



A LAZY GREEDY APPROXIMATION TO THE NP-HARD PROBLEM OF CHOOSING WHICH DRUGS AN ADVERSE DRUG EVENT CORPUS ANNOTATES

Luis Eduardo Muñoz Guerrero

PhD, Facultad de Ingenierías, Universidad Tecnológica de Pereira, Pereira, Risaralda, Colombia
Email: lemuno zg@utp.edu.co, <https://orcid.org/0000-0002-9414-6187>

Abstract

Background: Automated detection of adverse drug reactions from text is evaluated on a few corpora holding whatever drugs their sources discussed, yet the figures are read as general extraction performance.

Objective: To measure the therapeutic composition of a widely used adverse drug event corpus, and how much of the reaction space it reaches.

Methods: The CADEC corpus (1,250 forum posts, 5,796 reaction spans) was mapped to active ingredients and Anatomical Therapeutic Chemical classes. 51 quarterly US Adverse Event Reporting System extracts, 2004 to 2016Q3, gave 30,297,215 drug records, names normalised by a validated induced dictionary. Coverage is the share of an unseen drug's reaction mass inside the union of each corpus drug's 20 most frequent terms.

Results: The corpus holds 12 products but two ingredients, atorvastatin (80.0% of posts) and diclofenac, spanning 3 of 14 anatomical main groups. Against 1,526 profiled ingredients it covered 17.7% of an average unseen drug's reaction mass, 99.6% of drugs below 50%; normalisation reached 94.6% accuracy end-to-end. Reaching 80% took 208 ingredients chosen deliberately against over 500 at random, while the most reported drugs did worse than chance beyond two; conclusions held across 12 configurations. Even for its own ingredients only a quarter to a third of concepts aligned.

Conclusion: The reach of an adverse drug event corpus is bounded by the number of active ingredients it contains. Here that bound is structural, not a poor choice of the two, though choosing badly costs several fold. Performance should be reported as drug-specific, and therapeutic breadth treated as a design parameter.

Keywords: Pharmacovigilance, adverse drug reactions, spontaneous reporting systems, text corpora, therapeutic classification

Introduction

Adverse drug reactions are a substantial cause of hospital admission and of morbidity during admission, and their detection after a medicine reaches the market rests principally on spontaneous reporting systems [1,2]. Those systems are known to capture only a fraction of the reactions that occur, and the

fraction is neither constant across drugs nor across time [3]. The limitation has been described in this journal as a failing of national reporting infrastructure rather than a technical detail [4].

That gap has motivated a search for complementary evidence in text that patients and clinicians generate for other purposes: clinical narratives, medical case reports, and latterly consumer forums and social media [5,6]. Extracting reactions from such text at scale requires automated methods, and automated methods require annotated corpora on which to be trained and, more consequentially, on which to be evaluated. A small number of corpora have accordingly become standard reference points [7,8,9,10].

These resources share a feature that is rarely made explicit. Each was assembled from whatever source material was available, and the drugs it contains are a consequence of that availability rather than of deliberate sampling of the therapeutic space, a property measured here for one of them. A corpus built by harvesting posts about a particular medicine will contain that medicine's adverse reactions, and a system evaluated on it will be evaluated on that medicine's adverse reactions. Yet the resulting performance figures are reported, and read, as measures of adverse drug event extraction in general.

Whether that reading is warranted is a pharmacological question rather than a computational one. Adverse reaction profiles are drug-specific: musculoskeletal reactions are prominent in statin reporting, gastrointestinal and renal ones in reporting for non-steroidal anti-inflammatory drugs, and neither resembles the haematological profile of a cytotoxic agent.

How far apart two such profiles actually sit is measured below rather than assumed. If a corpus contains few ingredients, the reactions it can teach a system to recognise are bounded by those ingredients' profiles, and the bound is invisible to an evaluation whose held-out data are drawn from the same drugs. This study asks it. We take one widely used adverse drug event corpus, measure its therapeutic composition directly, compare it against a reference assembled from twelve and a half years of spontaneous reports, and generalise from the particular corpus to the design question it raises: how much of the adverse reaction space does a corpus of a given size reach, and how many ingredients would be enough?

Materials And Methods

CADEC is a corpus of consumer forum posts annotated for adverse drug events, drugs, diseases, symptoms and clinical findings [7].

It is distributed by the Commonwealth Scientific and Industrial Research Organisation under the CSIRO Data Licence; version 2 of the archive was used. The corpus was obtained from the CSIRO Data Access Portal rather than from a secondary mirror, because the mirror examined distributes preprocessed text with the drug identity removed, and drug identity is the variable of interest here.

The archive encodes the drug product in each filename, which makes therapeutic composition directly measurable. Each product was mapped to its active ingredients from its approved labelling and thence to an Anatomical Therapeutic Chemical code. Annotation layers were joined on character offsets, the only key consistent across them. Concept identity was taken from the SNOMED Clinical Terms layer, which carries readable labels, rather than the MedDRA layer, whose numeric identifiers need a licensed dictionary to resolve.

The United States Food and Drug Administration Adverse Event Reporting System publishes quarterly extracts of the reports it receives [11]. All 51 quarters from the first quarter of 2004 through the third Vol. 24 No.01 (2017): JPTCP (168-193)

quarter of 2016 were retrieved, yielding 30,297,215 drug records and 28,609,597 reaction records. The fourth quarter of 2016 was excluded because it was not published within the study period.

Two source characteristics required explicit handling: the report identifier changes name partway through the series, and the curated active ingredient field exists only in the later quarters.

The system records drug names as reporters wrote them, so one medicine appears under brand names, generic names, salt forms, strengths and dose forms in many combinations; the resulting variant burden has been quantified previously [12]. Names were reduced lexically by removing parenthetical glosses, strengths, dose forms, routes and salt, ester and hydrate suffixes. Removing salt suffixes is what merges diclofenac sodium with diclofenac potassium, the level at which a therapeutic class question is posed. No public crosswalk spans the whole of this series, so the mapping is induced rather than looked up in a standard vocabulary such as RxNorm [13].

Lexical reduction alone is insufficient, because brand names do not resemble their ingredients. The curated field of the later quarters was therefore treated as a labelled sample from which a surface-form-to-ingredient dictionary was induced, then applied to the whole series including the years that lack it. Write $n(s, i)$ for the number of curated records pairing surface form s with ingredient i , $N(s)$ for the total over all records naming an ingredient and $i^*(s)$ for the most frequent; Table 1 collects the notation used throughout, with the value each symbol takes here.

A form is mapped only where one ingredient dominates:

$$\iota_\tau(s) = i^*(s) \text{ if } \frac{n(s, i^*(s))}{N(s)} \geq \tau \text{ and } N(s) \geq 3; \quad (1)$$

$$\perp \text{ otherwise}$$

The threshold was set at $\tau = 0.80$. Abstention, written $\perp$, is a design decision rather than a failure: a wrong merge corrupts a profile silently, whereas an unmapped form falls through to its own lexical name, which costs precision rather than coverage and is what the attestation requirement below removes. The rule is therefore biased towards refusing to merge. The full sweep of τ is reported in Table 2.

The same field supplies a negative label: where an ingredient is not established the system writes an explicit placeholder, and the forms attached to it are drug classes and reporter shorthand. A further 602 were rejected on that basis rather than by extending the hand-written stop-list.

A reaction profile was built for each eligible ingredient, from reports in which that drug was a primary or secondary suspect alone. Concomitant medications are recorded for clinical context rather than as putative causes, and crediting a drug with reactions it was not suspected of causing would make the profiles describe co-prescription rather than pharmacology. The restriction is not the choice that flatters the argument: admitting concomitants lowers the coverage figures reported here rather than raising them, as the sensitivity analysis shows. Three conditions define the reference. An ingredient is profiled only if its name appears in the curated active-ingredient field of the modern quarters, so that a pack or device string surviving the normaliser as its own ingredient cannot enter as a drug; only if it is not a product containing a corpus ingredient, since Arthrotec and Caduet reach the reference as combinations and are corpus content rather than held-out data; and only if it had at least 200 suspect-drug reports,

giving 1,526 ingredients, 19,348 reaction terms and 39,607,431 ingredient-reaction pairings across reports. A profile is that ingredient's distribution over MedDRA preferred terms [14]. Writing $n(t, r)$ for the number of suspect-drug reports pairing ingredient t with term r , and U for the set of terms,

$$m(t, r) = \frac{n(t, r)}{\sum_{r' \in U} n(t, r')} \quad (2)$$

so each profile sums to one over U and reaction mass means probability under it, which is what puts a drug with hundreds of thousands of reports and a drug with a few hundred on one scale.

