two examples, have entered into the public consciousness due to publicised investigations relating the chemicals to adverse health effects, and prominent public health campaigns. To consider PFOS in particular, the chemical found use in a variety of industries, and entered into widespread use. PFOS proceeded relatively quickly from an unregulated status in Australia (National Industrial Chemicals Notification and Assessment Scheme, 2013) (and elsewhere), to being of concern, to being actively managed. This management included the recommendation that PFOS-based chemicals should be restricted to essential uses only. Guidelines for PFOS management are not yet finalised. This, combined with the concentration of PFOS in certain sites, has caused disruptions to infrastructure projects (as seen for the West Gate Tunnel project in Melbourne, Victoria) due to concerns around how to treat and dispose of contaminated soil (Jacks & Hatch, 2020).

The PFAS case demonstrates the limitations of taking management actions after a risk has been realised, which we may call a 'reactive' approach. Regulators would prefer to anticipate the emergence of chemicals which have (at least) the potential for adverse effects on human and environmental health. If a regulator could recognise such chemicals, it could prioritise these for further consideration. For example, a regulator could consider each chemical's possible adverse effects alongside its prevalence in (or potential to enter) the environment in ascribing a measure of risk to each chemical. Such risk assessments may allow the regulator to recognise emerging risks, and hence, may assist the regulator in applying necessary management actions in a timely manner.

However, such a ′proactive′ approach may be impeded when a regulator has an inadequate amount of information available. To take the example of PFAS:

> The majority of the many thousands of PFAS, including those in commercial use, have very limited or no toxicity data. This is a critical data gap in health effects information for PFAS. (Interstate Technology & Regulatory Council, PFAS Team, 2020, Page 97)

### 2.1.1. Project history

The New South Wales Environment Protection Authority (NSW EPA) has an ongoing interest in improving its capacity to proactively manage environmental impacts and health risks due to chemicals. This motivated NSW EPA to engage the Centre of Excellence for Biosecurity Risk Analysis (CEBRA)[1] at The University of Melbourne in 2019 in order to explore approaches to the problem.

CEBRA conducted a limited study, and the resulting report (Whyte & Robinson, 2020) sought to use "research interest" relating to a chemical in the scientific literature as a proxy for the concern of the scientific community towards that chemical. The report considered research interest for a small number of chemicals supplied by NSW EPA.[2] NSW EPA also supplied a list of alternative names for these chemicals. We determined that the database Web of Science (WoS) was suitable for our purposes. Briefly, the study obtained research interest by a “shallow scrape” of a WoS application programming interface (API) — the WoS API Lite. The shallow scrape involved a single chemical, and applied a query to the database for each (calendar) year in a nominated publication range. The query considered the content of publication titles, abstracts, and keywords. (The API Lite does not permit access to the full text of publications.) In order for a publication to register as a "hit", any of its associated features must contain at least one of the chemical name or its alternative names.

Query results allowed us to obtain counts of hits for a given year directly, and did not attempt to scrape any further bibliographic information. By obtaining hits over each year in a publication range for a given type of query, the report obtained the time series of research interest for each of a number of particular chemicals. Mostly the report sought to use a limited research interest series (five years) to predict research interest in chemicals in the near future. The project was largely successful in this task. Also, the report demonstrated the feasibility of using queries implemented in computer code to reproducibly access bibliographic information from a scientific publication database.

In response to a late draft of the report, NSW EPA made a new request. NSW EPA asked if it were possible to detect "emerging concern" (which we now term "emerging

[1] The author of this report was also the lead author of the CEBRA report.

[2] We shall only summarise those features of (Whyte & Robinson, 2020) which provide useful context for this related work. We direct the interested reader to the original report for details.

interest") relating to a chemical within its research interest over the studied publication range (from 1988-2018). We accommodated this request by adding new content (Section 6) to the report (Whyte & Robinson, 2020). Briefly, the new section explored a new use of our shallow scrape. We theorised that emerging interest relating to a chemical could appear as a sharp increase in research interest at a particular time point, after which research interest is sustained for some years. We sought to test this theory against chemicals which had a regulatory history in Australia. This choice allowed us to judge whether or not one could detect emerging interest in advance of Australian management actions. A positive response would demonstrate the value of persisting with our approach.

There was limited time available before submission of the final report. Hence, there was a need to develop and test new computer code, and choose chemicals with an appropriate Australian regulatory history, quite quickly. As such, there was time to consider only two chemicals. We selected PFOS and PFOA as suitable test chemicals. Results showed that, for both chemicals, one could indeed detect emerging interest in advance of management actions. Based on this, we judged our approach as having potential to assist the proactive management of risk due to chemicals. However, we noted that further exploration and validation of our ideas was required in order to produce a robust and reliable method for detecting emerging interest. We recommended pursuing this by application of the approach to a broader range of chemicals, from a range of families, in any subsequent study.

### 2.1.2. Recent developments

NSW EPA and CEBRA/CEER agreed that it is valuable to subject ideas developed for projects (or considered for future project work) to peer review. CEER has pursued this following submission of the final version of Whyte & Robinson (2020) to NSW EPA in January 2020. On February 15th 2020, we submitted an abstract to the International Environmental Modelling and Software Society Conference 2020. This abstract described an extension of the approach of Whyte & Robinson (2020) to detecting emerging interest in a research interest time series. A favourable reception by all three reviewers prompted the submission of a paper for peer review (Whyte, 2020). The paper (now accepted for publication) considered the detection of emerging interest for six PFAS. Judgements made of the chemicals were compared with the actions of regulators. Regulatory histories were drawn from reports by the National Industrial Chemicals Notification and Assessment Scheme (NICNAS), and other jurisdictions where necessary.

In preparation for paper writing, we strategically addressed issues that we anticipated would benefit further projects undertaken for NSW EPA (such as that we report on here). These include:

1. Gaining an increased familiarity with the conditions of Clarivate's Web of Science API Lite. This has enabled code optimisation, and hence, substantial reductions in the runtime required to obtain information from WoS. Accordingly, we can now consider a greater number of chemicals in a reasonable runtime.
2. Negotiating access to, and appropriate resources on, a cloud server. This has further improved the efficiency of our computer code. Of particular value is the faster internet transfer speed available on the cloud server. This has proven especially useful given our inability to access the University of Melbourne network

whilst being forced to work from home under the University's COVID-19 management plan.

3. A thorough investigation of sources of alternative names and identifiers for chemicals of interest. (Henceforth we refer to these as "synonyms".) This substantially expanded upon the use of alternative names in Whyte & Robinson (2020), and gives our process the potential to recognise more of the research interest associated with chemicals than we could previously. (We describe these chemical synonyms, which also include various types of Chemical Abstracts Service Registry Numbers, further in Section 3.1.2.)
4. An exploration of how to discern emerging interest in a time series of hits. We developed code for some sample tests. (We develop these further in Section 5.4.) These tests produced promising results. In turn, we were able to explore appropriate methods for interpreting and presenting results.

We shall see the benefits of these investigations later in this report.

### 2.1.3. Extensions sought in this project

Following discussion between NSW EPA and CEER, we agreed the main project goal:

> NSW EPA is seeking a way to screen chemicals to identify those where there is sufficient research interest into the potential for the chemical to cause environmental / health issues to prompt the question of "does this chemical need stricter regulation?" (email from NSW EPA, June 9 2020)

Henceforth, we refer to any such chemicals as 'chemicals of interest', or CoI (which may also indicate a single chemical, as necessary).

The NSW Government (Department of Planning, Industry and Environment) document entitled "Part D — Statement of Requirements" and dated April 20th 2020, listed the project objectives:

**Project Objective 1** To establish a robust algorithm for searching database(s) to identify and quantify research literature on chemicals and associations with adverse effects / harm.

**Project Objective 2** To determine a decision rule which can be used to identify emerging chemicals of concern.

**Project Objective 3** To validate the decision rule using chemicals that have previously emerged, are currently emerging or are unlikely to emerge.

**Remark 1** *As the project evolved, CEER became aware of inconsistency around the meaning of "currently emerging" chemicals. We informed NSW EPA that this made it impractical to consider such chemicals at this stage.*

## 2.2. Overview of inputs, methodology, and guiding principles

### 2.2.1. Inputs

Our methodology depends on various inputs. We shall begin with a discussion of those inputs requested in broad terms by CEER, refined and supplied by NSW EPA.

**Input 1** the chemicals (for brevity, henceforth we may use this to denote individual chemicals, as well as families[3] and subfamilies[4] of chemicals) for which we will obtain research interest.

CEER requested that Input 1 include some CoI, and some chemicals not of interest, so that we may attempt to ascertain features of research interest for each group. We intend to exploit these features later in devising our classification system.

**Input 2** some means of discerning between those chemicals or subfamilies supplied in Input 1 that are CoI, and those that are not.

In particular, NSW EPA allocated each supplied individual chemical or subfamily to one of three groups, indicating a perception of the current associated (environmental and/or health) risk. (Families of chemicals were not allocated to groups.) These groups are:[5]

**Group 1** Chemicals having a regulatory history such that previously unidentified risks surfaced and required regulatory intervention to protect people and the environment. (NSW EPA has supplied a regulatory history for these.)

**Group 2** Chemicals where there is evidence to suggest that the risks to people or the environment are higher than previously thought and regulatory intervention may be required.

**Group 3** Chemicals that have recently undergone rigorous assessment and/or are generally considered to be safe according to the regulator's risk appetite, and under current usage conditions.

Regarding Group 1 chemicals, NSW EPA noted that it was appropriate

> ...to ensure we use a representative set of chemicals, particularly in establishing the candidate decision rules, so that it's not biased towards one pathway of emergence. (NSW EPA, April 9 2020)

**Remark 2** *We note that, at some point in time, certain chemicals may be of concern to a regulator, only to have this concern dismissed later. It is inappropriate to give such chemicals a Group 1 label. We would prefer all chemicals labelled as Group 1 to have concern that is supported by subsequent action. However, we do not observe this in all cases.*

**Input 3** information on the activities of regulators towards each of the Group 1 chemicals or subfamilies supplied in Input 1.

We can illuminate the role of Input 3 by first considering the features we require in our classification system. We expect that, in general, the earlier some chemical is correctly judged as being of concern, the greater the judgement's value to a regulator. That is, the classification system extends the regulator's ability to anticipate potential

[3] We use the term 'family' to refer to related chemicals which share particular properties or functional groups that confer a characteristic feature (e.g. dioxins and furans, neonicotinoids).

[4] We use the term 'subfamily' to refer to chemicals that occur as heterogeneous mixtures of closely related compounds (e.g. isomers, congeners etc). For example, hexachlorodibenzo-p-dioxin is a subfamily of the dioxin family.

[5] Supplied by NSW EPA on June 3 2020.

risk. This anticipation may empower a regulator to investigate CoI and implement appropriate regulation in a timely manner so as to manage risks.

The benefits of timely regulation shaped our thinking around Input 3 and its use for each chemical. We agreed that it was appropriate to consider the year in which regulators (drawn from list having international membership, proposed by NSW EPA) first evinced concern that a chemical was associated with adverse health effects. More particularly, NSW EPA advised that:

> "Concern" indicates there is documented evidence that at least one key jurisdiction (of the EU, US, Canada, Australia, and international) was taking steps to evaluate the severity and significance of newly identified risks of the chemical. (NSW EPA email June 11 2020)

NSW EPA has provided the earliest of the years in which the key regulators recorded concern — 'International Earliest Concern' (IEC). IEC provides one means of evaluating our classification system.

We note that there are other ways in which we could judge the usefulness of our classification scheme. For example, we could consider when regulators applied regulation to a chemical. More specifically:

> "Regulation" indicates there is documented evidence that steps were taken to restrict or prohibit the use of the chemical. Note that this includes both legislative changes and voluntary action (by industry). (NSW EPA email, July 16 2020)

Thus, in an analogous manner to IEC, we may also use 'International Earliest Regulation' (IER) in evaluating our classification system. Further, we may observe the performance of our system by comparison with Australian equivalents of IEC and IER: AEC and AER, respectively.

**Remark 3** *Australia is included in the list of NSW EPA's "key jurisdictions". Thus, in considering some chemical, by definition the IEC (respectively, IER) will either precede or equal the corresponding AEC (respectively, AER). As such, given our interest in timeliness, a comparison of our results against IEC provides a more stringent test of our system than does comparison against IER, AEC, or AER.*

**Input 4** A list of modifiers that we may suitably adapt for use in 'near-hits' queries.[6]

The final input was developed by CEER.

**Input 5** A list of synonyms for each of the Group 1 and Group 3 chemicals supplied.[7]

### 2.2.2. Methodology

We shall address Project Objective 1 by building on earlier studies by CEBRA/CEER. Broadly, we intend to refine our previous approaches to obtaining research interest data in a "data acquisition" phase. We illustrate this with a schematic in Figure 2.1.[8] For

[6] Rather than introduce a substantial volume of technicalities here, we provide a detailed discussion of the modifiers in our discussion of queries in Section 3.2.

[7] We shall explain our approach to obtaining synonyms in Section 3.1.

[8] We provide a detailed discussion of the query types we employ in Section 3.2.

each of the Group 1 and Group 3 chemicals, we obtain measures of research interest over the publication range from 1980 to 2019.

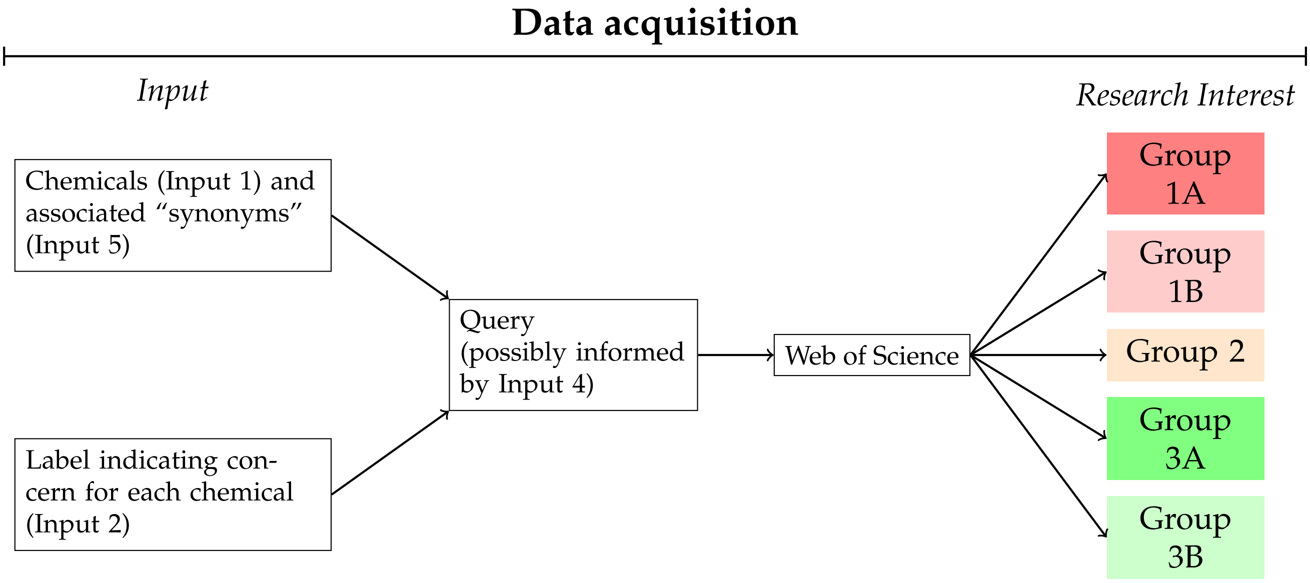

Data acquisition
Input
Research Interest
Chemicals (Input 1) and associated "synonyms" (Input 5)
Label indicating concern for each chemical (Input 2)
Query (possibly informed by Input 4)
Web of Science
Group 1A
Group 1B
Group 2
Group 3A
Group 3B

**Figure 2.1.:** A schematic of our process of acquiring research interest.

Given research interest data, we pursue Project Objective 2 by developing methods for recognising features in this data that may suggest emerging interest. We test the usefulness of these features by using them as inputs to a classification system, intending to classify chemicals as being CoI, or alternatively, not CoI.

We address Project Objective 3 by validating the accuracy (alternatively, timeliness) of these judgements subsequently by determining whether or not they are consistent with (alternatively, precede) the activities of regulators.

In preparation for devising a classification system, Group 1 and Group 3 were divided into groups of roughly equal size with distinct elements; Groups 1A and 1B, and Groups 3A and 3B, respectively. We shall refer to the collection of research interest for Group 1A and Group 3A as our "training set". We will use this in developing methods for discerning 'emerging interest' in the research interest associated with chemicals. We will achieve this by designing particular tests, oriented towards achieving certain results. The 'outcome' of a test will be the year in our studied publication range in which a test is satisfied, or a non-result. We interpret outcomes by use of a 'classification rule' that decides whether or not a chemical is CoI.

We expect that a useful classification system can reliably distinguish between Group 1A chemicals and Group 3A chemicals. As an initial means of evaluating the usefulness of our system, we shall compare its outcomes for each of the Group 1A chemicals against the relevant IEC. A useful system should produce a year that leads (that is, occurs before) the year of regulatory concern or action in the majority of cases.

Having shown that our classification system can exhibit a reasonable ability to correctly classify chemicals, we proceed to further evaluate the usefulness of our system. We achieve this by ascertaining the system's ability to anticipate the IEC for a "validation set" composed of the research interest for Group 1B and Group 3B chemicals. We derive some confidence in the general usefulness of our methodology if most of our judgements for Group 1B chemicals here also precede the relevant IEC. We apply a similar analysis by comparing classifications against IER, AEC, and AER.

We summarise the process of developing and applying a classification system in Figure 2.2.

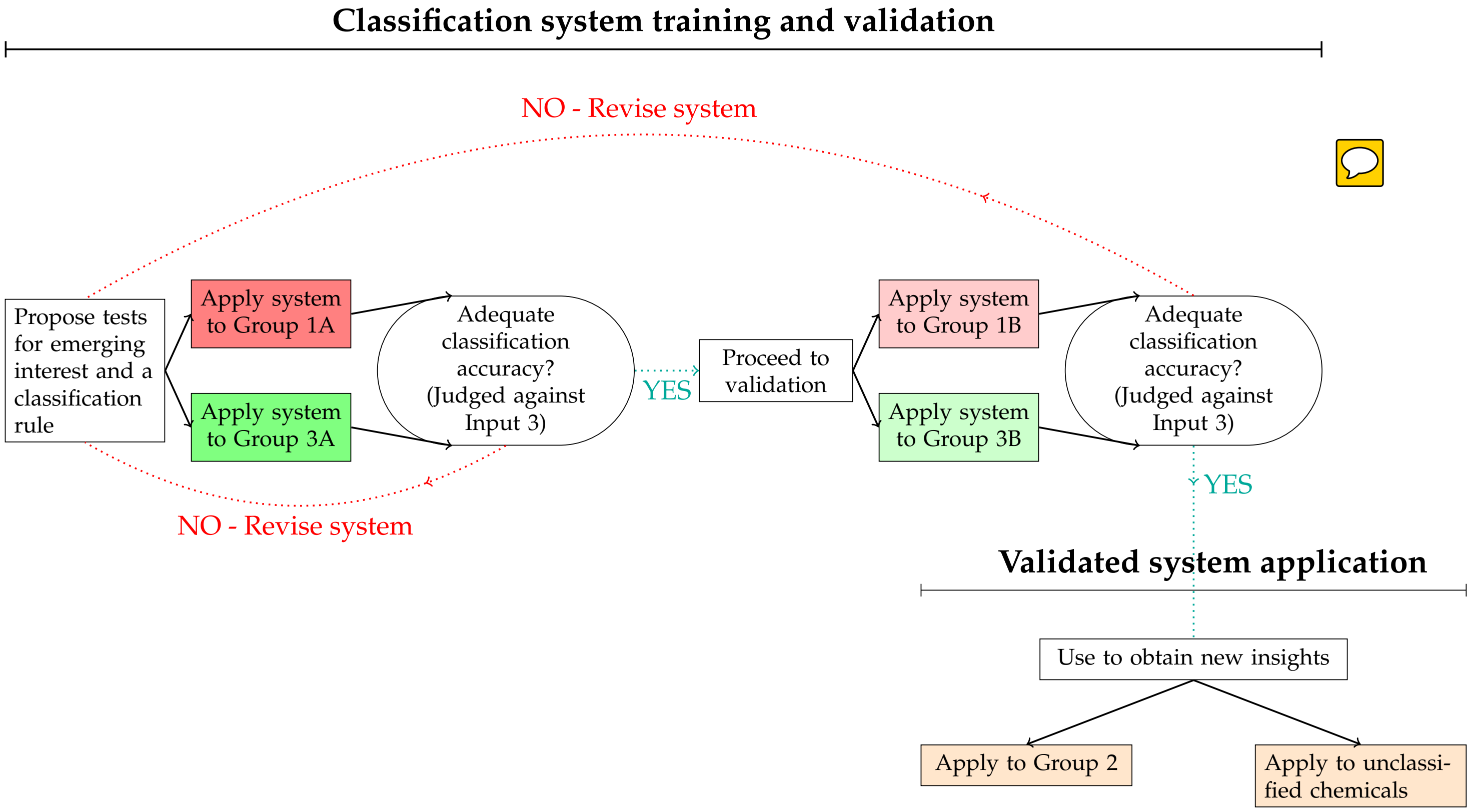

Classification system training and validation
NO - Revise system
Propose tests for emerging interest and a classification rule
Apply system to Group 1A
Apply system to Group 3A
Adequate classification accuracy? (Judged against Input 3)
YES
Proceed to validation
Apply system to Group 1B
Apply system to Group 3B
Adequate classification accuracy? (Judged against Input 3)
YES
NO - Revise system
Validated system application
Use to obtain new insights
Apply to Group 2
Apply to unclassified chemicals

**Figure 2.2.:** A schematic of our process of training and validating a classification system, and the subsequent use of a system judged as acceptably accurate.

### 2.2.3. Guiding principles

In undertaking this project, we aspire to develop an approach that respects certain key principles:

**Accuracy** It is appropriate to balance the correct judgement of CoI against the need to avoid excessively labelling chemicals as of concern when they are not. (We shall consider this further in Chapter 5.)

**Modularity** Code is constructed in a modular manner, enabling variation to a distinct section (say, queries) as requirements change, or as new information is acquired.

**Reproducibility** As much as possible, we seek to use computer code to automate the extraction of research interest and its subsequent processing. By doing so, we intend to create a robust process for recognising CoI that can be applied in the future with minimal need for modification. This minimizes the opportunity for errors to arise from manual processing.

**Timeliness** Insights gained from our system should provide a regulator with adequate time to act so as to mitigate the adverse effects of CoI.

## 2.3. Chemicals considered in this study

Towards Input 1 and Input 2, NSW EPA provided CEER with a list of chemicals belonging to each of Group 1A (14 chemicals), Group 1B (16 members; 12 chemicals, 4 subfamilies), Group 2 (20 chemicals), and Group 3 (14 chemicals). (We subsequently divided the Group 3 chemicals evenly between Group 3A and Group 3B.) Additionally, NSW EPA requested a secondary investigation of chemical families to which our studied chemicals or subfamilies belong.[9] These families are: Chlorinated Ethenes, Dioxins and Furans, Neonicotinoids, PBDE, PFAS, and Phthalates. Families are not used in the development of our classification systems.

We show chemicals (by their group, recall the definitions on Page 9), with their Chemical Abstracts Service Registry Number® (CAS RN®)[10] in Tables 2.1 (Group 1A), 2.2 (Group 1B), 2.4 (Group 3A), and 2.5 (Group 3B). (As a consequence of Remark 1, we did not study the Group 2 chemicals, shown in Table 2.3.) As a chemical subfamily does not have a CAS RN®, we use the subfamily's acronym as a short label instead. (This action prevents our processing code from producing an error when it encounters a blank field.)

**Table 2.1.:** Names and Chemical Abstracts Service Registry Numbers® (CAS RN®s) of Group 1A chemicals considered in this study. As TCDD can refer to an individual chemical or the family that contains this (see Table 2.2), henceforth we refer to the chemical TCDD shown in this table by one of its other synonyms: TCDBD.

| Name | CASRN |
|---|---|
| 2,2′,3,3′,4,4′,5,5′,6,6′-Decabromodiphenyl ether | 1163-19-5 |
| 2,2′,4,4′-Tetrabromodiphenyl ether (BDE 47) | 5436-43-1 |
| 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD/TCDBD) | 1746-01-6 |
| Bisphenol A | 80-05-7 |
| DEHP - di-(2-ethylhexyl) phthalate | 117-81-7 |
| Dihexyl phthalate | 84-75-3 |
| Endosulfan | 115-29-7 |
| Imidacloprid | 138261-41-3 |
| Musk xylene | 81-15-2 |
| PCDD - polychlorinated dibenzo-p-dioxin | PCDD |
| PFOA - Perfluorooctanoic acid | 335-67-1 |
| PFOS - Perfluorooctane sulfonic acid | 1763-23-1 |
| TCE – trichloroethylene | 79-01-6 |
| Thiamethoxam | 153719-23-4 |

[9] Given some issues which became apparent over the course of this study (see, for example, Section 4.2), it was not appropriate to conduct this investigation.

[10] We note that a chemical may also have "alternative" or "deleted" CAS RN®s.

**Table 2.2.:** Names and CAS RN$^{®}$s of Group 1B chemicals considered in this study.

| Name | CASRN |
|---|---|
| 1,1,2,2-tetrachloroethane | 79-34-5 |
| 1,2,5,6,9,10-Hexabromocyclododecane | 3194-55-6 |
| 2,2′,4,4′,5-pentabromodiphenyl ether (BDE-99) | 60348-60-9 |
| Clothianidin | 210880-92-5 |
| Dicofol | 115-32-2 |
| diethyl phthalate | 84-66-2 |
| Diisobutyl phthalate | 84-69-5 |
| Diisopentyl phthalate | 605-50-5 |
| Hexabromocyclododecane | 25637-99-4 |
| HpCDD - heptachlorinated dibenzo-p-dioxin | HpCDD |
| HxCDF - hexachlorinated dibenzofuran | HxCDF |
| Methylmercury (MeHg) | 22967-92-6 |
| OCDF - octachlorinated dibenzofuran | OCDF |
| PCE - tetrachloroethylene | 127-18-4 |
| PFNA - Perfluorononanoic acid | 375-95-1 |
| TCDD - tetrachlorinated dibenzo-p-dioxin | TCDD |

**Table 2.3.:** Names and CAS RN$^{®}$s of Group 2 chemicals supplied by NSW EPA.

| Name | CASRN |
|---|---|
| 1,3-butadiene | 106-99-0 |
| 1,4-dioxane | 123-91-1 |
| 2,4,4′-tribromodiphenylether | 41318-75-6 |
| 8:2 FTS - 8:2 Fluorotelomer sulfonic acid | 39108-34-4 |
| Acetamiprid | 135410-20-7 |
| antimony | 7440-36-0 |
| Carbamazepine | 298-46-4 |
| Ciprofloxacin | 85721-33-1 |
| Diclophenac | 15307-86-5 |
| disodium tetraborate, anhydrous | 1330-43-4 |
| Fluoroborate | 14874-70-5 |
| Galaxolide | 1222-05-5 |
| Glyphosate | 1071-83-6 |
| MTBE - Methyl tert-butyl ether | 1634-04-4 |
| PFBS - perfluorobutane sulfonate | 375-73-5 |
| PFHxA - perfluorohexanoic acid | 307-24-4 |
| PFHxS – Perfluorohexane sulfonic acid | 355-46-4 |
| TCC - Triclocarban | 101-20-2 |
| Tetrafluoroboric acid | 16872-11-0 |
| Vinyl Chloride / chloroethylene | 75-01-4 |

**Table 2.4.:** Names and CAS RN®s of Group 3A chemicals considered in this study.

| Name | CASRN |
|---|---|
| 1,3-Propanediol | 504-63-2 |
| Caffeine | 58-08-2 |
| Calcium chloride, hexahydrate | 7774-34-7 |
| Carbon dioxide | 124-38-9 |
| Ethinyl estradiol | 57-63-6 |
| Lithium | 7439-93-2 |
| Lithium 12-hydroxystearate | 7620-77-1 |

**Table 2.5.:** Names and CAS RN®s of Group 3B chemicals considered in this study.

| Name | CASRN |
|---|---|
| 2-Mercaptoethanol | 60-24-2 |
| ethyl acetate | 141-78-6 |
| hydrogen peroxide | 7722-84-1 |
| iron(III) chloride | 7705-08-0 |
| isopropanol | 67-63-0 |
| Nonanoic acid | 112-05-0 |
| sodium chloride | 7647-14-5 |

⋆ ∞ ⋆

The remainder of this report is organised as follows. In Chapter 3 we present preliminary information which informs our approach to harvesting research interest for the supplied chemicals. In Chapter 4 we present graphs of the time series of hits for chemicals arranged by group. We also make observations on features of research interest of groups and differences between groups. We draw on these insights in Chapter 5, where we define our tests seeking to recognise emerging interest in research interest, and the classification system we use to interpret the test outcomes. Results from the application of our system to near-hits data are presented in Chapter 6. We draw conclusions and make recommendations in Chapter 7.

