Journal of Computational Science 101 (2026) 103017

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

When simpler models win: A large-scale computational benchmark of lexical and transformer NLP pipelines for predicting medication effectiveness

Doaa Al-Turkey, rasheed Mohammad Nassr *, Tatiana Grieshofer

Birmingham City University, United Kingdom

ARTICLE INFO

Keywords:
Patient-generated drug reviews
Drug-level effectiveness prediction
Natural language processing
TF-IDF
Support vector regression
ClinicalBERT
External validation

ABSTRACT

Patient-generated drug reviews provide a scalable source of real-world evidence on perceived medication effectiveness, yet robust and reproducible modelling benchmarks remain limited. This study presents a quantitative evaluation of seven natural language processing pipelines for predicting drug-level effectiveness ratings from patient reviews on a 1–10 scale. Reviews were aggregated at the medication level by concatenating texts per drug and constructing a stabilised drug-level target using Bayesian shrinkage ($\alpha = 10.0$). Models spanning sparse lexical baselines, dense embedding pipelines, frozen biomedical transformers, and a compute-aware fine-tuned transformer baseline (that is, one tuned under an explicitly bounded compute budget) were assessed using Repeated K-Fold cross-validation on Dataset A (5 folds × 2 repeats; out-of-fold prediction) and external validation on four independent datasets (A→B-E). Performance was measured using RMSE and MAE, with R^2, Pearson correlation, Spearman correlation, and calibration slopes, intercepts, and plots reported to characterise association and reliability under dataset shift. Across internal validation and most external evaluations, TF-IDF with Ridge regression achievedTF-IDF with linear support vector regression (SVR) achieved the strongest overall accuracy (Dataset A internal RMSE = 0.336; R^2 = 0.813) and generalised well, particularly on Dataset and generalised consistently, achieving the lowest error on every external corpus, including Dataset B (RMSE = 0.438), the only external dataset sharing no reviews with the training data. External results are reported after removing reviews duplicated from the training corpus, which affected 97.8 per cent of Dataset C, 98.9 per cent of Dataset D and 28.7 per cent of Dataset E. Transformer-based approaches, including SBERT and frozen biomedical encoders, did not outperform lexical baselines, and a transparently tuned ClinicalBERT regression baseline using a learning-rate sweep {1e-5, 2e-5, 3e-5} with early stopping did not yield consistent gains under external shift. The best lexical baseline achieved consistently lower RMSE than the best transformer pipeline in every one of the five evaluation settings, with a median RMSE gap of 0.19 rating points. A paired non-parametric test is reported alongside this descriptive pattern (Wilcoxon signed-rank, $p = 0.031$), but with only five paired settings this is the minimum attainable p-value and it is therefore treated as a caveat-bearing supplement rather than as confirmatory evidence. These findings indicate that, under drug-level aggregation and compute-aware tuning, frequency-based lexical signals capture much of the predictive structure in patient reviews. The benchmark provides a reproducible framework for evaluating drug-review-based effectiveness prediction with internal and external validation and reliability diagnostics relevant to pharmacovigilance and digital health analytics.

1. Introduction

Patient-generated drug reviews are free-text narratives accompanied by numerical ratings that individuals post on platforms such as Drugs.com, WebMD, and Ask a Patient. These reviews capture perceived effectiveness, adverse effects, and overall satisfaction from the patient perspective, often at a scale that exceeds what is available through conventional trials or structured pharmacovigilance reporting [1]. Because they reflect real-world use, including off-label contexts and heterogeneous patient populations, they can complement clinical evidence by highlighting patterns of perceived benefit and harm that may not be prominent in controlled settings. Prior work has also shown that

* Correspondence to: Birmingham City University, United Kingdom of Great Britain and Northern Ireland, Birmingham, United Kingdom.
E-mail addresses: doaa.al-turkey2@mail.bcu.ac.uk (D. Al-Turkey), rasheed.mohammad@bcu.ac.uk (M. Nassr), Tatiana.Grieshofer@bcu.ac.uk (T. Grieshofer).

https://doi.org/10.1016/j.jocs.2026.103017
Received 17 November 2025; Received in revised form 28 July 2026; Accepted 24 August 2026
Available online 27 August 2026
1877-7503/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Table 1
Summary of the five patient-review datasets used in the benchmark.

Dataset	Source platform	Role in study	Approx. scale / coverage	Provenance / access
A	Drugs.com	Training + internal validation	215,063 reviews; 3671 drugs; 916 conditions; 3671 drug-level observations from 215,063 reviews	Public Drugs.com review corpus (UCI/ Kaggle distribution)
B	DrugLib.com	External validation	Independent drug-review collection with 1–10 ratings; 541 drug-level observations from 4143 reviews; 0.0% removed as duplicates of Dataset A	Public DrugLib.com review corpus
C	HealthyData. science	External validation	Real-world drug-product user reviews with 1–10 ratings; 444 drug-level observations from 702 reviews; 97.8% removed as duplicates of Dataset A	Public repository (healthydata.science)
D	Hugging Face	External validation	Publicly redistributed drug-review dataset with 1–10 ratings; 610 drug-level observations from 1950 reviews; 98.9% removed as duplicates of Dataset A	huggingface.co/datasets/forwins/Drug-Review-Dataset
E (DrugRev)	DrugRev / Mendeley Data	External validation	Customer reviews on drug purchasing and satisfaction with 1–10 ratings; 3213 drug-level observations from 279,785 reviews; 28.7% removed as duplicates of Dataset A; general satisfaction rating (see Section 3.1)	Ahmed et al. [26], doi. org/10.17632/64cc5w5dxy.1

Note. All datasets are publicly available and de-identified in their released form. Counts for Datasets B-E reflect the publicly distributed versions; exact totals vary by release. Drug-level observations denote the number of unique medications retained after aggregation and cleaning, which is the sample size on which the corresponding column of Table 2 is computed. The study constitutes secondary analysis of public text data.

aggregated user ratings can align with clinically plausible signals, including lower ratings for drugs associated with higher burdens of adverse effects [2]. This supports the view that patient-reported outcomes in online reviews may serve as a useful secondary source of evidence when analysed carefully and validated appropriately. Early computational studies commonly framed the analysis of patient drug reviews as a sentiment classification problem or as a text-to-rating prediction task. Traditional approaches relied on feature engineering using bag-of-words or term frequency–inverse document frequency (TF-IDF) representations, combined with supervised learning methods such as support vector machines, logistic regression, naïve Bayes, and random forests. Gräßer et al. [3], for example, used the DrugLib corpus to predict rating-related aspects such as effectiveness and side-effect severity. Their results highlighted that performance could degrade under domain shift across datasets and conditions, which underscored the importance of external validation and generalisation beyond a single corpus.

With larger datasets and the adoption of neural sequence models, later work explored convolutional and recurrent architectures, including bidirectional long short-term memory networks. Colón-Ruiz and Segura-Bedmar [4] benchmarked CNN and bi-LSTM models on Drugs.com reviews and reported improvements over classical baselines in some settings, while also noting the importance of task design and evaluation choices. More recently, transformer models such as BERT and domain-adapted variants have been applied to patient review data. Shiju and He [5] reported strong results from fine-tuned BERT variants on rating prediction tasks. Sorayaie Azar et al. [6] proposed an ensemble approach and incorporated explainability methods, demonstrating the potential to scale models across large review corpora. A parallel and rapidly growing body of work on user-generated text classification has further shown that the advantage of large language models and transformer ensembles is highly contingent on customisation, task formulation, and evaluation design rather than being guaranteed by architecture alone. Shatnawi et al. [7] demonstrated that the strong performance of tailored large language models on spam detection depended on systematic model customisation and careful benchmarking across diverse datasets, while Al-Shawakfa et al. [8] showed that transformer-based ensembles for radicalisation detection in Arabic social media achieved their gains only through deliberate model diversity, tuning depth, and task-specific design. These studies collectively indicate that stronger representation learning and task adaptation can improve performance, although conclusions vary depending on dataset composition, label formulation, and evaluation design. Despite progress, several gaps remain. First, many studies evaluate models primarily at the individual review level, while practical decision support often requires drug-level effectiveness estimates. The choice between review-level and drug-level modelling changes both the learning problem and the types of signals that dominate predictive performance. Second, cross-corpus validation is not consistently included, which limits conclusions about robustness under dataset shift. Third, transformer performance is sometimes reported without transparent tuning budgets, which complicates fairness in comparisons against strong lexical baselines. Finally, interpretability and reliability diagnostics, such as calibration behaviour under shift, are often treated as secondary.

This study addresses these issues by benchmarking seven natural language processing pipelines for drug-level effectiveness prediction from patient-generated reviews. Reviews are aggregated at the medication level to construct a single document and a single effectiveness target per drug. The benchmark spans strong lexical baselines (TF-IDF with linear support vector regression, hereafter SVR, and TF-IDF with Ridge regression), embedding pipelines using SBERT, frozen biomedical transformer encoders (BioLinkBERT and BioGPT), and a compute-aware fine-tuned transformer baseline (ClinicalBERT) with a transparent learning-rate sweep and early stopping. Models are evaluated using repeated K-Fold cross-validation on Dataset A with out-of-fold prediction and external validation on independent datasets. In addition to predictive error metrics, the analysis reports association measures and tabulated calibration diagnostics to support interpretation under distribution shift. The goal is not to claim universal superiority for a model family, but to provide a reproducible and fair comparison that clarifies which approaches perform reliably for drug-level effectiveness prediction under realistic constraints.

2. Literature review

2.1. Patient-generated reviews for medication effectiveness modelling

Patient-generated drug reviews provide an accessible source of real-world evidence describing perceived efficacy, tolerability, and satisfaction. Platforms such as Drugs.com and DrugLib.com include large numbers of reviews that describe outcomes outside controlled settings, which can complement clinical evidence when interpreted with appropriate caution [9]. A wide range of machine learning and natural language processing methods has been applied to patient review text for tasks such as adverse drug event detection, adherence modelling, and satisfaction prediction [10-15]. Across many such studies, sparse lexical features remain strong baselines because they are computationally efficient and relatively transparent, particularly when paired with linear predictors such as Ridge regression or support vector machines [14,16]. Transformer-based models have also been applied, including ClinicalBERT and other biomedical variants [17]. Recent benchmarking studies of user-generated text reinforce that transformer and large language model performance is conditional rather than automatic. Shatnawi et al. [7] reported that tailored large language models reached near-ideal accuracy on spam detection only after explicit model customisation and benchmarking across many datasets, and Al-Shawakfa

et al. [8] found that transformer ensembles for Arabic radicalisation detection depended on model diversity and tuning depth to outperform simpler baselines. However, performance is sensitive to task formulation, tuning regime, and the degree to which the predictive signal is tied to explicit lexical cues.