Divergence between profiles was measured by the Jensen-Shannon divergence [15]. With $\bar{m}$ the equal mixture of two profiles and H the Shannon entropy,

$$\text{JSD}(t, t') = H(\bar{m}) - \frac{H(m(t)) + H(m(t'))}{2} \quad (3)$$

Taking logarithms to base two bounds this in the unit interval, which is what allows divergences to be compared across ingredient pairs; it is also symmetric and, unlike the Kullback-Leibler divergence it is built from, finite where the two supports differ, which they routinely do here.

An initial definition of coverage, the share of population reaction mass falling on terms a drug set reports at all, proved vacuous and was discarded: a single common ingredient reaches 98.3% under it, because a drug with several hundred thousand reports has recorded almost every common term at least once.

The measure adopted instead reflects what a corpus can plausibly teach. Let $S(d)$ be the 20 most frequently reported terms of ingredient d . The vocabulary of a corpus spanning the ingredient set D is their union,

$$V(D) = \bigcup_{d \in D} S(d) \quad (4)$$

and held-out coverage of an ingredient t outside D is the share of that ingredient's own reaction mass falling inside it:

$$C(t, D) = \sum_{r \in V(D)} m(t, r) \quad (5)$$

The quantity answers the question a corpus user faces: given a system developed on these drugs, how much of a new drug's reaction profile has it plausibly encountered? Because it is evaluated only at t outside D , it cannot be inflated by adding documents about drugs already present.

Stated formally, choosing which drugs a corpus should contain is an instance of maximum coverage. Let P be a candidate pool of ingredients, E a disjoint set whose profiles define the objective and E' a further disjoint set withheld from the optimiser, each drug carrying the profile of Equation 2. The objective is mean coverage over E ,

$$f(D) = \frac{1}{|E|} \sum_{t \in E} C(t, D) = \frac{1}{|E|} \sum_{t \in E} \sum_{r \in V(D)} m(t, r) \quad (6)$$

and the corpus of a given size that maximises it is

$$D^* = \text{argmax}_{D \subseteq P, |D| \leq k} f(D) \quad (7)$$

with the same quantity over E-prime reported as the result, since a set chosen to maximise coverage of E scores optimistically on E. Write $\Delta(d | A) = f(A \cup d) - f(A)$ for the marginal gain of adding ingredient d to a set A.

The objective is monotone, since enlarging D can only enlarge V(D), and submodular:

$$\Delta(d|A) \geq \Delta(d|B) \text{ for all } A \subseteq B \subseteq P, d \notin B \quad (8)$$

which holds because each inner term is a weighted coverage function and submodularity survives non-negative linear combination. Maximum coverage contains set cover, one of the original twenty-one NP-complete problems [16], and no polynomial-time algorithm can guarantee better than a $(1 - 1/e)$ approximation unless P equals NP [17]. Greedy selection takes

$$d_{i+1} = \text{argmax}_{d \in P \setminus D_i} \Delta(d|D_i) \quad (9)$$

and is guaranteed at least that bound for monotone submodular objectives [18]; its behaviour on maximum k-coverage in particular is well characterised [19].

Two baselines are scored against it on the same withheld drugs. The first draws the corpus uniformly at random from P, which models opportunistic assembly. The second is the rule a corpus builder would apply deliberately, taking candidates in descending order of the reports R(d) each carries:

$$d_{i+1} = \text{argmax}_{d \in P \setminus D_i} R(d) \quad (10)$$

Its contrast with Equation 9 is the whole of it: the right-hand side does not mention the set already chosen, so nothing stops successive choices from covering the same terms.

Two consequences follow that a descriptive curve cannot give. Greedy selection supplies a principled notion of a well-chosen corpus against which opportunistic assembly can be compared. And because greedy attains at least $(1 - 1/e)$ of the optimum, rearranging the guarantee turns the value it reaches into a ceiling on every corpus of that size drawn from P, for as long as that ceiling lies below what the objective can reach,

$$f(D^*) \leq \frac{f(D_{\text{greedy}})}{1 - 1/e} \quad (11)$$

so the shortfall against the best corpus of that size is bounded, on the objective, rather than merely unknown. This quantity constrains resources which do not exist.

Algorithm 1: Lazy greedy selection of a corpus of k ingredients

input candidate pool P, evaluation set E, budget k

output ordered set D, $|D| = k$

1 $D \leftarrow \emptyset$; $Q \leftarrow$ empty max-priority queue

2 **for** $d \in P$: push d into Q with key $\Delta(d | \emptyset)$, stamp 0

3 **while** $|D| < k$:

4 pop the largest-key d from Q; let σ be its stamp

5 **if** $\sigma = |D|$: $D \leftarrow D + d$ [the key is current, so d is the true maximiser]

6 **else**: push d into Q with key $\Delta(d | D)$, stamp $|D|$

7 return D

Line 5 is the whole economy. A stale key was computed against a smaller D, so by Equation 8 it is an upper bound on the true gain, and a candidate whose upper bound already loses cannot win. Both forms make the same selections and differ only in evaluations spent: the naive one scores every surviving candidate at every step, the lazy one primes the queue and then pays whatever each promotion costs. The worst case is identical, so the gap is empirical.

The ingredients were split once, at random, into 1,221 candidates, 152 defining the objective and 153 withheld from the optimiser. Greedy maximises the objective, so reporting it as coverage would be an in-sample figure while the baselines, optimising nothing, would not be. The guarantee concerns the objective being maximised, so the upper bound stays on it rather than being transferred. The lazy variant is Minoux's accelerated greedy [20], the economy that matters at the scale coverage problems reach on real data [21]; it was run in full, while the naive form was run to a short budget and extrapolated, so its timing is an estimate.

The normalisation step is the one place where a systematic error would propagate into every profile, so it was validated rather than assumed. The nine curated quarters supply 9,077,610 gold pairs of surface form and active ingredient, over 41,190 distinct forms and 6,704 ingredients. The dictionary was induced on a training split and scored on held-out data under two splitting schemes. Induction on a split necessarily yields a smaller dictionary than deployment, which draws on every curated record: the record-split dictionaries carry about 18,260 entries and the surface-split ones about 14,612, against the 20,891 of the deployed one.

Splitting by record measures the deployed operating condition, in which a form seen during induction may recur. Splitting by whole surface form holds out every test form, so the dictionary abstains by construction and the lexical fallback carries the load alone. Three configurations were compared at each of seven dominance thresholds: lexical reduction alone, dictionary alone, and the deployed combination. A configuration answering a share c of records with precision π among those has end-to-end accuracy\

$$A = c \cdot \pi \quad (12)$$

which is what makes τ a trade-off rather than a correctness boundary: raising it moves mass from c to π . Deployment spans quarters with no curated field at all, so neither arm describes it alone. Writing ρ for the share of all records the dictionary actually resolves, measured over the full series rather than assumed, expected accuracy end to end is

$$\hat{A} = \rho \cdot \pi_{\text{dict}} + (1 - \rho) \cdot A_{\text{lex}} \quad (13)$$

where the first term is dictionary precision measured on the record split and the second is lexical accuracy on the surface-form split.

For the corpus's two ingredients, annotated concepts were compared with the corresponding reference profiles. The two use different terminologies, and resolving one into the other needs the licensed MedDRA dictionary, so matching used exact correspondence after normalisation plus a small documented table of clinical equivalences. No approximate matching was used, because a wrong pairing manufactures agreement that does not exist; the proportion matched is reported with every result derived from it, and conservative matching makes that proportion a lower bound on true

alignment. The rank correlations are computed on the terms that aligned and are conditional on them, in either direction.

The analysis is organised as ordered stages (Figure 1), each writing a columnar artefact the next consumes, so any stage can be re-run without repeating its predecessors. Acquisition streams one quarter at a time and reduces it before fetching the next, holding peak storage at roughly one archive. Downloads are validated by inspecting the archive header rather than the HTTP status, because a success code does not guarantee an archive.

The output is gated rather than inspected: the produced document is opened as an OOXML package and every human-visible text node scanned for a date later than the study cut-off, and the assembled content is rejected if a display item is numbered out of the order a reader meets it, or if a citation or equation reference resolves to nothing. A further consequence of the staged design is that the manuscript's numbers are interpolated from the analysis artefacts, so a re-run rewrites the text; the gate rejects any count of three digits or more that is typed into the prose instead.

