

How reliable are drug names in Wikidata across languages? A frozen-snapshot audit of 34,207 labels

EasyxLab · EasyByte Hub S. Coop. Mad. · Study S4

 Working draft Published 2026-10-02 · Updated 2026-10-03

Abstract

We froze a revision-pinned snapshot of the labels of all 3,768 Wikidata items carrying an ATC code (34,207 labels in 20 languages) and of 10,823 items carrying ICD-10/ICD-11 codes. Seven explicit rule-based detectors were applied, and a stratified random sample of 118 flags was reviewed one by one by the study agent (an LLM-based agent; see §3.1) to measure each detector's precision. Script (0.71 drugs, 0.93 diseases), brand (0.83), shared-label (1.00), ICD-10 format (13/13) and obsolete-ATC (5/5) flags were mostly correct; INN (0.19), salt/parent (0.26) and duplicate-ATC (0/4) were not; ICD-11 is undetermined (1/2).

After adjusting for precision, we estimate that about 0.9% of drug labels are wrong (≈ 312 of 34,207; 0.5% at the lower 95% CI bounds of precision), a lower bound with respect to recall. Errors concentrate in Urdu (39 per 1,000 labels), Persian (35), Russian (28) and Hindi (25); the major Latin-script languages have 2–4 per 1,000 (excluding English, see §4.2). Error types include brands, English text, misspellings, an unrelated personal name and labels naming a different substance.

In code data, 2.5% of ICD-10 values are invalid, excluding block ranges, and 44 of them follow a single corruption pattern. 3.3% of ATC statements are codes replaced (WHO alterations list) but still ranked as current. Six of the seven errors from our earlier review were still present; the seventh was only partly fixed. We publish the dataset (CC0), the detectors and 135 corrections, reviewed one by one by the study agent, as a QuickStatements proposal; we made no edits ourselves.

1. Why this matters

Wikidata is released under CC0 and is designed for reuse. Its life-science overview paper states that "Wikidata has items for over 150 thousand chemical compounds, including over 3500 items which are specifically designated as medications", that "Compound attributes are drawn from a diverse set of databases, including PubChem, RxNorm, the IUPHAR Guide to Pharmacology, NDF-RT, and LIPID MAPS", and that "Wikidata is also natively multilingual" (Waagmeester et al., *eLife* 2020, doi:10.7554/eLife.52614).

The WikiProject Medicine on Wikidata states its aim directly: "The goals of this WikiProject are to centralize data about medical topics. This data will make maintenance of Wikipedia easier and will be available to anyone in the world". It also states that "Translation of items (Q...) and properties (P...) enables us to translate statements into all languages" (Wikidata:WikiProject Medicine, accessed 2026-10-02).

Documented reuse paths we could verify:

- **French Wikipedia** — the drug infobox pulls fields from Wikidata. The source of `Modèle:Infobox Médicament` contains `{{Wikidata|P267|{{{ATC}}}}}` and `{{Wikidata|P3350|{{{DCI}}}}}`.
- **Spanish Wikipedia** — the drug infobox does the same. The source of `Plantilla:Ficha de medicamento` contains `{{Propiedad|p267|{{{Prefijo_ATC}}}} }`.
- **Pfundner et al., JMIR 2015** (doi:10.2196/jmir.4163) — "We set up exemplary implementations demonstrating how the DDI data we introduced into Wikidata could be displayed in Wikipedia articles in diverse languages."
- **TA2Viewer** (Halle, Kikinis & Neumann, *Clinical Anatomy* 2024, doi:10.1002/ca.24162) — an anatomy browser that "can optionally use unofficial synonyms from Wikidata to provide multilingual term searches in hundreds of languages". This is the same reuse pattern Wikidata's drug labels would serve.
- **Waagmeester et al. 2020** — "An identifier translation service is a simple and straightforward application of the biomedical content in Wikidata."

Within our budget we did **not** find a verifiable source naming a specific consumer drug app, search engine or language-model pipeline that ingests these labels. We therefore make no claim about such downstream use.

