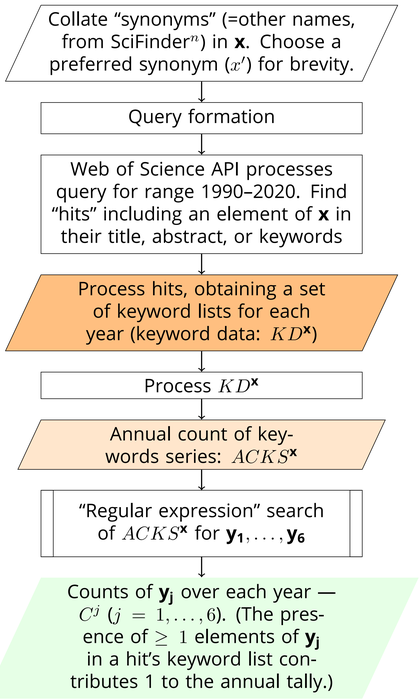
What's your poison (today)?

Anticipating chemical risk by publication keyword analysis

BACKGROUND: Years after a chemical's introduction, evidence of its adverse effects (on human or environmental health, AEs) may emerge. Hence, regulators often manage chemicals "reactively". They would prefer to anticipate risk and manage "proactively". Limited resources for assaying chemical properties make this difficult. An alternative approach may inform timelier regulation.

METHODS: For a particular chemical x and keyword themes of interest to regulators, $y_1, \ldots, y_6$:



We considered keyword data for:

- **Three per- and poly-fluorinated substances (PFAS):**

PFOS, PFOA (regulated following widespread exposure), GenX.

- **Two household chemicals:**

Ethylene glycol (antifreeze, ink), Sodium hydroxide (solvent).

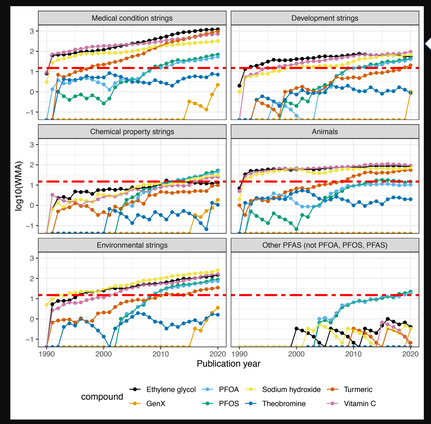
- **Three foodstuff chemicals:**

Theobromine (chocolate), Turmeric, Vitamin C.

We seek potentially instructive changes in C^j (j=1,..,6) by a "progressive" model fit – calculate measures using a data subset from 1990, include the next year if necessary. (Calculation on the whole series may provide insights too late!)

We validate our approach by comparing results against assessment reports by Australia's National Industrial Chemicals Notification and Assessment Scheme (NICNAS).

Keyword trends for chemicals may lead regulatory response. Could this hint at tomorrow's (unknown) poisons?



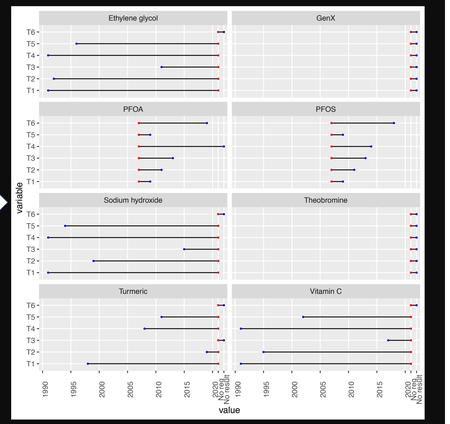
1

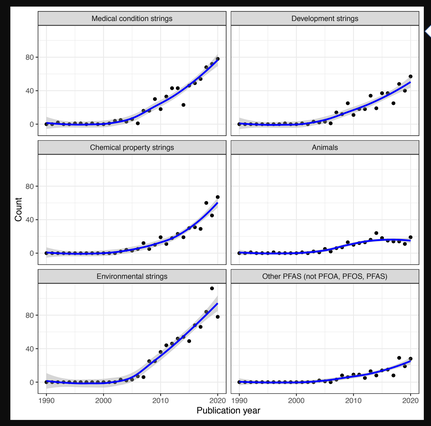
IDEA #1, Weighted moving average (WMA, *n*=5) of C^j
For eight chemicals and *j*=1,…,6 themes, find the earliest year of WMA > threshold (dashed red line=15).

Row "Tj" shows the first year WMA for C^j > threshold with 🔵 (or "No result" otherwise).
If a result is earlier than the year of a related Health report 🔴, further investigation ***may be*** appropriate.

("No reg" = a chemical does not have an assessment report.)

2





3

IDEA #2, Change-point analysis (CPA):
Consider PFOS — notorious for having AEs lead management. (Fire-fighting foams using PFOS were banned in NSW in 2021.)

Most count series seem to have change point → investigate!

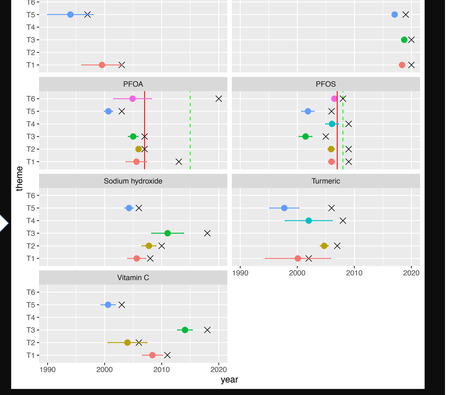
Trend with shaded area showing 95% conf. interval ([8]):
ggplot2::geom_smooth(formula= y ~ x , method="loess")

If the model fit ([1]) to a series subset is "acceptable", we see:
circle/bars: change point estimate/95% confidence interval,
×: the earliest data subset end year of an "acceptable" fit.