We consider some supplementary matters in our appendices. Following a request by NSW EPA, in Appendix A we show the time series of hits for Group 1 and Group 3 chemicals (which, unlike Group 2, were well-defined groups) by family (where families were supplied by NSW EPA).[11] Appendix B presents tables to provide an alternative view of results presented in figures elsewhere in this report. In Appendix C we show the results of a sample code validation exercise that demonstrates how our R code can obtain accurate values of research interest from WoS.

[11] These graphs are presented for illustrative purposes only. Our preference is to classify chemicals based on properties within groups, rather than within families. Given some matters discovered in this project, we do not recommend the use of either total-hits or near-hits time series in drawing broad conclusions around chemical families considered here, a task further complicated by the study of only a small number of chemicals for each family.

# 3. Preliminaries

We begin this chapter with an overview of features of data sources that led to their selection in Section 3.1. We explain features of the queries that utilise this data in Section 3.2. Following experimentation with the application of our queries, Section 3.3 notes some limitations of this study.

## 3.1. Data source considerations

### 3.1.1. Choice of a database of scientific publications

In preparation for the project that led to Whyte & Robinson (2020), we sought a source of publication data that we expected would:

- have a broad coverage of the scientific literature that we expected to be relevant,
- provide a flexible search syntax that would allow us to experiment with various types of queries,
- allow us to scrape a significant amount of bibliographic information in a manner that could be largely automated,
- allow us to readily conduct validation of the results obtained by our code against results obtained directly from some other method (e.g. a web browser) in a manner that did not require time-consuming manual processing.

For the previous project, we were granted access by Clarivate to the Web of Science (WoS) API Lite. This allowed us to pursue the project in a reliable and reproducible manner. We note that while an API associated with some other databases may have also been appropriate for our purposes, progress with these was impeded by limited access or slow responses to requests for further information or access.

In addition to the points listed above, some particularly useful features of the WoS API are:

- A “Topic Search", which allows the user to simultaneously search the fields “title”, “abstract”, “keywords” (supplied by author), and “keywords plus” (which Clarivate advises has content “...supplied by an algorithm that provides expanded terms stemming from the record’s cited references or bibliography”).
- A flexible search syntax. For example, the WoS API Lite allows queries such as “term1 NEAR/$n$ term2”, which can find occurrences of term1 within (some positive integer value) $n$ words of term2 in the fields relevant to a Topic Search.

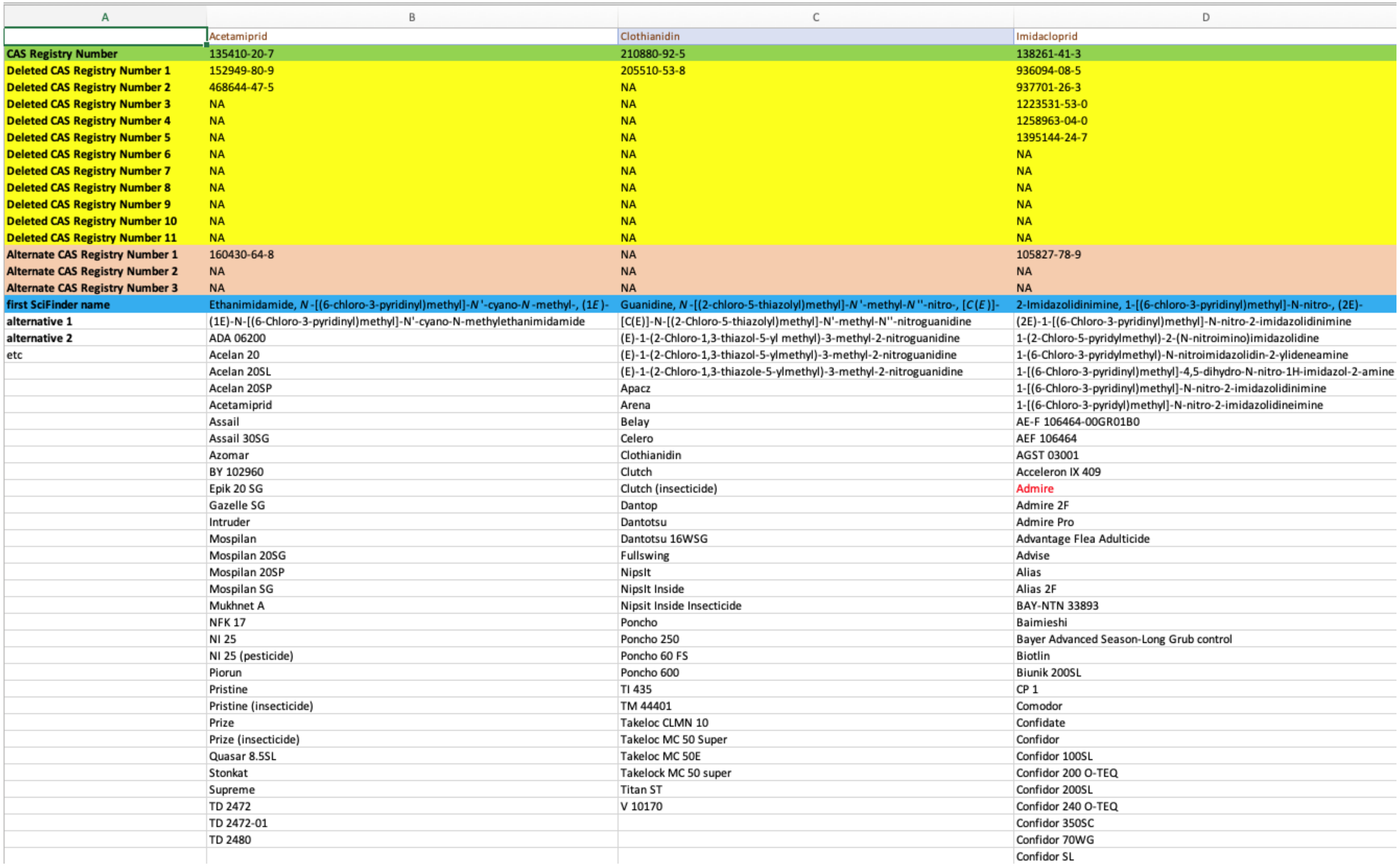
The WoS API Lite does not allow access to the full text of publications. However, as obtaining such access (via Clarivate’s Expanded API) would incur a substantial fee, we did not pursue this. Further, accessing and scrutinising publication full texts would require substantial new coding. This would lead to a certain delay in CEER’s ability to provide NSW EPA with results.

As this work is still in its early stages, we have not yet fully explored what may be achieved by applying our queries to titles, abstracts, and keywords of publications. CEER has maintained good relations with Clarivate, and negotiated (subject to conditions on what we can supply to NSW EPA) further access to the WoS API Lite. Given the body of code developed by CEER for scraping WoS in other projects, and the promising results there, we continue to use the WoS API Lite in this study. We will see that the results summarised in this report can provide NSW EPA with useful insights.

### 3.1.2. Choice of a source of alternative chemical names and identifiers

We sought certain features in a suitable source of alternative chemical names and identifiers ("synonyms"). We desired broad coverage of both common chemical names and names obtained from standard classification systems. We required the Chemical Abstracts Service Registry Number® (CAS RN®), and any associated “Alternate” or “Deleted” CAS RN®s. We determined that SciFinder® (Chemical Abstracts Service, 2019) was suitable for our purposes. However, one obtains information from SciFinder® via a web browser, which is not ideal. One must manually transfer chemical details (in our case, to an input data file) prior to commencing the web scraping. As we consider only 67 chemicals in this pilot study, the amount of associated manual processing of information is limited and tolerable.

Due to legal restrictions relating to the use of SciFinder® data, we cannot provide the lists of synonyms used in this study. However, for the purposes of illustration, we can provide a sample of the list of synonyms used for three of the neonicotinoid chemicals in Figure 3.1.



| A | B | C | D |
|---|---|---|---|
|  | Acetamiprid | Clothianidin | Imidacloprid |
| **CAS Registry Number** | 135410-20-7 | 210880-92-5 | 138261-41-3 |
| **Deleted CAS Registry Number 1** | 152949-80-9 | 205510-53-8 | 936094-08-5 |
| **Deleted CAS Registry Number 2** | 468644-47-5 | NA | 937701-26-3 |
| **Deleted CAS Registry Number 3** | NA | NA | 1223531-53-0 |
| **Deleted CAS Registry Number 4** | NA | NA | 1258963-04-0 |
| **Deleted CAS Registry Number 5** | NA | NA | 1395144-24-7 |
| **Deleted CAS Registry Number 6** | NA | NA | NA |
| **Deleted CAS Registry Number 7** | NA | NA | NA |
| **Deleted CAS Registry Number 8** | NA | NA | NA |
| **Deleted CAS Registry Number 9** | NA | NA | NA |
| **Deleted CAS Registry Number 10** | NA | NA | NA |
| **Deleted CAS Registry Number 11** | NA | NA | NA |
| **Alternate CAS Registry Number 1** | 160430-64-8 | NA | 105827-78-9 |
| **Alternate CAS Registry Number 2** | NA | NA | NA |
| **Alternate CAS Registry Number 3** | NA | NA | NA |
| **first SciFinder name** | Ethanimidamide, *N*-[(6-chloro-3-pyridinyl)methyl]-*N*'-cyano-*N*-methyl-, (1*E*)- | Guanidine, *N*-[(2-chloro-5-thiazolyl)methyl]-*N*'-methyl-*N*''-nitro-, [*C*(*E*)]- | 2-Imidazolidinimine, 1-[(6-chloro-3-pyridinyl)methyl]-N-nitro-, (2E)- |
| **alternative 1** | (1E)-N-[(6-Chloro-3-pyridinyl)methyl]-N'-cyano-N-methylethanimidamide | [C(E)]-N-[(2-Chloro-5-thiazolyl)methyl]-N'-methyl-N''-nitroguanidine | (2E)-1-[(6-Chloro-3-pyridinyl)methyl]-N-nitro-2-imidazolidinimine |
| **alternative 2** | ADA 06200 | (E)-1-(2-Chloro-1,3-thiazol-5-yl methyl)-3-methyl-2-nitroguanidine | 1-(2-Chloro-5-pyridylmethyl)-2-(N-nitroimino)imidazolidine |
| etc | Acelan 20 | (E)-1-(2-Chloro-1,3-thiazol-5-ylmethyl)-3-methyl-2-nitroguanidine | 1-(6-Chloro-3-pyridylmethyl)-N-nitroimidazolidin-2-ylideneamine |
|  | Acelan 20SL | (E)-1-(2-Chloro-1,3-thiazole-5-ylmethyl)-3-methyl-2-nitroguanidine | 1-[(6-Chloro-3-pyridinyl)methyl]-4,5-dihydro-N-nitro-1H-imidazol-2-amine |
|  | Acelan 20SP | Apacz | 1-[(6-Chloro-3-pyridinyl)methyl]-N-nitro-2-imidazolidinimine |
|  | Acetamiprid | Arena | 1-[(6-Chloro-3-pyridyl)methyl]-N-nitro-2-imidazolidineimine |
|  | Assail | Belay | AE-F 106464-00GR01B0 |
|  | Assail 30SG | Celero | AEF 106464 |
|  | Azomar | Clothianidin | AGST 03001 |
|  | BY 102960 | Clutch | Acceleron IX 409 |
|  | Epik 20 SG | Clutch (insecticide) | Admire |
|  | Gazelle SG | Dantop | Admire 2F |
|  | Intruder | Dantotsu | Admire Pro |
|  | Mospilan | Dantotsu 16WSG | Advantage Flea Adulticide |
|  | Mospilan 20SG | Fullswing | Advise |
|  | Mospilan 20SP | NipsIt | Alias |
|  | Mospilan SG | NipsIt Inside | Alias 2F |
|  | Mukhnet A | Nipsit Inside Insecticide | BAY-NTN 33893 |
|  | NFK 17 | Poncho | Baimieshi |
|  | NI 25 | Poncho 250 | Bayer Advanced Season-Long Grub control |
|  | NI 25 (pesticide) | Poncho 60 FS | Biotlin |
|  | Piorun | Poncho 600 | Biunik 200SL |
|  | Pristine | TI 435 | CP 1 |
|  | Pristine (insecticide) | TM 44401 | Comodor |
|  | Prize | Takeloc CLMN 10 | Confidate |
|  | Prize (insecticide) | Takeloc MC 50 Super | Confidor |
|  | Quasar 8.5SL | Takeloc MC 50E | Confidor 100SL |
|  | Stonkat | Takelock MC 50 super | Confidor 200 O-TEQ |
|  | Supreme | Titan ST | Confidor 200SL |
|  | TD 2472 | V 10170 | Confidor 240 O-TEQ |
|  | TD 2472-01 |  | Confidor 350SC |
|  | TD 2480 |  | Confidor 70WG |
|  |  |  | Confidor SL |

**Figure 3.1.:** A sample of the "synonyms" associated with a chemical's systematic name in SciFinder®. These three examples are drawn from an input file used in queries obtaining research interest for neonicotinoids. The examples demonstrate cases where, in addition to the CAS Registry Number® (CAS RN®), there are "Alternate" or "Deleted" CAS RN®s, and multiple alternative names.

## 3.2. Features of queries applied to Web of Science

Any individual query used in this study relates only to one particular chemical. We considered two types of queries, the second being more selective than the first. We anticipated that a comparison of results obtained for the two query types applied to our sample chemicals would allow us consider the importance of this selectivity.

Each query applied to WoS was composed of multiple conditions. We define a 'hit' as a publication which satisfies all conditions of a query. The simplest query sought to determine 'total-hits' for a chemical of interest. The component conditions of the query (combined with AND) are:

T1 a "Topic Search" (recall Page 16) for all alternative names or identifiers in an input list (combined with OR; a publication is retained as a possible hit if it contains at least one of these alternatives),

T2 a search of the "Publication Year" field for some value in a range, 1980–2019 inclusive.

We note that it is possible for total-hits to count unsuitable publications: those unrelated to an adverse effect on humans or the environment. As such, total-hits may not be the most appropriate measure of research interest for all situations. In order to address this, we developed a more restrictive query by requiring that any hit must match features T1 and T2 above, as well as satisfying further conditions (due to 'modifiers') in a Topic Search. We intended these modifiers to capture particular chemical properties and characteristics. These are associated with exposure to the chemical, and the chemical's ability to cause harm and/or adverse effects.

A query to determine 'near-hits' returns the count of hits that satisfied conditions T1 and T2, and the additional requirement:

T3 at least one modifier occurred at most three words from a chemical synonym in a field searched by a Topic Search.

We note that in (Whyte & Robinson, 2020, Table 4.1), the modifiers employed in near-hits queries were informed by a non-specialist reading of the literature. In this project, given the benefit of more development time, modifiers were developed in discussions with NSW EPA. These modifiers are presented in Table 3.1.

**Remark 4** *The choice of T3 was guided by our uses of this condition in our earlier project.*

> *We expect that the stricter requirement imposed by the near-hits search [compared to a search for total-hits] should eliminate a substantial proportion of spurious hits. Whilst allowing for up to three words between a chemical and a modifier may produce some false hits, this flexibility does allow for us to detect alternative means of reporting on a chemical (e.g. "...recorded harmful chemical X concentrations ...", "...after 12 months, X persists at toxic levels ..." so that true hits are not excluded. As such, given the importance of excluding false hits yet retaining true hits as much as possible, we deemed the counts of near-hits to be the most appropriate data for our purpose of predicting future mentions of harm caused by a chemical in the scientific literature. (Whyte & Robinson, 2020, Section 4.4)*

*It is possible that a near-hits search may return erroneous hits containing the negation of our modifiers (e.g. "non-toxic") in the fields searched by a Topic Search. In early 2020 we explored* *queries which sought to exclude publications having only negated modifiers from the hits returned. This substantially extended the length of queries. When this effect is combined with many chemical synonyms, the resulting query may reach a length such that it is rejected by the WoS API. (We comment further on this in Section 3.3.2.) We set aside this investigation so as to devote time to this current project. We may return to this matter in any later project.*

We shall now make some remarks on the implementation of queries. We ensured that queries were implemented in "url encoding" so as to be interpreted properly by the WoS API. In forming the template for queries, care was taken to ensure that a query would accurately represent our intent. We draw on WoS documentation (Clarivate, 2020) in noting other features of the search syntax that will assist understanding of our query implementation.

**Capitalization** is unimportant. For example, searches for Benzene and benzene are equivalent.

**Quotation marks** are used to denote an exact phrase. For example, a search for "adverse effect" will return hits that have exactly this phrase. The search will not return hits where only one of the words appears, or instances where the two words are not adjacent, or instances where the ordering of the words differs from the search phrase. Further, a search using a phrase (say for "toxic") will not return hits where the phrase appears with a non-hyphenated prefix (e.g. "nontoxic"). However, the search may return hits in which a word prior to the searched phrase changes its meaning (e.g. "not toxic").

**Hyphens, periods, commas** A search for a term including these will be interpreted as an exact phrase. The search will also look for a variant where the punctuation is omitted. For example, the search term **waste-water** will find records containing the exact phrase **waste-water** or the phrase "**waste water**". It will not match variants such as **water waste**, **waste in drinking water**, or **water extracted from waste**. A search for a phrase (e.g. "toxic") may return hits where the phrase appears with a hyphenated prefix (e.g. "non-toxic").

**Wildcards** The asterisk (*) represents any group of characters, including no character. For example, a Topic Search using a root with the wildcard, such as **carcinogen***, searches for variants including **carcinogen**ic and **carcinogen**s.
The dollar sign ($) represents zero characters or one character. For example, **phthalate$** searches for terms including **phthalate** and **phthalates**.
A wildcard may be used in a phrase. (However, there is limited support for multiple wildcards in a phrase.) For example, "**adverse effect***" is a valid search term that has the potential to return more hits than a search for "**adverse effect**".

**NEAR operator** Mentions of the uptake of chemicals to biota, food, or plants may occur in a variety of ways. This poses a particular challenge for the construction of near-hits queries. Our query must strike a balance

between capturing a variety of valid phrases, and not admitting a large number of spurious hits. After a modest amount of experimentation, we decided to employ "NEAR/7" between the chemical synonyms and text relating to chemical uptake. We can illustrate the effectiveness of this choice by considering some results for mercury uptake, which returned a range of paper titles including:
"Variation and range of **mercury uptake** into **plants** at a mercury-contaminated abandoned mine site"
"**PLANT UPTAKE** OF AIRBORNE **MERCURY** IN BACKGROUND AREAS"
"**MERCURY UPTAKE** BY **PLANTS** AND EFFECTS ON ROOT MEMBRANE PERMEABILITY"
"Assessment of **mercury uptake** routes at the soil-**plant**-atmosphere interface"
"Indicators of sediment and **biotic mercury** contamination in a southern New England estuary"
"Development of a **mercury** speciation, fate, and **biotic uptake** (biotranspec) model: Application to lahontan reservoir (Nevada, USA)"

**Table 3.1.:** Modifiers provided by NSW EPA for use in queries intending to return near-hits, and their implementation.

| Example required terms | Implementation |
|---|---|
| adverse effects | "adverse effect*" |
| adverse impacts | "adverse impact*" |
| bioaccumulate/bio-accumulation | bio$accumulat* |
| biomagnify/bio-magnification | bio$magnif* |
| carcinogen/carcinogenicity | carcinogen* |
| ecotoxicity/chronic ecotoxicity | ecotoxic* |
| endocrine (disruption/disrupter) | "endocrine disrupt*" |
| genotoxic | genotoxic* |
| (immuno/immuno-)suppressant | immuno$suppressant* |
| (multi/multi-)generational effect | "multi$generational effect*" |
| mutagenic/mutagenicity/germ cell mutagenicity | mutagen* |
| persistent/persistence | persisten* |
| reproductive effects | "reproductive effect*" |
| secondary poisoning | "secondary poison*" |
| (chronic/reproductive/target organ) toxicity | toxic* |
| mobile/mobility | mobil* |
| uptake into (biota/food/plants) | (uptake NEAR/7 (biot* OR food OR plant*)) |

## 3.3. Limitations of this study

In this section we note some limitations that pertain to two features of this study: the inputs, and our methodology. Where possible, we also outline how we acted to mitigate their influence, or alternatively, how we would seek to achieve this in any future project.

### 3.3.1. Undesirable features or limitations of inputs

CEER had some concerns around the input synonyms. We note two features:

**Non-unique synonyms:** Synonyms obtained for some chemicals may appear in the synonym lists of other chemicals. (For example, "Supreme" is a synonym of Acetamiprid, and is also associated with a brand of pool algaecide.) Also, a synonym may concern matters unrelated to chemicals. At this stage we do not know how this may have influenced hit counts.

Management: We expect that the near-hits queries will remove the effect of a common word that is unrelated to adverse effects on near-hits counts.

**Non-unique acronyms:** There are instances where an acronym is associated with both a family (or subfamily) and a specific chemical in this grouping.[1] Clearly there is no issue when such an acronym is used in searching for the grouping; it is appropriate that hits should include those for the specific chemical. However, when a search for a specific chemical also returns hits for other chemicals or groupings, it is conceivable that spurious hits will be included in the tally. (To take an extreme case, when the grouping's hits are substantially larger than those for a chemical, the latter's hit counts will be dominated by those of the former.) This is not ideal, and we would prefer to avoid such situations. It is possible that inflated hit counts will lead to some chemicals being classified as CoI when they are not (false positives). This could become problematic if there is a large number of false positives, and this requires a regulator to expend substantial effort in assessing these chemicals for no useful result.

NSW EPA were informed of the matter, and can tolerate the use of TCDD in queries at this stage.

Possible future management: Let us suppose that there are multiple instances of the problem described. This would motivate us to conduct a separate study to determine how many hits are associated with each of the individual chemical and its grouping. This may prompt investigations of how to refine queries so as to exclude hits for a grouping from those returned for an individual chemical. Implementing suitable query modifications will give us greater confidence in the appropriateness of hit counts returned for individual chemicals.

We also noted complexities in interpreting the actions of regulators.

**Regulator activities are not associated with particular chemicals:** For example, for 1,2,5,6,9,10-Hexabromocyclododecane (CAS RN® 3194-55-6), NSW EPA advised

[1] For example, TCDD represents the specific chemical with the CAS systematic name of Dibenzo[*b*,*e*][1,4]dioxin, 2,3,7,8-tetrachloro- (CAS RN® 1746-01-6), as well as the dioxin subfamily of tetrachlorinated dibenzo-p-dioxins.

(Excel spreadsheet, July 3 2020) that, both internationally and within Australia, "Concern and regulation for CAS 3194-55-6 is conflated with CAS 25637-99-4".

Possible management: On occasion it may be necessary to apply the regulatory history for some collection of chemicals to individual chemicals in the collection.

### 3.3.2. Limitations of our methodology

We also note some limitations of WoS. Some relate to the queries which can be processed by the WoS API Lite.

**Query length limitations:** For certain chemicals, the associated list of synonyms may be quite large, running to hundreds of elements. Such a lengthy list of synonyms (as seen for polyethylene glycol (PEG) and magnesium hydroxide) may lead to a total-hits query that is too large for the API, and is rejected. As such, no hits are returned. Further, even when the total-hits query is processed successfully, the addition of modifiers in forming a near-hits query can result in a lengthy query that is rejected. (Ascorbic acid provides one example of this.) Owing to this effect, we disregarded a small number of chemicals that NSW EPA proposed for consideration. As this issue arose quite late in the project, there was insufficient time available to investigate management strategies.

Possible future management: One practical solution involves rewriting the WoS scraping code. We would divide a chemical's synonyms into portions with a maximum size, and run separate queries for each portion. (We may require some experimentation to determine a maximum size that is appropriate for both total-hits and near-hits queries.) If all queries were processed successfully, we could then sum the hits returned for each publication year across the queries. This would yield an overall time series of hits for a single chemical that is equivalent to that obtained for a single chemical in this study.

An additional (possible) limitation of WoS arises from it being largely a database of documents written in English.

**Limited coverage of non-English publications in WoS:** A study of major databases such as Web of Science and Scopus (Vera-Baceta *et al.*, 2019) noted that a substantial number of non-English documents are omitted. The authors considered WoS publications with a publication year of 2018, and determined that the percentage of English-language documents was 95.37%. Spanish was the second most common language in 2018, comprising 1.26% of the year's publications.

Commentary: Our queries were designed to inspect documents written in English. To do otherwise would require translating our modifiers into other languages, and possibly a working knowledge of the grammar of these languages. This was considered impractical given the time allowed for this project, the relatively small proportion of non-English documents, and the diversity of languages (and alphabets) amongst these. (Twenty-five languages each comprised at least 0.01% of WoS publications (that is, 294 documents) in 2018 (Vera-Baceta *et al.*, 2019).)

We note that some synonyms (obtained from SciFinder$^{®}$) are from languages other than English. As such, these synonyms, and the numerical CAS RN$^{®}$s, may be found by total-hits queries. However, we expect that this benefit may be outweighed by the lack of discrimination in total-hits queries compared to near-hits queries. We expect this effect to be especially prominent for Group 3 chemicals, such as caffeine or sodium chloride, which have attracted substantial research interest that is unrelated to the aims of this project. As we are yet to thoroughly evaluate the usefulness of the English-language literature for NSW EPA's aims, we consider that a consideration of other languages is not justified at this point.

⋆ ∞ ⋆

We show the results of applying our two query types to WoS in Chapter 4.

# 4. Time series of hits and data exploration

In this chapter we give detailed consideration to the time series of total-hits and near-hits for chemicals across the groups that form our training set (Groups 1A and 3A) and validation set (Groups 1B and 3B).[1] We present plots of the hits time series in Section 4.1. We discuss features of these plots in Section 4.2. In Section 4.3 we demonstrate some methods used in exploring features of the data. We will draw on insights gained in formulating our classification system.

## 4.1. Time series of hits by groups

We will consider the time series of total-hits and near-hits by group for each of Group 1A, Group 1B, Group 3A, and Group 3B.

A comparison of results allowed us to gain an appreciation for how the query type influenced the order of magnitude of hit counts obtained for our sample chemicals. We will make further comment in Section 4.2.

[1] Various data limitations made it inappropriate to consider Group 2 chemicals in this report. Further, under these limitations, time series graphs of the research interest into Group 2 chemicals cannot convey any useful information. In order to respect our agreement with Clarivate to supply a limited number of summary graphs, we have confined our attention to showing plots which have value to this project.

### 4.1.1. Group 1A chemicals

Plots are shown in Figures 4.1 and 4.2.