2. Data

All data come from the Wikidata Query Service, retrieved 2026-10-02 between 15:24 and 15:29 UTC with 101 polite queries and no rate-limit responses. Every item is stored with its `lastrevid`.

- **Drugs.** All items with any P267 (ATC) statement: 3,768 items and 4,459 ATC statements. 363 of these items hold only class-level codes, i.e. they are ATC groups rather than substances.
- **ICD items.** All items with P494 (ICD-10) or P7329 (ICD-11 MMS): 10,823 items, 5,536 ICD-10 values and 8,193 ICD-11 values.
- **Diseases (for label analysis).** The 5,212 ICD items classed as diseases, syndromes or signs. Excluded: 799 taxa, whose Latin binomials are legitimate in every language; 161 templates and categories; and 4,651 other items, mostly anatomy carrying ICD-11 extension codes.
- **Labels and aliases** in en, es, fr, de, it, pt, pl, ru, uk, tr, ar, fa, ur, hi, bn, zh, ja, ko, sw and am.
- **Context fields:** WHO INN (P2275), sitelink titles in the same 20 languages, and P31/P279.

Reference lists:

- **RxNorm Prescribable subset** (public domain): 4,125 brand names after removing names that are also ingredient names.
- **WHO ATC alterations list:** 355 code changes.
- **DrugBank Open Data Vocabulary** (CC0) could not be used: its download answered HTTP 403 without registration.

Coverage is itself uneven. Labels exist for:

- nearly every item in English (3,759);
- 60–75% of items in fr, es, de and ar;
- 46% in ru, 42% in uk and 31% in ko;

- 8–10% in hi, bn and ur;
- 5% in sw;
- 0.2% in am (9 labels).

3. Method

Each detector applies one explicit rule (full specification in METHOD.md):

- **(a) script** — a label in a non-Latin-script language contains no letter of that script, or a Latin-script language label contains Cyrillic, Arabic, CJK, Indic or Ethiopic letters.
- **(b) brand** — the label equals an RxNorm Prescribable brand name that is not also an ingredient name.
- **(c) salt/ester vs parent** — a counter-ion or ester term (from a lexicon in 12 scripts) is present in the label and absent from the English label, or the reverse. Inorganic compounds are excluded.
- **(d) shared label** — the same label in the same language on two or more drug items.
- **(e) INN** — the label differs from the item's same-language WHO INN (P2275).
- **(f) ICD format** — the value fails Wikidata's own P1793 regex. A dedicated subtype catches a repeated code part (B2424.).
- **(g) ATC** — the code fails the format regex, is a code replaced according to the WHO alterations list but still ranked as current, or the same substance code appears on several items.

Precision was estimated on a stratified random sample: one fixed-seed RNG per subtype, 103 flags from (a)–(g) plus 15 disease-label script flags. Verdicts were TP (real error), FP (correct or accepted variant) or DOUBT. Informational subtypes were not sampled: script-mixed labels, development codes and ICD-10 block ranges. Precision is $TP/(TP+FP)$ with a Wilson 95% interval.

For each language we estimate errors as $\Sigma (\text{flags} \times \text{detector precision})$ over (a), (b), (c) and (e). This is a lower bound, because recall was not measured.

Terms. The *earlier review* is EasyxLab's exploratory check of 2026-10-02 (scripts in `scripts/legacy/`). It looked at 59 active ingredients used in chronic treatment and 34 chronic diseases, and reported the seven errors re-checked in §4.5. The *chronic subset* is those 34 diseases, resolved by English label; 29 of them resolved on the snapshot date.

3.1 Automation and review

Every step of this study was carried out by an LLM-based agent (the "study agent"): extraction, detector design, sampling, verdicts, the selection of corrections and the writing. The agent read each sampled flag together with the item's other labels, sitelink titles and INNs, and assigned TP, FP or DOUBT. The reviewer was not blind to the detector that produced a flag. The precision figures are one automated reviewer's judgement; the sample is published (`data/validation_*.csv`) so that it can be re-reviewed, ideally by native speakers of each language. The rules applied to every verdict are written down in METHOD.md §4.

