Constructing a Commercial Secrecy Dataset

Shirley Tang* Yufeng Xia†

September 23, 2026

Abstract

We are constructing a database of confidentiality claims filed under the U.S. Environmental Protection Agency's Chemical Data Reporting (CDR) program. The data cover the cycles of 2012, 2016, 2020, and 2024. The data identify detailed firms' confidentiality claims filed across multiple dimensions of a chemical, such as the identity of the substance, the attribution of production to a named company or site, the volume produced, or its allocation across uses. This data would contribute to the understanding of firms' commercial secrecy strategy.

1 Introduction

Empirical work on appropriability has long treated secrecy as a construct observed indirectly: survey responses about the effectiveness of secrecy relative to patents (Cohen et al., 2000), changes in the legal environment that raise or lower the returns to keeping information private (Png, 2017), and redactions in securities filings (Glaeser, 2018). These measures capture how much a firm relies on secrecy, not what it is keeping secret, because the object of protection is not recorded.

Mandatory regulatory reporting in the chemical industry offers an opportunity to observe how firms protect secrecy. When a firm files information to comply with EPA's regulation and simultaneously asks that particular entries be withheld from public release, it leaves a dated, field-specific record of a confidentiality decision. This paper describes a database of such records from CDR, the recurring program under section 8(a) of the Toxic Substances Control Act (TSCA) in which manufacturers and importers above reporting thresholds report chemical identity, site and parent company, production and import volumes, and the uses to which that volume is directed. Each reportable element carries its own confidentiality flag, and in the public microdata the protected entries appear as "CBI", so the claim itself, rather than an inferred propensity to be secretive, is the unit of observation.

Confidentiality here is worth measuring for two reasons. First, a CDR claim can protect an *association* rather than an independently secret fact: EPA permits claims over the link between a chemical and a company or site, so a firm whose existence is public and whose product is publicly named can still withhold the connection. Protection of substance identity, of attribution, and of scale are therefore separately observable and can be observed in combination. Different combinations leave different inferences available to a rival, so an aggregate count of confidential fields pools substantively different information environments.

Second, the rules changed within the sample period, and unevenly across fields. The 2016 Lautenberg amendments moved TSCA confidentiality from a largely self-certified regime toward active oversight, introducing substantiation requirements, EPA review of all claims over specific chemical identity once a substance is offered for commercial distribution and of a representative subset

*Boston University, Questrom School of Business, xtang1@bu.edu.
†Boston University, Questrom School of Business, xyf@bu.edu.

of at least 25 percent of other non-exempt claims, and a ten-year expiration for most claims. Information exempt from substantiation under section 14(c)(2), which includes marketing and sales information and specific production or import volumes, is also exempt from expiration under section 14(e)(1)(B). The 2020 CDR rule separately made several general processing and use descriptors ineligible for confidential treatment. The cost of maintaining confidentiality thus rose sharply for some objects of protection and barely at all for others, which is a rare source of variation in the price of secrecy that operates within firm and within chemical.

These features make the database useful to innovation research in two specific ways. First, secrecy becomes observable at the granularity at which patenting is already observed. Research on appropriability has had to treat secrecy as the unmeasured alternative to the patent (Hall et al., 2014), inferred from surveys or from legal shocks, so the two mechanisms have rarely been measured on a common unit. Here the decision to withhold is recorded for a named firm and a named substance in a given year. By merging a chemical's CDR record with its patents, we can observe how firms patent and use secrecy strategies to keep information confidential for the same chemical. This is the margin on which selective disclosure accounts turn on (Alexy et al., 2013). Second, the data reach commercialization rather than invention. Patents record what was invented; CDR records what is manufactured, in what volume, and for which uses, so the step from a protected invention to a product in commerce becomes measurable for an entire sector. Beyond innovation, the same measurement problem arises wherever reporting is mandatory but redaction is permitted (Floros et al., 2025), and the coding scheme is portable to those settings.

2 Context of the CDR Program

2.1 The reporting program

CDR collects information every four years from manufacturers and importers of TSCA Inventory substances whose volumes exceed reporting thresholds. A submission is organized by site and substance and contains three blocks: identifying information for the parent company, the site, and the substance; manufacturing information, including production, import, and domestic volumes, physical form, maximum concentration, and worker counts; and processing and use information, recording the applications to which volume is directed and the shares going to each. Thresholds and exemptions apply, so reporting is not a census of chemical commerce, but within its scope it brings a product and the activity surrounding it into one record. Confidentiality is asserted element by element at filing; where specific chemical identity is protected the substance appears under an accession number, which preserves the record while removing the identifier. Because section 14 of TSCA adopts the architecture of FOIA Exemption 4, a surviving claim is an administrative determination that EPA will continue to withhold the information rather than a judicial finding of trade secret status: the database measures commercial confidentiality under mandatory disclosure, not trade secret law in the doctrinal sense.