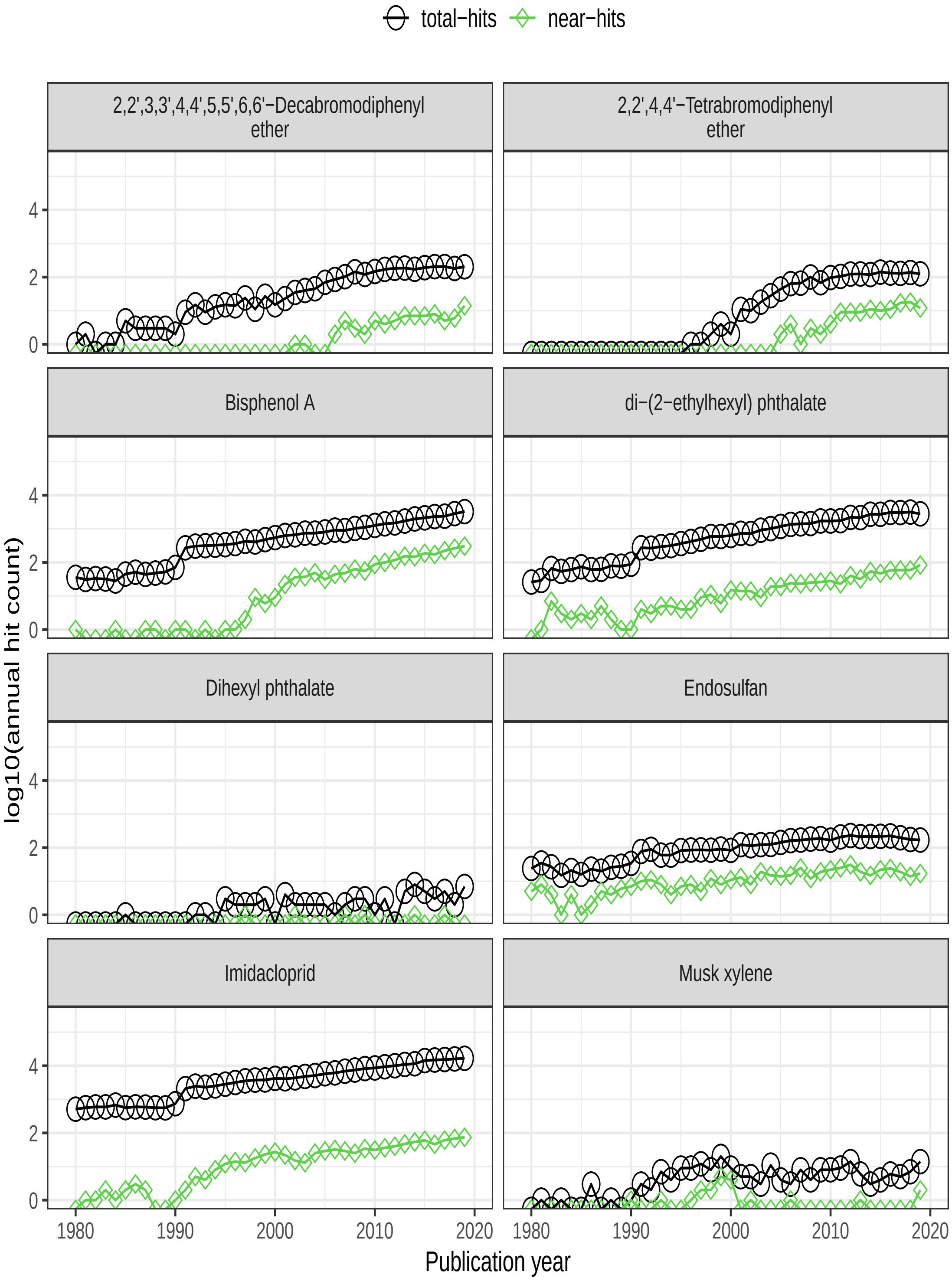


**Figure 4.1.:** Plots of hits time series obtained for Group 1A chemicals (page 1 of 2).

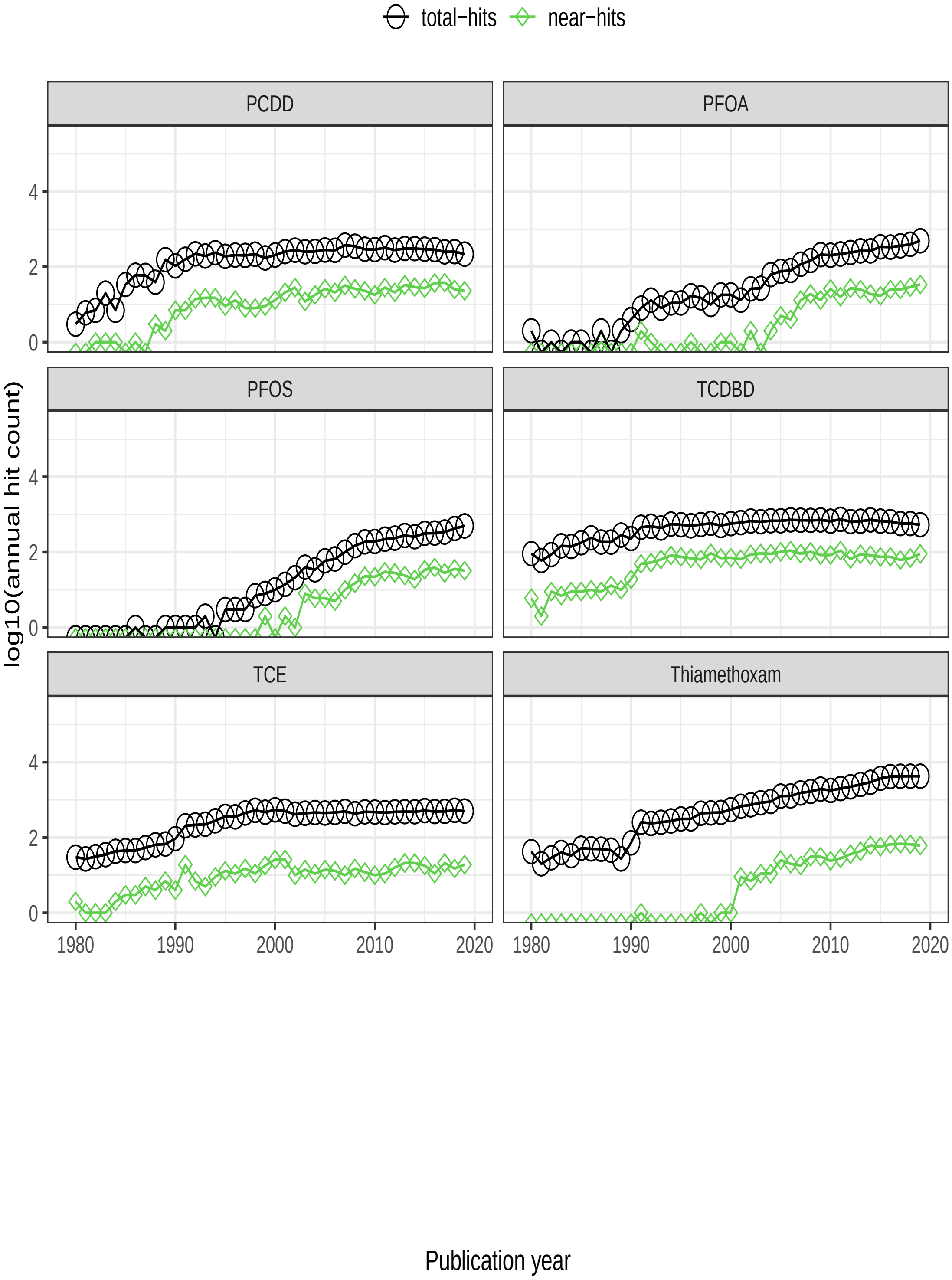


**Figure 4.2.:** Plots of hits time series obtained for Group 1A chemicals (page 2 of 2).

## 4.1.2. Group 1B chemicals

Plots are shown in Figures 4.3 and 4.4.

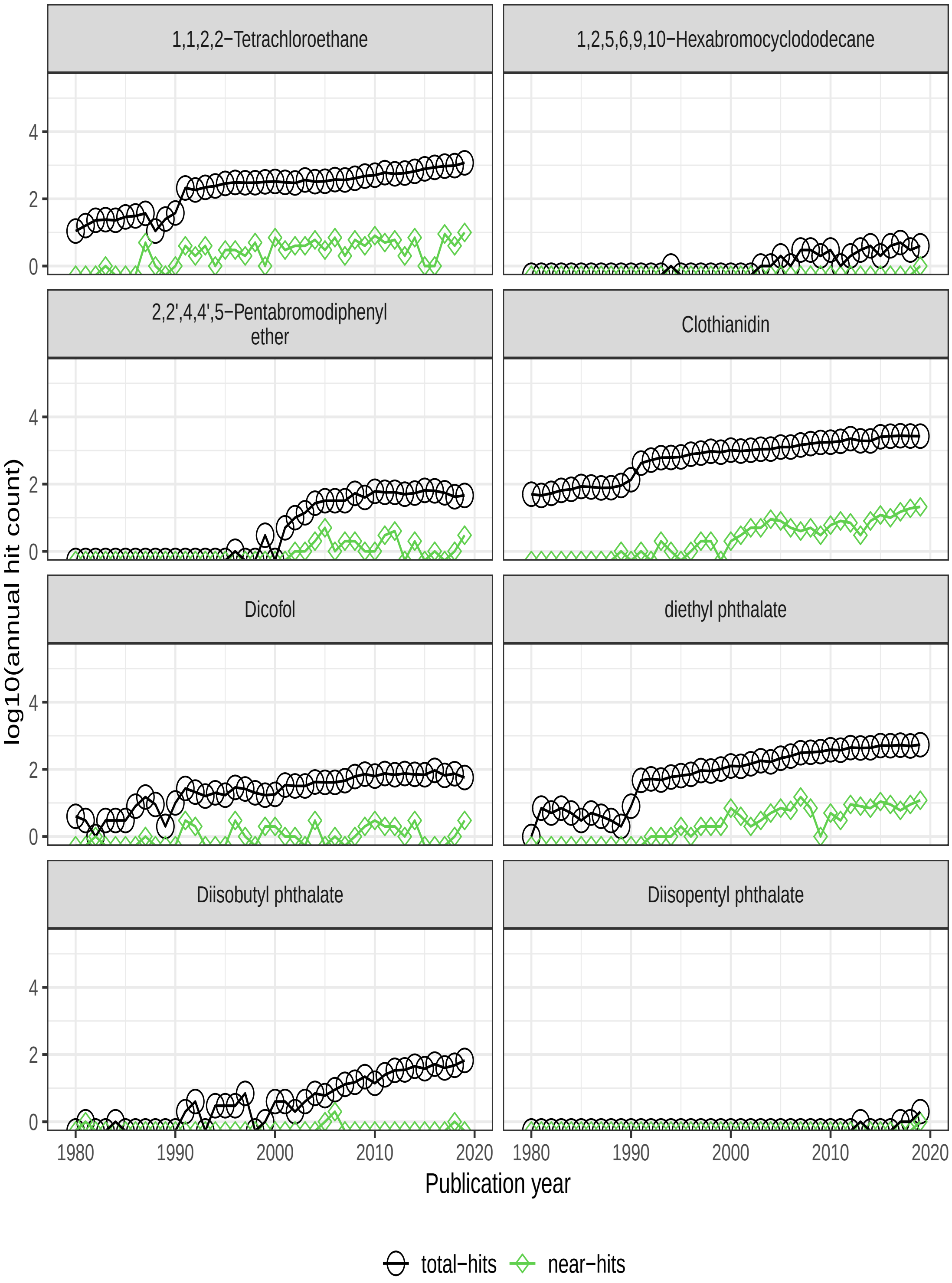


**Figure 4.3.:** Plots of hits time series obtained for Group 1B chemicals (page 1 of 2).

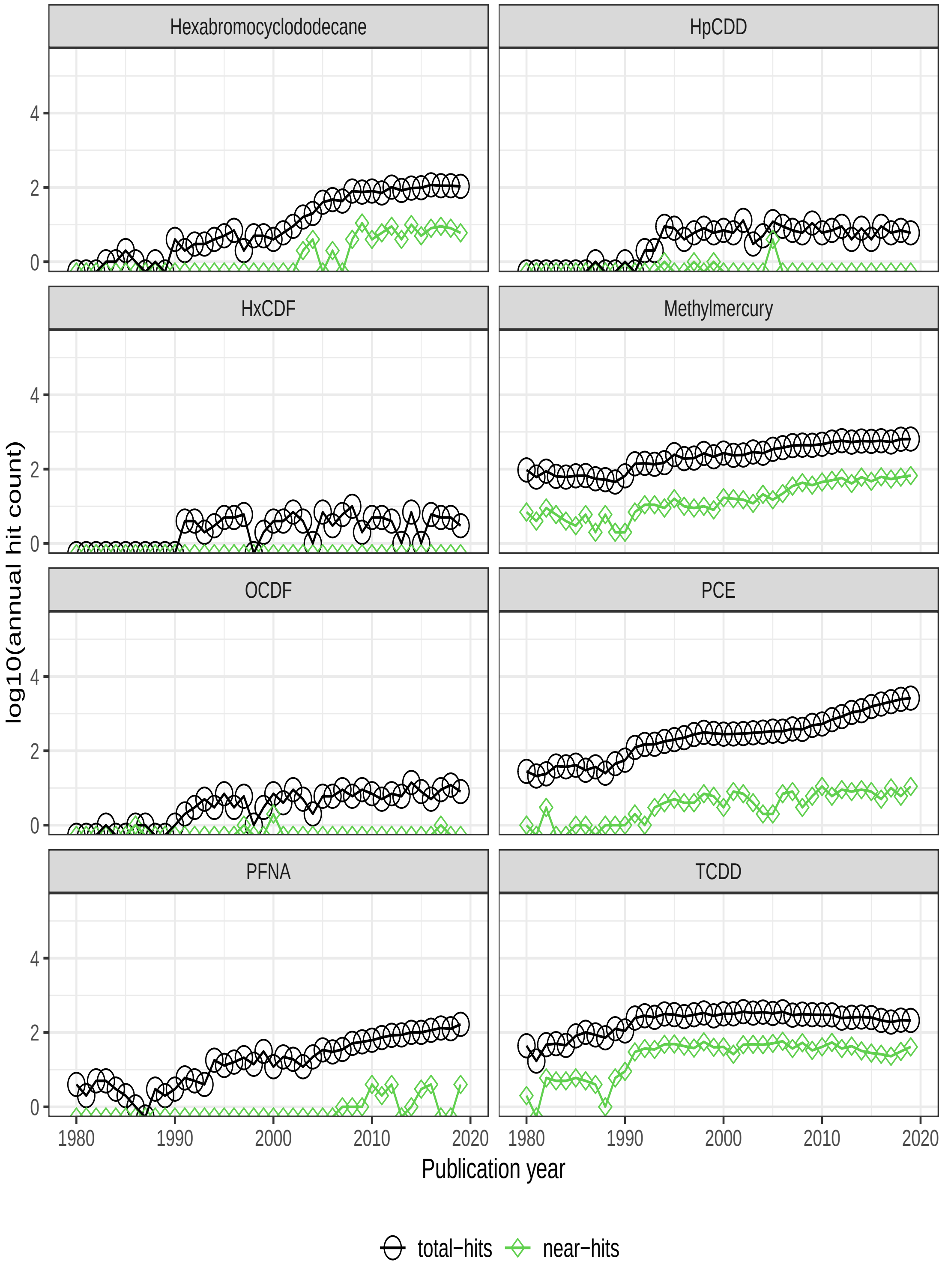


**Figure 4.4.:** Plots of hits time series obtained for Group 1B chemicals (page 2 of 2).

### 4.1.3. Group 3A chemicals

Plots are shown in Figure 4.5.

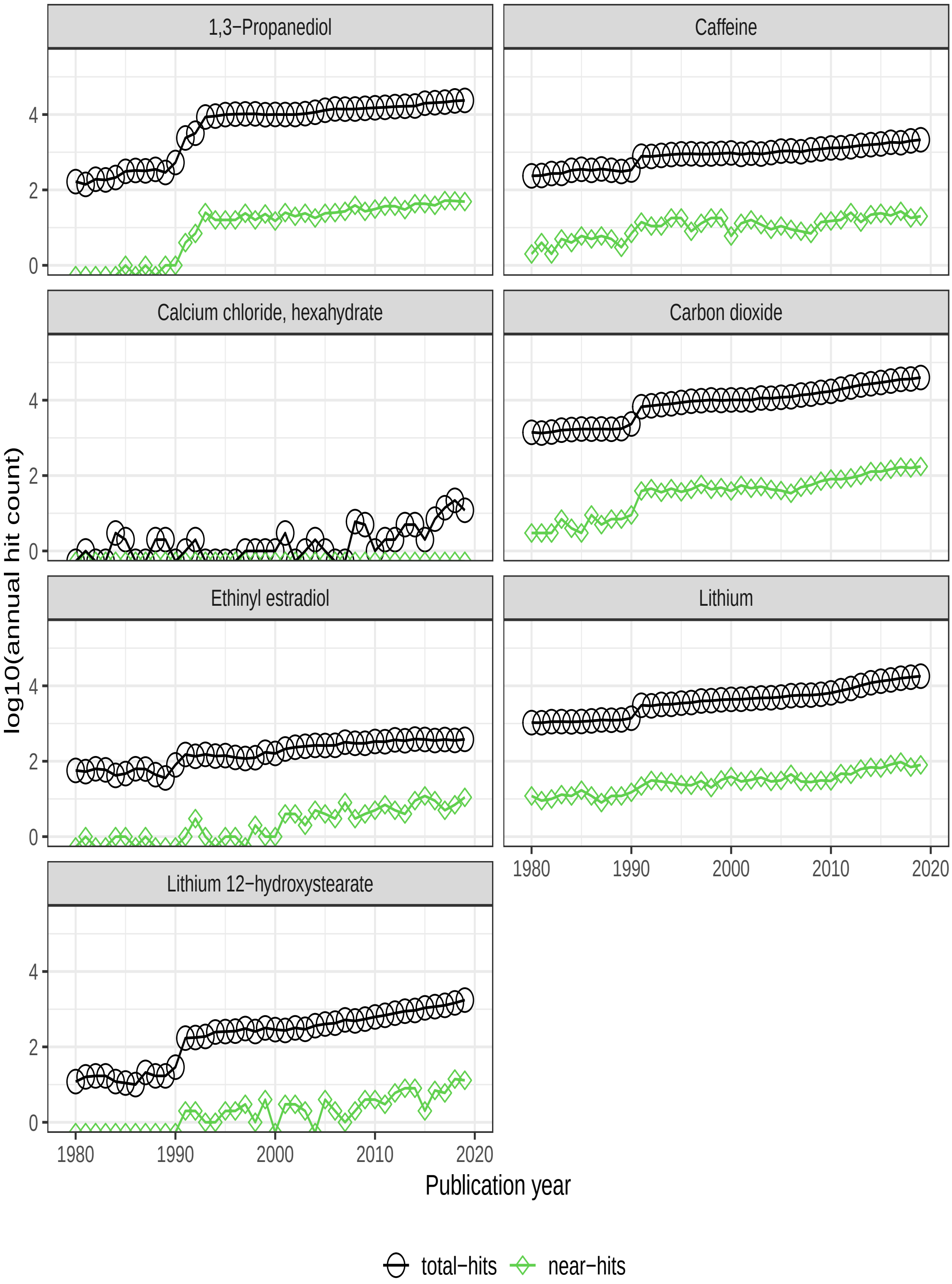


**Figure 4.5.:** Plots of hits time series obtained for Group 3A chemicals.

### 4.1.4. Group 3B chemicals

Plots are shown in Figure 4.6.

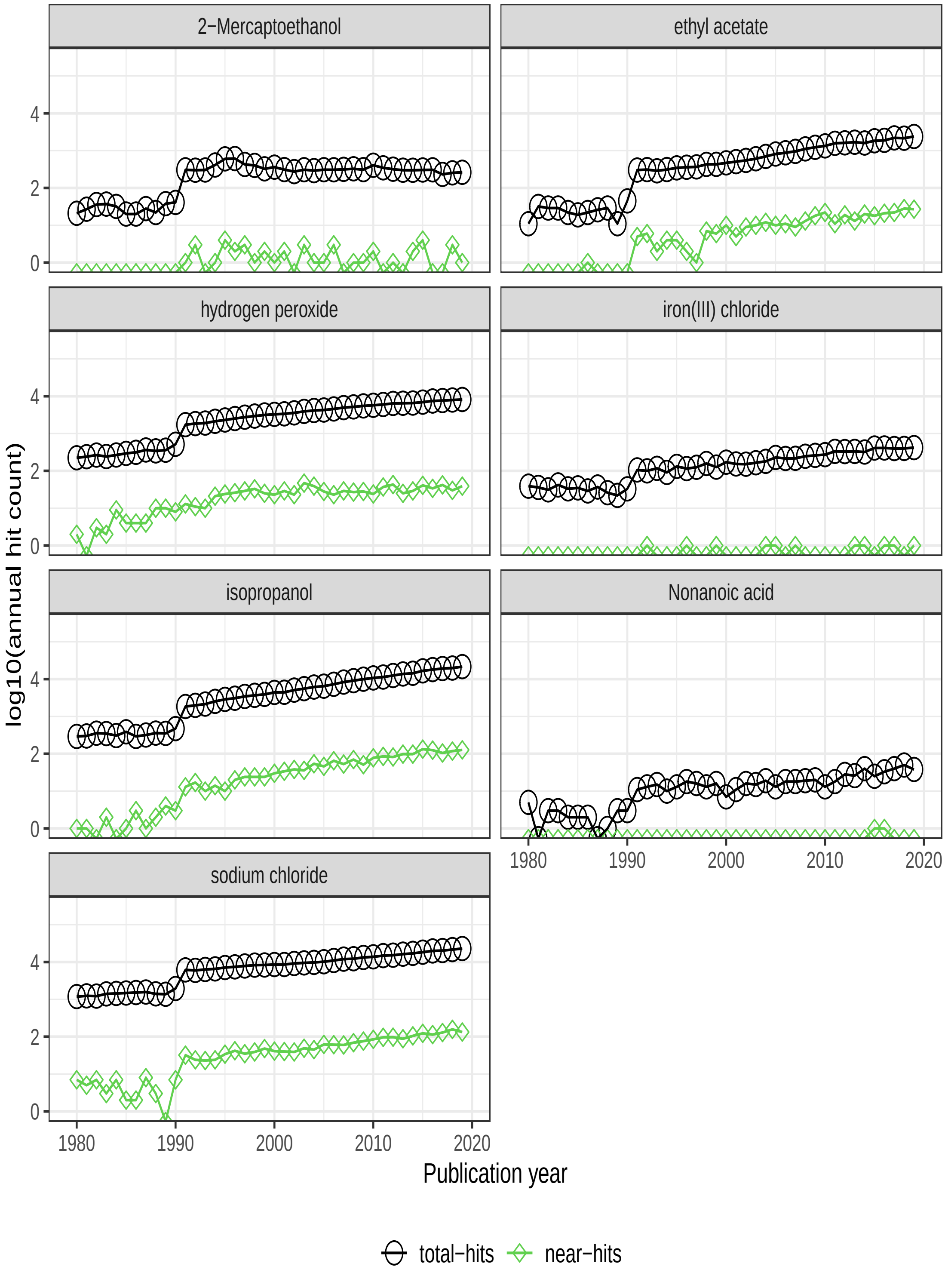


**Figure 4.6.:** Plots of hits time series obtained for Group 3B chemicals.

## 4.2. General remarks

As a total-hits query will find any mention of a chemical (or its synonyms), including those unrelated to adverse effects, we consider total-hits to be an inappropriate input to a classification system. (Following some discussion, we comment on this matter further in the next section.) Thus, we shall proceed with an investigation of near-hits time series in this report.

We note that in certain cases — say Thiamethoxam in Group 1A (Figure 4.2) and Clothianidin in Group 1B (Figure 4.3) — the near-hits values may be an order of magnitude (or more) smaller than the associated total-hits values. As such, we wonder whether further refinement of the near-hits query will capture more of the research interest that is relevant for this study. We find further motivation to revise the near-hits query by considering the results obtained for our Group 3 chemicals. For example, the near-hits series for Carbon dioxide (Figure 4.5) is comparable to that of some Group 1 chemicals. It may be that some parts of the near-hits query, such as that which finds mentions of carbon dioxide uptake into plants, has introduced hits that do not suit our purpose of relating chemicals to adverse effects.

Clearly, for certain chemicals, the near-hits totals are modest. This is of particular concern for the chemicals in Groups 1A and 1B, which we had intended to use in training and validating some (to be proposed) classification system. We may consider this data limitation more formally by considering the sums of near-hits across our publication range for each chemical in Groups 1A and 1B, as shown in Tables 4.1 and 4.2, respectively.

**Table 4.1.:** Group 1A chemicals presented in ascending order of their sum of near-hits over the publication range from 1980 to 2019.

| Name | Near-hits sum |
|---|---:|
| Dihexyl phthalate | 6 |
| Musk xylene | 21 |
| 2,2′,3,3′,4,4′,5,5′,6,6′-Decabromodiphenyl ether | 82 |
| 2,2′,4,4′-Tetrabromodiphenyl ether | 122 |
| PFOA | 308 |
| PFOS | 370 |
| TCE | 455 |
| Endosulfan | 479 |
| PCDD | 613 |
| Thiamethoxam | 678 |
| di-(2-ethylhexyl) phthalate | 749 |
| Imidacloprid | 880 |
| Bisphenol A | 2168 |
| TCDBD | 2419 |

**Table 4.2.:** Group 1B chemicals presented in ascending order of their sum of near-hits over the publication range from 1980 to 2019.

| Name | Near-hits sum |
|---|---:|
| HxCDF | 0 |
| 1,2,5,6,9,10-Hexabromocyclododecane | 1 |
| Diisopentyl phthalate | 1 |
| OCDF | 5 |
| Diisobutyl phthalate | 5 |
| HpCDD | 7 |
| PFNA | 25 |
| 2,2′,4,4′,5-Pentabromodiphenyl ether | 30 |
| Dicofol | 39 |
| Hexabromocyclododecane | 92 |
| 1,1,2,2-Tetrachloroethane | 132 |
| diethyl phthalate | 148 |
| Clothianidin | 168 |
| PCE | 178 |
| Triclosan | 229 |
| Methylmercury | 921 |
| TCDD | 1222 |

We see that 3 chemicals from Group 1A, and 10 from Group 1B, have a near-hits sum below 100, which is a very modest amount of research interest. This will make it difficult (or impossible in certain cases) to discern useful features of time series features for our classification system. As such, we shall exclude those Group 1 chemicals having a near-hits sum smaller than 100 from further consideration. This leaves us with 11 Group 1A chemicals for training and 7 Group 1B chemicals for validation. We note that each data set is quite small considering the classification task at hand. Also, quite inconveniently, the remaining chemicals in the two groups show quite different features in their near-hits totals. For example, the restricted Group 1A has only one chemical with fewer than 300 hits, whilst the restricted Group 1B has five such chemicals.

We proceed to give an overview of features of near-hits time series in the next section.

## 4.3. Overview of data features

It will assist our efforts in classifying chemicals as CoI or not if we can recognise that the near-hits time series for chemicals in a particular group tend to exhibit characteristic features. We may then exploit these features in our subsequent classification system training and validation.

In this section we provide examples of some methods employed in inspecting near-hits data. Results of such inspections have informed the tests we employ in our classification system.

Recall the near-hits time series for Group 1A and Group 1B chemicals, presented in Sections 4.1.1 and 4.1.2, respectively. We cannot expect to infer useful features from hit time series for chemicals having little research interest. As such, henceforth we confine our attention to chemicals which had a near-hits total of at least 100 over 1980 to 2019, as shown by Tables 4.1 and 4.2. This allows us to consider eleven chemicals from Group 1A (the training set), and seven from Group 1B (the validation set).

### 4.3.1. Spread of near-hits values

It is difficult to draw definitive conclusions given such small groups. However, obvious exceptions aside (which we will consider shortly), we may note some general features. Group 1 chemicals tend to exhibit a larger spread of annual near-hits values, and larger maxima, than Group 3 chemicals.

Regarding the exceptions, we note that certain Group 3 chemicals (such as hydrogen peroxide of Group 3B) have research interest that is comparable to Group 1 chemicals of moderate interest. Others, such as sodium chloride from Group 3B, can exhibit a spread of near-hits values which are comparable to that seen for the most-studied Group 1A chemicals. NSW EPA has advised (December 1st 2020) of possible reasons for the substantial research interest in sodium chloride, which we report (or paraphrase) as:

- "studies including hyper salinity",
- "epidemiology studies that use salt intake as as one factor when trying to tease out the effects of multiple chemicals",
- toxicity studies, which may examine interactions between a given chemical and cations, anions, or salts: "... this could result in pick up of salts in the literature - this means the salt could be accidentally picked up as toxic substance in the search, rather than the chemical which was examined."

We do not currently know how many publications relating to the above topics are unrelated to the interests of this project. However, suppose that the near-hits query in its current form cannot 'filter out' such publications. Then, the research interest returned will have the undesirable property of including spurious hits, which will adversely affect our attempts to correctly classify chemicals as CoI or otherwise.