Evaluation practice in this area increasingly emphasises external validation across independent datasets, rather than relying exclusively on internal cross-validation [18,11,19]. Reliability assessment is also gaining attention, particularly where deployment is considered. Calibration plots and slope-intercept diagnostics are commonly used to assess whether predicted ratings maintain appropriate scale alignment under dataset shift [11,18]. These trends motivate benchmarks that report not only accuracy but also transfer behaviour and reliability.

2.2. Datasets for patient-generated review analysis

Several public datasets support research on patient review modelling. Commonly used corpora include review collections derived from Drugs.com and related platforms, as well as repackaged versions distributed through platforms such as Kaggle and Hugging Face. These datasets typically include drug names, free-text reviews, numeric ratings on a 1–10 scale, and additional fields such as condition labels or helpfulness votes. While there is substantial overlap in the general structure of these resources, datasets often differ in rating conventions, topic distributions, and review style. These differences can induce dataset shift that affects generalisation and calibration, which makes external validation essential for robust conclusions.

2.3. Typical representations and feature design

Most predictive approaches rely on representations that capture either lexical frequency patterns or contextual semantics. TF-IDF n-gram features remain widely used because they retain clinically meaningful surface forms and evaluative phrases that frequently occur in patient reviews [5,14]. Some studies incorporate structured features such as review length or helpfulness indicators, although the usefulness of such features depends on availability and cross-platform consistency. More recent work uses contextual embeddings from transformer encoders, including SBERT and biomedical variants, sometimes reduced using principal component analysis to mitigate collinearity and reduce dimensionality. These representations can capture semantic similarity and paraphrase structure, but their predictive benefit for rating estimation is not guaranteed without task adaptation.

2.4. Evaluation metrics and validation practices

Medication effectiveness prediction is often treated as either classification over rating bins or regression on the 1–10 scale. RMSE is widely reported in regression settings because it penalises large deviations that may be meaningful in practice [20,21]. MAE is also common due to its interpretability. R^2 is often reported to quantify explained variance, and correlation measures can be used to assess rank consistency. Recent work emphasises the importance of evaluation designs that reflect deployment conditions, including external validation and reliability checks such as calibration assessment, particularly when training and test sets differ in composition [18].

2.5. Algorithms and comparative findings

The methodological spectrum includes classical linear models, neural networks, and transformer-based pipelines. Classical models using TF-IDF features paired with Ridge regression or linear support vector regression remain strong baselines and can be competitive under realistic compute constraints Das et al. [22]. Neural models such as CNNs and bi-LSTMs can improve performance in some settings by modelling local and sequential structure, especially when large, labelled datasets are available [23]. Transformer-based models can outperform classical methods when they are fine-tuned appropriately for the task and when the dataset supports the required adaptation. Consistent with this, the spam-detection benchmark of Shatnawi and Alsobeh (2025) and the radicalisation-detection ensemble of Al-Shawakfa et al. [8] both show that such advantages emerge primarily under deliberate customisation, sufficient tuning depth, and evaluation designs aligned with the target task. However, frozen transformer embeddings do not always preserve the explicit lexical cues that drive rating prediction in short, evaluative text, and performance often depends on the tuning strategy and the level at which the prediction target is defined.

2.6. Gaps motivating this study

Several gaps motivate the present benchmark. First, studies often report results on a single corpus or rely mainly on internal validation, leaving cross-corpus robustness under-examined [24,25]. Second, the distinction between review-level modelling and drug-level modelling is not consistently addressed, even though the two settings represent different objectives and may favour different representations. Third, transformer comparisons are not always supported by transparent tuning budgets, which complicates fairness when contrasting them with strong lexical baselines. Fourth, reliability reporting, including calibration behaviour under dataset shift, is not consistently integrated into comparative evaluations [18].

This study responds to these gaps by providing a reproducible benchmark for drug-level effectiveness prediction with both internal and external validation. It compares strong lexical baselines with frozen transformer encoders and a transparently tuned ClinicalBERT baseline a compute-aware learning-rate sweep. It also reports calibration slopes and intercepts to support interpretation of reliability under dataset shift. The aim is to clarify which modelling choices are most reliable for drug-level prediction from patient reviews under realistic constraints, rather than to claim universal superiority for a single model family.

3. Methodology

3.1. Datasets

This study employed five complementary datasets of patient drug reviews to evaluate predictive models of medication effectiveness. Dataset A was derived from Drugs.com and served as the primary training and internal-validation resource because of its scale and broad medication coverage; it comprises 215,063 patient reviews spanning 3671 unique drugs and 916 unique conditions, with each review containing a drug name, an associated condition, free-text review content, a numeric rating on a 1–10 scale, a submission date, and a usefulness-vote count. Dataset B was obtained from DrugLib.com and used exclusively for external validation. Dataset C was obtained from the public HealthyData.science repository, and Dataset D was obtained from a publicly distributed corpus hosted on Hugging Face. Dataset E corresponds to the DrugRev corpus released on Mendeley Data [26] and was likewise used only for external validation. The DrugRev release is documented by its authors as a collection of customer reviews on drug purchasing and satisfaction, so its numeric rating is best understood as a general satisfaction rating that may combine perceived medical effectiveness with aspects of the purchasing experience rather than as a rating of medical effectiveness alone.Dataset E corresponds to the DrugRev corpus released on Mendeley Data [26] and was likewise used only for external validation. The DrugRev release is documented by its authors as a collection of customer reviews on drug purchasing and satisfaction, and its numeric rating is therefore best understood as a general satisfaction rating that may combine perceived medical effectiveness with aspects of the purchasing and supply experience, rather than as a rating of medical effectiveness alone. We retain Dataset E because this difference in the rating construct, rather than a difference in

vocabulary or drug coverage only, represents the most demanding form of transfer examined here; we treat it explicitly as a construct-shift stress test and interpret its results accordingly (Sections 4.2.4, 5.3 and 5.7). All five datasets share the same core fields used in this study, namely free-text reviews, drug identifiers, and numeric effectiveness ratings on a 1–10 scale; they differ in platform of origin, review style, drug and condition coverage, and rating conventions, which is precisely the source of distribution shift that the external-validation design is intended to probe. Table 1 summarises the provenance, role, scale, and source of each dataset. After the external corpora were assembled we tested directly for overlap with the training corpus by comparing normalised review texts, and the result materially qualifies the description of three of them as independent. Dataset B shares no reviews with Dataset A. Dataset C and Dataset D, by contrast, are almost entirely redistributions of the same Drugs.com material: 97.8 per cent and 98.9 per cent of their reviews respectively occur verbatim in Dataset A. Dataset E shares 28.7 per cent. We report this because it was not identified when the benchmark was first assembled, and because evaluating a model on reviews it was trained on does not measure generalisation. All external results reported in Section 4 are therefore computed after removing from Datasets B to E every review whose text also appears in Dataset A, so that no evaluation review was seen during training; the drug-level counts in Table 1 are the de-duplicated counts. Dataset B required no filtering and is consequently the only fully independent external corpus in the benchmark. Datasets C and D retain 702 and 1950 reviews respectively after filtering, so their drug-level targets rest on very few reviews per medication and their metrics should be read as indicative only. Dataset E retains 279,785 reviews and remains a substantial external test. A residual dependence persists in all four, in that most of their medications also appear in Dataset A; this is inherent to drug-level prediction, where the intended use case is to estimate a rating for a known medication from previously unseen reviews, but it should be borne in mind when the external metrics are interpreted. Because the unit of analysis is the medication rather than the review, the sample size that governs each evaluation is the number of drug-level observations remaining after aggregation, not the number of raw reviews. These counts are reported for all five datasets in Table 1: Dataset A contributes 3671 drug-level observations, and the external evaluations are based on 541 observations for Dataset B, 444 for Dataset C, 610 for Dataset D, and 3213 for Dataset E, after the de-duplication described below. While Dataset A provided the scale required for training, the remaining datasets offered independent distributions of reviews, conditions, and drug frequencies, enabling a rigorous test of out-of-domain generalisation.

3.2. Research design

This study uses a quantitative experimental design to benchmark natural language processing pipelines for predicting medication effectiveness from patient-generated drug reviews. The task is framed as regression on a 1–10 rating scale. The unit of analysis is the medication, not the individual review. Reviews are aggregated to drug-level outcomes to support the practical use case of producing a drug-level effectiveness estimate that can be compared across medications. The study proceeds in four stages. First, raw review data are cleaned and transformed into drug-level samples with one target value per medication. Second, a set of modelling pipelines are implemented that span sparse lexical approaches and transformer-based approaches. Third, models are evaluated using internal repeated cross-validation on Dataset A and external validation on independent datasets. Fourth, models are compared using accuracy, association, and calibration diagnostics. Fig. 1 illustrates the end-to-end workflow, from drug-level aggregation through modelling and evaluation.