2.2 What is claimable, and when

Two sets of rules govern what may be withheld. Section 14(c)(2) of TSCA is the first: it exempts certain categories, including marketing and sales information and specific production or import volumes, from substantiation at the time of filing, and it applies to everything submitted to EPA

under TSCA. The CDR rule and the reporting instructions add a second, field-specific layer on top, and that layer changes from one collection to the next. In 2020, manufacturing information may generally be claimed, but upfront substantiation is not required for the three volume fields and is required for every other manufacturing element; in the processing and use sections, general descriptors such as process type, industrial sector, function code, and product category may not be claimed at all, while the quantitative fields trigger substantiation automatically. The 2016 rules differed: site- and chemical-identity claims triggered written substantiation, company-identity claims did not, and substantiation was required for all processing and use elements except production volume. Table 1 summarizes the mapping we encode, taken from EPA's reporting instructions and frequent-questions documents for each cycle, since it is not recoverable from the microdata. Storing it explicitly is what allows a change in the denominator of claimable fields to be distinguished from a change in claiming behavior.

Table 1: Field-level confidentiality treatment for CDR variables

Field	Claimable as CBI	TSCA 14(c)(2) upfront exemption	Upfront substantiation in CDR
Chemical identity	Yes	No, except pre-commercial identity under 14(c)(2)(G)	2016: yes; 2020: yes
Company identity and the company–chemical link	Yes	No	2016: no; 2020: yes
Site identity and the site–chemical link	Yes	No	2016: yes; 2020: yes
Technical-contact link	Yes	No	2020: yes
Annual production, imported, and domestically manufactured volume	Yes	Yes, under 14(c)(2)(F)	2016: no; 2020: no
Workers reasonably likely to be exposed	Yes	No	2016: yes if claimed; 2020: yes
Maximum concentration	Yes	No	2016: yes if claimed; 2020: yes
Physical form	Yes	No	2016: yes if claimed; 2020: yes
Manufacturing-side activity indicators	Yes	No	2016: yes if claimed; 2020: yes
Processing/use general descriptors (type of process or use, industrial sector, function code, product category, consumer or commercial indicator)	No, from 2020	Not applicable	2016: yes; 2020: not claimable
Processing/use quantitative fields (percentage of production volume, number of sites, number of workers, maximum concentration)	Yes	Partial, for volume percentages	2016: yes except volume; 2020: yes

Notes. Exemption from upfront substantiation under section 14(c)(2) is not exemption from later EPA review or disclosure, a distinction the updated 2016 database makes visible. The mapping is stored as a machine-readable rule file keyed on collection year and field name.

3 Constructing the Datasets

Figure 1 sets out what a CDR record contains and what each of its identifiers reaches. The record is a site by substance observation within a collection cycle, and three of its fields do double duty as join keys: the parent company name, the site name and address, and the chemical identity. Each of those keys is itself claimable, so a confidentiality claim does not merely blank a value, it detaches the record from an entire body of outside data. We build the CDR panel first and then attach external

sources through those three keys.

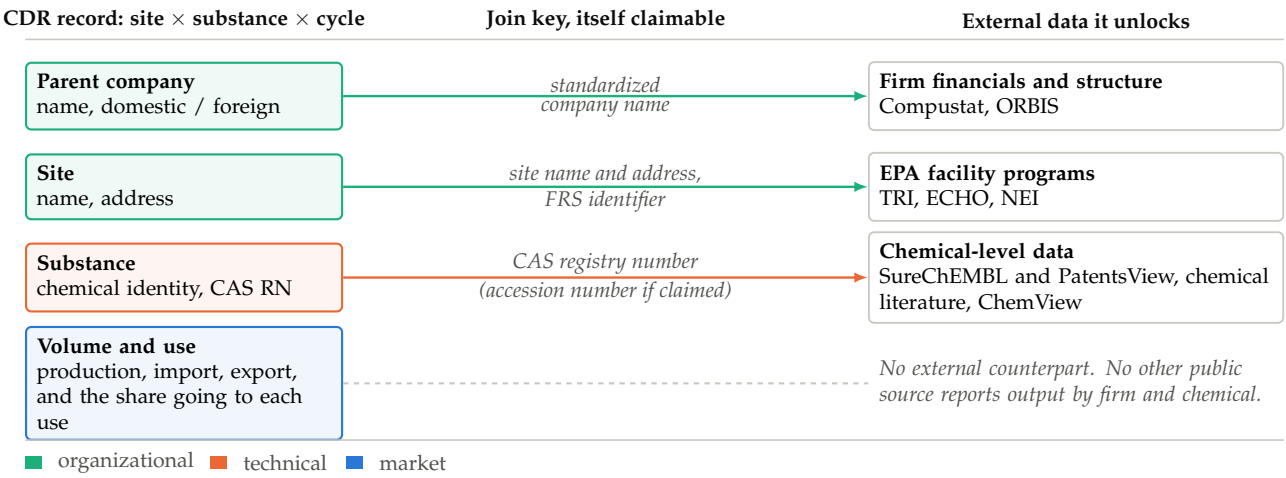


Figure 1: What a CDR record contains and what its identifiers link to

Notes. Each arrow is a join key that is itself eligible for a confidentiality claim. When the claim is granted the record stays in the data but the link is severed: the substance appears under an accession number, or the producer is unnamed. Colors mark the component the field belongs to, as classified in Section 3.3.