Similarly, the substantial research interest in Group 3B's isopropanol (linked to substance abuse, e.g. Aditi Sharma and Jonathan D. Morrow (2017)) may have a component unrelated to NSW EPA's interests. Alternatively, the research interest into a substance such as Group 3A's 1,3-Propanediol (with uses including cosmetics, toothpaste, and shampoo) may indicate that a Group 3 chemical has become associated with harmful effects due to some new interaction with other chemicals, or a change in its use.

We also see substantial research interest for lithium (Group 3A), which, being an element, is not a particularly specific search term. We expect that the associated time series of near-hits is the superposition of time series for a variety of lithium compounds (including Lithium 12-hydroxystearate, also of Group 3A). This could create a signal that differs materially from that associated with a specific compound. As such, we may find that classification of an element via consideration of its near-hits is inappropriate.

Given these observations and concerns, it may be appropriate to review certain groupings, or to consider how we should analyse the hits time series of chemicals which are in common use. This may encourage us to reconsider the construction of our near-hits query — recall the note on carbon dioxide (of Group 3A) in Section 4.2.

We show boxplots of the spread of near-hits values for Group 1 and Group 3 chemicals in Figure 4.7.

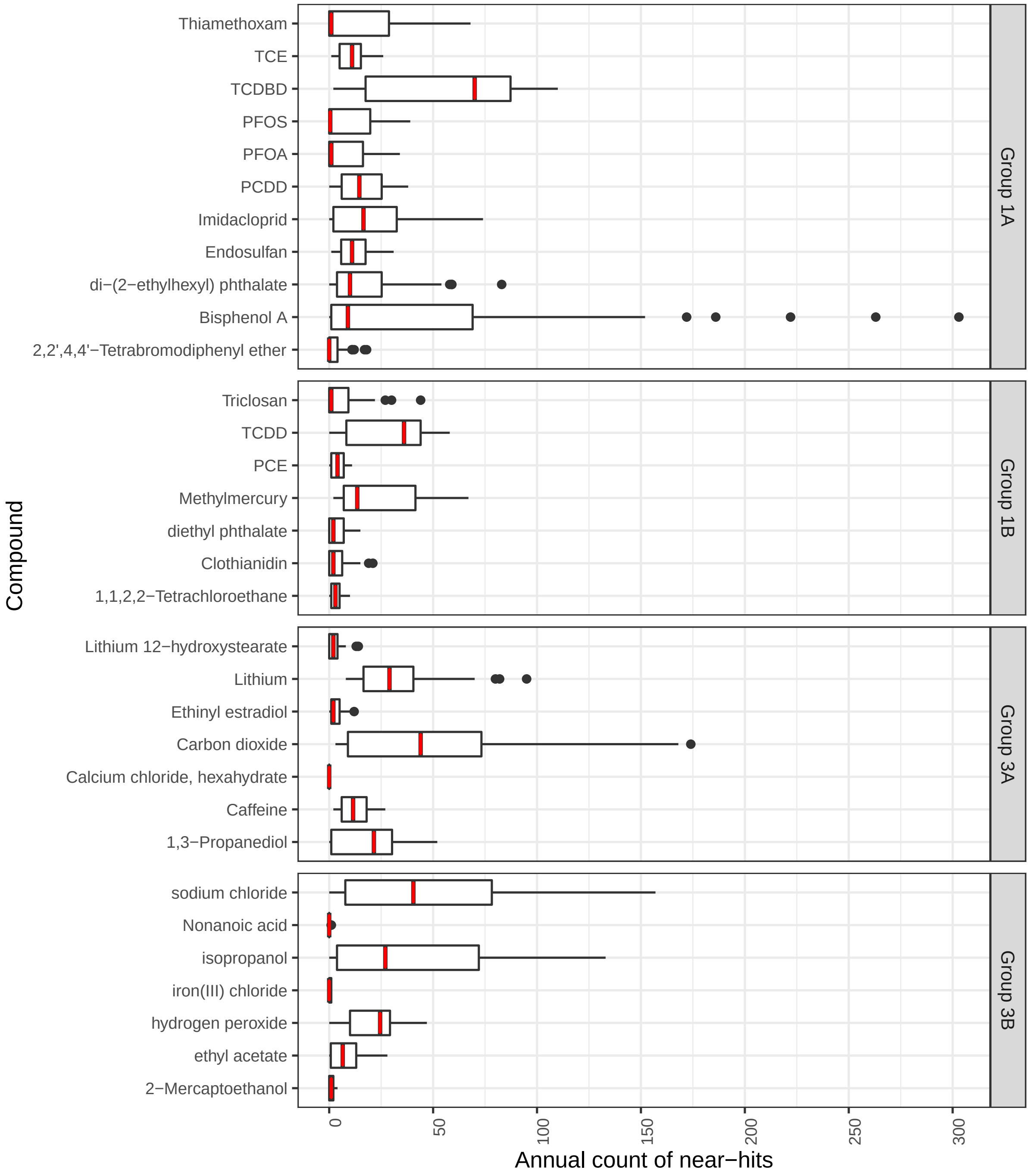


**Figure 4.7.:** Boxplots showing the spread of near-hits values (annual near-hits counts for each year from 1980-2019) for Groups 1A, 1B, 3A, and 3B. The red line on each boxplot indicates the median value. The box shows the "interquartile range" (IQR) from $Q1$ to $Q3$ — the middle 50% of the data. Values below $Q1$ (respectively, above $Q3$) to an extreme of, at most, $Q1-1.5\times$IQR (respectively, $Q3+1.5\times$IQR) are shown with horizontal lines. Any values more extreme are shown as points.

**Remark 5** *As a companion to Figure 4.7, we present boxplots of the (annual counts of) total-hits for chemicals from Group 1 and Group 3 in Figure 4.8. We show counts on a log10 scale to assist comparison of values within and between groups. We note that in a number of cases, total-hits values for Group 3 chemicals are far greater than those seen for chemicals in Group 1. Such a feature supports our choice to use near-hits as our measure of research interest, rather than total-hits.*

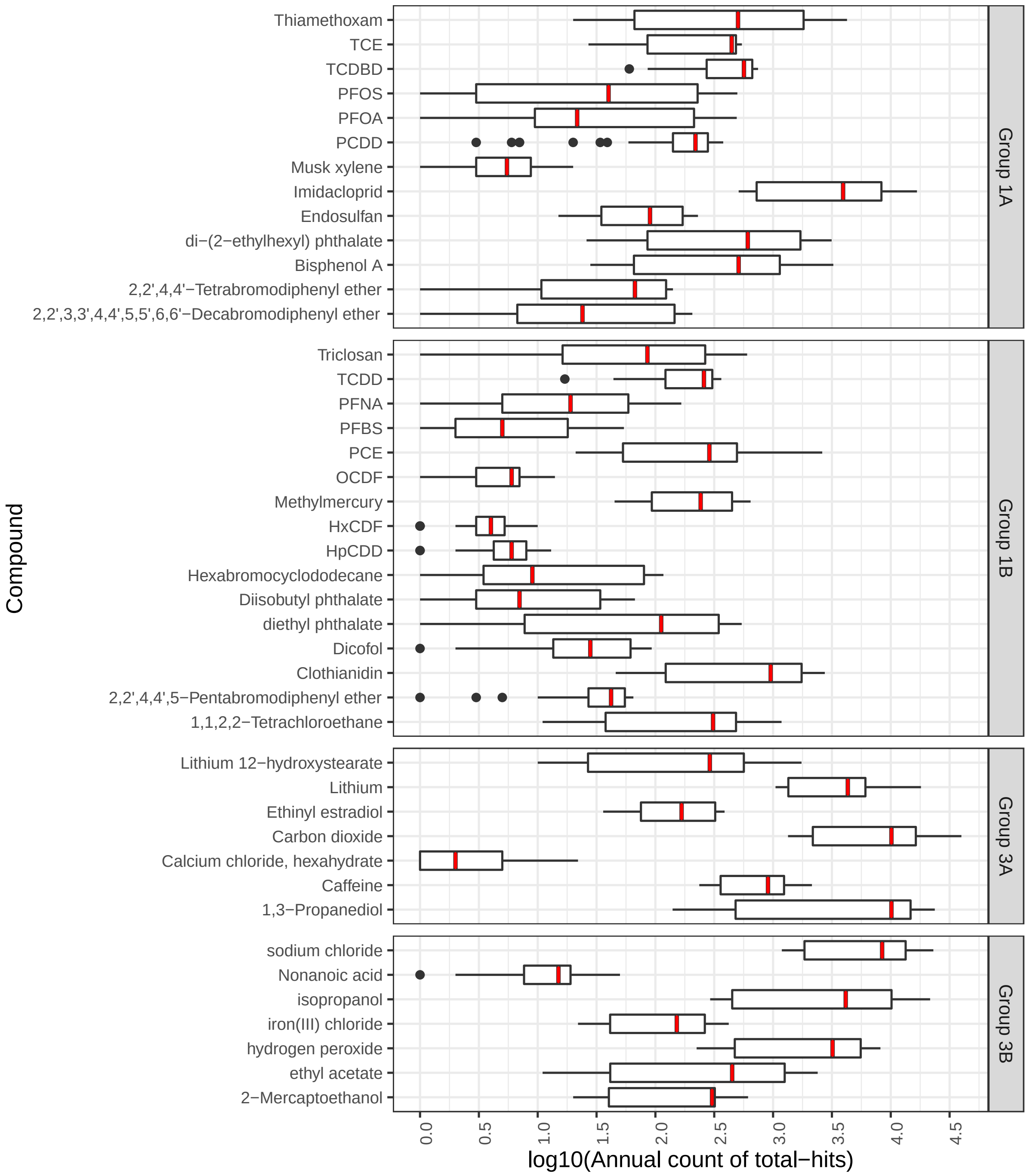


**Figure 4.8.:** Boxplots of log10-transformed annual counts of total-hits (from the publication range of 1980-2019) for all chemicals from Group 1 and Group 3.

We may also gain insights into features of near-hits time series by attempting to uncover any systematic features of the data.

### 4.3.2. Extraction of systematic features from near-hits data

Let us suppose that some near-hits time series we observe at time $t$, $y(t)$, is composed of a smooth systematic component, $s(t)$, and a random (error) component, $\varepsilon(t)$. Most simply, we may express this as:

$$y(t) = s(t) + \varepsilon(t)\,. \tag{4.1}$$

We expect that a systematic component will allow us to recognise features of research interest that are concealed by noisy data. By a suitable choice of method, we may estimate $s(t)$ for each chemical in a group. We may then attempt to use these results in characterising features of research interest for group members. So informed, we can aim to capture observed features by designing appropriate tests for use in our classification system.

**Remark 6** *Considering some near-hits series $y$, in this report we obtain the associated $s$ by fitting a (generalised linear) model to $\ln(y)$ that is a cubic (a third-degree polynomial) in the time variable $t$, representing years after 1980. That is, we determine*

$$f(t) = at^3 + bt^2 + ct + d \quad t = 0, \ldots 39, \textit{where} \tag{4.2}$$

$$s(t) = \exp(f(t))\,. \tag{4.3}$$

*The model coefficients in (4.2) ($a$, $b$, $c$, and $d$) are estimated from the data by minimizing a function of the squared errors between data and predictions.*

As a demonstration, in Figure 4.9 we present time series for the Group 1B chemicals, showing an estimated systematic component in each case. We shall make some observations on features of these systematic components.

**Observation 1** There may be a well-defined point after which the next year's near-hits shows a noticeable increase. (See, for example, the behaviour for TCDD after 1990.)

**Observation 2** In six of seven cases there is a well-defined point beyond which the series maintains an upward trend to the end of the series. For example, consider Triclosan after it reaches 10 hits in 2008. Following this, some curves (most noticeably Clothianidin and Triclosan) show a super-linear rate of increase, as shown by the increasing gradient of the tangent to the curve (rate of change of hits). (We observe similar behaviour in the near-hits for Group 1A chemicals, such as PFOS and Thiamethoxam, in Section 4.1.1. Each shows an upward trend after the year in which near-hits reaches 10, equivalently, 1 on the log10 vertical scale used for graphs.)

**Observation 3** In four of seven cases (1,1,2,2-Tetrachloroethane, Clothianidin, diethyl phthalate, and PCE) after an initial growth period, the systematic component may display a period of little growth, but does not decrease. Following this, the curve increases at a greater rate than that observed for the first growth phase.

**Observation 4** In one of the seven cases (TCDD), having reached a maximum, the systematic component shows a downward trend. However, the maximum rate of decline is smaller in magnitude than the maximum rate of increase.

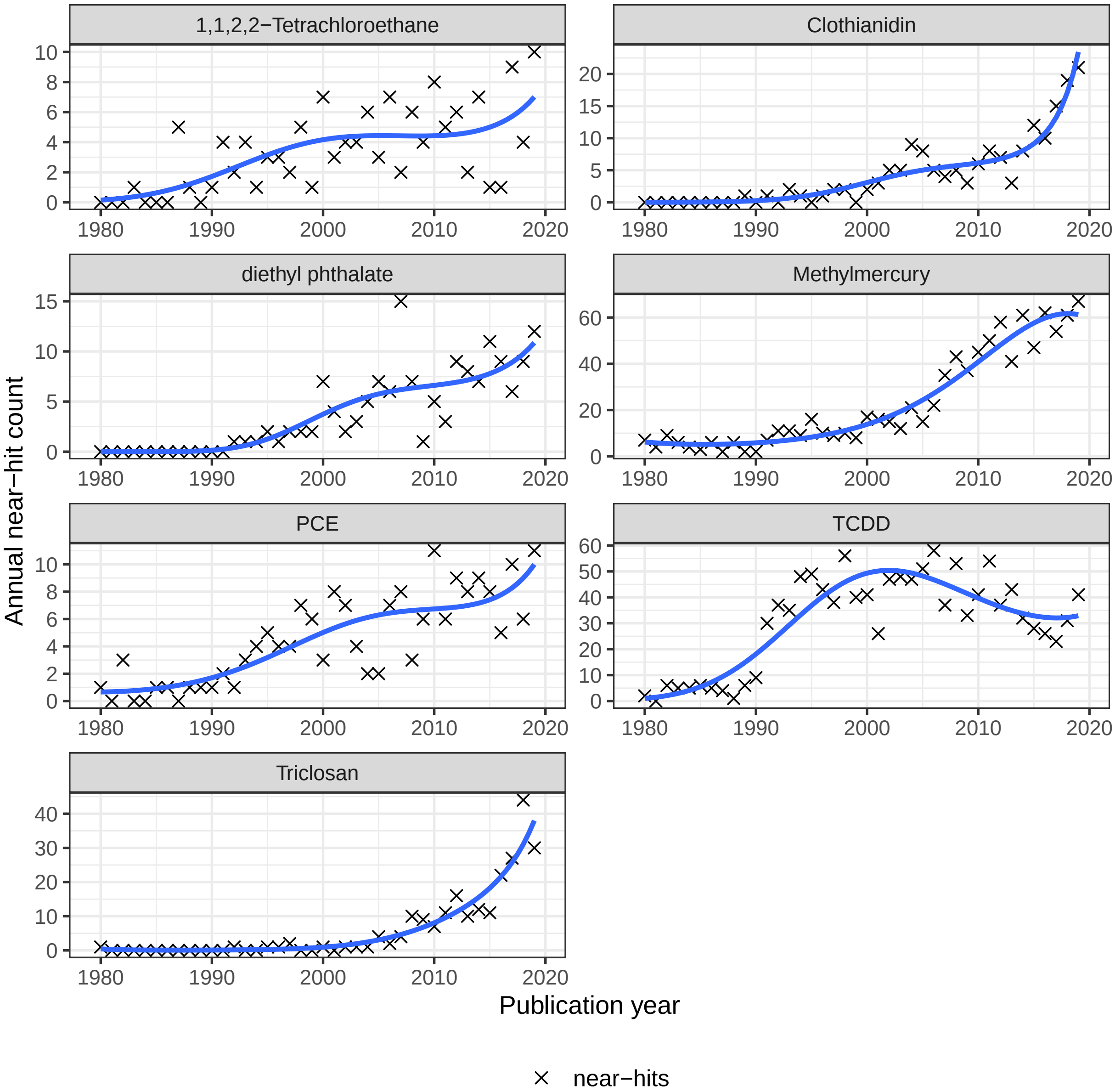


**Figure 4.9.:** Group 1B near-hits data (shown with crosses) with a smoothing line to show the fitted systematic component.

In order to provide a view of a different collection of chemicals, we present estimates of the systematic component of Group 3A chemicals in Figure 4.10. We may neglect calcium chlorate, hexahydrate as, with so little research interest, the systematic component is practically zero and offers no insight. There are some similarities to the systematic components for Group 1B chemicals, and also some differences. The curves shown for Group 3A chemicals exhibit longer periods of flatter behaviour compared to Group 1B chemicals. Also, we do not see Triclosan-like behaviour in the Group 3A chemicals.

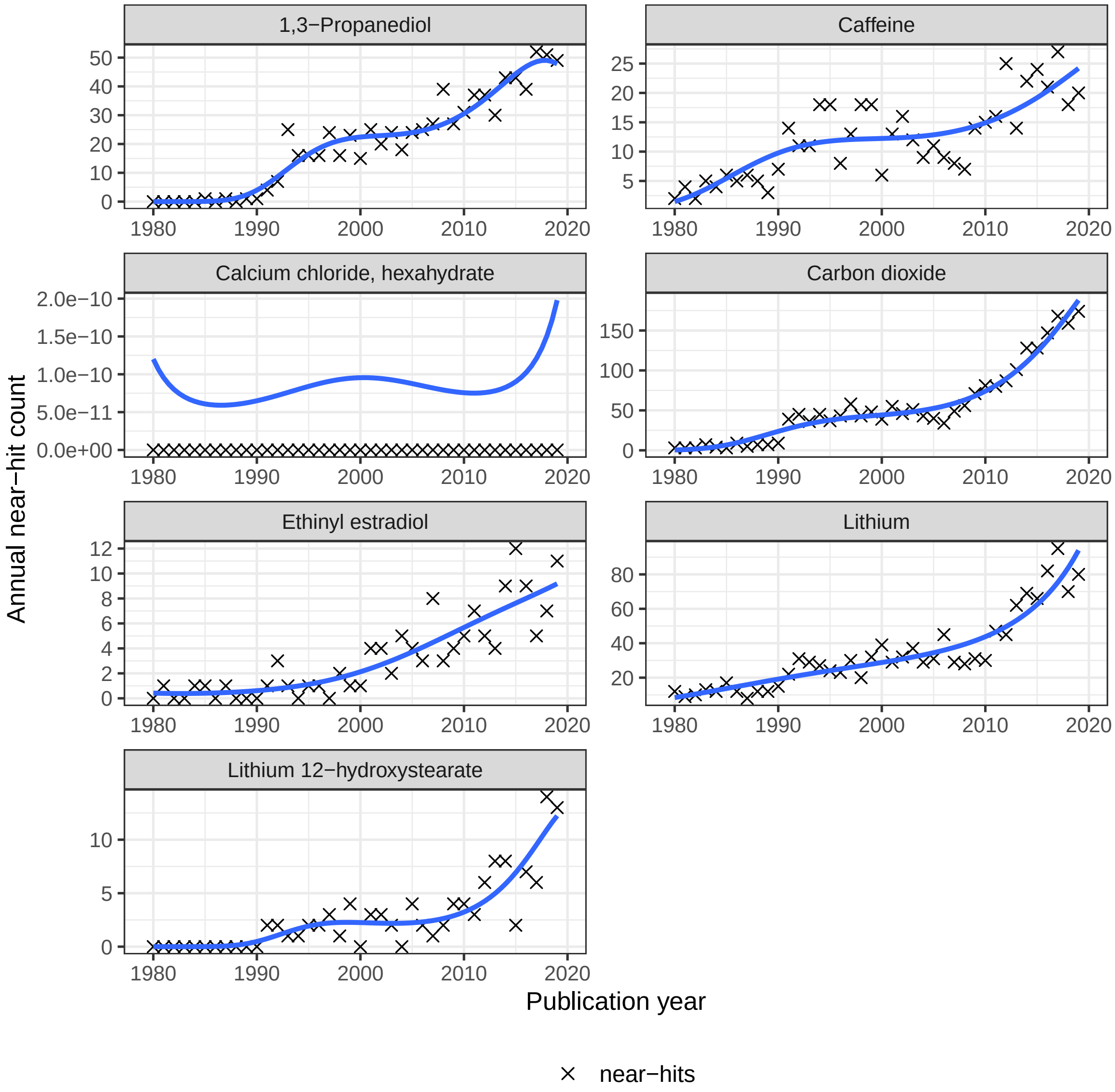


**Figure 4.10.:** Group 3A near-hits data (shown with crosses) with a smoothing line to show the fitted systematic component.

⋆ ∞ ⋆

In the next chapter we explore concepts of value to the classification system that we intend to apply to near-hits time series for a range of chemicals. We employ these concepts in formulating our system.

# 5. Preparation for the classification of chemicals into groups

We now turn our attention to the formulation of a classification system (henceforth, 'system') which we can apply to the research interest data associated with a chemical (here, a time series of hits returned by a query). Such a system should classify a chemical as CoI, or not a CoI, at a certain point time, given the data available up to that time. Recall the schematic summarising the development of our system in Figure 2.2. We shall expand upon that summary in this chapter.

We begin by noting some preliminary concepts in Section 5.1. We provide a discussion of desirable features and abilities of a classification system in Section 5.2. We proceed to outline certain key assumptions regarding features of the data and our approach to classification in Section 5.3. Finally, we define our classification system in Section 5.4.

## 5.1. Preliminaries

The development of our system is initially informed by scrutiny of the research interest data obtained for chemicals drawn from Group 1A (as designated by NSW EPA) and Group 3A (randomly selected from Group 3). Following the discovery of particular "features" in this "training set" (recall Section 4.3) we proposed tests aiming to recognise these features in a time series. Combining test results leads to a classification of each chemical. If the classification accuracy[1] is inadequate, we adjust features of the system so as to improve this. That is, we will train our classification system such that it delivers an adequate classification accuracy when applied to our training set.

Following this, we evaluate the usefulness of our trained system by applying it to a "validation set" of research interest data. This data is associated with chemicals from Group 1B (as designated by NSW EPA) and Group 3B (those chemicals from Group 3 not in Group 3A). We note that the training and validation sets relate to distinct collections of chemicals. That is, research interest for some chemical is not used in both training and validation.

**Remark 7** *Following advice from NSW EPA, our training and validation sets include the time series of hits associated with individual chemicals and subfamilies of chemicals in proposing our tests and outcomes. These data sets do not include research interest for families of chemicals. We emphasise that we choose to consider features of near-hits time series in this study.*

We proceed to present certain desirable attributes of a system in the next section.

[1] We describe a system's performance as "accurate" if the system correctly allocates a large proportion of input chemicals (but not necessarily all of them) to the correct group. We shall discuss this further in Section 5.2.

## 5.2. Desirable features and abilities of a classification system

A system suitable for our purposes must balance competing objectives:

**Objective 1** correctly classify as many CoI (true positives, that is, Group 1 chemicals) and non-CoI (true negatives, that is, Group 3 chemicals) as possible.

**Objective 2** limit the mis-classification of CoI (that is, false negatives) and non-CoI (that is, false positives).
A false negative conceals a CoI, whilst a false positive may cause a regulator to expend effort on an unnecessary chemical assessment.

It is appropriate for a regulator to decide on what it considers an acceptable balance between Objective 1 and Objective 2. We may judge a classification system as sufficiently accurate for a regulator's needs (and possibly, available resources) if it can achieve (at least) the regulator's acceptable balance when applied to a validation set. We understand that NSW EPA would use a classification system such as ours within a broader risk assessment system. As such, NSW EPA "would far prefer false positives than false negatives" (advice from NSW EPA staff) to result from our classifications. This preference will inform our approach to assessing our system.

**Remark 8** *In this study we have to negotiate the limitations of extremely small data sets — recall Groups 1B and 3A each contain only seven chemicals. In such cases, each single misclassification translates into a substantial loss of accuracy. We attempt to minimise the impediments posed by small data sets to our explorations. We manage this by requiring a modest minimum percentage of correct classifications for each of Group 1A, Group 1B, Group 3A, and Group 3C of 70%. This modest target allows us to misclassify two of seven chemicals and still obtain a satisfactory result for Group 1B and Group 3A.*

We consider a trained system as validated if it has the following ability:

**Ability 1** Given some validation set, the trained system can distinguish Group 1 chemicals from Group 3 chemicals with acceptable accuracy.

We may hope that a validated system has further abilities:

**Ability 2** Given research interest for Group 2 chemicals, the system can accurately classify these as belonging to either Group 1 or Group 3.

**Ability 3** Suppose we have the research interest for a set of 'novel' chemicals (not used in system training or validation), each of which actually belongs to either Group 1 or Group 3, and the grouping is unknown. The validated system can accurately classify these.

We will evaluate whether or not our system has Ability 1 in Chapter 6.[2]

Additionally, we expect that a useful classification system will deliver timely insights to a regulator. These insights should assist a regulator in implementing appropriate

[2] Owing to various matters relating to the supplied data quality and its impact on the performance of our system, it is not appropriate to consider Ability 2 and Ability 3 in this study. We may return to these matters in a future project.

actions so as to limit the adverse effects of chemicals. As such, we require that a classification system can produce some year (outcome) in which a classification is made. Then, one may judge the performance of the classification system on a training or validation set by comparing the outcome with the year of some regulatory awareness or action (recall Section 2.1.3).

In the next section we consider some assumptions that underpin our process of creating a system that can perform accurately.

## 5.3. Data, modeling, and queries: assumptions

Key assumptions underpin our pursuit of an accurate — and hence useful — system. These assumptions relate to four matters. The first matter concerns the classifications given to the chemicals which feature in our training and validation sets.

**Assumption 1** Recalling Section 2.1.3, the classifications of supplied chemicals as belonging to Group 1 (defined by the collection of Group 1A and Group 1B chemicals) or Group 3 (the collection of Group 3A and Group 3B chemicals) are generally accurate.

The second matter relates to the nature of the research interest (however this is defined) associated with the chemicals in our training and validation data sets.

**Assumption 2** As a general rule, the research interest for Group 1 chemicals (for example, as shown in Sections 4.1.1 and 4.1.2 respectively) exhibit features that are not seen in the analogous research interest for Group 3 chemicals (as shown in Sections 4.1.3 and 4.1.4).

If Assumption 1 or Assumption 2 is incorrect, we may focus on certain features in training and validating our system that are inappropriate predictors of a chemical's true group. As a result, we could not expect the system to display Ability 1, or Ability 2, or Ability 3.

The third matter relates to the research interest of novel chemicals to which we would like to apply our system.

**Assumption 3** Suppose we have the research interest of novel chemicals that belong to either Group 1 or Group 3 (the actual group to which each chemical belongs is unknown). As a general rule, the novel chemicals from Group 1 or Group 3 exhibit comparable features to chemicals from the same group that were used in system training and validation.

If Assumption 3 is incorrect, then, analogously to when Assumption 1 or Assumption 2 is incorrect, we cannot expect our system to display Ability 3.

**Remark 9** *We do not include Group 2 chemicals in our test cases as this is an inhomogeneous group; in the future some chemicals may become CoI, and others may be reclassified as Group 3 chemicals. In order to obtain accurate results, training data for a classification system should feature groups that are as homogeneous as possible. That is, a group designation for a chemical (or "label" in the terminology of Machine Learning) should have a particular meaning if it is to be useful for classification. This is illustrated by (with our bold face for emphasis):*

> *The labels used to identify data features must be* ***informative, discriminating and independent*** *to produce a quality algorithm. A properly labelled dataset provides a ground truth that the ML [Machine Learning] model uses to check its predictions for accuracy and to continue refining its algorithm.* Wigmore (2019)

The fourth matter (related to Assumption 2 and Assumption 3) arises from our choice to use near-hits time series as our research interest.

**Assumption 4** We may choose modifiers in our near-hits queries such that the research interest returned does not invalidate Assumption 2 and Assumption 3.

Suppose that an inspection of the near-hits time series returned for chemicals across Group 1 and Group 3 shows that Assumption 4 is not valid. It is then appropriate to revise the modifiers employed, to obtain research interest for the revised near-hits query, and then to reconsider the validity of Assumption 4. Experimenting in such a fashion, we may be able to produce results that satisfy our assumption, giving us confidence that it is appropriate to apply our system to data.