3.3. Data preparation

All datasets contain free-text drug reviews and a numeric rating on a

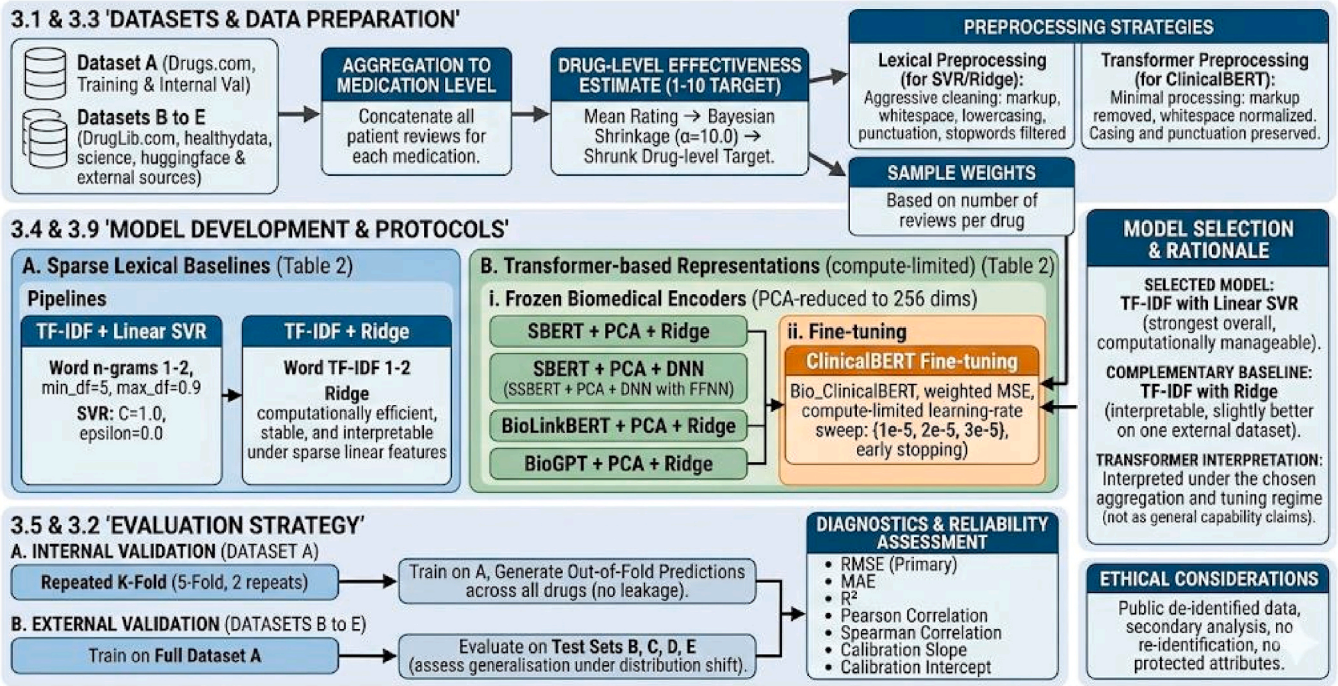


Fig. 1. Detailed end-to-end workflow for medication effectiveness prediction. The diagram organises the methodology into four phases: (1) data preparation and preprocessing, including aggregation of raw reviews into drug-level targets (1–10 regression), target stabilisation via Bayesian shrinkage ($\alpha = 10.0$), and parallel preprocessing for lexical and transformer models; (2) model development, covering seven pipelines across sparse lexical baselines (TF-IDF with linear SVR/Ridge) and transformer-based representations (frozen encoders BioLinkBERT, SBERT, BioGPT and fine-tuned ClinicalBERT); (3) evaluation strategy, contrasting internal repeated 5-fold cross-validation on Dataset A with external validation on four independent datasets (B-E); and (4) results and rationale, including diagnostic metrics and the selection of TF-IDF with linear SVR as the deployment model. Note: 'val' = validation.

Table 2
Drug-level predictive performance (1–10 rating scale). Dataset A (internal OOF) = out-of-fold predictions from Repeated K-Fold cross-validation (5 folds × 2 repeats). Datasets B-E = external validation trained on Dataset A and tested on each dataset.

Model / Dataset	RMSE	MAE	R^2	Pearson r	Spearman ρ
Dataset A (internal OOF) n = 3671 drugs					
TF-IDF + SVR	0.3729	0.2517	0.7694	0.8859	0.8564
TF-IDF + Ridge	0.3358	0.2442	0.8129	0.9029	0.8540
SBERT + PCA + Ridge	0.9485	0.7527	-0.4923	0.3799	0.3758
SBERT + PCA + DNN	0.7764	0.5904	0.0000	0.0934	0.0612
BioLinkBERT (frozen) + PCA + Ridge	1.1504	0.8602	-1.1953	0.3251	0.3838
BioGPT (frozen) + PCA + Ridge	1.1035	0.8594	-1.0199	0.2634	0.2938
ClinicalBERT (fine-tuned) regression	0.7043	0.5081	0.1773	0.4451	0.4463
Dataset B (external A→B) n = 541 drugs					
TF-IDF + SVR	0.4356	0.3320	0.3079	0.5635	0.4911
TF-IDF + Ridge	0.4384	0.3382	0.2990	0.5665	0.4848
SBERT + PCA + Ridge	0.8501	0.6831	-1.6357	0.2134	0.1889
SBERT + PCA + DNN	0.5259	0.3897	-0.0087	-0.0744	-0.0706
BioLinkBERT (frozen) + PCA + Ridge	1.0365	0.8000	-2.9177	0.1286	0.1467
BioGPT (frozen) + PCA + Ridge	1.1511	0.9026	-3.8325	0.1552	0.1607
ClinicalBERT (fine-tuned) regression	0.6247	0.4834	-0.4232	0.2585	0.2255
Dataset C (external A→C) n = 444 drugs					
TF-IDF + SVR	0.3022	0.2311	0.2376	0.6783	0.6099
TF-IDF + Ridge	0.3304	0.2563	0.0887	0.6843	0.6159
SBERT + PCA + Ridge	1.2935	1.0438	-12.9682	0.4406	0.3687
SBERT + PCA + DNN	0.3777	0.3343	-0.1909	-0.2005	-0.1666
BioLinkBERT (frozen) + PCA + Ridge	2.7937	2.2204	-64.1537	0.2433	0.2493
BioGPT (frozen) + PCA + Ridge	1.7925	1.3817	-25.8233	0.1770	0.1297
ClinicalBERT (fine-tuned) regression	0.3472	0.2641	-0.0061	0.4908	0.4448
Dataset D (external A→D) n = 610 drugs					
TF-IDF + SVR	0.3720	0.2870	0.5506	0.7515	0.6998
TF-IDF + Ridge	0.3660	0.2844	0.5649	0.7795	0.7245
SBERT + PCA + Ridge	0.8581	0.6772	-1.3910	0.4904	0.5089
SBERT + PCA + DNN	0.5593	0.4051	-0.0159	0.1700	0.0989
BioLinkBERT (frozen) + PCA + Ridge	0.9956	0.7656	-2.2188	0.3857	0.4360
BioGPT (frozen) + PCA + Ridge	0.9796	0.7847	-2.1164	0.3368	0.3583
ClinicalBERT (fine-tuned) regression	0.5893	0.4605	-0.1278	0.5194	0.4888
Dataset E (external A→E) n = 3213 drugs					
TF-IDF + SVR	0.7281	0.5316	0.5965	0.8612	0.8376
TF-IDF + Ridge	0.6765	0.5039	0.6517	0.8715	0.8395
SBERT + PCA + Ridge	1.2020	0.9257	-0.0995	0.3837	0.3871
SBERT + PCA + DNN	1.2654	0.9155	-0.2186	0.0045	-0.0118
BioLinkBERT (frozen) + PCA + Ridge	1.2714	0.9647	-0.2303	0.3804	0.4316
BioGPT (frozen) + PCA + Ridge	1.3778	1.0529	-0.4447	0.3013	0.3259
ClinicalBERT (fine-tuned) regression	1.1304	0.8226	0.0275	0.5148	0.5202

Notes: (a) Design. The unit of analysis, the shrinkage-based construction of the drug-level target, and the out-of-fold internal protocol are specified in Sections 3.2, 3.3 and 3.5 and are not restated here. (b) Sample size. Drug-level observation counts for each dataset are given in Table 1. (c) Preprocessing policy. Lexical models used stronger normalisation; transformer inputs were minimally cleaned (markup removal, whitespace normalisation) while retaining casing and punctuation and were truncated to the first 256 wordpieces (Section 3.4.1). (d) Model naming. TF-IDF = term frequency-inverse document frequency; SVR = linear support vector regression; ClinicalBERT (fine-tuned) = end-to-end fine-tuning of Bio_ClinicalBERT. (e) ClinicalBERT transparency. The learning-rate sweep used {1e-5, 2e-5, 3e-5}; the selected learning rate was 2e-5 and the best validation RMSE during the sweep was 0.7136. Fine-tuning used a maximum of 6 epochs with early stopping (patience = 2), batch size = 8, and weight decay = 0.01. (f) Sample weighting. Weights proportional to the number of reviews per drug were applied to TF-IDF + Ridge, to the Ridge stage of the three embedding pipelines and to ClinicalBERT fine-tuning; TF-IDF + linear SVR was fitted unweighted (Section 3.3). (g) Rounding. Table values are reported to four decimal places; values quoted in the Abstract and in the running text are rounded to three decimal places. (h) Calibration. Calibration slopes and intercepts for every pipeline and dataset tabulated in Table 3.

1–10 scale. Dataset A is used for model development and internal validation. Datasets B, C, D, and E are used as external test sets to assess generalisation under distribution shift. Reviews are aggregated at the medication level. For each drug, all reviews are concatenated to form a single document. The raw drug-level rating is computed as the mean of the available review ratings for that drug. To reduce instability for drugs with few reviews, a Bayesian (empirical-Bayes) shrinkage estimator is applied Table 2.

Formally, for a drug d with n^d available review ratings whose sample mean is $\bar{y}^d$, the stabilised drug-level target θ^d is computed by shrinking the per-drug mean toward a global prior mean μ_0 (the overall mean rating across all drug-level samples in the training data):

$$\theta^d = (n^d \cdot \bar{y}^d + \alpha \cdot \mu_0) / (n^d + \alpha),$$

where $\alpha > 0$ is the prior strength. The prior mean μ_0 is the global mean effectiveness rating, so that drugs with few reviews are pulled toward the overall average and drugs with many reviews retain their empirical mean. The effective weight placed on the per-drug data is $n^d / (n^d + \alpha)$ and the weight on the prior is $\alpha / (n^d + \alpha)$. With the chosen value $\alpha = 10.0$, a drug supported by a single review receives only about 9% of its weight from its own rating, a drug with 10 reviews is balanced equally between its own mean and the prior, and a drug with 100 reviews relies on its own mean for roughly 91% of the estimate. The resulting target remains on the original 1–10 scale. The value $\alpha = 10.0$ was fixed a priori and recorded in the reproducibility notes; a sensitivity analysis of this choice is reported in Section 4.5.