Compare results against assessments (found for PFOS, PFOA):
Red (solid) vertical line: "Health"
Green (dashed) vertical line: "Environment"

Looking horizontally: × left of a vertical line is encouraging.

4



5

RESULT HIGHLIGHTS:
PFOS and PFOA:

- **Idea #1:** T1, T5 are satisfied two years after the Health report. (PFOS was "restricted to essential uses" (2007, [4]), and for PFOA "industry should seek alternatives" or similar in 2007, 2015 ([4,5]).
  Idea #2: T3, T5 are satisfied earlier than the Health report year in 3 of 4 cases.

Other chemicals:

- Sizeable interest in **Turmeric**'s *positive* effects ([2]); can we readily dismiss this chemical?
- Little research into **GenX**; it's too early to make a firm judgement based on CPA.
- **Sodium hydroxide** reports note " … existing level of control had not been sufficient, noting lye water accidental ingestions and injuries had continued to occur since the mandating of CRCs [Child-resistant closures]." (2016, [3]), and "… consumer protection against improper use of sodium hydroxide is insufficient." (2007, [6])
- **Ethylene glycol** is a case of disposal changes in Canada: " … reported releases to water have increased fivefold since 2000." ([7])

CPA may yield a snapshot of concern around unknown effects, making it instructive.

6

SUMMARY:

As PFAS show, stricter chemical control may occur some years after first concern. Comparing results against early reports is a stringent test, yet results show promise.

These results encourage further tuning:

- theme refinement (e.g., with input from contaminant chemists),
- method development (e.g., use regression models suitable for count data),
- consideration of how to combine methods,
- application to more chemicals (regulated and other): regulators have concerns around GenX given its similarity to other PFAS. Does this suggest it's appropriate to search for groups of structurally similar chemicals?

Investigation of these points may lead to advance warning of tomorrow's poisons.

Details, details …

KEYWORD THEMES:

- **T1, Medical condition strings:** terms relating to diseases or organs.
- **T2, Development strings:** terms relating to reproduction, embryos, etc. intended to not completely overlap with T1.
- **T3, Chemical property strings:** undesirable chemical properties.
- **T4, Animals:** from a corpus in R; these may suggest environmental exposure or lab work.
- **T5, Environmental strings:** aspects of geography (e.g. rivers) and location (e.g. biosolids).
- **T6, Other PFAS:** a check on validity of results. We ONLY expect a strong signal for PFAS keyword acronyms in a PFAS search.

IDEA #1 WMA: The weighted moving average of C^j over *n* years calculated for year i ($i = n, \ldots, N$) is:

$$WMA_i^j = \frac{n \cdot c_i^j + (n-1) \cdot c_{i-1}^j + \cdots + 1 \cdot c_{i-(n-1)}^j}{n + (n+1) + \cdots + 2 + 1}.$$

(Note the greater weight on more recent data.)

IDEA #2, CPA: assumes that a change point indicates the start of substantial growth in interest. We wish to detect this ASAP.

Fit a linear change-point model ($y=a+bx$) to the first five data points. An "acceptable" fit has:

- adjusted-$R^2 > 0.85$,
- post-change $b > 1.25$ x pre-change b.

Otherwise, add the next year to the data subset, repeat until an acceptable fit, or all data is used.

ACKNOWLEDGMENTS:
Early project work was funded by the Centre of Excellence for Biosecurity Risk Analysis (CEBRA), School of BioSciences, University of Melbourne.

Gratitude to 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).

REFERENCES:

[1] Vito M. R. Muggeo, "segmented: an R Package to Fit Regression Models with Broken-Line Relationships". R News, 8/1, 20-25, (2008). (https://cran.r-project.org/doc/Rnews/)
[2] Hewlings, Susan J., and Douglas S. Kalman. "Curcumin: a review of its effects on human health." *Foods* 6, no. 10 (2017): 92. (772 Google Scholar citations as of 21/7/2021.)
[3] Australian Government, Therapeutic Goods Administration, "Final decisions and reasons for decisions by delegates of the Secretary to the Department of Health", Section 1.17 Potassium Hydroxide and Sodium Hydroxide, Jun 2016.
[4] NICNAS, "PFC derivatives and chemicals on which they are based". Tech report, Jan 2013.
[5] NICNAS, "Perfluorooctanoic acid (PFOA) and its direct precursors: Environment tier II assessment". Tech report, Sydney, Apr 2015.
[6] European Commission DG Joint Research Centre, Institute of Health and Consumer Protection, Toxicology and Chemical Substances (TCS) European Chemicals Bureau, "European Union Risk Assessment Report - sodium hydroxide", Volume 73, EUR 23040 EN, 2007.
[7] Environment Canada, Health Canada, "Canadian Environmental Protection Act, 1999 Priority Substances List Assessment Report, Follow-up to the State of Science Report, 2000 Ethylene Glycol", Ch. 1, Apr 2010.
[8] H. Wickham, "ggplot2: Elegant Graphics for Data Analysis", Springer-Verlag New York, 2016. (https://ggplot2.tidyverse.org)

Jason M. Whyte
jason.whyte@unimelb.edu.au