Cost is modest and the method is therefore reproducible on a single workstation with 16 GB of memory. Table 3 gives the measured cost of each stage. Excluding acquisition, whose duration measures the regulator's server rather than this method, the whole analysis completed in 13 min on 8 physical / 8 logical cores with a peak resident set of 11.1 GB. Cost is dominated by the repeated passes of the sensitivity and validation stages rather than by the single pass over the 30,297,215 drug records, the shape expected of a design that re-runs a cheap analysis under many settings rather than an expensive one once.

This study used only publicly available, de-identified data: spontaneous reporting extracts published by the regulator, and public forum posts distributed under an open research licence by their custodian. No individually identifiable information was accessed and no intervention was applied to any person, so institutional review board approval and informed consent were not required. The work was conducted in accordance with the principles of the Declaration of Helsinki.

Rank agreement between corpus and reference term frequencies was assessed with the Spearman coefficient over the terms that could be aligned.

Uncertainty on coverage and divergence is sampling variability over which drugs a corpus happens to contain, not sampling error in the reference, which is observed in full. It was obtained by drawing ingredient sets of the stated size without replacement, and ingredient pairs for the divergence null, under a fixed seed. Intervals are the 2.5th and 97.5th percentiles of 200 such draws and are not bootstrap confidence intervals; the divergence null rests on 2,000 sampled pairs. The intervals should be read to no finer a resolution than a percentage point.

The manuscript follows the TRIPOD recommendations on transparent reporting where they apply [22]. This is not a prediction-model study, but its subject is the data such models are developed on, and the TRIPOD requirements on describing the data source map onto it directly. Analyses were conducted in Python; all code and derived tables form a reproducible pipeline.

Results

The corpus contains 1,250 posts carrying 8,404 annotated spans, of which 5,796 are typed as adverse reactions and resolve to 789 distinct concepts. Those posts are distributed across 12 named drug products. Mapping the products to their active ingredients reduces them to two: atorvastatin, in 1,000

posts (80.0%), and diclofenac, in 250 posts (20.0%), of which 145 are a fixed-dose combination with misoprostol. Table 4 gives the full breakdown and Figure 2 displays it.

The two ingredients occupy only 3 of the 14 Anatomical Therapeutic Chemical anatomical main groups. Eleven groups, including anti-infectives, antineoplastic and immunomodulating agents, nervous system agents and respiratory agents, are absent. A widely used adverse drug event corpus is, in four fifths of its content, a single statin.

Normalisation reduced 30,297,215 drug records to a usable ingredient in the large majority of cases: lexical reduction alone collapsed 600,375 distinct raw strings to 288,312 surface forms, and the induced dictionary of 20,891 entries resolved 92.3% of records. Applying the 200 report threshold left 1,526 ingredients over 19,348 terms.

The resulting profiles recover known pharmacology without being told it. Atorvastatin's most reported reaction is type 2 diabetes mellitus, ahead of a musculoskeletal group led by myalgia, pain in extremity, muscle spasms, arthralgia, rhabdomyolysis and muscular weakness. Diclofenac's profile is led by administrative rather than clinical terms, drug administered at inappropriate site and drug ineffective, with acute renal failure, vomiting and nausea behind them. Figure 3 shows both against the population profile.

Administrative terms are a real feature of spontaneous reporting rather than noise to be filtered, and the most reported term across the whole database is drug ineffective.

They nonetheless enter the corpus vocabulary, where this corpus's narratives never match them, so the coverage figures below are upper bounds on what such a corpus could reach.

On held-out records the induced dictionary covered 97.6% of drug mentions at 99.7% precision, and the deployed configuration of dictionary with lexical fallback reached 97.9% accuracy (Table 2). Lexical reduction alone reached only 33.3%, and on surface forms never seen during induction it reached 32.6%. The dictionary is therefore doing most of the work rather than decorating a rule-based system.

Weighting the two arms by the rate at which records actually resolve through the dictionary, 92.3%, gives an expected end-to-end accuracy of 94.6% across all 51 quarters, including the 42 that carry no curated field. The dominance threshold governs a trade-off: raising it from 0.50 to 0.99 lifts precision from 99.0% to 100.0% while coverage falls from 99.5% to 94.7%, and the value used here sits near the knee of that curve.

A statin and a non-steroidal anti-inflammatory drug might be expected to have near-disjoint reaction profiles, in which case a two-drug corpus would at least span two distant regions of the space. The Jensen-Shannon divergence between atorvastatin and diclofenac is 0.259 bits against a median of 0.579 for pairs drawn at random from the whole reference, placing the corpus pair among the 1% most similar (Figure 3a).

That comparison is confounded, and the confounding runs one way. Divergence from the population profile falls as an ingredient's report count rises, at a rank correlation of -0.77: a drug with a few hundred reports has a sparse profile that diverges from everything, and 57% of the reference has fewer than two thousand. Atorvastatin and diclofenac are among the most reported drugs in the database, so a null drawn uniformly asks whether they resemble each other more than obscure drugs resemble anything, which they must.

Restricting the null to ingredients reported between a third as often as the less reported corpus drug and three times as often as the more reported one, 2,000 pairs over 289 ingredients, gives a median of 0.415

and places the corpus pair among the 7% most similar. The two drugs are therefore somewhat more alike than comparable pairs rather than strikingly so, and the headline similarity of the unrestricted comparison is mostly an artefact of report volume. What follows about coverage does not depend on either figure.

The union of the two ingredients' 20 most reported terms is a vocabulary of 35 reaction terms. Across the remaining 1,524 profiled ingredients, that vocabulary covers a mean of 17.7% of each drug's own reaction mass, with a median of 16.7%. Some 99.6% of ingredients fall below 50% coverage (Figure 4). This is not a consequence of an unlucky choice of drugs.

Randomly drawn pairs of ingredients achieve a mean of 13.9% (95% range 8.4% to 17.4%), so the corpus performs at least as well as a typical two-drug corpus rather than worse. That sits alongside the pair being unusually alike without contradicting it: similarity between the two makes the corpus narrower than a count of two implies, while both being heavily reported drugs whose prominent reactions are common ones makes their vocabulary reach marginally further than an average pair's. Their net effect is small. The limitation belongs to the size of the drug set, not to its membership.

Nor is it an artefact of thinly reported drugs, whose profiles are the least stable and which make up much of the reference. Coverage rises only weakly with report volume (Spearman rho +0.14) and the mean across volume quintiles spans 1.9 percentage points, from 16.7% for ingredients with fewer than 426 reports to 18.6% for those reporting between 2,279 and 7,925, with the most reported fifth just behind at 18.5%. The drugs carrying the most evidence are covered slightly better and still below a fifth. Table 5 and Figure 5 give the coverage curve. Coverage rises steeply at first and then flattens: 31.9% at ten ingredients, 41.0% at twenty, 61.2% at one hundred and 76.7% at five hundred. Eighty per cent coverage requires between 750 and 1000 ingredients, the curve crossing near 793. Ninety per cent is not reached at all within the reference set: 84.0% at 1,500 ingredients, by which point almost every drug with a stable profile has been included.

Treating composition as a maximum-coverage problem separates what is achievable from what opportunistic assembly delivers. Scored on the 153 drugs neither strategy was optimised against, greedy selection dominated random selection at every size tested (Figure 6). It reached 14.2% with a single drug against 8.9% for a random choice, and 73.6% with one hundred against 62.1%.

Expressed as the effort required rather than the coverage obtained, the gap widens with ambition; the greedy figures are exact, while the random ones are interpolated between the sampled sizes. Reaching 50% coverage took 18 deliberately chosen ingredients against 39 chosen at random; 70% took 72 against 220, a factor of 3.1. At 80% the comparison ends: greedy needed 208 ingredients and random selection did not reach the target within the budget of 500 at all.

The third arm is the one a corpus builder would actually use, and it loses to both. Ranking candidates by report volume falls below the mean of random selection from the 3rd ingredient onward, and from the 20th it sits below the 2.5th percentile of random outcomes at every size tested (Figure 6). Reaching 70% took 395 ingredients chosen that way against 220 chosen at random and 72 chosen deliberately; 80% it did not reach within the budget at all.