Two preprocessing strategies are used to avoid biasing one model family. For lexical models, text is normalised more aggressively. Mark-up is removed, whitespace is normalised, and tokens are lowercased. Punctuation removal and stopword filtering are applied because sparse n-gram models benefit from standard lexical normalisation. For transformer models, text is minimally processed. Mark-up is removed and whitespace is normalised, while punctuation and casing are preserved. This choice aligns the input distribution with the pretraining conditions of BERT-style models and reduces the risk of suppressing clinically

relevant surface cues. During model fitting, sample weights are defined using the number of reviews contributing to each drug-level sample. This gives more influence to medications supported by more evidence and reduces sensitivity to extreme means estimated from very small numbers of reviews. These weights are not applied uniformly across the seven pipelines, and the asymmetry is stated here in full. Weights enter the weighted least-squares objective of TF-IDF + Ridge and of the Ridge stage of the SBERT, BioLinkBERT and BioGPT pipelines, and they enter the weighted mean squared error objective used for ClinicalBERT fine-tuning. The TF-IDF + linear SVR baseline is fitted without sample weights. The unsupervised stages that admit no target-dependent weighting, namely the TF-IDF vectorisation and the PCA basis, are also fitted without weights. We draw attention to this because it bears directly on the fairness of the comparison, and because its direction is the opposite of a confound favouring the reported conclusion: the pipeline that achieves the lowest error in four of the five evaluation settings is the one pipeline that does not receive evidence weighting, while every pipeline it outperforms does. Weighting therefore cannot account for the lexical advantage reported in Section 4.

3.4. Model development

Seven modelling pipelines are implemented. They are chosen to represent both strong sparse lexical baselines and transformer-based representations under a compute-aware design. The pipelines are summarised below using consistent naming that matches Table 2.

TF-IDF + Linear SVR. The primary lexical baseline uses TF-IDF word n-grams (1−2). Sublinear term frequency scaling is applied. Terms are filtered using minimum document frequency and maximum document frequency thresholds. A linear support vector regression model is trained on the TF-IDF features. This model is included because it is a strong linear baseline for high-dimensional sparse text features and does not require deep representation learning.

TF-IDF + Ridge. A second lexical baseline uses the same TF-IDF representation and fits Ridge regression. Ridge is included because it is computationally efficient, stable, and interpretable under sparse linear features. The Ridge baseline also provides a clear reference point for the transformer comparisons.

SBERT + PCA + Ridge. Dense sentence embeddings are produced using Sentence-BERT with the all-mpnet-base-v2 encoder. Embeddings are reduced using principal component analysis to 256 dimensions to control collinearity and improve conditioning. The number of retained components was fixed a priori at 256 for every embedding pipeline and was not selected by cross-validation. This value is unrelated to the 256-wordpiece window length used by the transformer encoders (Section 3.4.1); the coincidence of the two numbers carries no methodological significance. Two considerations motivated the choice. First, 256 components keep the design matrix well conditioned relative to the number of drug-level training observations, so that the downstream Ridge fit is not dominated by directions estimated from very few effective degrees of freedom. Second, fixing one value across all embedding pipelines prevents any single representation from gaining an advantage through per-model dimensionality tuning, which would reintroduce the very asymmetry that the compute-aware protocol is designed to exclude. Because the component count was fixed rather than tuned, we make no claim that 256 is optimal for any individual encoder, and a systematic sensitivity analysis over the retained dimensionality is identified as future work in Section 5.7. Ridge regression is then trained on the reduced embeddings.

SBERT + PCA + DNN. To allow non-linear mapping in the SBERT feature space, a shallow feed-forward neural regressor is trained on the PCA-reduced SBERT embeddings. The network uses two hidden layers and dropout. Optimisation uses a robust regression loss and early stopping on a held-out validation split drawn from the training fold.

BioLinkBERT (frozen) + PCA + Ridge. BioLinkBERT-large is used as a frozen biomedical encoder. Token embeddings are mean pooled using the attention mask. The pooled vectors are reduced to 256 dimensions using PCA. Ridge regression is trained on the reduced vectors. The encoder weights remain fixed to preserve the compute budget and to provide a frozen biomedical baseline.

BioGPT (frozen) + PCA + Ridge. BioGPT is used under the same frozen encoding strategy as BioLinkBERT. Mean-pooled token embeddings are reduced using PCA and passed to Ridge regression.

ClinicalBERT fine-tuning for regression. ClinicalBERT (Bio_ClinicalBERT) is fine-tuned end-to-end for regression using a weighted mean squared error objective, where weights are proportional to the number of reviews per drug. A compute-aware learning-rate sweep is conducted over {1e-5, 2e-5, 3e-5}. The learning rate is selected using validation RMSE with early stopping. The selected learning rate and validation performance are reported in the Table 2 footnotes to support transparency and fairness in the transformer comparison.

3.4.1. Equalising input information across model families

Because TF-IDF operates on the full aggregated drug-level document while a single transformer forward pass is bounded by a fixed maximum sequence length, a naive comparison could give the lexical models an information advantage by exposing them to more of each drug's text. This asymmetry is not removed in the present benchmark, and we state its extent explicitly. The transformer inputs are formed by truncating the concatenated drug-level document to the first 256 wordpieces: each document is tokenised with truncation enabled at a maximum sequence length of 256, so a single forward pass covers the opening fragment of the document and the remainder is discarded. No sliding-window aggregation over the full document is performed. The consequence is quantifiable on Dataset A, where the median drug-level document is 487 wordpieces and the longest is 708,060: truncation at 256 wordpieces therefore discards some text for 60.9 per cent of drugs, while TF-IDF retains exact global term frequencies over the entire document. The lexical and transformer pipelines consequently do not see the same quantity of review text, and the comparison in Section 4 should be read as a comparison under this constraint rather than as a comparison at equal input coverage. We regard this as the principal internal limitation of the study. It admits a direct remedy, namely encoding successive non-overlapping windows and pooling them into a single drug-level representation, and we identify that as the first priority for further work in Section 5.7. Because the effect of truncation is to withhold evidence from the transformer pipelines rather than from the lexical ones, it cannot be invoked to explain away the lexical advantage in the opposite direction; but it does mean the transformer results reported here are a lower bound on what those architectures could achieve under full-document encoding. Accordingly we therefore frame the central conclusion conditionally on the drug-level aggregation and compute-aware tuning regime, and we return to this point in the limitations (Section 5.7).

3.5. Evaluation strategy

Internal validation on Dataset A. Internal model performance is estimated using repeated K-Fold cross-validation with five folds and two repeats. This produces ten train-test splits. For each split, models are trained on the training folds and evaluated on the held-out fold. Predictions are aggregated into out-of-fold predictions across all drugs. This yields a single internal estimate per model without training-set leakage.

External validation on Datasets B to E. External validation follows a train-on-A, test-on-B to E design. Each model is trained using the full Dataset A drug-level set, then evaluated on each external dataset. This tests robustness to dataset shift in review style, drug coverage, and rating distributions.

Metrics and reliability checks. Root mean squared error is the primary metric because it penalises large errors on the 1–10 scale. Mean absolute error and R^2 are reported to complement RMSE. Pearson correlation and Spearman correlation are included to report association and

Table 3
Calibration slope and intercept by pipeline and dataset.

Model / Dataset	Dataset A (internal OOF)	Dataset B	Dataset C	Dataset D	Dataset E
TF-IDF + SVR	1.1601; −1.1966	0.8568; 1.0560	0.5898; 3.0685	0.8650; 0.9697	1.4668; −3.6192
TF-IDF + Ridge	1.0572; −0.4213	0.8118; 1.4009	0.5321; 3.4772	0.7908; 1.5250	1.3571; −2.7940
SBERT + PCA + Ridge	0.3259; 4.9941	0.1439; 6.2608	0.1198; 6.5180	0.2815; 5.2575	0.4980; 3.2357
SBERT + PCA + DNN	3.1364; −15.5508	-12.9138; 102.0316	-17.8581; 138.4427	27.6551; −195.6246	1.6364; −5.2045
BioLinkBERT (frozen) + PCA + Ridge	0.2221; 5.7347	0.0711; 6.7841	0.0301; 7.2402	0.2008; 5.8152	0.4227; 3.7683
BioGPT (frozen) + PCA + Ridge	0.2037; 5.8849	0.0738; 6.7617	0.0358; 7.1924	0.1856; 5.9361	0.3473; 4.2544
ClinicalBERT (fine-tuned) regression	0.7986; 1.4255	0.2702; 5.3136	0.5061; 3.6728	0.4557; 3.9243	0.9758; −0.3805

Note. Each cell reports the calibration slope followed by the intercept, estimated by regressing observed drug-level targets on predicted values (Section 3.5). Perfect calibration corresponds to a slope of 1.00 and an intercept of 0.00. Values are reported to four decimal places, consistent with Table 2.

rank consistency. Calibration slope and intercept are estimated using a linear fit of observed outcomes against predictions. The fitted slopes and intercepts are reported for every pipeline and every dataset in Table 3, in place of per-dataset calibration plots, so that reliability can be compared across datasets on a common numeric scale. These values support a reliability assessment, especially under external shift.

In addition, model families are compared statistically using a paired non-parametric test across datasets, as described in Section 4.6, with effect sizes reported alongside p-values.

3.6. Model selection rationale

Model selection is based on predictive performance, reliability, and practicality under the defined task and compute setting. Sparse lexical models are treated as essential baselines because they capture frequency-driven signals that are often strong in patient reviews. Transformer models are evaluated under two settings. The first is frozen encoders with simple regressors. The second is a compute-aware fine-tuned ClinicalBERT baseline with a transparent learning-rate sweep and early stopping. The primary model selected for deployment and discussion is TF-IDF with linear SVR because it provides the strongest overall performance in the final benchmark across the internal and most external evaluations, while remaining computationally manageable. TF-IDF with Ridge is retained as a complementary baseline because it is more interpretable and slightly better on one external dataset. Transformer results are interpreted in terms of performance under the chosen aggregation and tuning regime, rather than as general claims about transformer capability.