There was little time available for such experimentation in this pilot study. At this stage we have noted some concerns around the validity of Assumption 4. However, we do not have conclusive evidence that the assumption is unreasonable. Thus, for the purposes of this study, we proceed assuming (at least initially) that our modifiers ensure that Assumption 1 and Assumption 2 are valid. These assumptions are necessary for our system training and validation. We shall critically examine Assumption 2 later following the application of our system to research interest.

In the next section we present the components of our classification system.

## 5.4. Outline of our classification system

The system we employ in this project shares similarities with, and follows on from, ideas developed in Whyte (2020). The system consists of two parts. The first is a set of tests, where each test seeks to recognise the presence (or absence) of certain features in the research interest for a given chemical. (Data exploration leads us to propose these features, and we intend these to relate to emerging interest.) Each test is associated with an outcome: the year in which the test is first satisfied, or a non-result otherwise. The second part is a 'classification rule' that combines outcomes for each chemical to produce a classification.

Following inspection of the data, we proposed particular tests, which depend on particular parameters. After some experimentation, we have chosen parameter values which allow our system to exhibit acceptable classification accuracy on the training set. As such, the system presented below is our trained classification system.

### 5.4.1. Proposed tests

Inspection of the near-hits time series (outlined in Section 4.3) suggested some specific tests for investigation. We shall present some prospective tests (and their associated outcomes) that we will apply to our training and validation sets.

We expect that a system useful to a regulator would employ "online processing" of data as it becomes available. That is, we expect that a regulator would apply a system to the data available up to the present, and, lacking exact predictions of the future, would choose to reapply the system when more data becomes available. This would allow a regulator to update its classification of a given chemical, and then to react accordingly. We mimic this situation in designing our tests. Although we have a near-hits time series over the publication range of 1980–2019, we do not begin by considering the entire data set. Instead, we start from the earliest year of the publication range under consideration and use subsets of the data that include later years only if necessary. At no point do we behave as if we can 'predict the future' and access data beyond the range of the subset we currently consider. This gives us the ability to judge if we can detect features of near-hits via our online processing. (This principle is shown most clearly in Tests 5 and 6 below.)

**Remark 10** *The alternative to online processing is to use the entire data set from 1980–2019 ("batch processing"). Although we may obtain more accurate estimates of features of near-hits time series (e.g. by fitting some model to the entire data set), we can only claim that a test for these features is satisfied in the data set's final year. Clearly this cannot suit our aim of providing a regulator with timely insights.*

Our first three tests directly relate to properties of the near-hits values.

**Test 1** Sustained initial interest:
Starting from 1980, consider overlapping intervals of consecutive years, each of a three-year duration. Is there any interval for which the near-hits meets or exceeds 5 hits in each year?

**Outcome 1** If yes, for the earliest interval that satisfies the test, record the final year. If no, record 'No result'.

**Test 2** Volume of interest:
Starting from 1980, does the cumulative sum of near-hits reach at least 150?

**Outcome 2** If yes, record the earliest year in which this occurs. If no, record 'No result'.

**Test 3** Year of substantial interest:
Starting from 1980, is there a year in which the near-hits reaches at least 10?

**Outcome 3** If yes, record the earliest year in which this occurs. If no, record 'No result'.

Tests 4, 5, and 6 consider features of the changes or trend in near-hits. We may explore this initially by considering the first-order difference in near-hits values between some year and the year immediately prior. To explain, we use $y(t)$ to represent the near-hits value in year $t \geq 1980$. Then, we represent the first-order difference in $y$ at time $t'$ as

$$\Delta(t') = y(t') - y(t'-1), \quad t' \in \{1981, \ldots, 2019\} . \tag{5.1}$$

When $\Delta(t') > 0$ (alternatively, $\Delta(t') < 0$) the near-hits is increasing (alternatively, decreasing) relative to the previous year's value. When $\Delta(t') = 0$, there is no change in near-hits from year $t' - 1$ to year $t'$.

Recalling Observation 1, we propose a test which aims to capture the event of an abrupt increase in near-hits.

**Test 4** Abrupt increase in interest:
Starting from 1981, is there some year $t$ in which $\Delta(t) \geq 5$?

**Outcome 4** If yes, record the earliest year that satisfies the test. If no, record 'No result'.

As an extension of Test 4, we may investigate near-hits to determine whether or not there is a sustained increase. Recall Observation 2, which is seen in various figures in Section 4.1 as approximately linear growth in near-hits on the log10 vertical scale, possibly after an initial sharp increase. We intend to design a test (Test 5 below) that can detect this behaviour over an interval of four consecutive years.

**Test 5** Sustained increase in interest:
Process the near-hits time series such that, if it contains any zero values, we retain only the data which occurs after the latest zero. Calculate the log10 transform of near-hits (our dependent variable) to yield a 'processed' data set. Starting from the earliest year of the processed data, consider every (overlapping) interval of four consecutive years. For each interval:

- Scale the years so that this (independent) variable now takes values $t = 1, 2, 3, 4$.
- Calculate the line of best fit by a linear regression of the independent variable against the dependent variable. The equation of the regression line is $\log_{10}(y_t) = m \cdot t + c + \epsilon_t$, where the slope, $m$, and the $y$-intercept of the line, $c$, are calculated from the data. As we do not expect an exact linear relationship, we use $\epsilon_t$ to model the random error at time $t$, and we assume its mean value is zero.
- Record $m$, and associate this with the final year of the interval.

Starting from the earliest year of the processed data set for which there is a calculated $m$ (this is the data set's earliest year plus three, given the four-year intervals considered), is there some year in which $m \geq 0.2$?

**Outcome 5** If yes, record the the earliest year that satisfies the test. If no, record 'No result'.

**Remark 11** *We note the near-hits time series may show a generally increasing trend that has some variability; some years may show a decrease in hits. As such, a test for uniform growth in near-hits cannot capture the (potentially instructive) generally increasing trend. As Test 5 employs a line of best fit, it is able to recognise a sustained — but not necessarily constant — growth of research interest.*

Tests 1–5 only consider the short-term behaviour of a near-hits time series, which limits the features of research interest behaviour we can recognise. Suppose that a Group 3 chemical exhibits behaviour similar to that of a Group 1 chemical early in its publication history, but that research interest over the longer term is quite dissimilar to that of a Group 1 chemical. A test which considers only the short-term behaviour of research interest may fail to distinguish a Group 3 chemical from a Group 1 chemical. As such, a classification system which considers only short-term research interest may produce misleading results. Hence, it is appropriate for us to also consider the longer-term behaviour of research interest in case this is useful to our classification task.

A consideration of longer-term research interest allows us to experiment with a type of test that has more complicated conditions than those presented earlier. This should allow more flexibility for 'tuning' the test (as in system training), which may allow us to improve the test's ability to accurately recognise chemicals from a particular group. Some motivation for experimentation with tests is provided by Observation 3, relating to features of the near-hits for Group 1B chemicals. The observation noted instances where research interest progressed through three distinct phases: a period of sustained growth, a relatively flat period, and finally a second growth period of a larger rate of increase than was seen in the first growth period. We will shortly propose a test able to recognise such behaviour (Test 6, below). The test will consider the systematic component of a near-hits time series (recall Section 4.3.2), as this allows us to recognise the phases of near-hits behaviour noted above more readily than does the raw data. The test considers the systematic component over an interval of at least 10 consecutive years. (Such an interval length gives us the opportunity to observe sustained behaviour, rather than short-term movement that is not sustained.)

**Test 6** Well-separated times of growth in the systematic component of near-hits: Suppose we calculate the systematic component of a time series, $s(t)$, on some time interval. We are interested in the behaviour of the first-order differences in $s(t)$, $\Delta_s(t)$. Our test has two conditions, where the first must be satisfied in order to proceed to checking the second. First, we use $\Delta_s(t)$ to establish whether or not $s(t)$ can exhibit at least some minimum rate of increase on our interval of interest. If so, we consider if there is some time at least a minimum number of years ahead on this same interval after which $s(t)$ increases at a greater rate.

In this test we specify inputs:

- A window length $w > 0$ which defines an interval from 1980. (Initially we set $w = 10$ years, this value may change as the algorithm proceeds.)
- Two (fixed) minimum values of $\Delta_s(t)$: $m_1$ and $m_2$ (units of hits/year), where $m_2 > m_1$. (We use $m_1 = 0.5$ and $m_2 = 2$.)
- Some (fixed) minimum time difference $d$ between when $\Delta_s(t)$ first reaches $m_1$ and the year at which we commence a search for when $\Delta_s(t)$ first reaches $m_2$. (We use $d = 5$ years.) This allows us to ignore the short-term behaviour of $s(t)$ after $\Delta_s(t)$ first exceeds some minimum threshold, as this may not be indicative of longer-term behaviour.

The algorithm proceeds as follows:

Step 1 Considering the time interval $T_w = \{1980, \ldots, 1980 + w - 1\}$, estimate $s(t)$. Use this in calculating $\Delta_s(t)$ for all years in $T'_w = \{1981, \ldots, 1980 + w - 1\}$. Proceed to Step 2.

Step 2 Determine the earliest year $\tau_1 \in T'_w$ in which $\Delta_s(\tau_1) \geq m_1$.

a) If such a $\tau_1$ exists, proceed to Step 3.

b) If $\tau_1$ does not exist:

i. If $1980 + w - 1 < 2019$: we have more possibilities to search. Increment $w$ and return to Step 1.

ii. If $1980 + w - 1 = 2019$: incrementing $w$ would cause us to exceed the span of the data available, so exit.

Step 3 a) If $t = \tau_1 + d$ is within $T'_w$, proceed to Step 4.

b) Otherwise,

i. If $1980 + w - 1 < 2019$: increment $w$ and return to Step 1.
ii. If $1980 + w - 1 = 2019$: exit.

Step 4 a) If there exists some $\tau_2 \geq \tau_1 + d$ (which must belong to $T'_w$) such that $\Delta_s(\tau_2) \geq m_2$, record $w$ and exit.
b) If $\tau_2$ does not exist:
i. If $1980 + w - 1 < 2019$: increment $w$ and return to Step 1.
ii. If $1980 + w - 1 = 2019$: exit.

**Outcome 6** If the algorithm exits having recorded some $w$, we use this in recording the latest year of data needed for the judgement that the test was satisfied — $t = 1980 + w - 1$. Otherwise record 'No result'.

**Remark 12** *In applying Test 6 to our data, in each case where Outcome 6 was a year (not a 'No result') we checked the output of the model fitting process to verify that it had "converged" to a minimum of the error function, as we require. As such, we can trust that the outcome returned is legitimate. If the fitting process did not converge in some case, we would not accept any associated outcome year, and would need to adjust and repeat the fitting process so as to achieve convergence. We may still be able to refine the process in any future project. We have noted convergence checking as an issue of importance when considering a larger set of chemicals, for which manual inspection of results and intervention in the model fitting process might not be feasible.*

### 5.4.2. A proposed classification rule

Following some experimentation, we propose a classification rule.
If we obtain a 'No result' for at least one of Test 1, 5, or 6 for a chemical, classify it as not CoI. Otherwise, the chemical is a CoI.

⋆ ∞ ⋆

We shall consider the usefulness of our trained system by applying it to near-hits time series in Chapter 6.

# 6. Application of our trained classification system

In this chapter we examine the value of our trained classification system (as outlined in Section 5.4). We begin in Section 6.1 by showing that the system can classify chemicals from Group 1A and Group 3A with sufficient accuracy, as defined in Remark 8. (It is for this reason that we refer to the system as "trained".) We show the test outcomes graphically, and for Group 1A chemicals we also show elements of the regulatory histories (concern or action of regulatory bodies). We proceed to present another view of our results which allows us to readily appreciate whether the test outcomes for each chemical in Group 1A lead (occur before) or lag (occur after) aspects of the associated regulatory history. We follow a similar approach in Section 6.2 where we validate our system by application to the research interest of chemicals from Groups 1B and 3B.

As we have quite small collections of chemicals in our training and validation sets, extensive formal analysis of results is not justified. We shall limit ourselves to a consideration of the percentage of chemicals correctly classified by our system, and a discussion of the usefulness of our proposed tests. We note that it is usual to evaluate the performance of a classifier against the entirety of the training (or, as appropriate, validation) set. Here, given NSW EPA's expressed desire to avoid false negatives, we break with convention. In evaluating the accuracy of our classification system when applied to the training set, we consider the CoI (Group 1A) and other chemicals (Group 3A) individually. Similarly, regarding our validation set, we consider our system's performance on Group 1B and Group 3B individually. We provide some discussion of results in Section 6.3.

For ease of reference, we summarise the functions of the tests we shall apply to the research interest time series (defined in Section 5.4.1) in Table 6.1.

## 6.1. Results for application of our trained system to chemicals from Groups 1A and 3A

The outcomes obtained by application of Tests 1–6 to Group 1A chemicals are shown (with appropriate numerals) in Figure 6.1. The figure also shows (with symbols) features of Australian and International regulatory histories. By considering the figure's column for each chemical individually, we obtain a first view of those tests able to produce an outcome which leads one (or more) features of a chemical's regulatory history.[1] (For an alternative view of these results, see Appendix B, Table B.1.) By applying our classification rule to the recorded test outcomes, we classify ten of eleven chemicals as CoI (true positives), and only TCE as not CoI (a false negative). This yields a classification accuracy above 90%.

[1] We will provide a discussion of the performance of tests on Group 1A chemicals in Section 6.1.1.

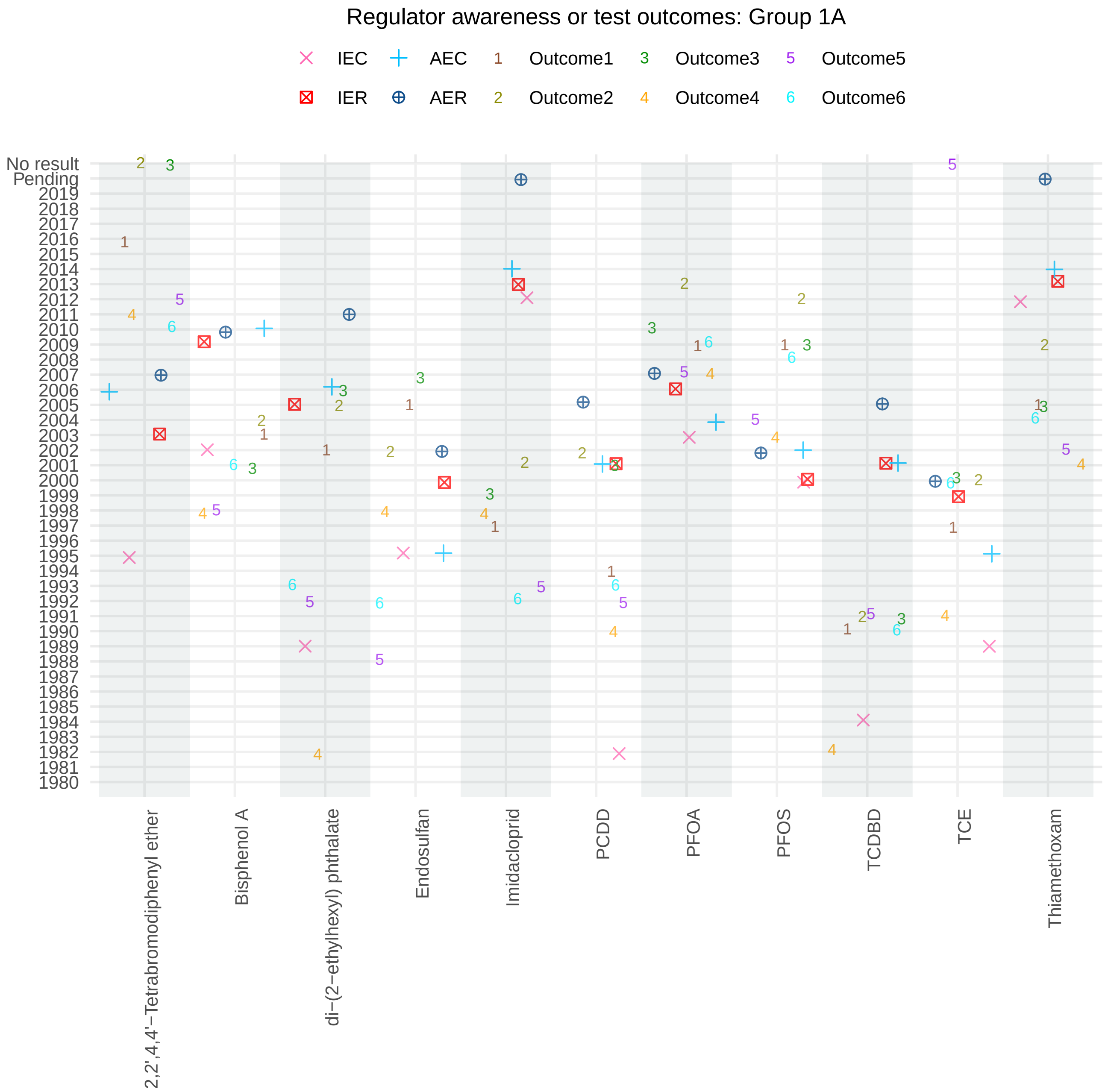


**Figure 6.1.:** Group 1A chemicals: summary of regulatory histories and test outcomes. Key: IEC = international earliest concern, IER = international earliest regulation, AEC = Australian earliest concern, AER = Australian earliest regulation.

**Table 6.1.:** A brief summary of the function of tests applied to near-hits time series. If a test is not satisfied, its outcome is "No result". Otherwise, the test outcome is the first year from the start of the studied publication range in which all test conditions are satisfied.

| Test | Summary |
|---|---|
| Test 1 | **Sustained initial interest:** Starting from 1980, is there any three-year interval for which the near-hits meets or exceeds five hits in each year? |
| Test 2 | **Volume of interest:** Starting from 1980, does the cumulative sum of near-hits reach at least 150? |
| Test 3 | **Year of substantial interest:** Starting from 1980, is there a year in which the near-hits reaches at least 10? |
| Test 4 | **Abrupt increase in interest:** Starting from 1981, is there some year $t$ in which the increase in near-hits counts from years $t-1$ to $t$ is at least 5? |
| Test 5 | **Sustained increase in interest:** disregard any part of the near-hits data which occurs before the first non-zero value. Transform the retained data. Does the slope of a line of best fit applied to the transformed data from an interval of four consecutive years exceed some minimum (positive) threshold? |
| Test 6 | **Well-separated times of interest growth:** it is necessary to transform the near-hits series so as trends are more readily apparent. Does the transformed near-hits series show at least some specified growth within some interval, and then larger growth in some subsequent interval? |

We present the Group 3A test outcomes in Figure 6.2. We classify two of seven chemicals (1,3-Propanediol and Carbon dioxide) as CoI (false positives). We classify the remaining five chemicals as not CoI (true negatives). Test 1 contributes to these (appropriate) classifications by returning ‘No result’ in three of five cases. Test 5 performs equally well, enhancing the classification process by recognising two cases that are not detected by the other tests which contribute to our classification rule. Test 6 contributes a ‘No result’ in two of the five cases. Overall, our system shows a classification accuracy above 71%.

As the results for our training set exceed our minimum requirement, we proceed to apply our (trained) system to the validation set in Section 6.2. However, first we shall present our Group 1A results in a form that aids the comparison of test outcomes for each chemical against particular elements of the chemical’s associated regulatory history.

**Remark 13** *Although the classification accuracy for Group 3A chemicals is not especially good, recall Remark 8 on the difficulties posed by small data sets. By choosing a modest classification accuracy target in this pilot study, and achieving this target on the training set, it is appropriate for us to proceed to validating our trained system so that we may continue the exploration of our system’s strengths and weaknesses.*

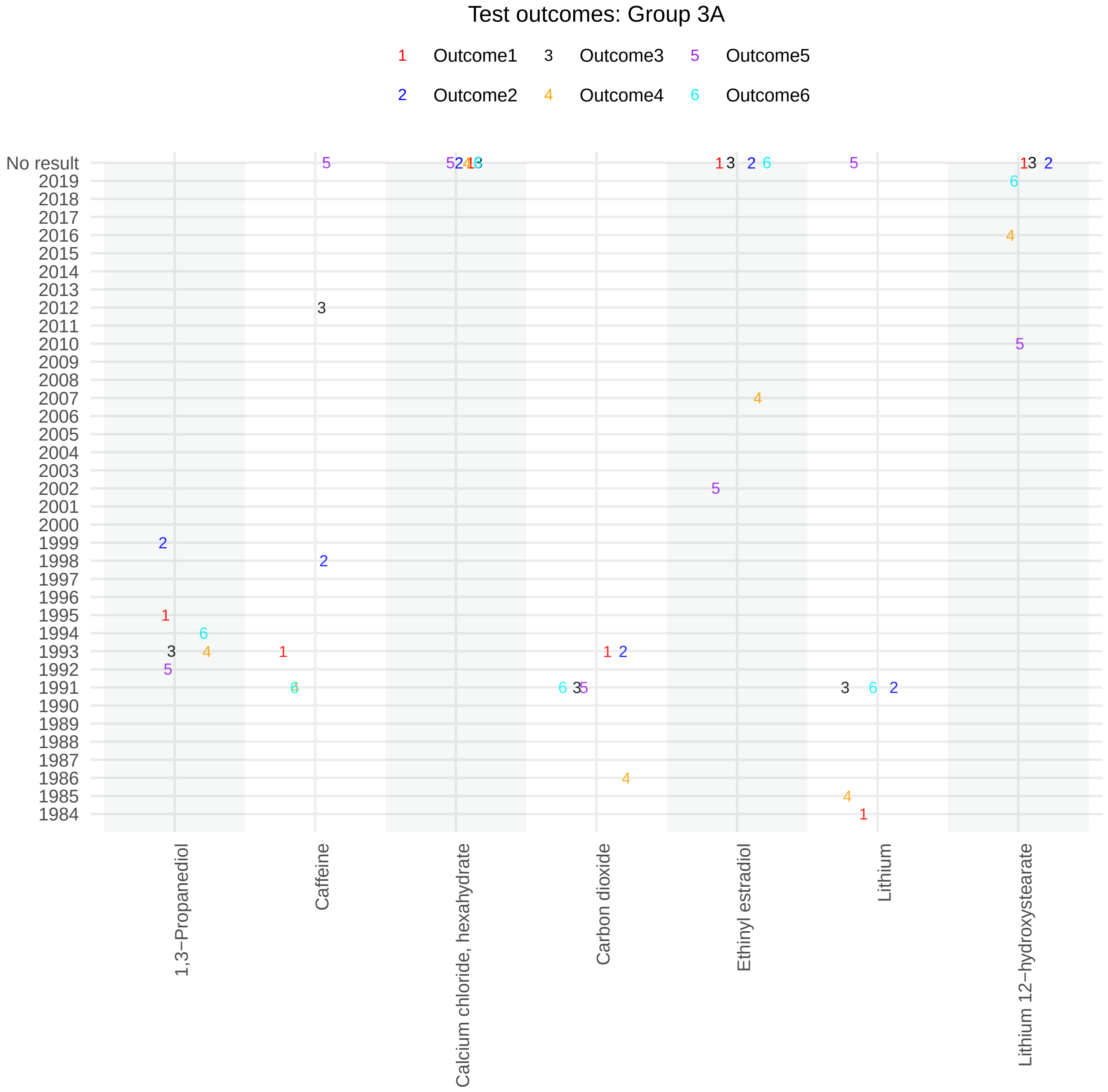


**Figure 6.2.:** Group 3A chemicals: test outcomes.

## 6.1.1. Comparisons of Group 1A test outcomes against regulatory histories

Recall from Remark 3 that a comparison of test outcomes for Group 1 chemicals against associated IEC was our most stringent test of system performance. To assist the determination of whether test results for Group 1A chemicals lead (occur before, useful) or lag (occur after, not useful) the associated IEC, we present differences between these values in Figure 6.3. The label Diff1 indicates the year of IEC minus Outcome 1 if this exists, otherwise Diff1 is set to ‘No result’. Diff2 to Diff6 are defined similarly. The figure shows that in six of eleven cases (seen reading the figure from left to right):

1. Bisphenol A,
2. di-(2-ethylhexyl) phthalate,
3. Endosulfan,
4. Imidalacloprid,
5. TCDBD,
6. Thiamethoxam,

each a true positive, at least one test outcome leads the IEC. Notably, Outcome 4 leads the IEC five of these six cases. In two of the six cases, Outcome 4 is the only outcome to lead the IEC. Test 5 and Test 6 also show some value; in four of the six cases, both Outcome 5 and Outcome 6 lead the IEC. This suggests that there may be some value in further experimentation with tests that consider the rate of change of a near-hits time series.

We present a similar view of the ability of Group 1A test outcomes to anticipate IER in Figure 6.4. (For this figure we redefine a label such as Diff1 such that it represents the quantity IER minus Outcome 1.) We exclude TCE from further discussion as our system has judged this as a false negative. Noting this exclusion, the figure shows that in seven of eleven cases, the cases 1–6 listed above, and

7. PCDD,

at least three test outcomes lead the IER. Outcome 4 features in each of these seven cases, and in five cases Outcome 4 is the earliest (or equal earliest) year returned. Outcome 5 and Outcome 6 also lead the IER in each of the seven cases.