4. Results

4.1 Detector precision

DETECTOR	REVIEWED	TP	FP	DOUBT	PRECISION	95% CI
a script (drug labels)	14	10	4	0	0.71	0.45–0.88
a script (disease labels)	15	14	1	0	0.93	0.70–0.99
b brand (RxNorm)	12	10	2	0	0.83	0.55–0.95
c salt vs parent	20	5	14	1	0.26	0.12–0.49
d shared label	14	14	0	0	1.00	0.79–1.00
e INN discrepancy	16	3	13	0	0.19	0.07–0.43
f ICD-10 invalid (non-range)	13	13	0	0	1.00	0.77–1.00
f ICD-11 invalid	2	1	1	0	0.50	0.09–0.91
g obsolete ATC (complete change)	5	5	0	0	1.00	0.57–1.00
g obsolete ATC (split)	1	0	1	0	0.00	0.00–0.79
g ATC format	2	2	0	0	1.00	0.34–1.00
g same ATC on several items	4	0	4	0	0.00	0.00–0.49

The false positives are informative:

- **(a), 4 FP:** all four were vaccine code names (Ad5-nCoV, Gam-COVID-Vac) or international abbreviations (Tc-99m-MIBI). These are language-neutral designations.
- **(b), 2 FP:** both were "Adrenalin", which is a US brand but also the everyday German and Turkish word for epinephrine.

- **(c)** failed for two reasons:
 - **Lexicon matches inside stems:** «استات» inside the Persian *simvastatin*, and «戊酸» (valeric acid) inside 丙戊酸 (*valproic acid*).
 - **Defensible labels:** the item for a salt is often the only item for the drug, so naming the parent is acceptable.
- **(e)** mostly flags accepted synonyms: USAN or BAN forms (mechlorethamine/chlormethine, meclizine/meclozine), stereodescriptors ((RS)-baclofen) and common names (rutin/rutoside). Its three true positives are:
 - a missing accent in an official INN form (es *edoxabán*, fr *tiuxétan*);
 - a label naming the radionuclide instead of the medicine: es «Lutecio 177» for lutetium (¹⁷⁷Lu) oxodotreotide, whose English label is also misspelled («lutenium»);
 - an INN in the wrong script (ru «tozinameran»).
- **(g) duplicate codes** are not errors: brand and product items (Actrapid, Caregyl) legitimately share their substance's code.

4.2 Error rates by language (drug labels)

LANG	LABELS	COVERAGE	FLAGS A	FLAGS B	FLAGS C	FLAGS E	EST. ERRORS	PER 1,000 LABELS
ur	387	10%	20	0	3	0	15.1	38.9
fa	2,589	69%	109	3	40	0	90.8	35.1
ru	1,752	46%	56	3	19	5	48.4	27.6
hi	322	9%	10	0	3	0	7.9	24.6
ko	1,157	31%	18	0	9	0	15.2	13.2
uk	1,577	42%	18	0	16	0	17.1	10.8
en	3,759	100%	0	0	—	170	32.0	8.5*
zh	2,482	66%	9	0	41	6	18.3	7.4
bn	311	8%	2	0	1	0	1.7	5.4
ja	2,468	66%	7	0	28	0	12.4	5.0
ar	2,462	65%	4	0	20	11	10.2	4.1
pt	1,851	49%	0	1	22	0	6.6	3.6

LANG	LABELS	COVERAGE	FLAGS A	FLAGS B	FLAGS C	FLAGS E	EST. ERRORS	PER 1,000 LABELS
it	2,182	58%	0	2	21	0	7.2	3.3
fr	2,811	75%	0	4	16	9	9.2	3.3
tr	1,262	34%	0	2	8	0	3.8	3.0
es	2,275	60%	0	3	11	5	6.3	2.8
de	2,276	60%	1	2	11	0	5.3	2.3
pl	2,101	56%	1	0	15	0	4.7	2.2
sw	174	5%	0	0	0	0	0	0
am	9	0.2%	0	0	0	0	0	0

*The English figure comes almost entirely from detector (e), whose precision is low and uncertain, so it should not be compared with the script-driven rates of other languages.