What the optimiser selects instead explains it (Table 6). The intuitive reading, that heavily reported drugs are poor material, is wrong: scored one at a time against an empty corpus, a candidate's own contribution rises with its report count (Spearman rho +0.33 over 1,221 candidates), and the best single choice, Cephalixin, ranks 473rd by volume rather than first. What sinks the strategy is overlap. A drug

is heavily reported partly because its reactions are common ones, so the prominent terms of any two such drugs are largely the same terms: 20 drugs with 20 terms each would hold 400 between them if none overlapped, and the 20 most reported hold 187, fewer than 20 drawn at random (200), against 272 for 20 chosen deliberately. A coverage objective forces each addition to be unlike what is already held, which pushes the later choices down the volume distribution: the median of the first twenty ranks 934th of 1,526. Report volume measures the evidence a drug has accumulated, which is what detecting a signal needs; it does not measure what that drug's vocabulary adds to what a corpus already covers.

The ordering does not depend on how deep each drug's vocabulary is cut (Table 7). Repeated at 10, 30 and 50 terms per drug, the volume rule stays behind an unguided choice at every sampled size from 20 ingredients up. What depends on depth is how much the choice is worth: at 10 terms an unguided choice needed 2.8 times as many ingredients as a deliberate one to reach half the reaction space, and at 50 terms only 1.5 times as many. Annotating each drug more deeply buys part of what choosing it well would have bought.

The overlap offered above as the explanation is itself a property of deep vocabularies. At 50 terms the 20 most reported hold 365 distinct terms between them against 413 for 20 drawn at random, and the gap closes as the cut gets shallower until at 10 terms the volume rule's vocabulary is no smaller than an unguided choice's (118 against 115), and it nonetheless loses by its widest margin there, needing 1.8 times the ingredients. Where every drug's most frequent terms are generic, what separates the strategies is which terms a corpus holds rather than how many.

Because greedy attains at least $(1 - 1/e)$ of the optimum for a monotone submodular objective, its value bounds what any corpus of the same size could achieve on the objective it maximises. At twenty ingredients the bound is 82.7%, so even a perfectly chosen twenty-drug corpus could not exceed that. The bound binds only up to about 20 ingredients, beyond which it exceeds 83.2%, the whole mass any set of these drugs can reach, and states nothing (Figure 6). That range is the one the question lives in: the corpus measured here holds two ingredients.

Submodularity holds by construction, so gains must be non-increasing; what the reference determines is the rate, and it is steep. The first ingredient adds 14.0% of the reaction space, the 13th adds a tenth of that, and by the 464th a further ingredient adds nothing distinguishable from zero (Figure 7a).

The objective saturates at 83.2%, which is not complete coverage: the remaining 16.8% lies on reactions no candidate reports often enough for them to enter its vocabulary.

Exploiting it paid. The lazy form spent 7,868 marginal-gain evaluations against an estimated 485,750 for the naive one, a speedup of 49 times in wall time. Priming the queue accounts for 1,221 of them; after the first selection the median step costs 2 (Figure 7b).

Three parameters were fixed by judgement: vocabulary depth, the profiling threshold and the restriction to suspect drugs. Varying each across a plausible range over 12 configurations (Table 8) left the conclusion intact. A two-drug corpus covered between 9.4% and 27.3% of an unseen drug's reaction mass, and the number of ingredients needed for 80% ranged from 200 to 1000, with 3 configurations not reaching it: 2 at the largest size tried and 1 where the reference set was too small to go further.

Vocabulary depth was the most consequential parameter, as expected, since it sets directly how much each drug contributes.

The setting most favourable to the corpus, counting each drug's fifty most reported terms rather than twenty, still left it at 27.3% and still required 200 ingredients to reach 80%. Admitting concomitant

medications lowered coverage rather than raising it, because it pulls every profile toward the population mean: the two drugs' combined vocabulary contracts from 35 terms to 26 as their most reported reactions converge on the common ones.

Even for its own two ingredients, the corpus and the reference describe overlapping but distinct sets of events. Conservative matching aligned 25.6% of atorvastatin concepts and 33.8% of diclofenac concepts to a reported term, covering 49.7% and 44.6% of annotation mass respectively.

Among terms that did align, rank agreement was moderate: Spearman rho 0.51 over 182 terms for atorvastatin and 0.49 over 77 for diclofenac, though 57% of the aligned atorvastatin concepts carry one or two mentions, so the coefficient rests largely on ties. The corpus recalls 70% of atorvastatin's twenty most reported terms and 45% of diclofenac's (Table 9).

Where both sources carry a term with appreciable mass, they still weight it differently, by up to 7-fold in the corpus's favour and 4-fold in the reference's (Figure 8). The divergences are specific to each drug rather than following a single axis, and they rest on 30 and 10 terms respectively, so they are observations about two medicines and not a general property of consumer text. An earlier analysis appeared to show a systematic split between diagnosed and experiential events; it did not survive the correction of two matching faults described in the limitations, and is not claimed here.

Table 1: Notation used throughout, with the value each symbol takes in this study. Equation numbers refer to the definition.

Symbol	Meaning	In this study
U	Set of MedDRA preferred terms	$ U = 19,348$
$n(t, r)$	Suspect-drug reports pairing ingredient t with term r	2,139,761 non-zero pairs
$m(t, r)$	Reaction profile of ingredient t (Equation 2)	sums to 1 over U
S(d)	The most reported terms of ingredient d	20 terms per drug
D	Ingredient set a corpus spans	$ D = 2$ for the corpus studied
V(D)	Corpus vocabulary (Equation 4)	$ V = 35$ for that corpus
C(t, D)	Held-out coverage of an unseen ingredient (Equation 5)	mean 17.7%
P	Candidate pool of ingredients	$ P = 1,221$
R(d)	Reports naming ingredient d (Equation 10)	median 1,378 over P
E, E'	Sets defining the objective and withheld from it	$ E = 152$, $ E' = 153$
f(D)	Objective, mean coverage over E (Equation 6)	monotone submodular; results on E'
$\Delta(d A)$	Marginal gain of adding d to A (Equation 8)	non-increasing in A
k	Cardinality constraint on the corpus	budget 500
D*	Best corpus of size k drawn from P (Equation 7)	not computable; bounded by Equation 11
τ	Dominance threshold for induction (Equation 1)	0.80 deployed, 0.50 to 0.99 swept

Symbol	Meaning	In this study
ρ	Share of records the dictionary resolves (Equation 13)	92.3%

Table 2: Accuracy of the drug-name normalisation on held-out records from the curated quarters. Coverage is the share of records for which the configuration produces an answer; precision is accuracy among those. The dominance threshold governs the trade-off between the two.

Configuration	Dominance	Dictionary entries	Coverage (%)	Precision (%)	Accuracy (%)
Lexical reduction only	-	-	100.0	33.3	33.3
Induced dictionary	0.50	18,632	99.5	99.0	98.6
Induced dictionary	0.60	18,524	98.8	99.4	98.2
Induced dictionary	0.70	18,381	98.3	99.5	97.9
Induced dictionary	0.80	18,260	97.6	99.7	97.4
Induced dictionary	0.90	18,074	96.3	99.9	96.2
Induced dictionary	0.95	17,950	95.8	99.9	95.7
Induced dictionary	0.99	17,830	94.7	100.0	94.7
Dictionary with lexical fallback	0.80	18,260	100.0	97.9	97.9

Table 3: Cost of each analysis stage, measured rather than estimated. Acquisition is excluded, since its duration measures the regulator's server rather than this method. Peak memory is the largest resident set observed during the stage.

Stage	Scale	Wall time (s)	Peak memory (GB)
Parse the annotation corpus	1,250 posts, 8,404 spans	1.0	0.1
Normalise drug names	30,297,215 drug records	101.4	7.5
Reaction profiles and coverage curve	39.6M ingredient-reaction pairs	128.6	5.9
Corpus against the reference	2 ingredients, 359,531 reaction rows	11.8	3.7
Render figures	8 figures	10.0	0.4

Stage	Scale	Wall time (s)	Peak memory (GB)
Validate the normaliser	9.08M gold pairs, 7 thresholds x 2 splits	180.4	4.6
Maximum-coverage optimisation	1,221 candidates, budget 500	32.3	6.1
Sensitivity analysis	12 configurations	314.6	11.1
Total	8 physical / 8 logical cores	780.1	11.1

Table 4: Composition of the annotation corpus by drug product, active ingredient and Anatomical Therapeutic Chemical code. Arthrotec is a fixed-dose combination; its misoprostol component is a third active ingredient and is counted with diclofenac throughout because the annotations describe the product rather than either constituent.