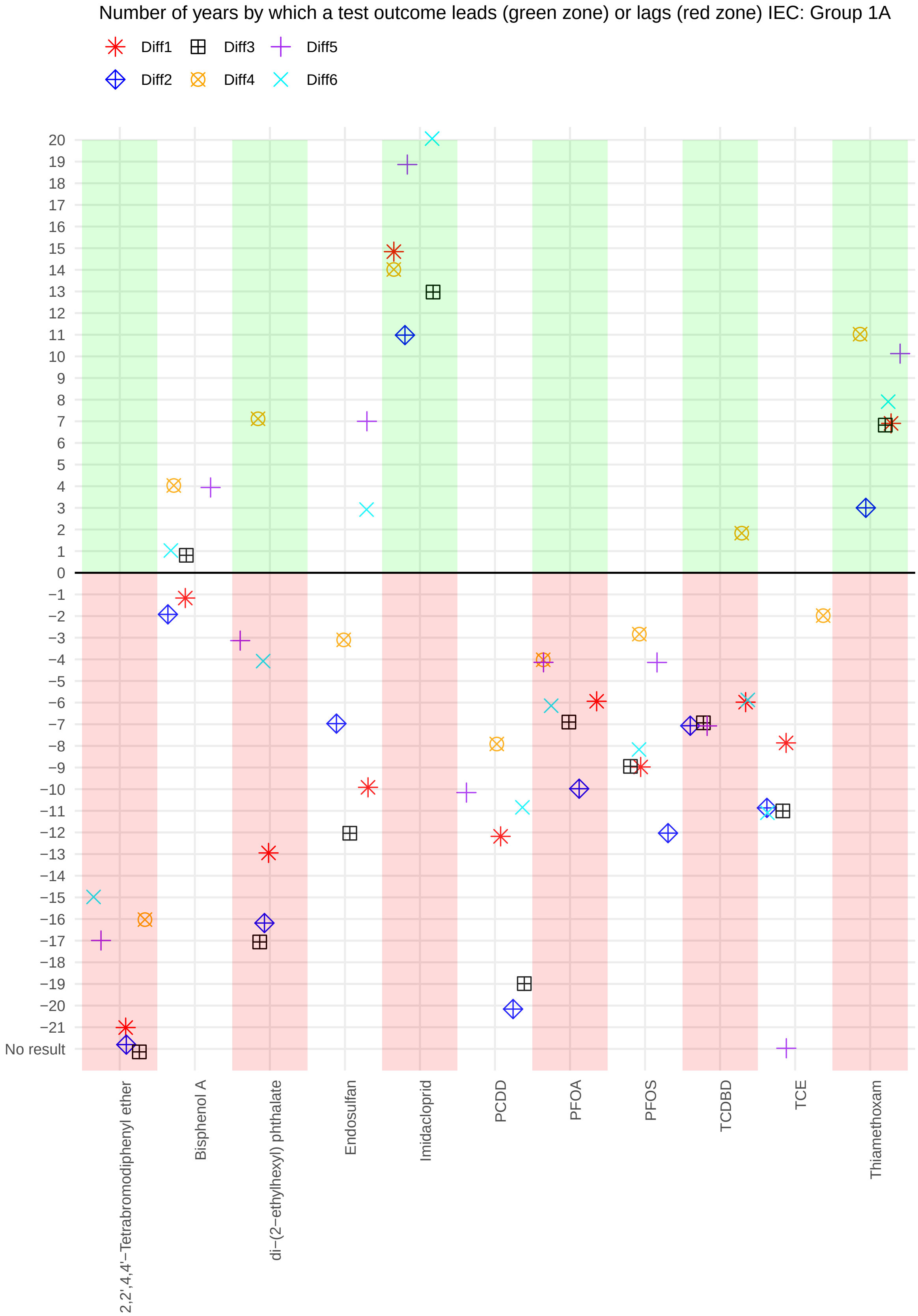


**Figure 6.3.:** Group 1A chemicals: year differences between IEC and test outcomes.

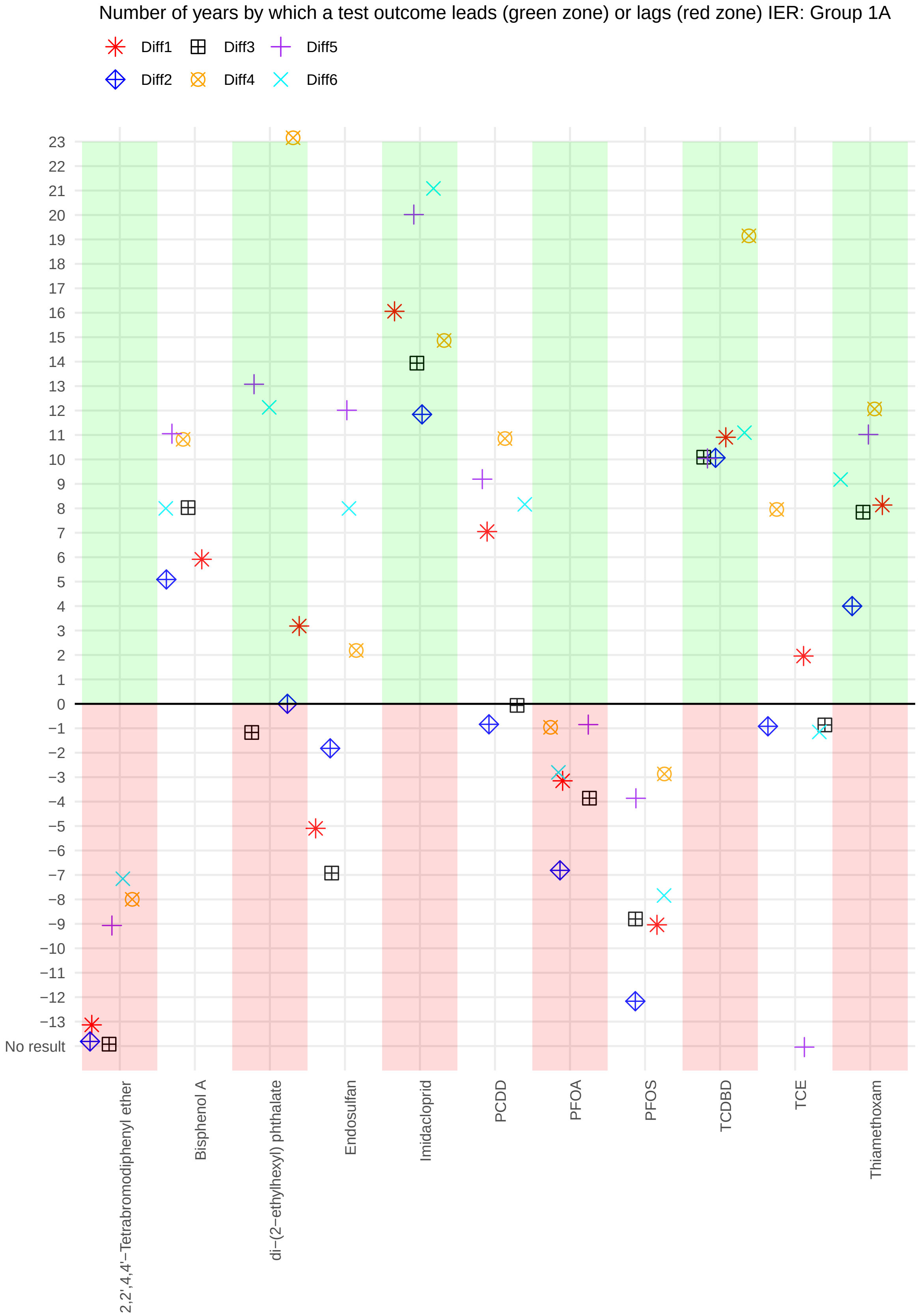


**Figure 6.4.:** Group 1A chemicals: year differences between IER and test outcomes.

Although it is not a focus of this report, in the interests of exploring information supplied to CEER by NSW EPA, we will briefly consider the ability of our test outcomes to lead AEC or AER. We show comparisons of test outcomes against AEC or AER in Figures 6.5 and 6.6, respectively. (In each case, Diff1 to Diff6 are defined with reference to either AEC or AER, as appropriate. Also, recall that the AER for a Group 1 chemical cannot occur prior to the AEC.) Figure 6.5 shows that at least two outcomes lead the AEC in seven cases (the true positives, shown as cases 1–7 on Page 53) out of eleven cases. Figure 6.6 may appear to contradict our expectations, as it shows that at least three outcomes lead the AER in only five of eleven cases, seemingly a worse result than seen for the AEC comparison. We can explain this as two chemicals, Imidacloprid and Thiamethoxam, do not have an associated AER. As such, we cannot define terms such as Diff1, and must show these in the 'No result' row of the graph.

To give some idea of the value of our results, we note that Figure 6.5 shows Outcome 5 leading AEC for all seven true positives, by 7 to 21 years. This suggests that, even though this is a limited study, our approach has the potential to anticipate Australian concern. Based on this, we expect further development of our approach to positively influence Australian regulatory practices.

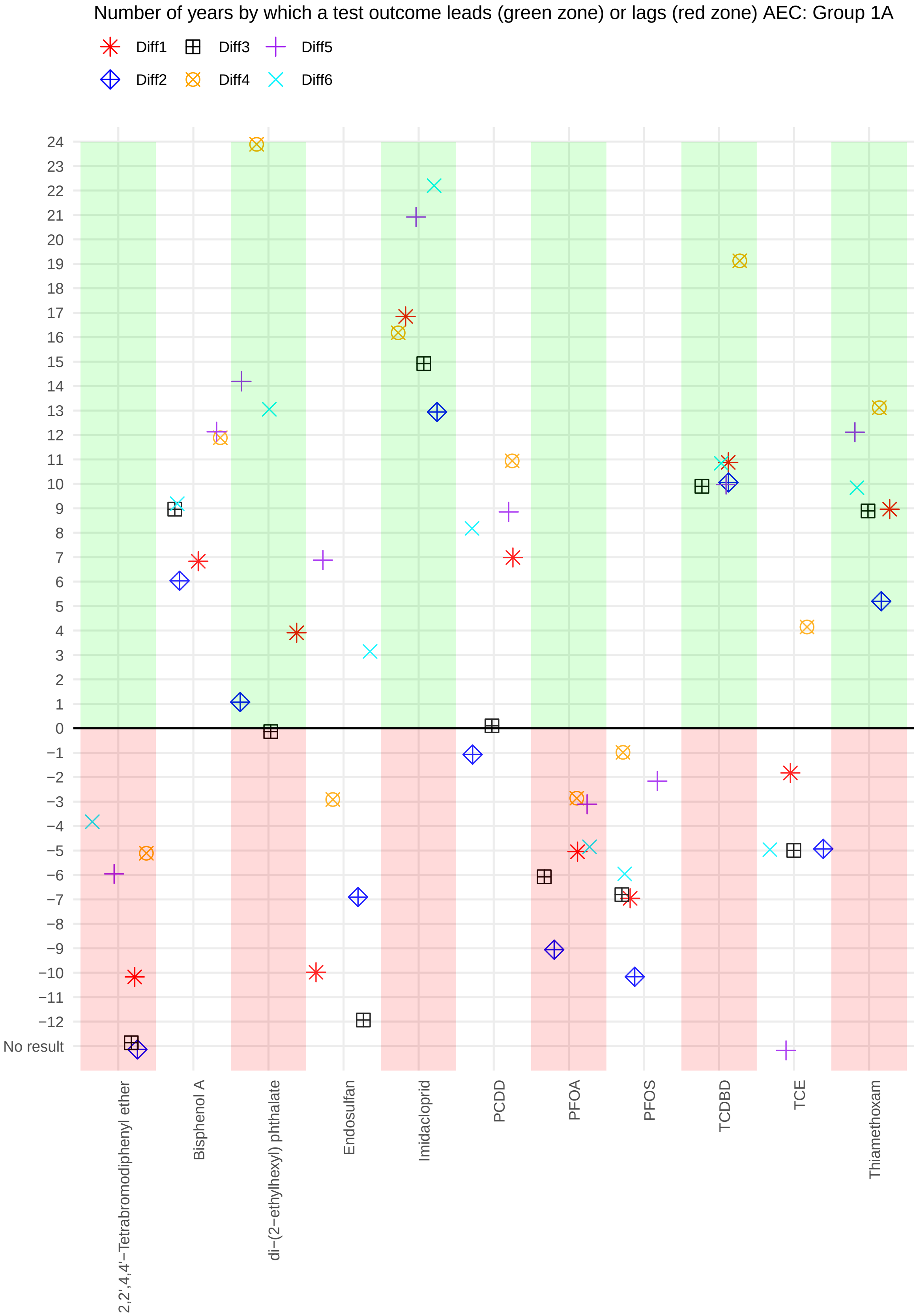


**Figure 6.5.:** Group 1A chemicals: year differences between AEC and test outcomes.

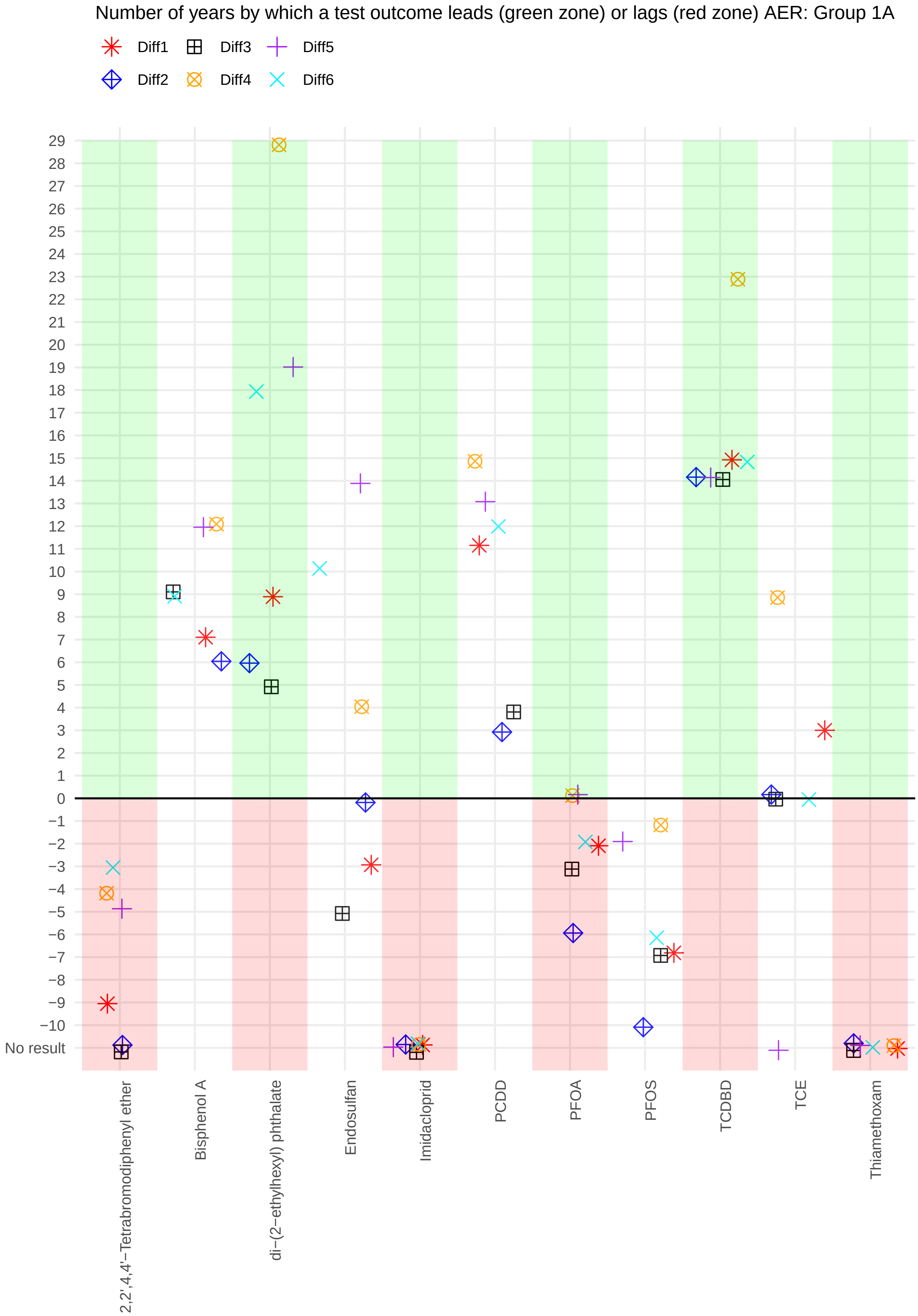


**Figure 6.6.:** Group 1A chemicals: year differences between AER and test outcomes.

## 6.2. System validation on chemicals from Groups 1B and 3B

In considering the results for our validation set, we proceed in a similar manner to that employed in discussing the training set results in Section 6.1. Outcomes of tests applied to the research interest of Group 1B chemicals, with attending regulatory histories, are shown in Figure 6.7.[2] (For an alternative view, see Appendix B, Table B.2.) Using our classification rule, we classify four of seven chemicals as not CoI (false negatives). We see these by reading from left to right in Figure 6.7: 1,1,2,2-tetrachloroethane, Clothianidin, diethyl phthalate, and PCE. We classify the remaining three chemicals:

1. Methylmercury,
2. TCDD,
3. Triclosan

as CoI (true positives). This yields a classification accuracy of approximately 42%, which is inadequate for our purposes. As such, we have not validated our system. However, in the interests of completing our investigations, we will continue to consider system performance on the research interest of Group 3B chemicals.

[2] We provide further comment on the performance of our tests in Section 6.2.1.

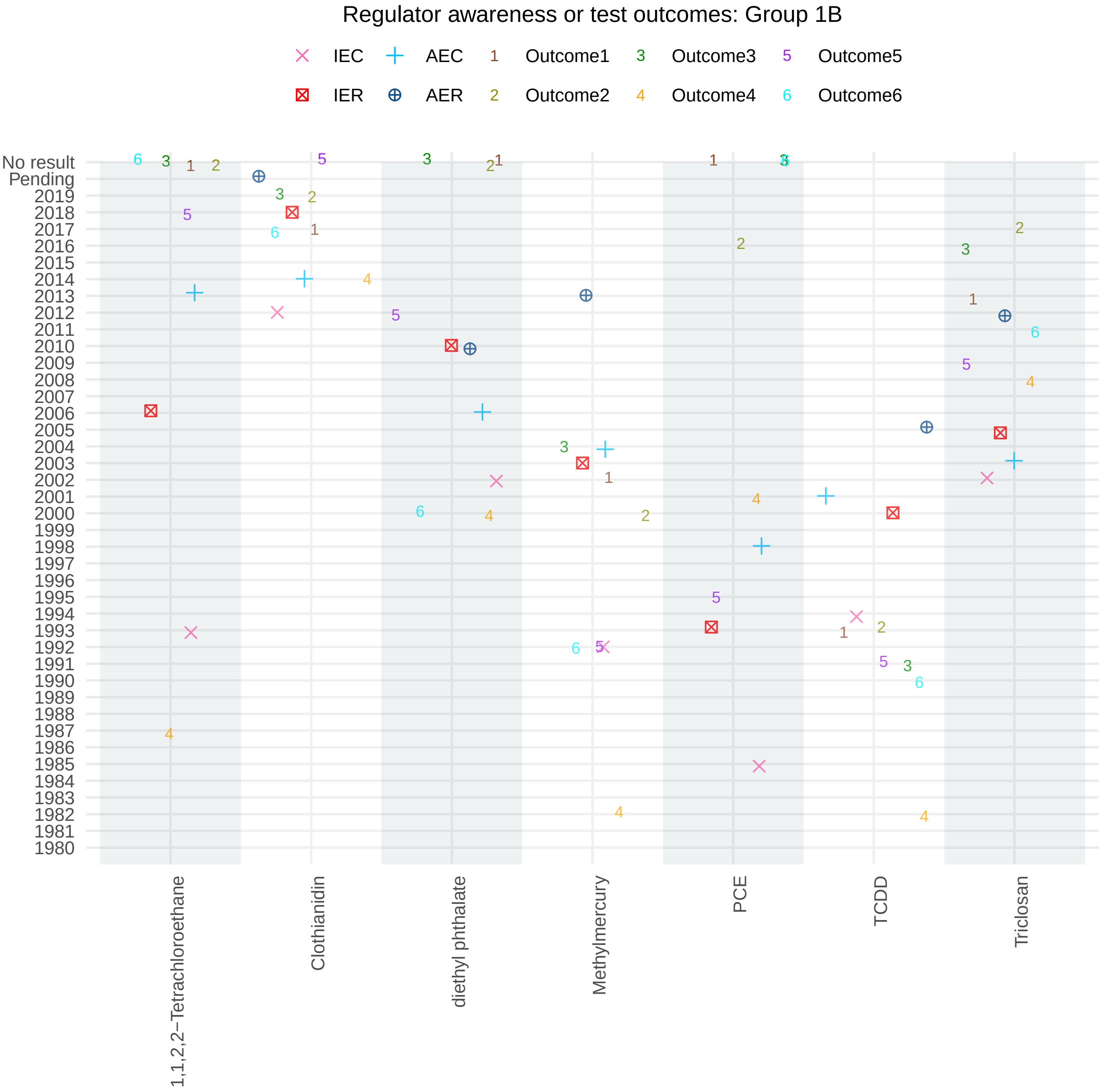


**Figure 6.7.:** Group 1B chemicals: summary of regulatory histories and test outcomes.

We present the results of our tests for Group 3B chemicals in Figure 6.8. Using our classification rule, we classify five of seven chemicals as not CoI (true negatives), and two as CoI (false positives). This yields an adequate classification accuracy of approximately 71%. This result is equivalent to that seen for the classification of Group 3A chemicals.

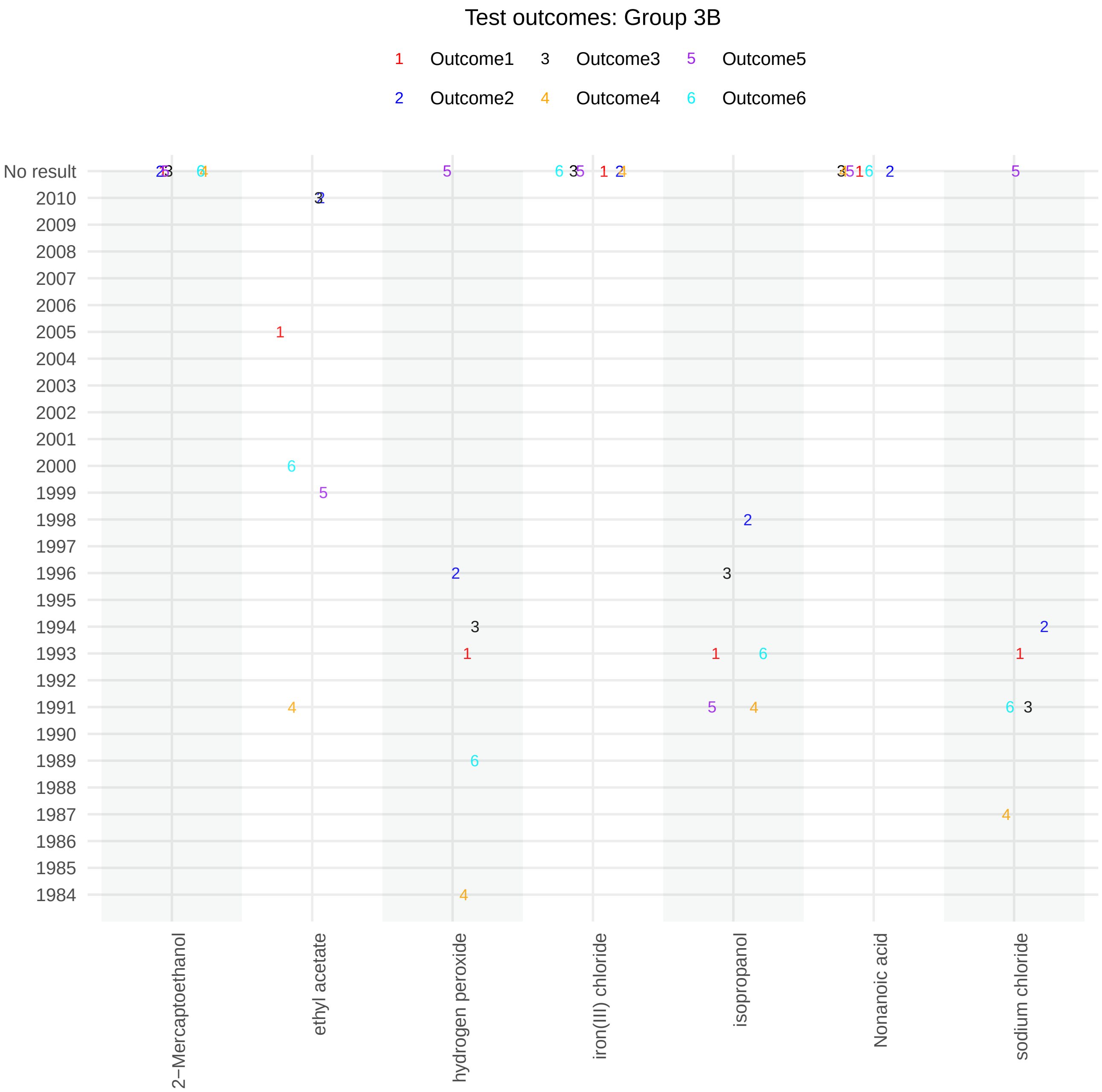


**Figure 6.8.:** Group 3B chemicals: test outcomes.

### 6.2.1. Comparisons of Group 1B test outcomes against regulatory histories

As the Group 1B classification accuracy was inadequate, we shall present only a brief discussion of results. We show differences between test outcomes and IEC in Figure 6.9. Recall our true positives, cases 1–3 on Page 60. In two of these cases (Methylmercury and TCDD) Outcome 4 leads the IEC by some years (at least 10 years), as we observed for Group 1A results. Suppose we disregard the misclassifications of chemicals. Then, we see that Outcome 4 leads the IEC in four of seven cases. This observation supports our view that there may be value in further experimentation with tests investigating features of near-hits trends.

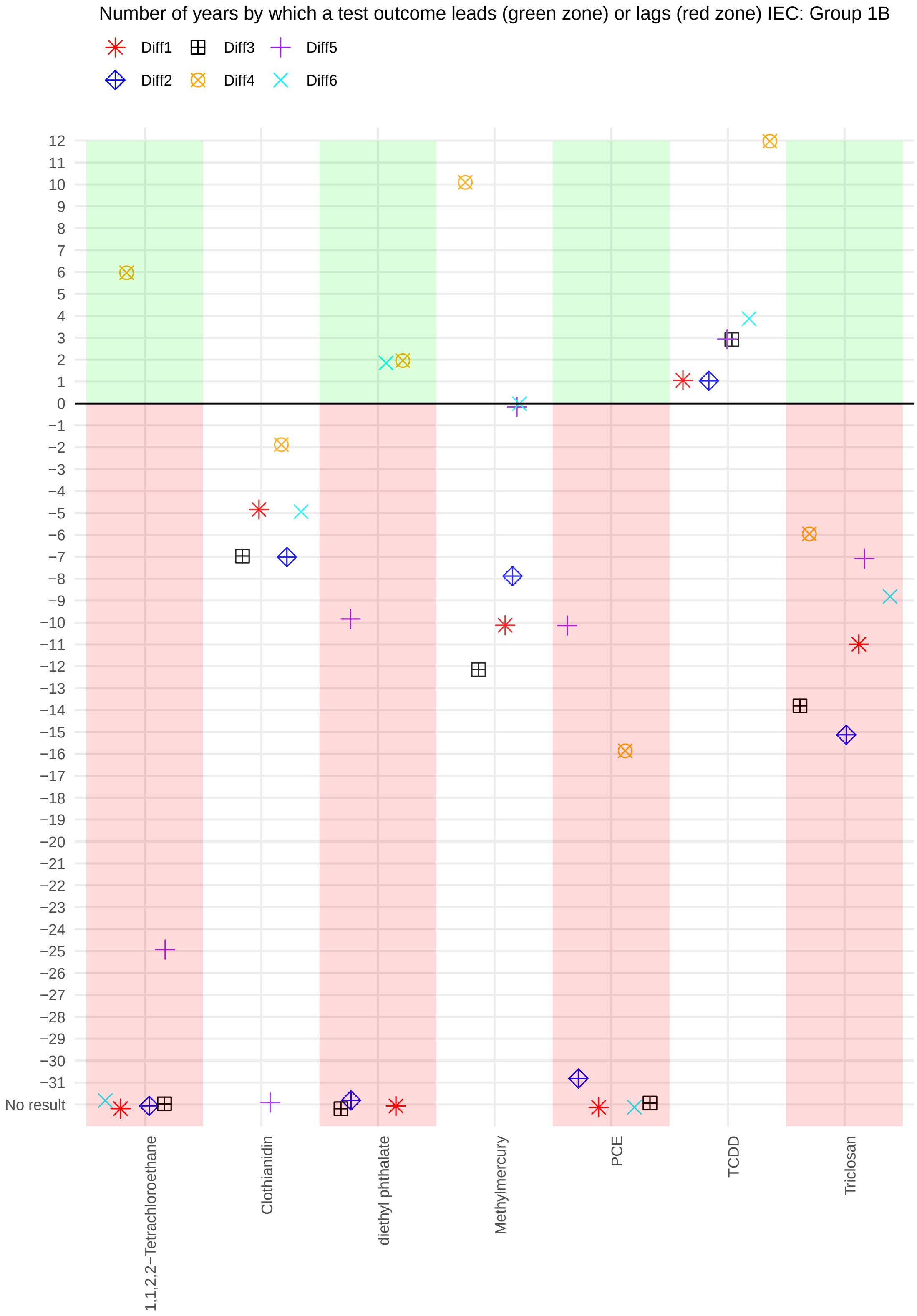


**Figure 6.9.:** Group 1B chemicals: year differences between IEC and test outcomes.

We present a similar view of the ability of our test results for Group 1B chemicals to anticipate IER in Figure 6.10. The figure shows that in five of seven cases (which does include some misclassified chemicals), at least one outcome leads the IER. In these cases, Outcome 4 leads the IER by at least 4 years, ranging up to 21 years.