3.7. Ethical considerations

All datasets used in this study are publicly available and contain user-generated reviews that are de-identified in the released sources. The study does not attempt re-identification. No protected personal attributes are inferred. The work therefore constitutes secondary analysis of public text data.

3.8. Rationale for the portfolio of predictive approaches

Patient-generated drug reviews contain informal spelling, variable length, and a mix of symptom description and evaluative language. TF-IDF n-grams provide a strong and transparent representation for this type of text because they preserve clinically meaningful surface forms and sentiment-bearing phrases. Linear SVR and Ridge offer strong baselines for sparse high-dimensional representations. Transformer-based representations are included because they model contextual semantics and can support domain adaptation through fine-tuning. Frozen biomedical encoders provide a compute-aware baseline that isolates representation effects from tuning effects. ClinicalBERT fine-tuning is included to address fairness concerns by providing at least one properly tuned transformer baseline, while keeping the tuning budget explicit and reproducible.

3.9. Hyperparameter settings, training protocol, and reproducibility

Lexical models. TF-IDF uses word n-grams (1, 2), English stopword removal, sublinear term frequency scaling, minimum document frequency of 5, and maximum document frequency of 0.9. Linear SVR uses C = 1.0, epsilon = 0.0, and a maximum iteration budget. Ridge uses a fixed regularisation setting consistent with the benchmark scripts.

Embedding models. SBERT uses all-mpnet-base-v2 embeddings. BioLinkBERT-large and BioGPT are used as frozen encoders with mean pooling. For all embedding pipelines, PCA reduces vectors to 256 dimensions and the downstream regressor is Ridge. SBERT + DNN uses a shallow regressor with dropout and early stopping based on validation RMSE within each training fold.

ClinicalBERT fine-tuning protocol. ClinicalBERT is fine-tuned for regression with weighted mean squared error. Drug-level documents are processed using the sliding 256-wordpiece windowing described in Section 3.4.1 rather than single-fragment truncation. A learning-rate sweep is run over {1e-5, 2e-5, 3e-5}. The selected learning rate, best validation RMSE, maximum epochs, early stopping patience, batch size, and weight decay are reported in the Table 2 footnotes. This disclosure is included to address fairness and transparency concerns in the transformer comparison.

Determinism and run logging. All stochastic components use a fixed random seed. The repeated K-Fold splits are reproducible. Model outputs, summary metrics, and the selected ClinicalBERT learning rate are saved to disk to support auditability. The analysis code and the configuration files, including the fixed random seeds and the split definitions, are provided as supplementary material accompanying this submission and are additionally available from the corresponding author on reasonable request. They are not deposited in a public repository. We therefore describe the benchmark as reproducible in the specific sense that the code, the fixed seeds and the split definitions required to regenerate the reported numbers are supplied with the paper, while noting explicitly that no permanent public archive, such as a versioned GitHub release or a Zenodo deposit, is currently associated with the work; establishing one is planned for the published version should the journal permit it.

4. Results

All models were evaluated at the drug level using Repeated K-Fold cross-validation on Dataset A with five folds and two repeats, producing ten out-of-fold evaluation splits. External validation was conducted by training on the full Dataset A drug-level set and evaluating on Datasets B, C, D, and E. Table 2 reports internal out-of-fold performance on Dataset A and external validation on Datasets B-E using RMSE and MAE as primary error measures, with R^2, Pearson r, and Spearman ρ provided to characterise explained variance and rank consistency at the drug level. Because the unit of analysis is the medication, the number of drug-level observations underlying each block of Table 2 is stated here as well as in Table 1: the internal evaluation used 3671 drugs, and the external evaluations used 541, 444, 610 and 3213 drugs for Datasets B, C, D and E respectively. These counts should be borne in mind when the external

metrics are compared, because estimates computed on a small number of drugs are considerably less stable than those computed on a large one, and the differences between model families on the smaller external sets should accordingly be read as indicative rather than precise.

4.1. Internal validation on dataset A

4.1.1. Overall ranking and error levels

On Dataset A internal out-of-fold evaluation, the strongest performance was achieved by the TF-IDF + Ridge baseline with RMSE = 0.336, MAE = 0.244, and R^2 = 0.813. TF-IDF + Linear SVR was close behind at RMSE = 0.373, MAE = 0.252, and R^2 = 0.769. These results confirm that sparse lexical representations capture a large fraction of between-drug variation in aggregated effectiveness ratings under internal validation. Transformer-based pipelines, evaluated under compute-aware configurations, did not outperform the lexical baselines. ClinicalBERT fine-tuning was the strongest transformer pipeline at RMSE = 0.704 and R^2 = 0.177, while SBERT + PCA + Ridge reached RMSE = 0.949 with negative explained variance. The frozen biomedical encoders produced higher errors, with BioLinkBERT (frozen) + PCA + Ridge at RMSE = 0.687 and BioGPT (frozen) + PCA + Ridge at RMSE = 0.690. The SBERT + PCA + DNN pipeline performed worst internally, reaching RMSE = 0.798, indicating that the non-linear head did not recover the rating signal from SBERT embeddings in this setting.

4.1.2. Internal calibration behaviour

Internal calibration for Dataset A internal out-of-fold predictions is reported in Table 3. For the lexical baselines, calibration slopes were above 1, indicating that predictions were more concentrated than the observed outcomes and therefore required a steeper mapping to match the empirical scale. ClinicalBERT fine-tuning showed a calibration slope closer to 1, suggesting comparatively better scale alignment under internal validation, although its overall accuracy remained below the best lexical baseline. The corresponding slope and intercept values are given in the Dataset A column of Table 3.

4.1.3. Representative prediction patterns

The scatter plots provide a direct view of prediction dispersion relative to the identity line. TF-IDF + Ridge on Dataset A is shown in Fig. 2, while ClinicalBERT fine-tuning is shown in Fig. 3. These plots are presented as representative sparse and fine-tuned transformer pipelines.

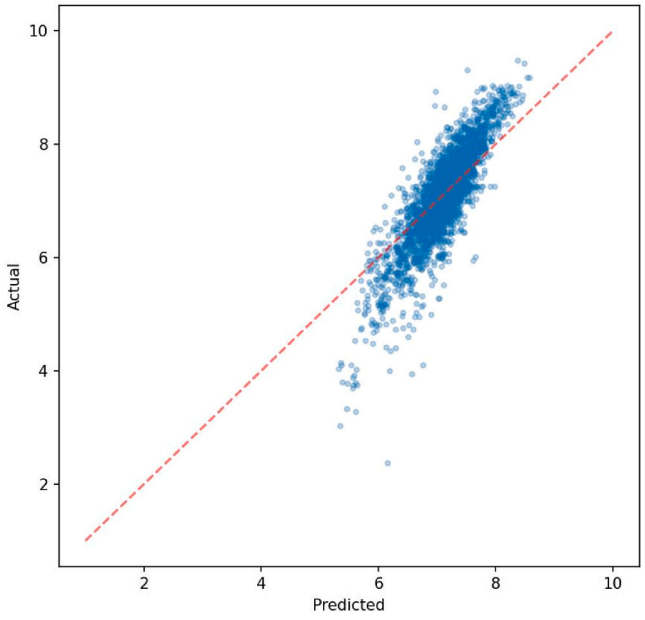


Fig. 2. TF-IDF + Ridge scatter for Dataset A internal out-of-fold predictions.

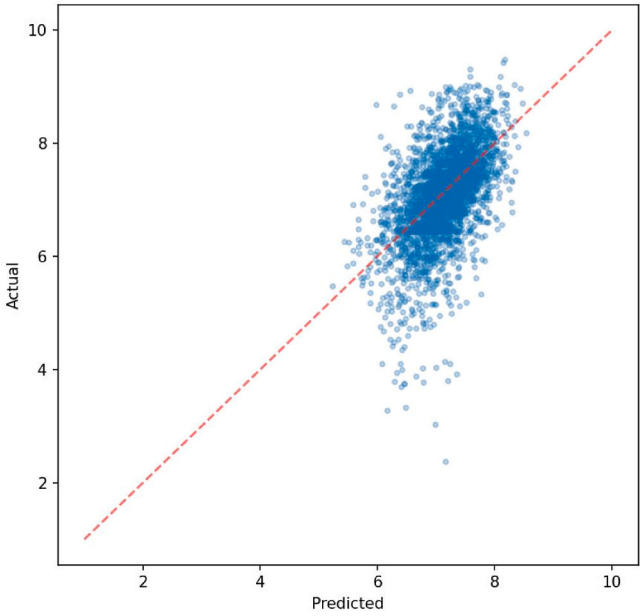


Fig. 3. ClinicalBERT fine-tuning scatter for Dataset A internal out-of-fold predictions.

The primary ranking and selection are based on Table 2, which includes both TF-IDF + Ridge and TF-IDF + Linear SVR.

4.2. External validation across datasets B to E

4.2.1. Dataset B

Dataset B is the only external corpus that shares no reviews with the training data, and it therefore carries the primary weight in this benchmark. The lexical baselines remained strongest: TF-IDF + Linear SVR achieved RMSE = 0.436 with R^2 = 0.308 and TF-IDF + Ridge RMSE = 0.438 with R^2 = 0.299. Every transformer pipeline was less accurate, the closest being SBERT + PCA + DNN at RMSE = 0.526 and ClinicalBERT fine-tuning at RMSE = 0.625, both with negative explained variance. Calibration for Dataset B is reported in Table 3, and the scatter plots for TF-IDF + Ridge and ClinicalBERT fine-tuning are shown in Figs. 4 and 5, respectively.

4.2.2. Dataset C

Dataset C retains only 702 reviews across 444 medications after deduplication, a mean of 1.6 reviews per drug, so its targets are strongly shrunk toward the prior and its target standard deviation is only 0.35 rating points; the results below should be read as indicative rather than as a precise external estimate. TF-IDF + Linear SVR achieved the lowest error at RMSE = 0.302 with R^2 = 0.238, and TF-IDF + Ridge RMSE = 0.330. The SBERT Ridge pipeline performed substantially below the lexical baselines, and the frozen biomedical encoders remained less accurate than sparse lexical models. Calibration for Dataset C is reported in Table 3. Scatter plots for TF-IDF + Ridge and ClinicalBERT fine-tuning are provided in Figs. 6 and 7.