Drug product	Active ingredient	ATC code	Posts	% of corpus
Lipitor	atorvastatin	C10AA05	1,000	80.0
Arthrotec	diclofenac + misoprostol	M01AB55	145	11.6
Voltaren	diclofenac	M01AB05	46	3.7
Voltaren Xr	diclofenac	M01AB05	22	1.8
Cataflam	diclofenac	M01AB05	10	0.8
Diclofenac Sodium	diclofenac	M01AB05	7	0.6
Zipsor	diclofenac	M01AB05	5	0.4
Cambia	diclofenac	M01AB05	4	0.3
Pennsaid	diclofenac	M02AA15	4	0.3
Diclofenac Potassium	diclofenac	M01AB05	3	0.2
Solaraze	diclofenac	D11AX18	3	0.2
Flector	diclofenac	M02AA15	1	0.1
Total	2 ingredients after combination	3 groups	1,250	100.0

Table 5: Mean held-out coverage of the adverse reaction space as a function of the number of active ingredients in a corpus. Coverage is the share of an unseen drug's own reaction mass falling within the corpus vocabulary, averaged over all remaining ingredients. Ranges are the 2.5th and 97.5th percentiles over randomly drawn ingredient sets of each size, and describe which drugs a corpus happens to contain rather than sampling error in the reference.

Ingredients in corpus	Vocabulary size (terms)	Mean coverage (%)	95% range (%)
1	20	8.9	4.3 to 12.4
2	36	13.9	8.4 to 17.4
5	76	23.6	19.8 to 26.4
10	124	31.9	29.1 to 34.7
20	199	41.0	38.5 to 43.7
50	349	52.9	50.6 to 54.9

Ingredients in corpus	Vocabulary size (terms)	Mean coverage (%)	95% range (%)
100	515	61.2	59.4 to 62.7
200	733	68.6	67.4 to 69.9
500	1,134	76.7	75.9 to 77.5
750	1,351	79.7	78.9 to 80.4
1,000	1,521	81.6	80.9 to 82.2
1,500	1,791	84.0	82.0 to 85.7

Table 6: The first twenty active ingredients that maximum coverage selects, each with the marginal gain it contributes to the objective and its rank among the 1,526 profiled ingredients by number of reports. The median selection ranks 934th, because each addition has to be unlike the ones already chosen rather than prominent in its own right. Gains are in-sample; the coverage they translate into on withheld drugs is Figure 6.

Order	Active ingredient	Reports	Rank by reports	Marginal gain (%)
1	Cephalexin	3,594	473	13.95
2	Oxazepam	3,078	508	5.75
3	Roflumilast	2,529	570	4.71
4	Fusidic Acid	620	1,073	3.81
5	Glimepiride Rosiglitazone	2,434	583	2.43
6	Danaparoid	629	1,063	2.30
7	Phentermine Hydrochloride Topiramate	2,004	662	2.22
8	Atracurium	1,185	829	2.09
9	Magnesium Sulfate Potassium Sulfate Sodium Sulfate	600	1,085	2.06
10	Tafluprost	544	1,111	1.91
11	Gadoteridol	2,399	587	1.47
12	Oxacillin	445	1,202	1.46
13	Trichlormethiazide	312	1,350	1.22
14	Zanamivir	3,470	478	1.21
15	Retapamulin	677	1,038	1.04
16	Ambrosia Artemisiifolia Pollen	224	1,475	0.97
17	Diacetylmorphine	4,439	419	0.97
18	Insulin Lispro	42,547	50	0.91
19	Toremifene	272	1,397	0.89
20	Deflazacort	208	1,505	0.87

Table 7: Robustness of the three-way comparison to how deep each drug's vocabulary is cut. Ingredients needed is the corpus size at which a strategy reaches 50% mean coverage of drugs it was not chosen against. Terms in the first 20 is the distinct reaction terms its first twenty

ingredients hold between them, out of twenty times the depth were none shared. The deployed setting, 20 terms, is what Figure 6 plots.

Measure	10 terms	20 terms	30 terms	50 terms
Ingredients needed, maximum coverage	31	18	13	8
Ingredients needed, random selection	88	39	23	12
Ingredients needed, most reported first	161	58	31	17
Terms in the first 20, maximum coverage	157	272	357	563
Terms in the first 20, random selection	115	201	275	413
Terms in the first 20, most reported first	118	187	238	365

Table 8: Sensitivity of the two headline quantities to the three analytical choices, varied one at a time with the others held at their default value. The last column gives the largest size tried rather than a bare 'not reached', because a configuration with fewer profiled ingredients cannot be tested as far up the grid; the rest simply do not reach 80% within it.

Choice	Setting	Ingredients profiled	Two-drug coverage (%)	Ingredients for 80%
Vocabulary depth	top-10	1,526	9.4	above 1,500
Vocabulary depth	top-20	1,526	17.7	1,000
Vocabulary depth	top-30	1,526	21.8	500
Vocabulary depth	top-50	1,526	27.3	200
Profiling threshold	50 reports	2,155	17.2	750
Profiling threshold	100 reports	1,798	17.5	750
Profiling threshold	200 reports	1,526	17.7	1,000
Profiling threshold	500 reports	1,147	18.0	1,000
Profiling threshold	1000 reports	883	18.1	above 750
Causal role	PS only	1,192	19.2	750
Causal role	PS + SS (default)	1,526	17.7	1,000
Causal role	PS + SS + concomitant	2,225	15.3	above 1,500

Table 9: Agreement between corpus annotations and spontaneous reports for the two ingredients the corpus contains. Match rates are lower bounds, since cross-terminology alignment was deliberately conservative. Rank agreement is dominated by ties, most aligned concepts being annotated once or twice, so the coefficient is indicative rather than precise.

Measure	Atorvastatin	Diclofenac
Annotated reaction mentions	4,989	807

Measure	Atorvastatin	Diclofenac
Distinct annotated concepts	712	228
Reports in the reference	76,202	25,339
Distinct reference terms	5,743	4,873
Concepts aligned to a reference term	182 (25.6%)	77 (33.8%)
Annotation mass aligned	49.7%	44.6%
Spearman rho on aligned terms	0.51 (n = 182)	0.49 (n = 77)
Recall of the 20 most reported terms	70%	45%

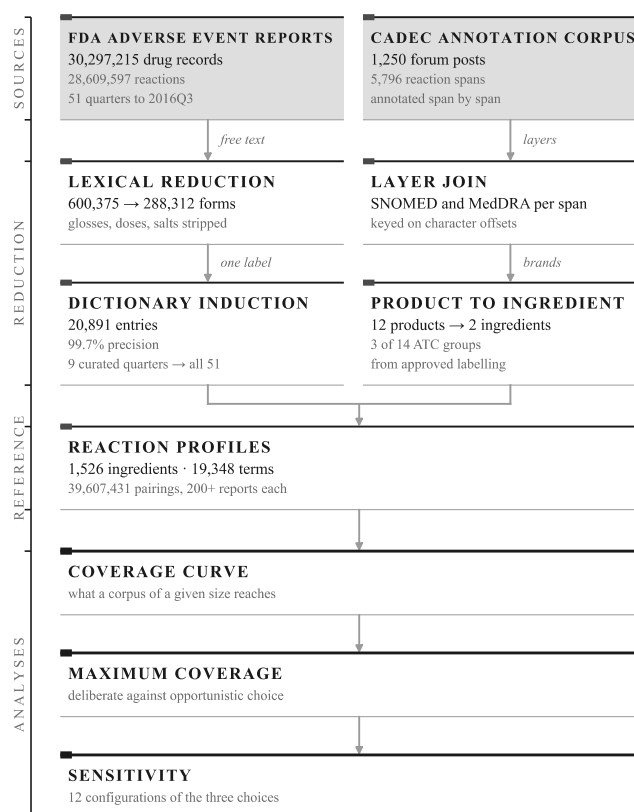


Figure 1: The analysis pipeline. Two public sources are reduced along separate paths and converge on a single set of reaction profiles, from which three analyses follow. Counts are those obtained in this study.