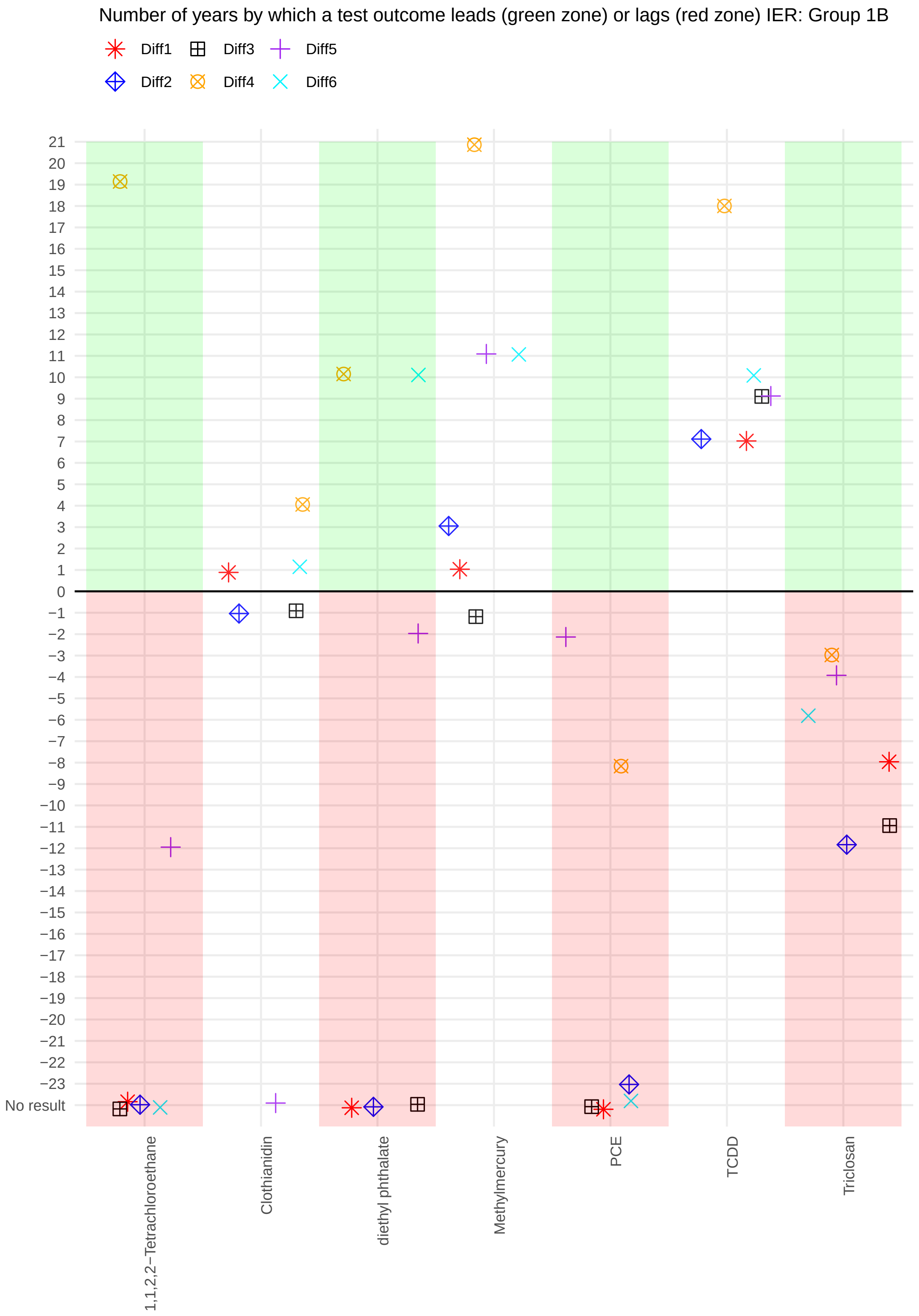


**Figure 6.10.:** Group 1B chemicals: year differences between IER and test outcomes.

## 6.3. Discussion

The results obtained encourage us to critically assess our assumptions (presented in Section 5.3) as they relate to our near-hits data. Assumption 2 is illuminated by a consideration of the mis-classifications made by our system.

We first consider the mis-classification of chemicals from Groups 1A and 1B (false negatives). TCE from Group 1A has a near-hits time series that is relatively flat for longer than other Group 1A chemicals. This may explain why Test 5 (relating to the gradient of a line of best fit) returned the 'No result' outcome which caused the non-CoI classification.

Of the four groups considered, we noted the worst classification accuracy for Group 1B chemicals. Recall Figure 4.7, which shows that the near-hits spread (and maxima) of Group 1B's false negatives are amongst the lowest of Group 1. This suggests that, as we expected, our system cannot produce reliable results when applied to chemicals having a modest amount of research interest. We may seek improvement by considering further tests for such cases (or cases like TCE noted above) by refining existing tests. For Group 1B we note that a 'No result' for Outcome 1 contributes to three false negative classifications. Also, a 'No result' obtained for Outcome 6 contributes to two of these three cases. As such, further experimentation with Tests 1 and 6 may improve our classification accuracy for Group 1B. However, this is not certain to deliver an overall benefit given the balance we intend to strike between Objective 1 and Objective 2.

We expect that there is a limit on what can be achieved by only experimenting with tests. We suspect that it will be more productive to review the near-hits modifiers (as ventured in Section 4.2) to determine if changes lead to the discovery of a greater volume of research interest for Group 1 chemicals. In turn, this may assist us in the discovery of characteristic features of research interest of chemicals in this group.

We obtain a different view of our system's performance by considering the classifications of Group 3 chemicals. Our system performed consistently and adequately for chemicals from Group 3A and 3B. However, it is useful to consider the mis-classifications here (false positives). Some of these results are not entirely surprising given the particularly large volume of research interest noted for these chemicals in Section 4.3.1. In particular the false positives from Group 3A (1,3-Propanediol and carbon dioxide), and another (isopropanol) from Group 3B, have near-hits that are similar to that of a Group 1 chemical (recall Figure 4.7). We could expect that a classification scheme such as ours, having relatively few chemicals available for training, could mis-classify some number of such cases. Also, we may wonder (following the discussion of Section 4.2) if the near-hits counts are inflated for some Group 3 chemicals as a consequence of query modifiers that are not sufficiently selective.

At this stage we are faced with the inconvenience of finding Assumption 2 insufficiently justified. We may be able to remedy this situation by reviewing the near-hits modifiers, harvesting near-hits for our chemicals (and preferably, for a number of extra chemicals so that we have a larger data set), and analysing the new research interest data. This may allow us to draw a sharper distinction between the features of research interest for Group 1 and Group 3 chemicals, giving us more confidence in the validity of Assumption 2.

It is also appropriate to revisit Assumption 1. Recall that in Remark 2 we noted the potential for some inconsistency in the use of the Group 1 label. For example, certain Group 1 chemicals did not experience regulation (or even concern) in the Australian setting. This may either reflect a lack of use in Australia, or that Australian regulators are yet to impose regulation applied elsewhere. Alternatively, a Group 1 chemical may have an indefinite status with respect to concern or regulation, such as "pending". In cases such as these, it appears that a chemical is labelled as Group 1 based on the concern of an international regulator, which may (or may not) be followed by regulation. This may make some Group 1 labelling somewhat dubious, raising concerns around the validity of Assumption 1. (Also recall our concern around the potential for inhomogeneous groups to impede accurate classifications in Remark 9.) Further, for such cases, we cannot conduct an analysis as we did for IEC or IER (as in Sections 6.1.1 and 6.2.1), as we have no numerical value for a regulatory history against which we can compare test outcomes. It will be useful to devise a convention that dictates how we should analyse chemicals labelled as Group 1 which have a partial (or non-existent) regulatory history.

$\star \quad \infty \quad \star$

We draw conclusions and present some recommendations in the final chapter.

# 7. Conclusions and recommendations

It is now appropriate to consider the lessons learned from this pilot project, and how we may continue the progress achieved here.

In Section 7.1 we critically assess features of our processes for data collection and the classification of chemicals into groups by use of associated research interest. Related to this, in Section 7.2 we offer recommendations that we expect to benefit the continuation of this research project.

## 7.1. Conclusions

In this project we had limited time available for exploring the research interest associated with what would become the final list of chemicals for inspection. Consequently, after recognising potential input problems, we could not implement management strategies in all cases. We also had quite limited time for devising and refining our classification system. It is quite likely that further development time would have yielded better results than those achieved here. Further, system training and validation were limited by having a small number of chemicals available for this purpose. (This issue was particularly pronounced for Group 1 chemicals, as we were forced to disregard a number of these which had little associated research interest.) Despite these matters, the results presented in Chapter 6 are promising. This suggests that in at least some cases, the goal of using a (classification) system to anticipate the emergence of CoI by scrutiny of published research interest is achievable.

We note that since the first approach to the problem of recognising CoI in Whyte & Robinson (2020) (where this was a minor task):

- We have overcome various significant technical matters. For example, code optimisation has substantially reduced runtime, making it possible to obtain the research interest for a greater number of chemicals in a reasonable time. Further, code optimisation made it feasible for us to expand our publication range of interest to a span of 40 years quite late in the project (June 16 2020).
- The process of determining research interest associated with a chemical has expanded to include new inputs (synonyms for chemical names). We expect this to assist our acquisition of the relevant research interest relating to a chemical (or subfamily of chemicals).
- We have further refined robust and reliable code that can obtain research interest from Web of Science and analyse this in a reproducible manner.
- New investigations of data in this project have led to a working prototype of a classification system that has scope for improvement given suitable data.
- Discussions with NSW EPA over the course of the project have led to a shared understanding of project priorities and data requirements.

The progress achieved in this project is demonstrated by a consideration of the accuracy of our classification system (recall Ability 1). We note that — although the groups of chemicals available were smaller than we would have liked — in three of the four groups under consideration, Groups 1A, 3A, and 1B, we achieved sufficient accuracy.

## 7.2. Recommendations

We expect that further refinement of our processes for obtaining research interest, (including the quality of inputs, recall Section 2.2.1), will improve our system's ability to make accurate classifications.

**Recommendation 1.** *NSW EPA may wish to review the limitations of WoS queries presented in Section 3.3.2, to decide if addressing any of these is a particular priority.*

There may be benefits in further exploration of the tests applied to research time series as part of our classification system. For example, consider estimation of the systematic component of a time series, as required by Test 6 (and recalling Remark 6). This process may require substantial experimentation with statistical models for $s(t)$ and $\varepsilon(t)$, and also with methods for fitting a model to data so as to obtain a satisfactory fit. There was insufficient time for a thorough exploration of these matters here.

**Recommendation 2.** *We expect to find value in further exploring tests for features of research interest in any subsequent project.*

At present the process of allocating chemicals to groups (Group 1, 2, or 3) appears to be quite subjective. (We noted various concerns in Section 6.3.) These concerns were recently validated. For one particular example, carbon dioxide was originally allocated to Group 3 by NSW EPA, and this was used in the project. Recent advice from NSW EPA staff (December 21st, 2020) indicated that this chemical should have been allocated to Group 1. We noted in Section 4.2 that carbon dioxide research interest had features of a Group 1 chemical, which was inappropriate given the supplied original grouping. As such, the (appropriate) allocation of carbon dioxide to Group 1 during the study would have aided our efforts to distinguish between groups by features of their research interest. The appropriate allocation would have also improved our classification results.

**Recommendation 3.** *It is appropriate for NSW EPA to review its process for allocating chemicals to Group 1, Group 2, or Group 3.*

Our approach depends on capturing most of the research interest concerning a given chemical. We expect this interest to depend on the collection of synonyms used in queries. Other sources of alternative chemical names may provide novel synonyms that could also be included in queries. However, we must take care to not use synonyms (e.g. trade names) that are associated with multiple chemicals, so as to guard against spurious hits. If trade names are widespread, it may be appropriate to experiment with query conditions so as to make them more restrictive. For example, we may need to consider a new type of query that simultaneously searches for a trade name and a chemical subfamily. We would intend such a query to restrict hits to those relating to a chemical of interest, excluding hits relating to any other chemical which uses the same trade name.

**Recommendation 4.** *NSW EPA may wish to consider its stance on searching for non-scientific chemical names (e.g. trade names) in scientific publications in any future project.*

In this project we obtained our chemical synonyms only from SciFinder®. Whilst NSW EPA approved the use of this source, it requires manual harvesting of synonyms. If a future study intends to consider a much larger number of chemicals, such manual processing will become impractical. This problem is exacerbated if there is a need to collate synonyms over time as new names appear. Also, such manual processing is likely subject to human error. As of July 16 2020, we have access to SciFinder$^{n®}$, which is an enhanced version of SciFinder®, now retired. Whilst we have not yet evaluated SciFinder$^{n®}$, we were assured that this would improve our ability to efficiently harvest synonyms.

We have noted concerns (say in Section 5.3) that the modifiers employed in near-hits queries may need further refinement, for two reasons. The first is to exclude spurious hits from the research interest for Group 3 chemicals. The second is to capture more of the relevant research for Group 1 chemicals. We expect that further consideration of modifiers may assist in sharpening the distinction between groups of chemicals. This will allow us to exploit particular features of research interest within groups, thereby improving the accuracy of our classification system.

**Recommendation 5.** *Further consideration of, and experimentation with, the modifiers employed in near-hits queries is appropriate.*

Towards Recommendation 5, we propose two broad approaches to the task of obtaining suitable modifiers. The first approach is for NSW EPA to access expert knowledge it considers suitable. Such a panel should include a range of expertise, as otherwise matters that lead to inflated hit counts for Group 3 chemicals, such as deliberate chemical misuse (recall Section 4.3.1), may not be recognised. The expert panel, with input from NSW EPA, may also advise on (what NSW EPA has called) the "routine tests" conducted on chemicals. We would aim to use provided keywords and phrases in revising queries so that hits do not include those publications which refer to routine tests applied to chemicals, but which do not record harmful properties or adverse effects. However, the use of an expert panel has a potential limitation. A panel may be unaware of emerging issues that are yet to attract substantial attention from regulators or the scientific community. (This may lead to certain modifiers not being included in a near-hits query, causing an underestimate of the near-hits for Group 1 chemicals.)

We expect that the second approach, relating to CEER's mining of Web of Science, will deliver results in a reasonable time, permit some rounds of experimentation into which NSW EPA may have input, and complement the first option. We have formed this impression following a preliminary study, used to inform the Project 2 pitch document presented to NSW EPA. The document considered harvesting keywords from publications and forming them into networks, showing co-appearances of keywords. As such, the process may find connections between keywords that the expert panel are not yet aware of. It should be possible to adapt the methodology so as to harvest keywords in an iterative manner. We would aim to make a complete version of the process as automated as possible, although some human oversight may prove useful in development.

Towards our final recommendation, we note that having groups (of chemicals) with few members is not ideal for a classification project. Here, the small number of chemicals does not allow us to survey a range of research interest time series so as to determine similarities within (or differences between) groups. Further, even a single misclassification (say in a group of seven chemicals) has a substantial effect on the classification accuracy.

**Recommendation 6.** *Following the development of theory and computer code in this pilot study, a future study should use larger groups of chemicals in the training and validation of a classification system. This, combined with the recommendations above, may lead to a sufficiently large collection of results such that it becomes feasible to employ methods from machine learning in the classification of chemicals into groups.*

At this stage, various complexities of (and unknowns in) our classification problem make it difficult for us to make specific statements regarding the data required by future work on our problem. However, having consulted some recent machine learning literature, we can offer some general comments. In a future project we would require research interest for sufficiently large number of chemicals in a given group so that we have examples of the different types of time series we may see in a group. We would like this to be "high-quality data", that is, data that has accurate labels, and which does not have missing data, or contain errors. As a general rule of thumb, many machine learning techniques require a training set of at least 500 samples (where each is a research interest time series in our problem), and a larger set of values is often preferred. (In machine learning, a data set composed of samples in the order of thousands is considered a "medium sized" data set.) Furthermore, certain techniques expect upwards of 50 000 samples in the training set. These techniques are likely to produce poor classification accuracy if given much smaller data sets.

We appreciate that it may require a substantial (even impractical) amount of time and effort for NSW EPA to provide CEER with 500 labelled chemicals. It may be that a training set of 40 chemicals will be adequate (if not ideal) for our needs if this is high-quality data. We note that such a small data set will limit our choice of classification method. However, based on the results obtained so far, we expect that an exploration of the classification methods applicable to smaller data sets will improve on the results produced here. Furthermore, these methods will assist us in further automating our classification system.

# Bibliography

Aditi Sharma and Jonathan D Morrow (2017) Isopropyl Alcohol Swabs as a Preferred Substance of Abuse. *Journal of Psychoactive Drugs*, **49**, 258–261.

Chemical Abstracts Service (2019) SciFinder, Version 42.060116. Data accessed April to June 2020.

Clarivate (2020) Web of Science Core Collection Help: Search Rules. URL https://images.webofknowledge.com/images/help/WOS/hs_search_rules.html.

Interstate Technology & Regulatory Council, PFAS Team (2020) PFAS Technical and Regulatory Guidance Document and Fact Sheets PFAS-1. Tech. rep., Washington, D.C. URL https://pfas-1.itrcweb.org/.

Jacks T, Hatch P (2020) West Gate Tunnel soil dumping late, pushing out project timeline. URL https://www.theage.com.au/national/victoria/west-gate-tunnel-soil-dumping-late-pushing-out-project-timeline-20200504-p54ppv.html.

National Industrial Chemicals Notification and Assessment Scheme (2013) PFC derivatives and chemicals on which they are based. Alert fact sheet. Tech. rep., Sydney, Australia.

National Industrial Chemicals Notification and Assessment Scheme (n.d.) The public Australian Inventory of Chemical Substances. URL https://www.nicnas.gov.au/chemical-inventory.

Vera-Baceta MA, Thelwall M, Kousha K (2019) Web of science and scopus language coverage. *Scientometrics*, **121**, 1803–1813. doi:10.1007/s11192-019-03264-z. URL https://doi.org/10.1007/s11192-019-03264-z.

Whyte JM (2020) On Using 'Emerging Interest' in Scientific Literature to Inform Chemical Risk Prioritisation. In: *10th International Congress on Environmental Modelling and Software* (ed. Ann van Griensven, Jiri Nossent, and Daniel P Ames). International Environmental Modelling and Software Society. URL https://scholarsarchive.byu.edu/iemssconference/2020/. Accepted, to appear in the conference proceedings.

Whyte JM, Robinson AP (2020) Outcome 2: Automating the extraction of chemical prevalences from a bibliographic database to estimate 'emerging concern' — a pilot study. Unpublished report by the Centre of Excellence for Biosecurity Risk Analysis (CEBRA), School of BioSciences, University of Melbourne.

Wigmore I (2019) Data Labeling. URL https://whatis.techtarget.com/definition/data-labeling.

# Appendices

# A. Graphs of hits time series presented by family

## A.1. Chlorinated Ethenes

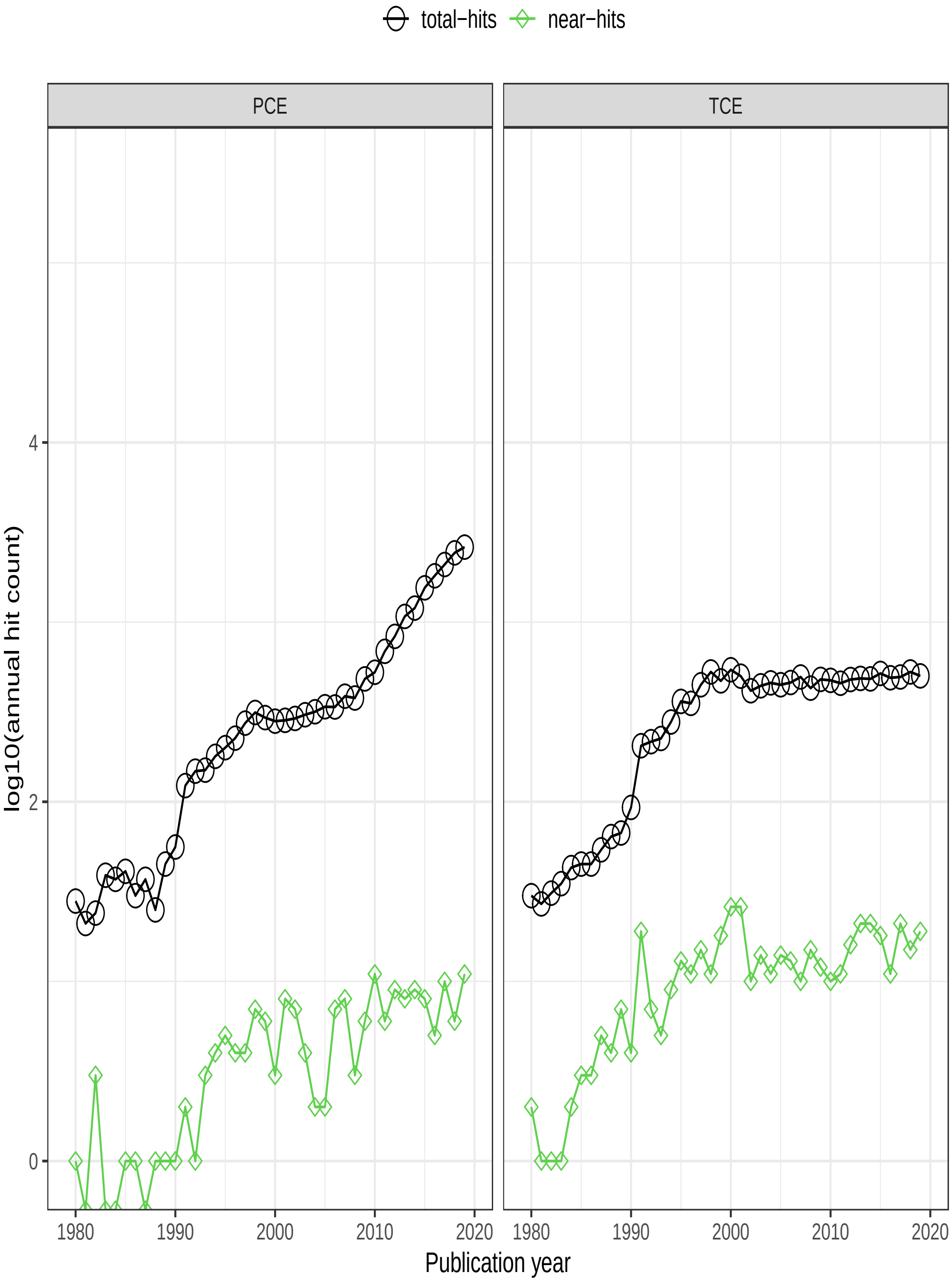


**Figure A.1.:** Time series plots for members of the Chlorinated Ethenes family studied in this report.

## A.2. Dioxins and Furans

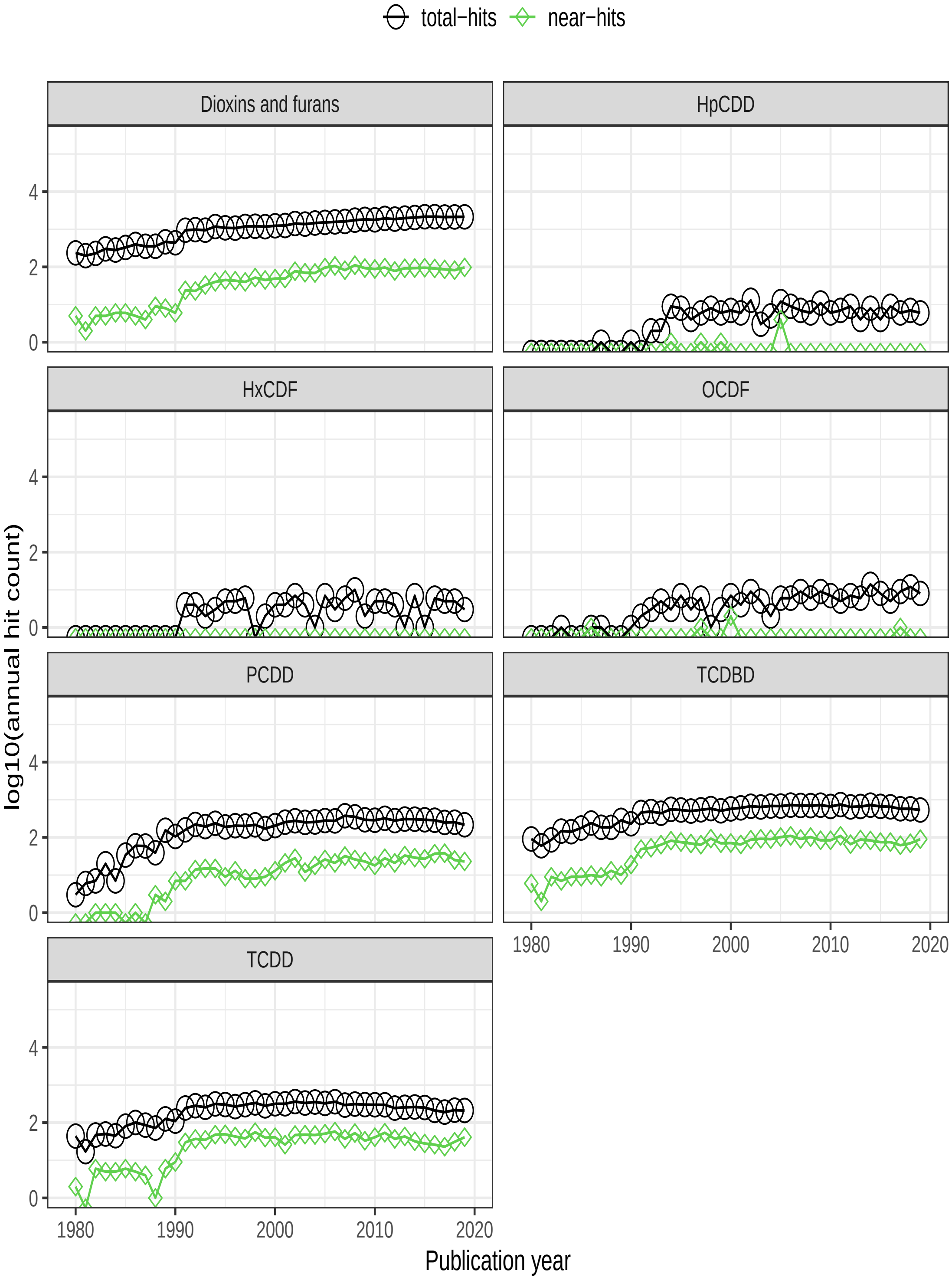


**Figure A.2.:** Time series plots for the phrase "Dioxins and Furans" and the members of this family studied in this report.