3.1 The CDR microdata

The primary inputs are the public microdata EPA releases after each collection, for 2012, 2016, 2020, and 2024. Each release is a set of tables covering the three reporting blocks, with confidentiality carried as literal "CBI" entries, as suppressed values, or as an accession number in place of a chemical name. The current assembly contains **[N]** submission records covering **[N]** distinct reported substances and **[N]** identified parent companies.

The 2016 collection needs separate handling because it exists in two public vintages. EPA published the initial database in May 2017, withholding some aggregate and all site-specific production volumes, and released an updated one in May 2020 containing information previously treated as confidential, after a substantiation and verification process prompted by the Lautenberg amendments. The two releases describe the same filings and differ because of regulatory review, so we retain both. Their difference is an observable quantity, the claims asserted that did not survive review, and to our knowledge this comparison is not available in any other confidentiality setting.

3.2 Harmonization and coding

Cross-cycle alignment cannot rely on field names, which are not stable: parent company identity is a single field in 2012 and 2016 and separate domestic and foreign fields in 2020, and confidential chemical identity is carried by an accession number in some cycles and a confidential inventory identifier in others. We construct a crosswalk that maps each raw column to a harmonized concept, attaches the rule in force for that concept in that cycle, and flags concepts that are absent or non-claimable rather than silently coding them as unclaimed. EPA's published claim percentage tables serve as an independent check: our field-level counts should reproduce them.

For each concept in each record we store whether it was reportable, whether it was claimable in that cycle, and whether it was claimed; we do not impute suppressed values. Claimable concepts

are assigned to three components that differ in what an outsider can infer: *technical* (substance identity and composition-adjacent attributes), *market* (scale and allocation), and *organizational* (who is producing and where). From these we compute the share of claimable fields claimed within each component, an overall claiming intensity, and a concentration measure. Two analysis files follow, since the ideal firm by chemical by cycle panel is not constructible when firm identity is itself claimable: a firm-observable file conditioning on non-confidential parent identity, which supports firm fixed effects at the cost of selection correlated with the outcome, and a chemical-cycle file that avoids that attrition. We also ship a record of the attrition, since the records that drop out are substantively interesting rather than a nuisance.

3.3 External data

Patents and the chemical literature. Where chemical identity is public, the CAS registry number joins a CDR record to SureChEMBL, which links chemical structures to the patent documents disclosing them, and through the patent number to PatentsView for assignee, inventor, and citation data. The same key reaches the chemical literature and EPA's ChemView record of regulatory action on the substance. This is the link that puts a firm's invention and its confidentiality decision on the same molecule.

Firm accounts and ownership. Where parent identity is public, a name-standardized key joins the record to Compustat and ORBIS, which supply research spending, size, and ownership structure for the reporting firm.

Facility-level operations. Site name and address, with the EPA facility identifier where available, join the record to the Toxics Release Inventory, ECHO, and the National Emissions Inventory, which report releases, compliance history, and emissions for the same plant.

Each link ships as a separate crosswalk file with match scores rather than merged into the core data, so a user can adopt one linkage without inheriting the match decisions behind the others.

4 The Plans for Dissemination

We are currently working on cleaning and harmonizing CDR datasets from 2012 to 2020. Once it is done, we will release the project in I3 group's website and the authors' personal websites, including three parts. The first part is the cleaned datasets and related codebook. The cleaned record-level datasets would be in CSV, dta, and Parquet formats. We will consult the industry expert and draft the codebook, which explains every harmonized variable and its source. The second part is the crosswalk for linking the CDR dataset with external public datasets, such as scientific publications and patents. The third part is the codes. We will release the codes in cleaning these CDR datasets.

References

Oliver Alexy, Gerard George, and Ammon J Salter. Cui bono? the selective revealing of knowledge and its implications for innovative activity. *Academy of management review*, 38(2):270–291, 2013.

Wesley M Cohen, Richard Nelson, and John P Walsh. Protecting their intellectual assets: Appropriability conditions and why us manufacturing firms patent (or not), 2000.

Ioannis V Floros, Shane A Johnson, and Wanjia Zhao. Redaction as cross-regulatory disclosure avoidance. *Journal of Accounting Research*, 63(2):807–855, 2025.

Stephen Glaeser. The effects of proprietary information on corporate disclosure and transparency: Evidence from trade secrets. *Journal of Accounting and Economics*, 66(1):163–193, 2018.

Bronwyn Hall, Christian Helmers, Mark Rogers, and Vania Sena. The choice between formal and informal intellectual property: a review. *Journal of Economic Literature*, 52(2):375–423, 2014.

Ivan PL Png. Law and innovation: Evidence from state trade secrets laws. *Review of Economics and statistics*, 99(1):167–179, 2017.