Across all languages: ≈ 312 estimated errors in 34,207 labels (0.9%; ≈ 174 or 0.5% at the lower 95% CI bounds of precision), plus 168 labels in 62 shared-label groups (each sampled group contained at least one wrong label or duplicate items). 508 drug items have at least one flag from (a), (b), (c) or (e).

The script detector is the main driver in fa, ru, ur, hi, ko and uk. In Persian, 92 of the script flags come from a single pattern: ATC class items carrying English labels («ATC code R03»), which mirror English-titled pages on Persian Wikipedia. The same class items share the generic label «ATC» in Polish and «ATC代码» in Chinese, which accounts for most of the shared-label groups in pl and zh.

4.3 What the errors look like

The study agent's review of the sampled flags and of the 86 drug script flags whose item has a sitelink title in the same language that differs from the label (`data/flags.csv` : subtype `a_latin` , reference non-empty and `value`) turned up the following:

- **Brand names in place of the substance.**
 - Latin-script brands in non-Latin languages: fa «seroquel» (quetiapine), «Lyricea», «differin», «selexid»; ru «avodart», «naftin», «Arcoxia», «behicar» (olmesartan), «lobenox» (enoxaparin); zh «cialis tadalafil»; ko «macrofantina», a Spanish-market brand.
 - Brands in Latin-script languages: fr «dexedrine», «Truvada», «Coartem», «mycoster»; es «focalin»; it «angiomax»; de «CAPEX»; tr «esbriet».
 - Many of these brands are not in the US-centred RxNorm list. They were caught only because they are also in the wrong script.
- **A different substance.**

- fa «pregabalin» on flumequine;
- fa «pinaverium» on cimetropium bromide;
- fa «امپرازول» (omeprazole) on esomeprazole;
- fa and ar «doxepin» on cidozepin;
- **Unrelated personal name:** zh «阿米巴殺腸腔藥物» ("luminal amoebicide drug") on diloxanide furoate. fa «mohamad» on belimumab.
- **Salt on the parent.**
 - zh «双氯芬酸钠» (diclofenac sodium), «氯沙坦鉀» (losartan potassium), «鹽酸左西替利嗪» (levocetirizine hydrochloride);
 - fa «دانتروлін سدیم» (dantrolene sodium), «والپروات سدیم» (valproate sodium on valproic acid);
 - it «Bambuterolo cloridrato».
- **English or misspelled text.**
 - Misspelled: hi «valporic acid», «cetrazine»; fa «stanazol», «fintolimod»; ru «bromezepam».
 - English in Urdu: «Aspirin», «Urea», «Chloroform», «Isopropyl alcohol».
- **Wrong script in a Latin-script language:** Cyrillic «Азеластин» in Polish and «Децинон» in German.

Disease labels show the same script problem. Flags, and estimated errors after adjusting by the disease-label script precision (0.93):

LANG	FLAGS	EST. ERRORS	DISEASE LABELS	EST. RATE
hi	30	28.0	665	4.2%
ko	16	14.9	1,419	1.1%
ru	22	20.5	2,241	0.9%
bn	5	4.7	557	0.8%
fa	9	8.4	1,758	0.5%

The flagged disease labels are mostly English text, sometimes misspelled («portal hypertansion») or mistaken: hi «wbc» on leukocytosis, and a Bengali-script label on a Hindi slot.

4.4 Codes

ICD-10 (P494), 5,536 values:

- 294 values (5.3%) fail Wikidata's own format constraint.
- 156 of those are clean block ranges (I10-I15). Ranges are valid ICD notation and are reported only as constraint violations.
- The remaining 138 (2.5%) are invalid, with precision 13/13 on the sample:
 - 44 show the same corruption: the code part is repeated and often followed by a dot. 41 of them are on disease items, e.g. «B2424.» (HIV/AIDS), «I10-I1515.» (arterial hypertension), «D5656.»