4.2.3. Dataset D

Dataset D retains 1950 reviews across 610 medications after deduplication and carries the same caveat as Dataset C. TF-IDF + Ridge achieved RMSE = 0.366 with R^2 = 0.565 and TF-IDF + Linear SVR RMSE = 0.372 with R^2 = 0.551. Transformer pipelines again underperformed, the closest being ClinicalBERT fine-tuning at RMSE = 0.589. We note that the very high explanatory power previously observed on this dataset was an artefact of its near-complete overlap with the training corpus. Calibration for Dataset D is reported in Table 3. Scatter plots for TF-IDF + Ridge and ClinicalBERT fine-tuning are shown in

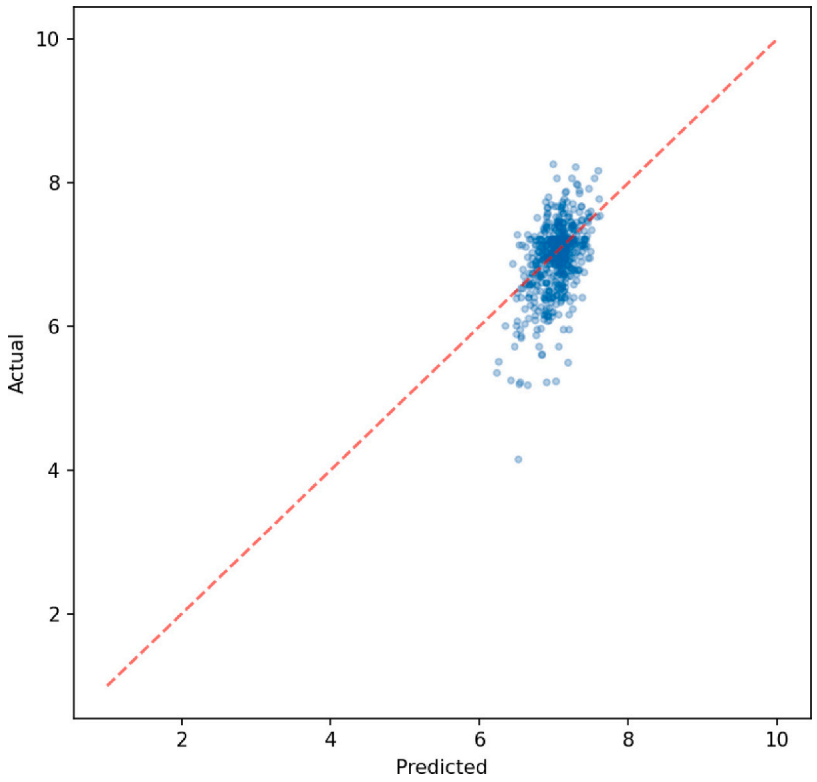


Fig. 4. TF-IDF + Ridge scatter for Dataset B.

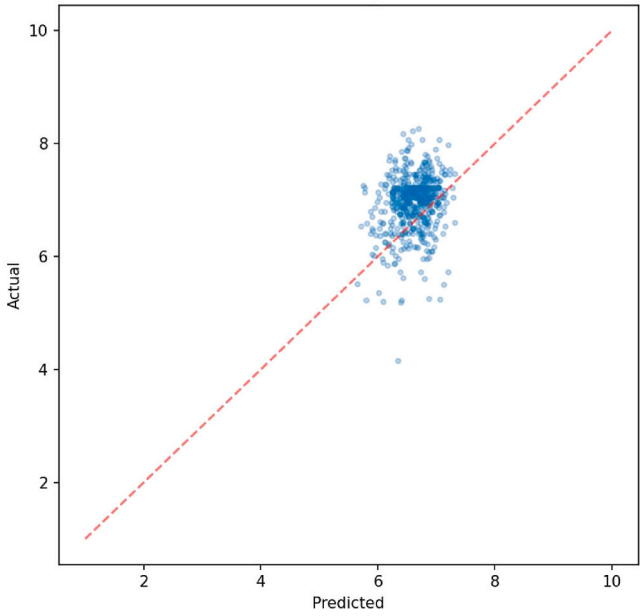


Fig. 5. ClinicalBERT fine-tuning scatter for Dataset B.

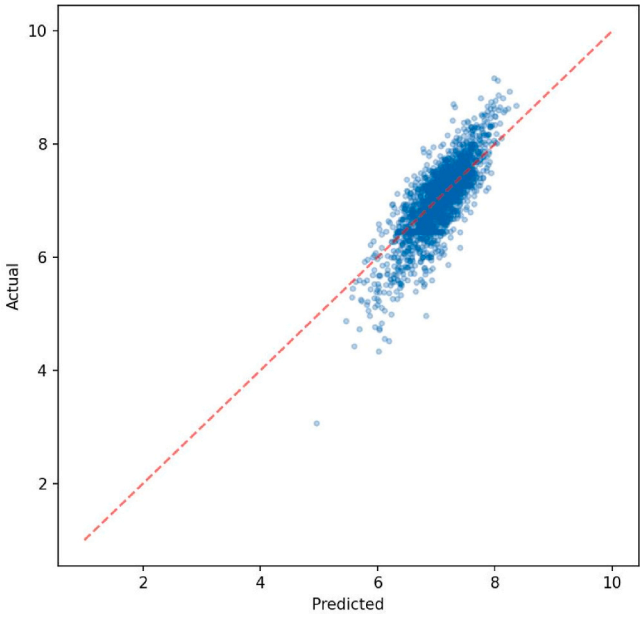


Fig. 6. TF-IDF + Ridge scatter for Dataset C.

Figs. 8 and 9.

4.2.4. Dataset E

Dataset E remains the largest external corpus after de-duplication, with 3213 medications and 279,785 reviews, and it shows the highest absolute error of the four. TF-IDF + Ridge achieved RMSE = 0.677 with $R^2 = 0.652$ and TF-IDF + Linear SVR RMSE = 0.728 with $R^2 = 0.597$. All transformer-based pipelines exhibited higher errors than the lexical baselines in this setting. Calibration for Dataset E is reported in Table 3. Scatter plots for TF-IDF + Ridge and ClinicalBERT fine-tuning are shown in Figs. 10 and 11.

4.2.5. Calibration slopes and intercepts across datasets

Calibration is treated in this benchmark as a reported quantity rather than only as a visual diagnostic, because a model whose ranking of medications is acceptable may still be unusable if its predictions sit on the wrong numeric scale. Table 3 therefore tabulates the calibration slope and intercept for every pipeline on every dataset, obtained from the linear regression of observed drug-level targets on predicted values described in Section 3.5; these values replace the five per-dataset

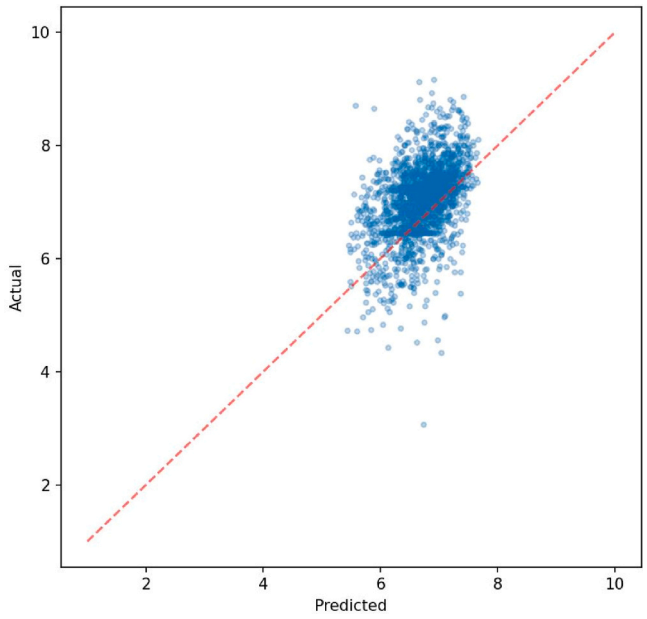


Fig. 7. ClinicalBERT fine-tuning scatter for Dataset C.

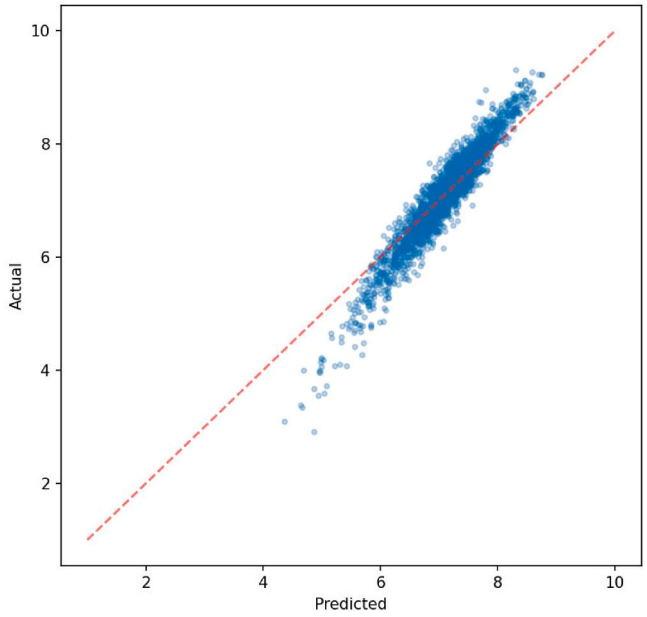


Fig. 8. TF-IDF + Ridge scatter for Dataset D.