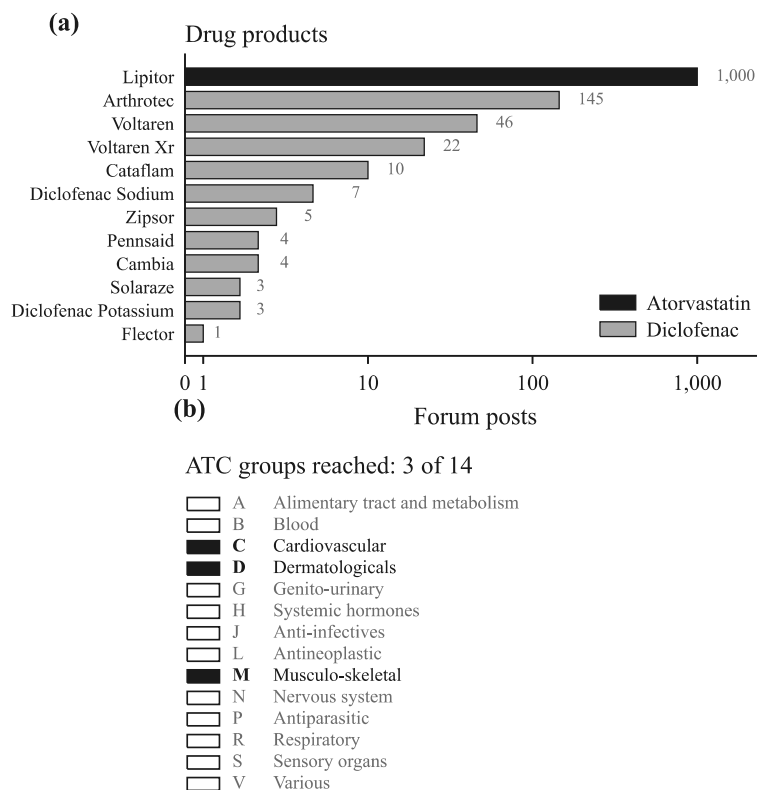


Figure 2: Composition of the annotation corpus. (a) Forum posts per drug product on a symmetric-log scale, linear below ten posts, shaded by active ingredient. (b) The Anatomical Therapeutic Chemical anatomical main groups the corpus reaches, filled, against those it does not.

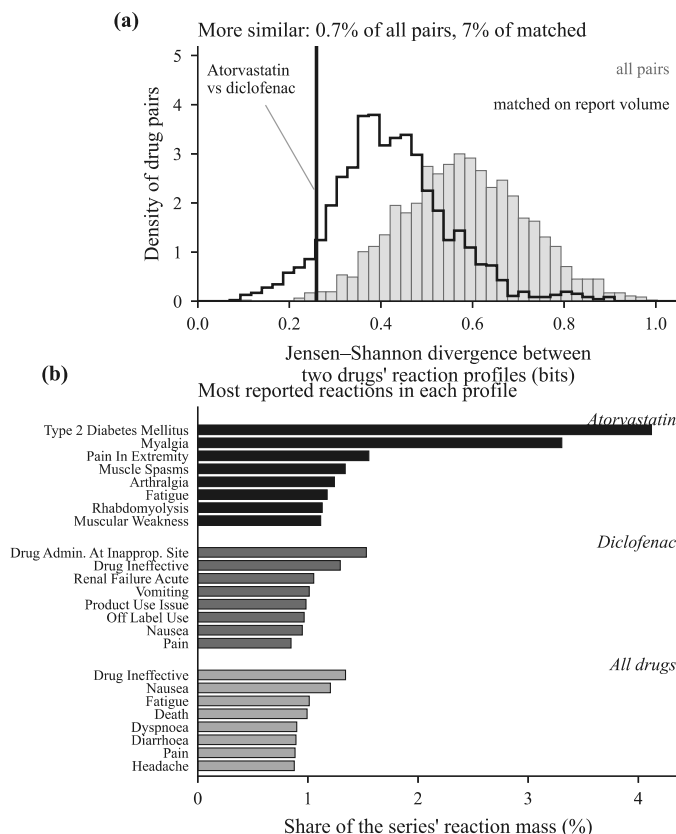


Figure 3: Reaction profiles. (a) Jensen-Shannon divergence between pairs of active ingredients, with the corpus pair marked. The shaded distribution is over pairs drawn from the whole reference; the outline is over pairs matched on report volume, which is the comparison the text relies on, because divergence falls as report count rises. (b) The most frequently reported reactions for each corpus drug and for all drugs combined, as a share of that series' reaction mass.

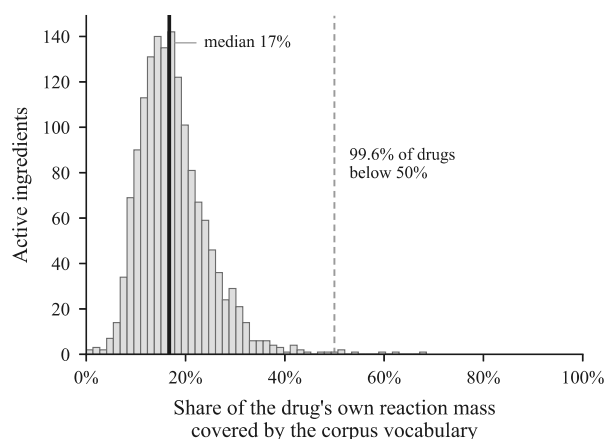


Figure 4: Distribution across the active ingredients the corpus does not contain, of the share of each drug's own reaction mass covered by the corpus vocabulary.

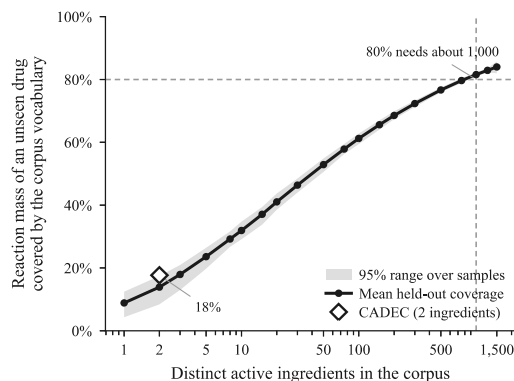


Figure 5: Mean held-out coverage of the adverse reaction space against the number of distinct active ingredients in a corpus. The diamond marks the corpus examined here; shading gives the 95% range across sampled ingredient sets.

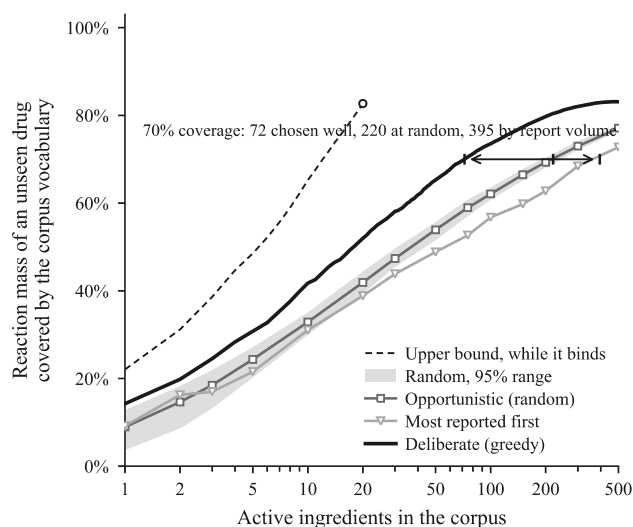


Figure 6: Three ways of choosing which drugs a corpus contains, all scored on drugs none of them was optimised against. Greedy selection maximises marginal coverage at each step; random selection is the mean over sampled sets, with its 95% range shaded; most-reported-first takes the candidates in descending order of report count, which is the strategy a corpus builder would reach for and the one that falls below the shaded range beyond twenty ingredients. The dashed line is the upper bound on the optimisation objective implied by the $(1 - 1/e)$ guarantee, the marker is where it stops constraining anything.