(thalassemia), «C3333.-C3434.» (lung cancer), «A01.001.0» (typhoid) and «F95.295.2.» (Tourette syndrome).

- The others: ICD-10-CM or ICD-10-NA extension codes (G44.801), lists packed into one value («A04.8,A28.2»), ICD-9 procedure codes (3241) and free text.
- 160 template items also carry ICD codes.

In the 29-disease chronic subset, 3 diseases (10%) carry a malformed code:

DISEASE	VALUE
HIV/AIDS	«B2424.»
glaucoma	«H4040.-H4242.»
thalassemia	«D5656.»

Bipolar disorder's ICD-11 value «6A6» fails the format regex but is the stem of an ICD-11 block (6A60–6A6Z). It is not counted, which is the same rule we applied to the ICD-11 chapter number «17» (judged FP) and to ICD-10 block ranges.

ICD-11 (P7329): 15 of 8,193 values (0.2%) fail the format regex.

ATC (P267), 4,459 statements:

- 149 (3.3%) are codes that the WHO alterations list marks as moved. They are still ranked normal and carry no end date. 132 of the items lack the new code entirely, e.g. leflunomide L04AA13 (moved to L04AK01 in 2024), crizotinib, vismodegib and the 2021 re-organisation of L01.
- The detector correctly separates partial "split" changes, where the old code remains valid (e.g. methotrexate L01BA01).
- Two veterinary ATCvet codes (QB03AC91, QB03AC) sit in the human ATC property.

4.5 Errors from the earlier review, re-checked

REPORTED ERROR	STATUS ON 2026-10-02
warfarin fr «Coumaphène» (pesticide name; INN <i>warfarine</i>)	still present on Q113368879, rev. 2492907989. A second item for the same racemate exists (Q407431, «(RS)-warfarin»), and the French Wikipedia page linked to Q113368879 is titled «Coumaphène»
quetiapine fa «seroquel»	still present
valproate fa «valproato sódico»	partly fixed: now «والپروات سدیم» (Persian script, still the sodium salt)
valproate hi «valporic acid»	still present

REPORTED ERROR	STATUS ON 2026-10-02
aspirin ur in Latin script	still present («Aspirin»)
levothyroxine zh as sodium salt, traditional script	still present («左旋甲狀腺素鈉»)
ICD-10 «B2424.» , «I10-I1515.»	still present

5. Proposed corrections

`corrections.qs` holds a **proposal that has not been applied**: 177 QuickStatements command lines implementing 135 corrections, each reviewed one by one by the study agent:

- **90 labels**, each taken from the same item's sitelink title in that language or from its P2275 INN;
- **40 ICD-10 repairs**: 39 undoubled values and one missing dot. Q520127's packed value «K05.205.2,K05.305.3» becomes two separate statements, K05.2 and K05.3;
- **1 ICD-11 repair**: 6B6.Z → 6B6Z;
- **4 additions of the current ATC code**, citing the WHO alterations list.

`corrections.csv` gives each correction's justification and the `lastrevid` it was reviewed against. Two warnings apply:

- **References and qualifiers are lost.** Each ICD repair removes the original statement (`-Qxx P494 "old"`) and adds the repaired value, which drops the references and qualifiers of the original. The editor has to restore them by hand; this warning is also in the header of the `.qs` file.
- **Five repairs stay out of the batch.** Their repaired value is a block range (I10-I15, H40-H42, C33-C34, M33.0-M33.1, L20-L30), which would still violate the P494 format constraint. They are listed in `corrections.csv` as manual actions, because choosing between a range and individual codes is an editorial decision.

Six further actions are left to a Wikidata editor's judgement:

- **The warfarin French label.** The French INN form is *warfarine*, but neither the item's sitelink («Coumaphène») nor P2275 (empty) supports it, and it depends on the merge decision.
- **Two possible merges**: warfarin and aloxiprin.
- **Deprecating the replaced ATC codes.**
- **Removing the ATCvet values.**
- **Relabelling the generic «ATC» class items.**

Corrections without an independent source were left out (e.g. ru «orencia», for which no Russian sitelink or INN exists), and replaced brand names were not added as aliases.