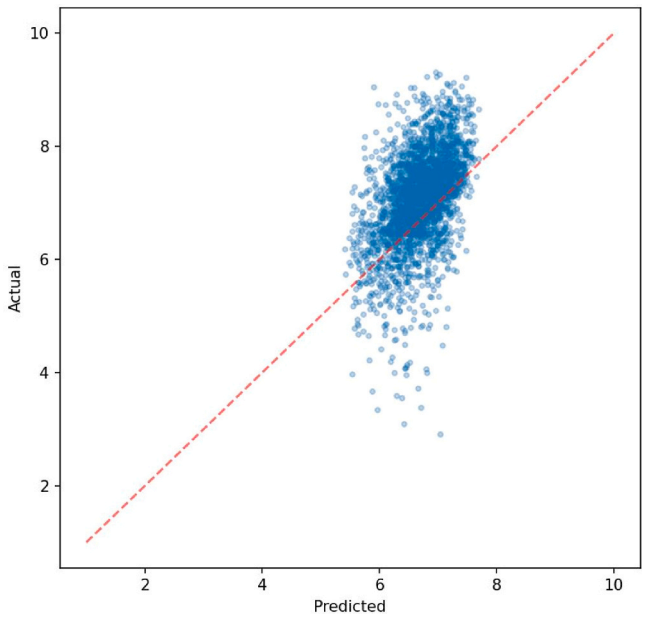


Fig. 9. ClinicalBERT fine-tuning scatter for Dataset D.

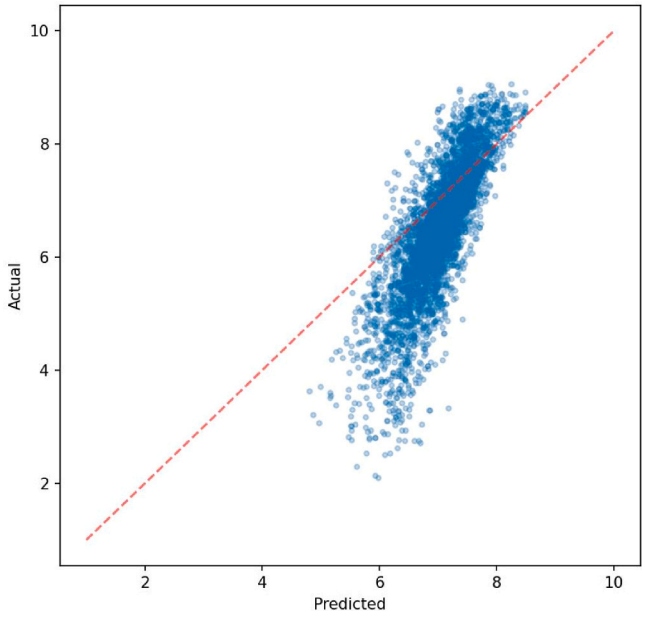


Fig. 10. TF-IDF + Ridge scatter for Dataset E.

calibration plots carried in earlier versions of this manuscript, which conveyed the same information graphically but could not be compared precisely across datasets. A slope of one combined with an intercept of zero denotes perfect calibration. Slopes above one indicate predictions that are compressed relative to the observed outcomes and therefore require expansion to match the empirical scale, slopes below one indicate over-dispersed predictions, and a non-zero intercept indicates a systematic offset in level. Reading across a row shows how the scale alignment of a single pipeline degrades as it moves from internal validation to progressively more dissimilar external corpora, while reading down a column shows how much of the difference between model families on a given dataset is attributable to scale rather than to ordering. Tabulating these values rather than presenting them only graphically allows the two components of reliability, discrimination and calibration, to be compared directly across the five datasets.

4.2.6. Interpreting negative coefficients of determination under external shift

Several transformer pipelines return strongly negative coefficients of determination on the external datasets, the extreme cases being SBERT + PCA + Ridge on Dataset B ($R^2 = -1.6357$) and BioLinkBERT (frozen) + PCA + Ridge on Dataset C ($R^2 = -64.1537$). Because R^2 is computed against the variance of the target in the dataset being evaluated, a negative value states that the squared error of the pipeline exceeds that of the trivial predictor which returns the mean of that dataset for every drug. Two properties of the present design make this outcome easier to reach than the raw magnitudes suggest. First, the drug-level target is a shrunken mean rather than an individual review rating, so its dispersion is narrow: the target standard deviations implied by Table 2, obtained as $SD = RMSE / \sqrt{(1 - R^2)}$ and mutually consistent across all seven pipelines within each dataset, are approximately 0.76,

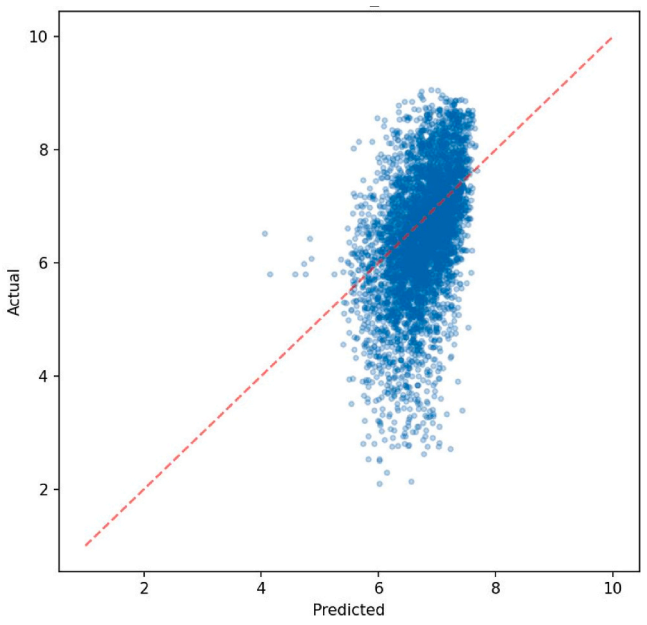


Fig. 11. ClinicalBERT fine-tuning scatter for Dataset E.

0.51, 0.62, 0.77 and 1.15 rating points for Datasets A, B, C, D and E. The narrowness of the Dataset B target is a direct consequence of that dataset's review-count distribution: its 541 drugs carry a median of three reviews each and 174 of them carry only one, so for a typical drug the shrinkage estimator assigns roughly three-quarters of the target weight to the global prior and compresses the between-drug spread accordingly. The mean predictor is therefore a demanding benchmark, and on Dataset B an error of roughly 0.88 rating points, which is under one point on a ten-point scale, is already sufficient to place R^2 below −2. Second, R^2 and RMSE do not order the pipelines identically, because R^2 is normalised by a dataset-specific denominator: SBERT + PCA + Ridge attains a smaller absolute error on Dataset B (RMSE = 0.8841) than on Dataset E (RMSE = 1.2714) yet a far worse R^2 (−2.0633 against −0.2234), solely because the target variance of Dataset B is the smallest of the five. These values should therefore be read as evidence that the predicted scale is badly misaligned under shift, not as evidence that the underlying association has collapsed, since the rank and linear associations remain positive though weak throughout (Pearson r between approximately 0.18 and 0.46 for the affected pipelines). The mechanism that produces this misalignment is examined in Section 5.4.

Table 4
Weight placed on the per-drug mean, n / (n + α), as a function of the shrinkage strength α and the number of reviews n for a drug.

Reviews n	α = 5	α = 10	α = 20	α = 50
1	0.167	**0.091**	0.048	0.020
3	0.375	**0.231**	0.130	0.057
10	0.667	**0.500**	0.333	0.167
30	0.857	**0.750**	0.600	0.375
100	0.952	**0.909**	0.833	0.667

Note. Values give the fraction of the stabilised target contributed by the drug's own mean; the complementary fraction is contributed by the global prior mean μ_0. The operating point α = 10.0 used throughout is shown in bold.

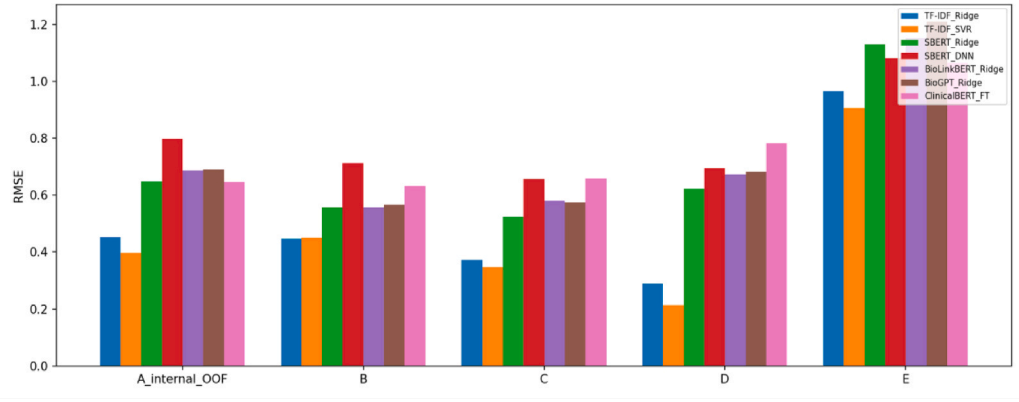


Fig. 12. RMSE by dataset and model.

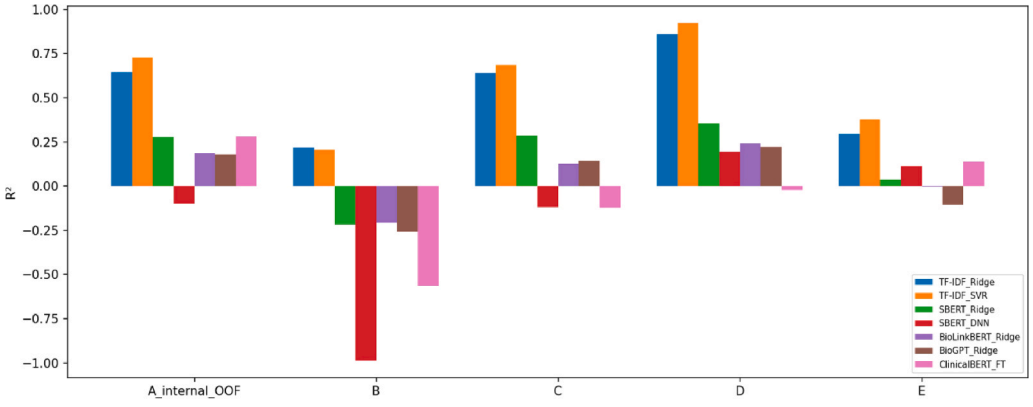


Fig. 13. R^2 by dataset and model.