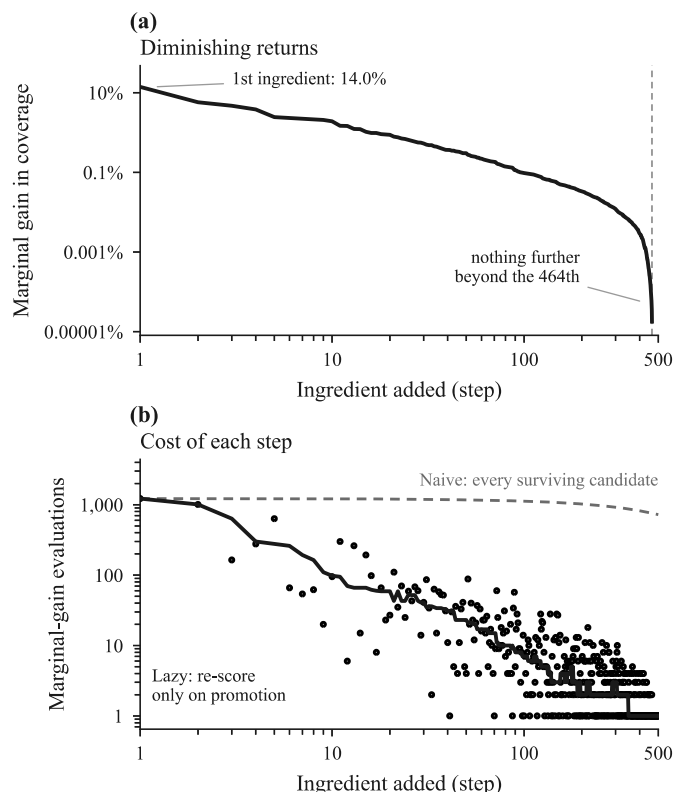


Figure 7: Structure of the objective and the cost of exploiting it. (a) Marginal gain of each ingredient the optimiser adds, on the objective it maximises, with the step beyond which a further ingredient adds nothing distinguishable from zero. Submodularity requires the decline; what the data determine is how steep it is. (b) Marginal-gain evaluations spent on each step, as points with a running median, against the cost a naive implementation would pay, which falls by one at each step as the winner leaves the pool.

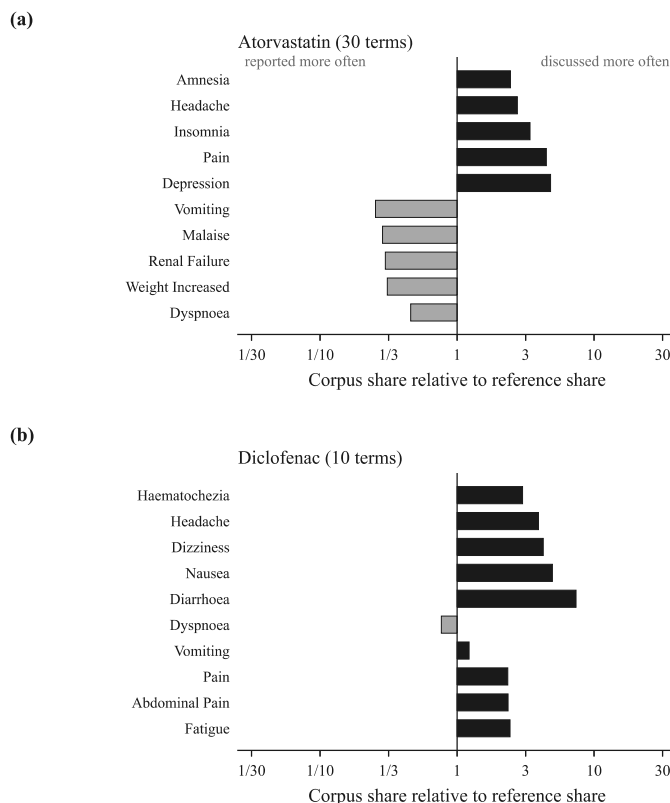


Figure 8: How differently the two sources weight the terms they share. For each aligned reaction term, the share of corpus mentions it takes divided by the share of reference reaction mentions it takes, for the five highest and five lowest ratios among terms with at least five corpus mentions and 0.2% of that drug's reported reaction mentions. The filter is two-sided by necessity: ranking on the ratio alone puts every term annotated once at the bottom whatever the sources agree about.

(a) Atorvastatin. (b) Diclofenac.

Discussion

A widely used adverse drug event corpus contains two active ingredients spanning 3 of 14 therapeutic anatomical groups, and its vocabulary reaches 17.7% of an average unseen drug's reported reactions. Reaching 80% would take roughly 793 ingredients sampled at random from the reference, and 208 even when they are chosen to maximise coverage, either way two orders of magnitude beyond the two this corpus contains.

The two ingredients present are somewhat more alike than drugs of comparable reporting volume, though their vocabulary still reaches marginally further than an average pair's; the limitation is the count of two rather than the choice. Composition is also a trap for the obvious heuristic: taking the most reported drugs first covers less than taking them at random from the third ingredient on, because at the vocabulary depth used here the prominent terms of prominent drugs are largely the same terms.

The finding is structural rather than a criticism of how any particular corpus was assembled. What the coverage curve shows is that the adverse reaction space is not compressible into a few drugs: reaction profiles are specific enough, and the long tail of terms broad enough, that coverage accumulates slowly and saturates well below completeness.

The practical consequence concerns how performance is described rather than how systems are built. A system reported as achieving a given F score at adverse drug event extraction has achieved it on the reactions of a few medicines, and nothing in an evaluation conducted inside the corpus can reveal the bound, because the held-out data inherit the training data's drugs.

This is the problem that spectrum bias poses in diagnostic accuracy research [23], transposed from the patients a test is evaluated in to the drugs a corpus is built from. The closer analogy is not confounding, which concerns a causal contrast; there is no exposure-outcome estimand here, only a sample of drugs that does not represent the population the instrument is used on.

Breadth of drugs may not be the only axis on which such a corpus is narrow.

Even for the two medicines it does contain, only a quarter to a third of annotated concepts align with a reported term at all, and agreement on relative prominence among those is moderate.

The two sources describe overlapping but distinct sets of events, which is reason to treat consumer narrative as a complement to spontaneous reporting rather than a substitute for it, at least for the two medicines measured here.

Whether the divergence follows a systematic direction is a question this study cannot answer; doing so properly needs a licensed terminology and more than two drugs.

One incidental result generalises beyond this study. Greedy selection seeks the drugs least like anything already covered, and normalisation errors produce exactly that signature: a mis-parsed name yields a small, idiosyncratic profile attached to an entity that does not exist. The optimiser's first selections were consequently a list of the reference's defects.

The first pass surfaced a lexical fault, a salt word used as a descriptor being stripped as though it were a counterion. After that correction the selection still led with pack and device strings, which traced to two further faults: a name reduced past its last counterion, so that saline and antacids entered the reference as the ingredients sodium and calcium, and unmapped surface forms passing through as ingredients in their own right. Requiring the name to be attested in the regulator's curated field removed the second class without a hand-written stop-list. The first twenty selections of the corrected run are ordinary medicines.

Over this reference, submodular maximisation was therefore a cheap audit, concentrating attention on the records where normalisation had failed and surfacing three faults in succession, and worth re-running after each correction rather than once.

Work on the quality of adverse drug event corpora has largely addressed annotation: inter-annotator agreement, span conventions and normalisation of lay expressions to controlled terminology [7,24,25]. Those ask whether the annotations are correct; the question here is prior to them, and concerns whether the drugs are representative given that they are. That a biomedical corpus should be designed against the distribution it will be applied to was argued early [26], and the dependence of shared-task resources on data availability rather than clinical representativeness has been noted since [27].

Curated derivatives of spontaneous reporting data [28] and inference over them [29] have made the reference side of this comparison progressively easier to assemble. The pharmacovigilance literature has long held that different data sources capture different fractions of the reaction space, and that signal detection should therefore be triangulated across them [30,31]. This study gives one quantitative expression of that principle, for a text corpus measured against spontaneous reports, and the fraction is small.

Neither vocabulary is strictly normalised to preferred terms: the reference carries muscle cramp and muscle spasms as separate strings, and the corpus annotates both joint pain and arthralgia. Matching by exact string first, and consulting the equivalence table only on failure, therefore defeated the table where it was needed, pairing a concept with a rare literal term while the common term it is identical to went unconsulted, and admitting one concept twice under two names.

Both vocabularies are now folded onto canonical terms first. A conclusion about the direction of disagreement between the channels had been drawn from the uncorrected output and is withdrawn.

Cross-terminology matching is the principal limitation. The corpus is annotated in SNOMED Clinical Terms and the reference in MedDRA, and mapping between them properly needs a licensed dictionary, so the match rates are lower bounds and the rank correlations are conditional on which terms aligned. The coverage results do not depend on this step, being computed entirely within the reference terminology.

Spontaneous reporting data carry well-known biases: reporting is voluntary and incomplete, subject to notoriety and stimulation, and the denominator of exposed patients is unknown [3,30]. The reference profiles therefore describe reported rather than incident reactions, which is acceptable here because the comparison is between what a corpus contains and what the reporting system records, and both are reporting artefacts.

Reports were deduplicated at the level of the individual report rather than the case, because the case identifier is absent from the earlier extracts. Follow up versions of a case may therefore be counted more than once, a documented feature of this source [32], which would inflate the weight of drugs whose cases attract more follow-up. The coverage results compare relative profile shapes rather than absolute counts and should be little affected; the volume-ranked arm and the entry threshold depend on report counts directly, and the size of that effect was not estimated.