## A.3. Neonicotinoids

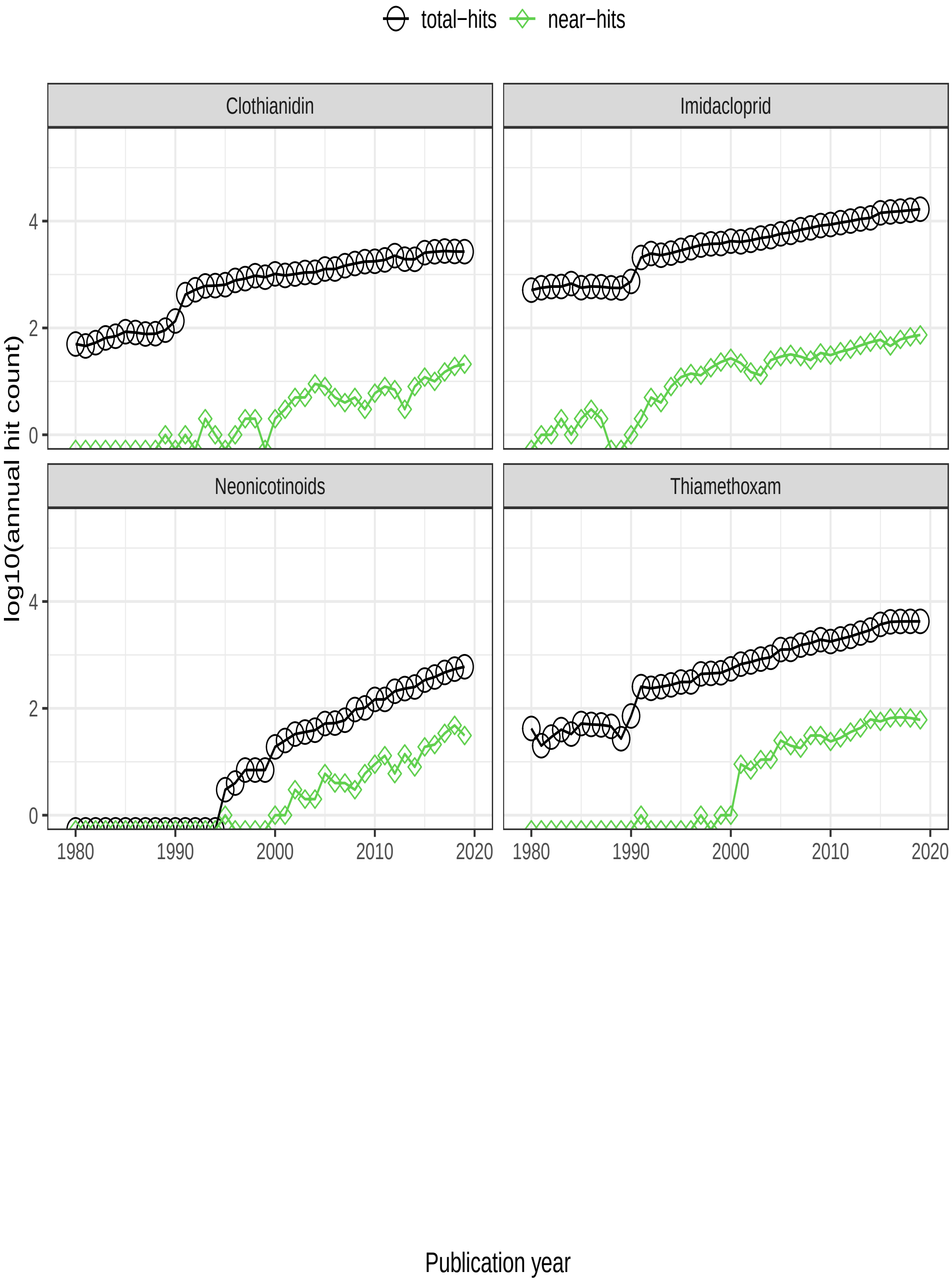


**Figure A.3.:** Time series plots for Neonicotinoids and the members of this family studied in this report.

## A.4. PBDEs

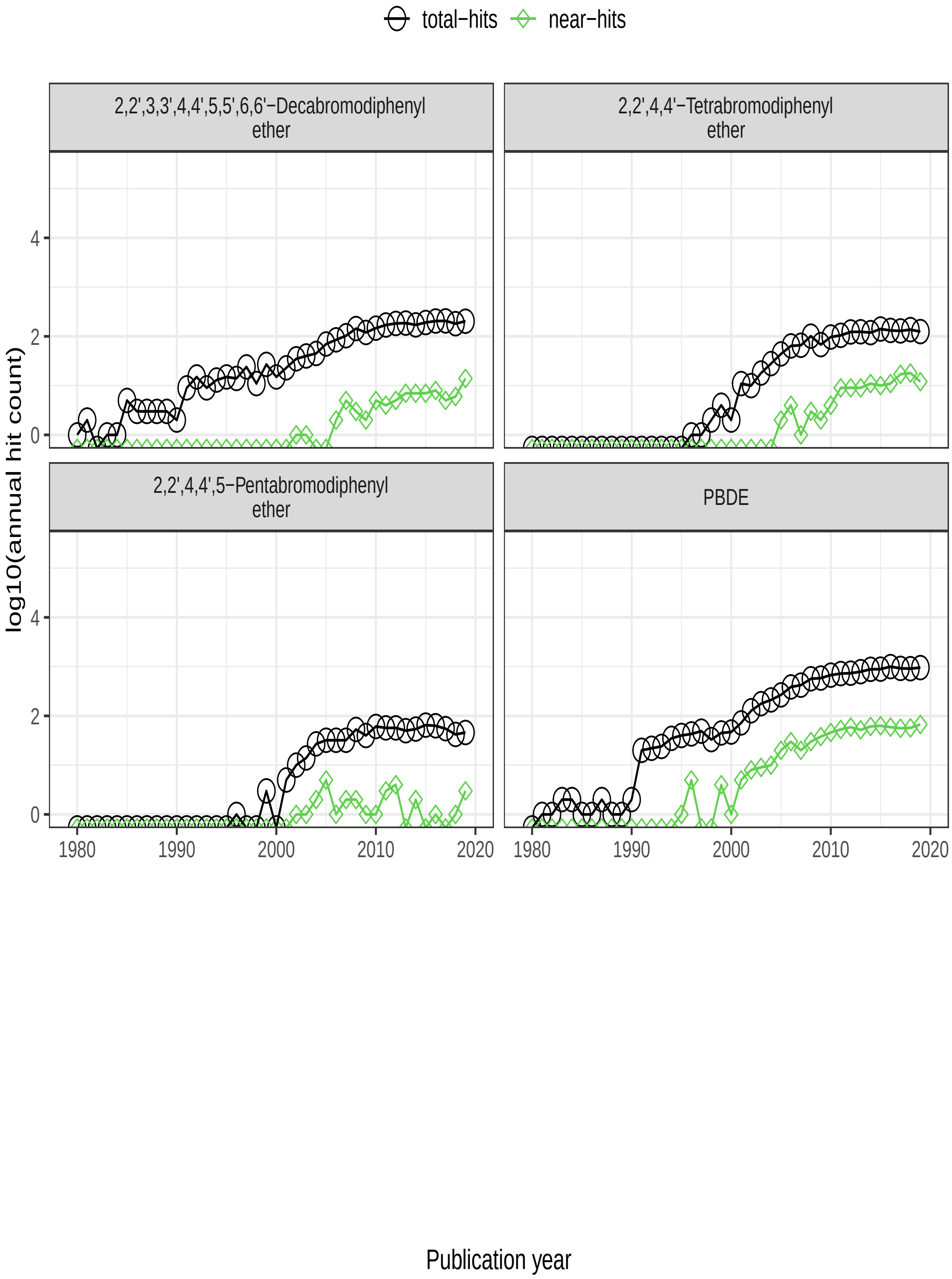


**Figure A.4.:** Time series plots for PBDE and the members of this family studied in this report.

## A.5. PFAS

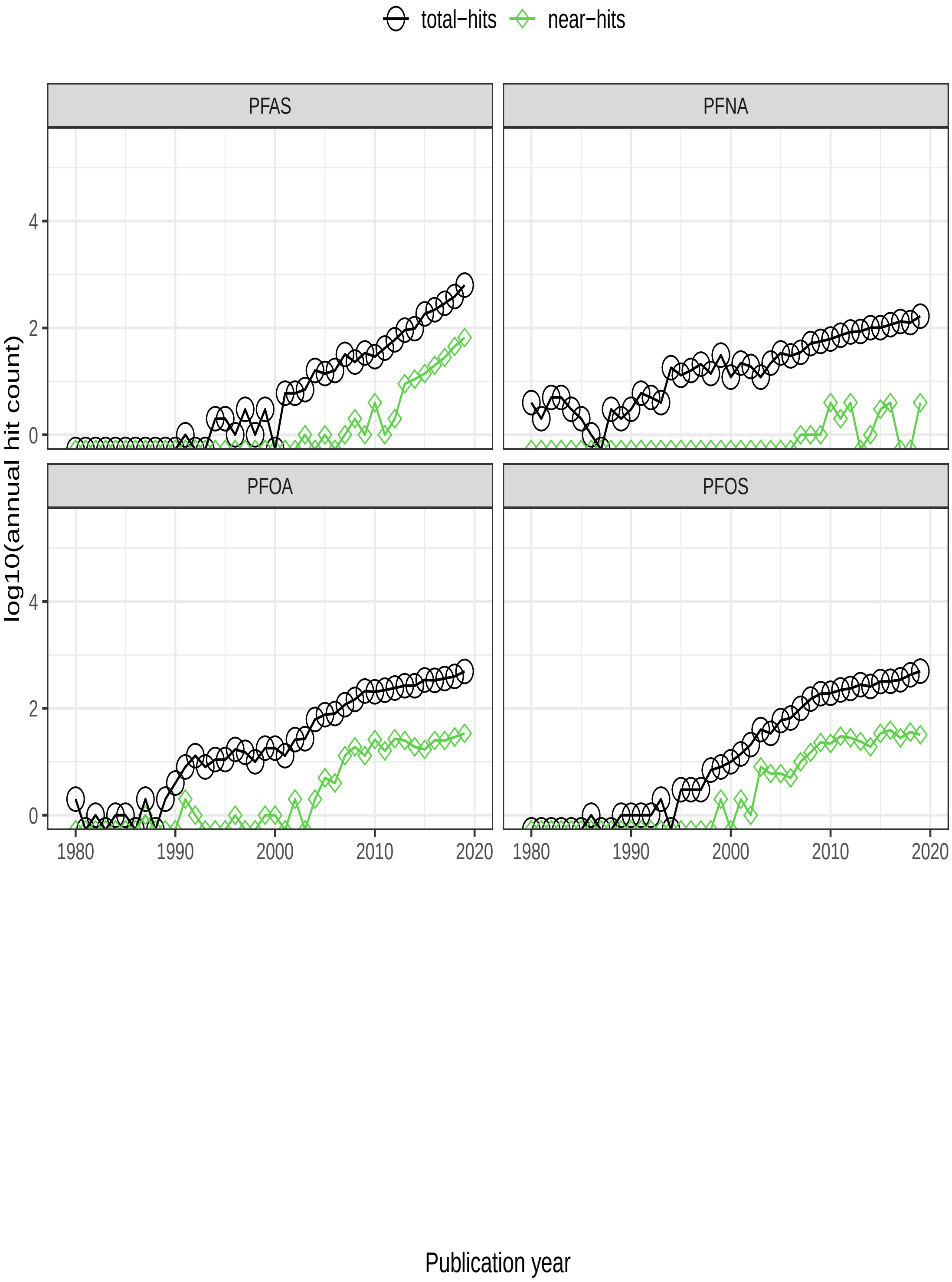


**Figure A.5.:** Time series plots for PFAS and the members of this family studied in this report.

## A.6. Phthalates

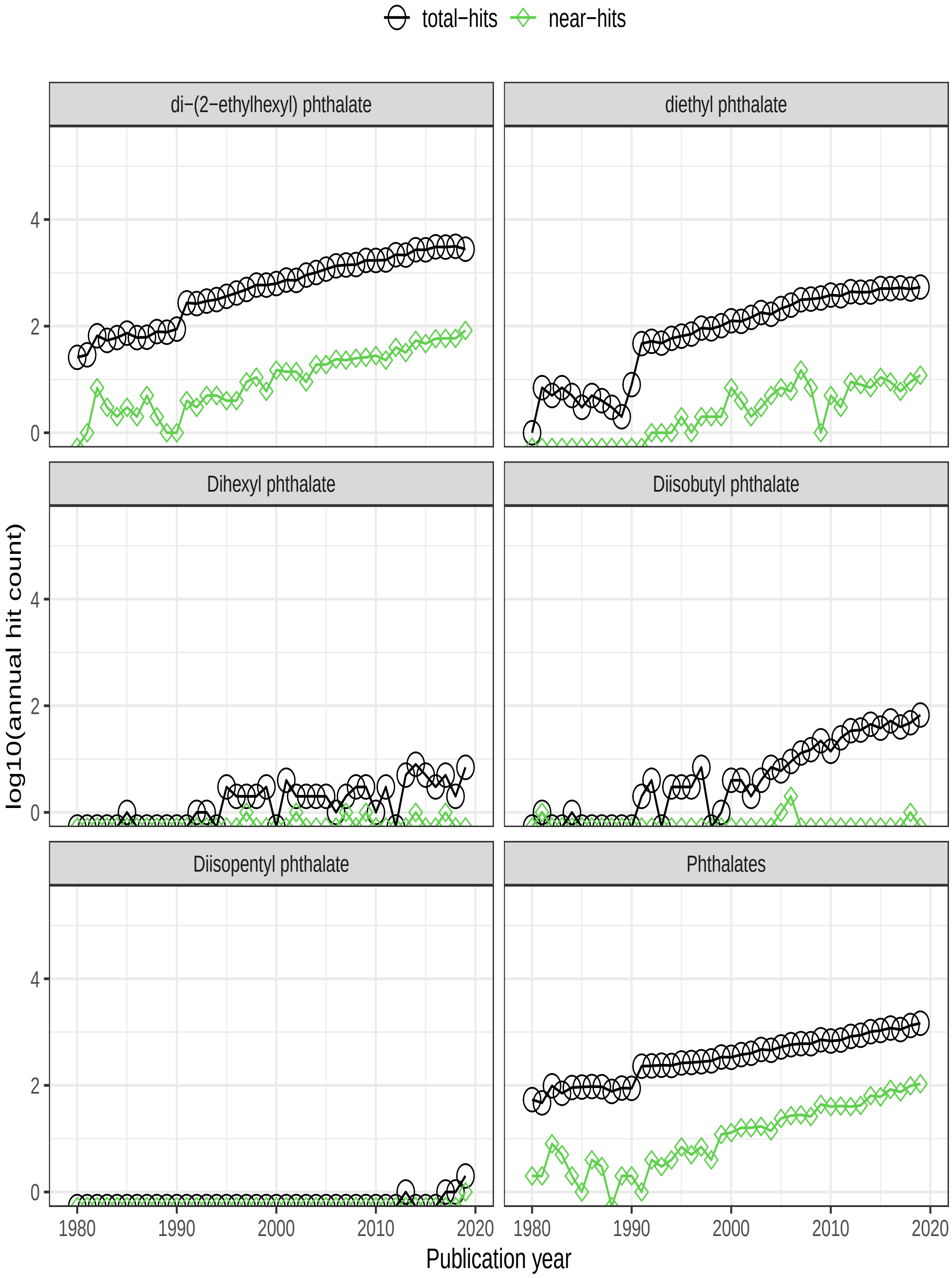


**Figure A.6.:** Time series plots for Phthalates and the members of this family studied in this report.

# B. Tables of results

**Table B.1.:** Group 1A chemicals: comparison of the supplied regulatory histories and years in which tests were satisfied (outcomes). The outcome of Test 1 is labelled as O1 (outcome 1), outcomes of Tests 2–6 are labelled similarly.

| **Name** | **IEC** | **IER** | **AEC** | **AER** | **O1** | **O2** | **O3** | **O4** | **O5** | **O6** |
|---|---|---|---|---|---|---|---|---|---|---|
| 2,2′,4,4′-Tetrabromodiphenyl ether | 1995 | 2003 | 2006 | 2007 | 2016 | No result | No result | 2011 | 2012 | 2010 |
| Bisphenol A | 2002 | 2009 | 2010 | 2010 | 2003 | 2004 | 2001 | 1998 | 1998 | 2001 |
| di-(2-ethylhexyl) phthalate | 1989 | 2005 | 2006 | 2011 | 2002 | 2005 | 2006 | 1982 | 1992 | 1993 |
| Endosulfan | 1995 | 2000 | 1995 | 2002 | 2005 | 2002 | 2007 | 1998 | 1988 | 1992 |
| Imidacloprid | 2012 | 2013 | 2014 | Pending | 1997 | 2001 | 1999 | 1998 | 1993 | 1992 |
| PCDD | 1982 | 2001 | 2001 | 2005 | 1994 | 2002 | 2001 | 1990 | 1992 | 1993 |
| PFOA | 2003 | 2006 | 2004 | 2007 | 2009 | 2013 | 2010 | 2007 | 2007 | 2009 |
| PFOS | 2000 | 2000 | 2002 | 2002 | 2009 | 2012 | 2009 | 2003 | 2004 | 2008 |
| TCDBD | 1984 | 2001 | 2001 | 2005 | 1990 | 1991 | 1991 | 1982 | 1991 | 1990 |
| TCE | 1989 | 1999 | 1995 | 2000 | 1997 | 2000 | 2000 | 1991 | No result | 2000 |
| Thiamethoxam | 2012 | 2013 | 2014 | Pending | 2005 | 2009 | 2005 | 2001 | 2002 | 2004 |

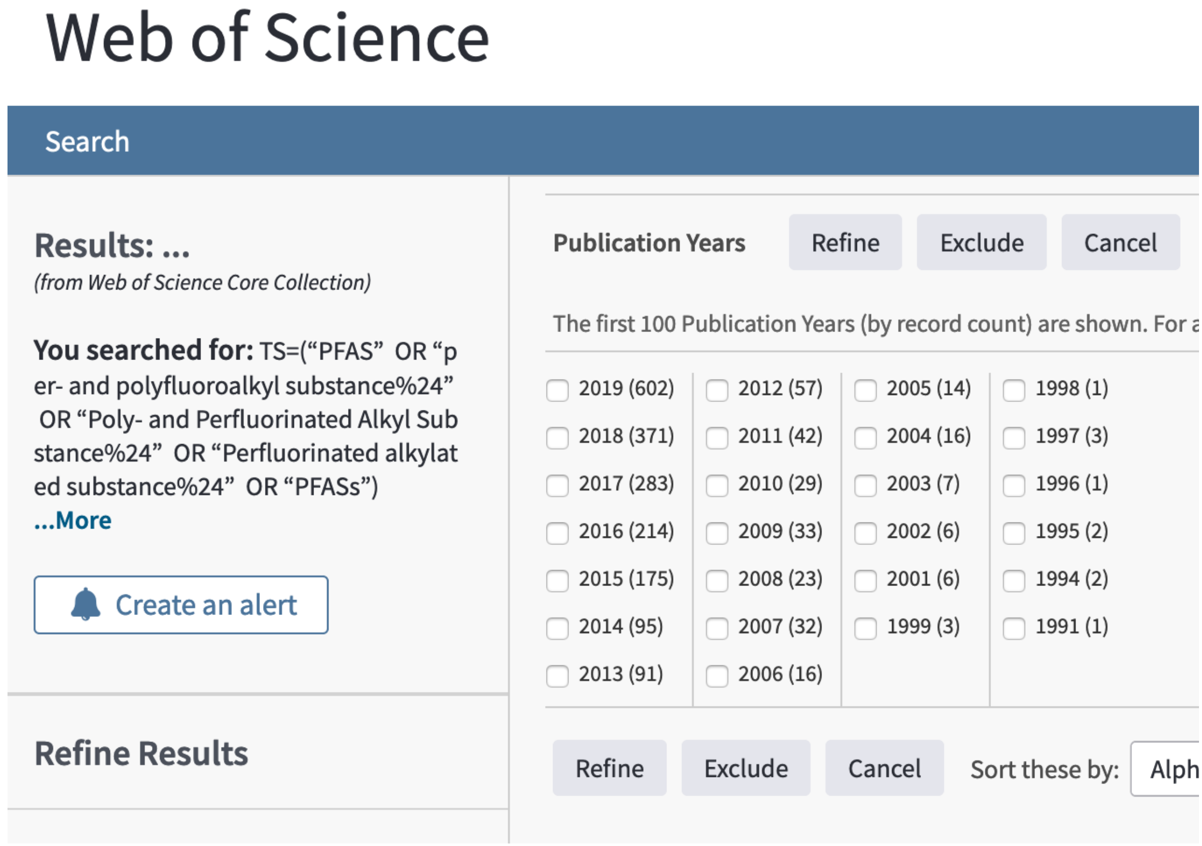
**Table B.2.:** Group 1B chemicals: comparison of the supplied regulatory histories and years in which tests were satisfied (outcomes). The outcome of Test 1 is labelled as O1 (outcome 1), outcomes of Tests 2–6 are labelled similarly.

| **Name** | **IEC** | **IER** | **AEC** | **AER** | **O1** | **O2** | **O3** | **O4** | **O5** | **O6** |
|---|---|---|---|---|---|---|---|---|---|---|
| 1,1,2,2-Tetrachloroethane | 1993 | 2006 | 2013 | None | No result | No result | No result | 1987 | 2018 | No result |
| Clothianidin | 2012 | 2018 | 2014 | Pending | 2017 | 2019 | 2019 | 2014 | No result | 2017 |
| diethyl phthalate | 2002 | 2010 | 2006 | 2010 | No result | No result | No result | 2000 | 2012 | 2000 |
| Methylmercury | 1992 | 2003 | 2004 | 2013 | 2002 | 2000 | 2004 | 1982 | 1992 | 1992 |
| PCE | 1985 | 1993 | 1998 | None | No result | 2016 | No result | 2001 | 1995 | No result |
| TCDD | 1994 | 2000 | 2001 | 2005 | 1993 | 1993 | 1991 | 1982 | 1991 | 1990 |
| Triclosan | 2002 | 2005 | 2003 | 2012 | 2013 | 2017 | 2016 | 2008 | 2009 | 2011 |

# C. Sample validation of R code

Total-hits and near-hits results for PFAS (and some synonyms) obtained from queries of WoS using a web interface are shown in Figure C.1 and Figure C.2 respectively.

Total-hits and near-hits obtained by applying the same queries to WoS using our R code are shown in Table C.1.



**Results: ...**

*(from Web of Science Core Collection)*

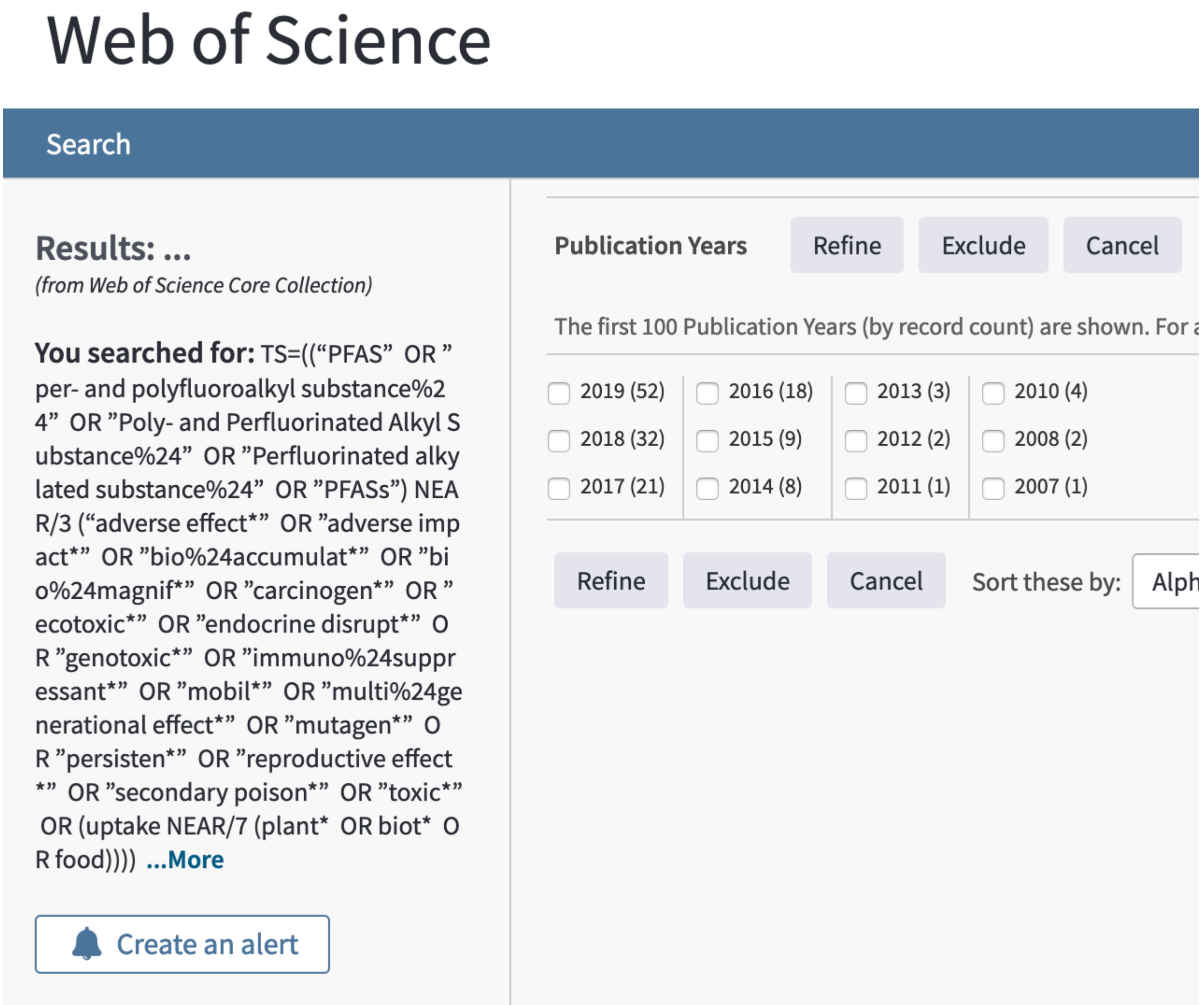
**You searched for:** TS=("PFAS" OR "per- and polyfluoroalkyl substance%24" OR "Poly- and Perfluorinated Alkyl Substance%24" OR "Perfluorinated alkylated substance%24" OR "PFASs")

...More

The first 100 Publication Years (by record count) are shown. For a

□ 2019 (602)
□ 2018 (371)
□ 2017 (283)
□ 2016 (214)
□ 2015 (175)
□ 2014 (95)
□ 2013 (91)
□ 2012 (57)
□ 2011 (42)
□ 2010 (29)
□ 2009 (33)
□ 2008 (23)
□ 2007 (32)
□ 2006 (16)
□ 2005 (14)
□ 2004 (16)
□ 2003 (7)
□ 2002 (6)
□ 2001 (6)
□ 1999 (3)
□ 1998 (1)
□ 1997 (3)
□ 1996 (1)
□ 1995 (2)
□ 1994 (2)
□ 1991 (1)

**Figure C.1.:** Results of a total-hits query applied to Web of Science via web browser for PFAS and its synonyms. Results are quite close to the total-hits results obtained by our R code applied to WoS. Small differences may be due to the different avenues accessing slightly different forms of WoS, e.g. different databases may or may not be included.



Web of Science

**Results: ...**
*(from Web of Science Core Collection)*

**You searched for:** TS=((“PFAS” OR ”per- and polyfluoroalkyl substance%24” OR ”Poly- and Perfluorinated Alkyl Substance%24” OR ”Perfluorinated alkylated substance%24” OR ”PFASs”) NEAR/3 (“adverse effect*” OR ”adverse impact*” OR ”bio%24accumulat*” OR ”bio%24magnif*” OR ”carcinogen*” OR ”ecotoxic*” OR ”endocrine disrupt*” OR ”genotoxic*” OR ”immuno%24suppressant*” OR ”mobil*” OR ”multi%24generational effect*” OR ”mutagen*” OR ”persisten*” OR ”reproductive effect*” OR ”secondary poison*” OR ”toxic*” OR (uptake NEAR/7 (plant* OR biot* OR food)))) **...More**

**Publication Years**

The first 100 Publication Years (by record count) are shown. For a

| | | | |
|---|---|---|---|
| □ 2019 (52) | □ 2016 (18) | □ 2013 (3) | □ 2010 (4) |
| □ 2018 (32) | □ 2015 (9) | □ 2012 (2) | □ 2008 (2) |
| □ 2017 (21) | □ 2014 (8) | □ 2011 (1) | □ 2007 (1) |

Sort these by: Alph

**Figure C.2.:** Results of a near-hits query applied to Web of Science via web browser for PFAS and its synonyms. Results are quite close to the near-hits results obtained by our R code applied to the WoS. Small differences may be due to the different avenues accessing slightly different forms of WoS, e.g. different databases may or may not be included.

**Table C.1.:** Results obtained from R code applying total-hits and near-hits queries for PFAS and its synonyms over the publication year range of 1980–2019 (inclusive) to the WoS API Lite.

| compound | prevalence | year.range | total.hits | near.hits |
|---|---|---:|---:|---:|
| PFAS | family | 1980 | 0 | 0 |
| PFAS | family | 1981 | 0 | 0 |
| PFAS | family | 1982 | 0 | 0 |
| PFAS | family | 1983 | 0 | 0 |
| PFAS | family | 1984 | 0 | 0 |
| PFAS | family | 1985 | 0 | 0 |
| PFAS | family | 1986 | 0 | 0 |
| PFAS | family | 1987 | 0 | 0 |
| PFAS | family | 1988 | 0 | 0 |
| PFAS | family | 1989 | 0 | 0 |
| PFAS | family | 1990 | 0 | 0 |
| PFAS | family | 1991 | 1 | 0 |
| PFAS | family | 1992 | 0 | 0 |
| PFAS | family | 1993 | 0 | 0 |
| PFAS | family | 1994 | 2 | 0 |
| PFAS | family | 1995 | 2 | 0 |
| PFAS | family | 1996 | 1 | 0 |
| PFAS | family | 1997 | 3 | 0 |
| PFAS | family | 1998 | 1 | 0 |
| PFAS | family | 1999 | 3 | 0 |
| PFAS | family | 2000 | 0 | 0 |
| PFAS | family | 2001 | 6 | 0 |
| PFAS | family | 2002 | 6 | 0 |
| PFAS | family | 2003 | 7 | 1 |
| PFAS | family | 2004 | 16 | 0 |
| PFAS | family | 2005 | 14 | 1 |
| PFAS | family | 2006 | 16 | 0 |
| PFAS | family | 2007 | 32 | 1 |
| PFAS | family | 2008 | 23 | 2 |
| PFAS | family | 2009 | 34 | 1 |
| PFAS | family | 2010 | 29 | 4 |
| PFAS | family | 2011 | 42 | 1 |
| PFAS | family | 2012 | 60 | 2 |
| PFAS | family | 2013 | 91 | 9 |
| PFAS | family | 2014 | 97 | 11 |
| PFAS | family | 2015 | 183 | 14 |
| PFAS | family | 2016 | 219 | 20 |
| PFAS | family | 2017 | 289 | 28 |
| PFAS | family | 2018 | 387 | 45 |
| PFAS | family | 2019 | 633 | 66 |