No edit was made to Wikidata. `corrections.qs` is a proposal and has not been applied: re-check each `lastrevid` and discuss at WikiProject Medicine first. Whether to apply any of it, item by item, is for a

6. Limitations

1. **Recall is unknown.** The figures are lower bounds. Detector (b) only knows US brand names in Latin script.
2. **A single automated reviewer.** All verdicts and corrections were produced by the LLM-based agent that also wrote the detectors, without blinding. Strata are small (1–20) and the confidence intervals are wide. Labels in Urdu, Hindi, Bengali, Korean and Persian were judged against sitelink titles and general knowledge.
3. **The precision of (c) and (e) depends on the definition of "error".** We counted accepted synonyms and salt forms on de-facto drug items as correct. A stricter clinical definition (e.g., "the label must equal the INN") would raise both.
4. **The per-language estimate applies one precision per detector to all languages.** Script-error precision may differ between, say, Persian and Korean.
5. **The snapshot is a single day.** Wikidata changes continuously, and every correction must be re-checked against the item's current revision.
6. **"ICD-10" means the WHO edition**, as Wikidata's P494 constraint does. National extensions count as invalid by design.
7. **The detectors were tuned on unsampled flags before the sample was drawn.** The tuning covered the salt lexicon, the inorganic exclusion, the taxon and template split, and the doubled-code regex. They were frozen afterwards, but tuning on the same population can inflate precision.
8. **INN coverage is scarce.** P2275 exists for 1,271 items in English and for at most 35 in each of fr, es, ru, ar and zh, so detector (e) is almost entirely an English-label measure.
9. **Downstream impact is documented only for Wikipedia infoboxes and research tools.** We did not verify reuse by commercial apps or AI systems.

7. Data and code

- **Snapshot:** `data/frozen/` , CCO, with `lastrevid` per item.
- **Detector flags:** `data/flags.csv` .
- **Sample and verdicts:** `data/validation_*.csv` .
- **Metrics:** `data/metrics.json` .
- **Pipeline:** `scripts/run.sh` (stdlib Python).
- **Earlier exploratory scripts:** `scripts/legacy/` .

References

- Waagmeester A, et al. Wikidata as a knowledge graph for the life sciences. *eLife* 2020;9:e52614. doi:10.7554/eLife.52614

- Pfundner A, Schönberg T, Horn J, Boyce RD, Samwald M. Utilizing the Wikidata system to improve the quality of medical content in Wikipedia in diverse languages. *J Med Internet Res* 2015. doi:10.2196/jmir.4163
- Halle MW, Kikinis R, Neumann PE. TA2Viewer: a web-based browser for Terminologia Anatomica and online anatomical knowledge. *Clin Anat* 2024. doi:10.1002/ca.24162
- Wikidata:WikiProject Medicine. https://www.wikidata.org/wiki/Wikidata:WikiProject_Medicine (accessed 2026-10-02)
- WHO Collaborating Centre for Drug Statistics Methodology. ATC alterations (cumulative). https://atcddd.fhi.no/atc_ddd_alterations__cumulative/atc_alterations/ (accessed 2026-10-02)
- U.S. National Library of Medicine. RxNav Prescribable API. <https://rxnav.nlm.nih.gov/> (accessed 2026-10-02)

Cite this study

CITATION

EasyxLab (2026). Multilingual quality of drug names in Wikidata. Study S4.
EasyByte Hub S. Coop. Mad.
<https://github.com/easybytehub/easyxlab/tree/main/studies/s4-wikidata-drug-labels>

BIBTEX

```
@techreport{easyxlab_s4,  
  title      = {Multilingual quality of drug names in Wikidata},  
  author     = {{EasyxLab}},  
  institution = {EasyByte Hub S. Coop. Mad.},  
  number     = {S4},  
  year       = {2026},  
  url        = {https://github.com/easybytehub/easyxlab/tree/main/studies/s4-wikidata-drug-labels}  
}
```