Requiring attestation removes pack and device strings but inherits whatever the curated field itself contains, and it contains some crude entries: calcium, magnesium, potassium and water appear as active ingredients in their own right, for mineral supplements and sterile water products. They were left in rather than filtered on judgement.

Two recommendations follow. Adverse drug event corpora should report their therapeutic composition, by active ingredient and by therapeutic class, as routine documentation; this costs nothing and lets readers calibrate the performance figures computed on them. And performance claims derived from such corpora should be stated with the drugs attached, as a diagnostic accuracy study states the spectrum of patients in which it was conducted.

A third follows from the arm that lost. Where a corpus is assembled around a therapeutic question the drugs are given, but where they are not, the default of starting from the most reported drugs is actively worse than drawing them at random, and costs 3 to 5 times as many ingredients, depending on the coverage aimed at, as choosing them against a coverage objective. The reference needed to do better is public, and the selection takes under a minute to compute.

The coverage curve also offers a design target. A corpus assembled to span the reaction space deliberately reaches a given coverage with far fewer ingredients than one assembled opportunistically. Whether class-stratified sampling improves held-out coverage relative to the random sampling modelled here is a direct extension of this work, testable with the same reference data.

Conclusion

The reach of an adverse drug event corpus is bounded by the number of distinct active ingredients it contains. This corpus's two cover 17.7% of an average unseen drug's reported reactions, and reaching 80% would require on the order of 793 ingredients drawn at random, or 208 chosen against a coverage objective. The bound is structural, invisible to evaluation drawn from the corpus's own drugs, and readily measurable against public spontaneous reporting data. Therapeutic breadth deserves to be treated as a design parameter of these resources, and as a stated limitation of the performance figures obtained from them.

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Abbreviations: ATC, Anatomical Therapeutic Chemical; CADEC, CSIRO Adverse Drug Event Corpus; CSIRO, Commonwealth Scientific and Industrial Research Organisation; MedDRA, Medical Dictionary for Regulatory Activities; PS, primary suspect; SS, secondary suspect; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

Conflict of Interest: The author declares no potential conflicts of interest with respect to research, authorship and/or publication of this article.

Ethical Approval: Not required. The study analysed only publicly available, de-identified data and involved no human participants. The work was conducted in accordance with the principles of the Declaration of Helsinki.

Data Availability: All data analysed are publicly available. The spontaneous reporting extracts are published quarterly by the United States Food and Drug Administration. The annotation corpus is distributed by the Commonwealth Scientific and Industrial Research Organisation under the CSIRO Data Licence. The acquisition, normalisation and analysis code, together with all derived tables, is available from the author on request.

References

1. Pirmohamed M, James S, Meakin S, et al. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ*. 2004;329(7456):15-9. doi:10.1136/bmj.329.7456.15
2. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA*. 1998;279(15):1200-5. doi:10.1001/jama.279.15.1200
3. Hazell L, Shakir SA. Under-reporting of adverse drug reactions: a systematic review. *Drug Saf*. 2006;29(5):385-96. doi:10.2165/00002018-200629050-00003
4. Rawson NSB. Canada's adverse drug reaction reporting system: a failing grade. *J Popul Ther Clin Pharmacol*. 2015;22(2):e167-72.
5. Sarker A, Ginn R, Nikfarjam A, et al. Utilizing social media data for pharmacovigilance: a review. *J Biomed Inform*. 2015;54:202-12. doi:10.1016/j.jbi.2015.02.004
6. Golder S, Norman G, Loke YK. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. *Br J Clin Pharmacol*. 2015;80(4):878-88. doi:10.1111/bcp.12746
7. Karimi S, Metke-Jimenez A, Kemp M, Wang C. Cadec: a corpus of adverse drug event annotations. *J Biomed Inform*. 2015;55:73-81. doi:10.1016/j.jbi.2015.03.010

8. Gurulingappa H, Rajput AM, Roberts A, et al. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J Biomed Inform.* 2012;45(5):885-92. doi:10.1016/j.jbi.2012.04.008
9. Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. *J Biomed Inform.* 2013;46(5):914-20. doi:10.1016/j.jbi.2013.07.011
10. Li J, Sun Y, Johnson RJ, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)*. 2016;2016:baw068. doi:10.1093/database/baw068
11. Sakaeda T, Tamon A, Kadoyama K, Okuno Y. Data mining of the public version of the FDA Adverse Event Reporting System. *Int J Med Sci.* 2013;10(7):796-803. doi:10.7150/ijms.6048
12. Wong CK, Ho SS, Saini B, et al. Standardisation of the FAERS database: a systematic approach to manually recoding drug name variants. *Pharmacoepidemiol Drug Saf.* 2015;24(7):731-7. doi:10.1002/pds.3805
13. Nelson SJ, Zeng K, Kilbourne J, et al. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc.* 2011;18(4):441-8. doi:10.1136/amiajnl-2011-000116
14. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). *Drug Saf.* 1999;20(2):109-17. doi:10.2165/00002018-199920020-00002
15. Lin J. Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory.* 1991;37(1):145-51. doi:10.1109/18.61115
16. Karp RM. Reducibility among combinatorial problems. In: Miller RE, Thatcher JW, editors. *Complexity of Computer Computations*. New York: Plenum Press; 1972. p. 85-103. doi:10.1007/978-1-4684-2001-2_9
17. Feige U. A threshold of $\ln n$ for approximating set cover. *J ACM.* 1998;45(4):634-52. doi:10.1145/285055.285059
18. Nemhauser GL, Wolsey LA, Fisher ML. An analysis of approximations for maximizing submodular set functions—I. *Math Program.* 1978;14(1):265-94. doi:10.1007/BF01588971
19. Hochbaum DS, Pathria A. Analysis of the greedy approach in problems of maximum k-coverage. *Nav Res Logist.* 1998;45(6):615-27. doi:10.1002/(SICI)1520-6750(199809)45:6<615::AID-NAV5>3.0.CO;2-5
20. Minoux M. Accelerated greedy algorithms for maximizing submodular set functions. In: Stoer J, editor. *Optimization Techniques, Part 2. Lecture Notes in Control and Information Sciences*, vol 7. Berlin: Springer; 1978. p. 234-43. doi:10.1007/BFb0006528
21. Cormode G, Karloff H, Wirth A. Set cover algorithms for very large datasets. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. New York: ACM; 2010. p. 479-88. doi:10.1145/1871437.1871501
22. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med.* 2015;162(1):55-63. doi:10.7326/M14-0697
23. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med.* 1978;299(17):926-30. doi:10.1056/NEJM197810262991705
24. Metke-Jimenez A, Karimi S. Concept identification and normalisation for adverse drug event discovery in medical forums. In: *Proceedings of the First International Workshop on Biomedical Data Integration and Discovery (BMDID 2016)*, co-located with the 15th International Semantic Web Conference; 2016 Oct 17; Kobe, Japan. *CEUR Workshop Proceedings*, vol. 1709.
25. Limsopatham N, Collier N. Normalising medical concepts in social media texts by learning semantic representation. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics; 2016. p. 1014-23. doi:10.18653/v1/P16-1096

26. Cohen KB, Fox L, Ogren PV, Hunter L. Corpus design for biomedical natural language processing. In: Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics. Detroit, MI: Association for Computational Linguistics; 2005. p. 38-45.
27. Chapman WW, Nadkarni PM, Hirschman L, et al. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc.* 2011;18(5):540-3. doi:10.1136/amiajnl-2011-000465
28. Banda JM, Evans L, Vanguri RS, et al. A curated and standardized adverse drug event resource to accelerate drug safety research. *Sci Data.* 2016;3:160026. doi:10.1038/sdata.2016.26
29. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. *Sci Transl Med.* 2012;4(125):125ra31. doi:10.1126/scitranslmed.3003377
30. Bate A, Evans SJ. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol Drug Saf.* 2009;18(6):427-36. doi:10.1002/pds.1742
31. Harpaz R, DuMouchel W, Shah NH, et al. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther.* 2012;91(6):1010-21. doi:10.1038/clpt.2012.50
32. Hauben M, Reich L, DeMicco J, Kim K. 'Extreme duplication' in the US FDA Adverse Events Reporting System database. *Drug Saf.* 2007;30(6):551-4. doi:10.2165/00002018-200730060-00009