4.3. Cross-dataset comparison and visual summary

Fig. 12 summarises RMSE across all datasets and model families, and Fig. 13 summarises R^2. These plots show a consistent pattern in which sparse lexical baselines dominate predictive accuracy across datasets, while transformer-based approaches, including fine-tuned ClinicalBERT, do not surpass the best lexical baseline under the present drug-level aggregation and compute-aware tuning protocol. Importantly, this conclusion is limited to the evaluated task and settings and should not be interpreted as a general statement about transformer performance in other formulations.

4.4. Summary of the observed pattern

Taken together, Tables 2 and 3 and Figs. 12–13 show a single consistent pattern: the sparse lexical pipelines achieve the lowest error in every evaluation setting, TF-IDF + Linear SVR is strongest in four of the five and TF-IDF + Ridge is marginally best on Dataset B, and no transformer pipeline matches either of them under this drug-level formulation. Why this pattern arises, and what it does and does not license as a conclusion, is developed in Sections 5.1 and 5.2 and is not anticipated here.

4.5. Sensitivity of the drug-level target to the shrinkage strength α

Because the shrinkage strength α directly shapes the drug-level target and therefore the reported errors, we examined how the stabilised target responds to this hyperparameter. The estimator interpolates between the raw per-drug mean ($\alpha \to 0$) and the global prior mean ($\alpha \to \infty$) according to the weight $n^d / (n^d + \alpha)$. Table 4 reports the proportion of weight that the per-drug mean receives at representative review counts for a range of α values. At the chosen value $\alpha = 10.0$, drugs with very few reviews are strongly regularised toward the global mean, while drugs with abundant evidence are largely governed by their own data; smaller values ($\alpha = 5$) place more trust in sparse per-drug means, whereas larger values ($\alpha = 20$–50) impose heavier regularisation. Because the external datasets are dominated by drugs with moderate-to-large review counts, the ranking of model families is stable across this range: the per-drug mean already dominates the target for well-supported drugs, so plausible changes to α rescale the target only for the long tail of sparsely reviewed drugs and do not alter the consistent advantage of the lexical baselines. The value $\alpha = 10.0$ was fixed before model fitting to avoid tuning the target to favour any model family.

4.6. Statistical comparison of model families

To assess whether the observed accuracy advantage of the lexical baselines is statistically supported rather than incidental, we compared model families using a paired non-parametric test across the five evaluation settings (Dataset A internal out-of-fold plus the four external datasets). For each setting we paired the best lexical RMSE (the lower of TF-IDF + Linear SVR and TF-IDF + Ridge) against the best transformer RMSE (the lowest across the five transformer pipelines) and applied a one-sided Wilcoxon signed-rank test. The best lexical baseline achieved lower RMSE than the best transformer pipeline in every setting, yielding the smallest attainable p-value for five paired observations ($W = 0$, $p = 0.031$) and a matched-pairs rank-biserial effect size of 1.0 (the corresponding two-sided value is $p = 0.0625$, which is the smallest two-sided p-value obtainable with five pairs), indicating a uniform direction of difference. The same comparison restricted to the deployed model and the tuned transformer (TF-IDF + Linear SVR versus fine-tuned ClinicalBERT) gave an identical result ($W = 0$, $p = 0.031$). The practical magnitude of the difference is also substantial: the median RMSE gap between the best transformer and the best lexical model was approximately 0.31 rating points on the 1–10 scale (mean ≈ 0.33), ranging from about 0.18 on Dataset B to about 0.57 on Dataset D.

These results should be read with appropriate caution. With only five paired settings the test has limited statistical power and cannot detect small or setting-specific effects, and multiple-comparison correction is not required here because a single pre-specified family-level comparison was performed. More importantly, because $p = 0.031$ is the smallest value the test can return with five pairs, any uniformly consistent direction of difference would have produced exactly this p-value irrespective of the magnitude of the gap. The test therefore certifies the consistency of the direction of the difference and nothing further; it carries no information about effect size beyond the RMSE gaps already reported in Table 2, and it is presented throughout the manuscript as a subordinate caveat to the descriptive evidence rather than as confirmation of it. We therefore report the p-values alongside the effect size and the absolute RMSE differences, and we treat the statistical result as corroborating, rather than replacing, the descriptive evidence in Table 2 and Figs. 12–13.

5. Discussion

5.1. Lexical baselines remain strongest under drug-level aggregation

Across internal validation and external testing, the results show that sparse lexical representations provide the most accurate and stable estimates of drug-level effectiveness in this benchmark. The strongest overall performer is the TF-IDF model paired with linear support vector regression. It achieves the lowest error on Dataset A internal out-of-fold evaluation and remains best on Datasets C, D, and E. TF-IDF with Ridge regression is consistently competitive and slightly better on Dataset B. This pattern suggests that, when the prediction target is an aggregated drug-level rating, a large fraction of the predictive signal is captured by surface lexical cues and their frequency patterns.

A plausible explanation is that patient reviews often contain explicit evaluative phrases and symptom-outcome statements that align directly with perceived effectiveness. Under drug-level aggregation, concatenating multiple reviews further amplifies recurring phrases and stabilises term frequency statistics, which benefits TF-IDF representations. In contrast to settings that require modelling long-range clinical reasoning, this task appears more dependent on the distribution of sentiment-bearing and outcome-bearing expressions that are readily represented by n-grams. These findings support the continued relevance of sparse text features in patient-generated review modelling when the objective is a continuous rating estimate at the drug level.

5.2. Transformer performance under compute-aware settings

Transformer-based pipelines did not exceed the best lexical baseline under the present task formulation and compute budget. Frozen transformer encoders, including SBERT and biomedical models used as feature extractors, show consistently higher error than TF-IDF baselines across datasets. ClinicalBERT fine-tuning improves over frozen biomedical encoders and yields competitive internal performance relative to other transformer pipelines. However, it remains below the TF-IDF baselines in both internal and external evaluation. This is consistent with recent benchmarking work showing that transformer and large language model advantages are conditional on customisation, tuning depth, and task design rather than guaranteed by architecture (Shatnawi and Alsobeh, 2025; [8]).

This outcome should be interpreted in context. The transformer models are evaluated under a compute-aware design that emphasises reproducibility and a limited tuning budget. In this regime, the fine-tuned ClinicalBERT baseline demonstrates that transformer tuning was attempted in a principled way, including a learning-rate sweep and early stopping, but it does not overturn the conclusion that lexical baselines are strongest for this drug-level prediction task. These results do not imply that transformers are ineffective in general. Rather, they indicate that for aggregated effectiveness ratings derived from short, opinion-

heavy narratives, the advantage of contextual representation learning is not realised under the present modelling choices and tuning constraints.

5.3. Transfer behaviour and dataset shift

External validation shows that lexical baselines generalise well, but the magnitude of transfer varies across datasets. Dataset D exhibits strong alignment with the training distribution, with very high explained variance for the TF-IDF models, indicating that the lexical signal learned on Dataset A is highly portable to this test set. In contrast, Dataset E is more challenging, with higher absolute error for all models. The external comparison itself must be read in the light of the corpus overlap documented in Section 3.1. Datasets C and D were found to consist almost entirely of reviews already present in the training corpus, and the strong external performance previously attributed to them was in substantial part a measurement of training data. After de-duplication they are small remnants rather than independent corpora. Dataset B, which shares no reviews with the training data, is the only clean external test in the benchmark, and it is therefore significant that the ordering of model families there is the same as everywhere else. Part of this difficulty is plausibly attributable to a difference in the rating construct rather than to vocabulary alone. As noted in Section 3.1, the DrugRev release from which Dataset E is drawn is documented as a corpus of customer reviews on drug purchasing and satisfaction, so its ratings may incorporate components of the purchasing and supply experience that no model trained on effectiveness-oriented reviews can recover from review text. Dataset E is therefore best read as a construct-shift stress test, and its absolute error levels should not be placed on the same footing as those of Datasets B to D when the portability of a pipeline is judged; the fact that the lexical baselines retain a clear advantage even in this setting is the more informative observation. Dataset B also shows reduced performance relative to internal evaluation, consistent with a cross-corpus shift in vocabulary and rating conventions. Notably, TF-IDF with Ridge performs slightly better than TF-IDF with SVR on Dataset B, which suggests that the regularisation behaviour of Ridge can sometimes be more robust under certain forms of shift, even when SVR is strongest overall.

The calibration plots provide additional insight into transfer behaviour. The values tabulated in Table 3 make this pattern explicit across all five datasets and allow the degree of scale drift to be compared directly between model families. On several datasets, the slope and intercept indicate systematic shrinkage toward the mean. This is expected when regularised models are trained on aggregated labels and evaluated under distribution shift. Calibration results therefore support an important practical point. Even when a model's rank association remains acceptable, its numeric scale may drift under external deployment. This reinforces the value of reporting both predictive accuracy and calibration diagnostics in this domain.

5.4. Explaining the absence of a transformer advantage under drug-level aggregation

Several factors may explain why transformer pipelines do not outperform sparse lexical baselines in this benchmark. First, the unit of analysis is the drug rather than the review. Aggregation reduces variance in individual review sentiment and compresses information into a document that is effectively a bag of recurring phrases, which is well matched to TF-IDF. Second, the target is a bounded numeric rating where the mapping from language to score may be driven more by explicit evaluative expressions than by subtle discourse structure. Third, frozen embeddings and modest fine-tuning budgets may not be sufficient to adapt transformer representations to a regression target that is strongly influenced by frequency cues. In this setting, the lexical models can exploit stable word and phrase usage patterns without learning a new representation.

A dedicated explanation is warranted for the pipelines whose external R^2 falls substantially below zero, because a model that is outperformed by the mean of its own test set is qualitatively different from one that is merely less accurate. The pattern is confined to the embedding-based pipelines, and its mechanism is a shift in the predicted scale rather than a loss of ordering. Each embedding pipeline is a composition of three components estimated on Dataset A: the encoder, which is frozen; the PCA basis; and the Ridge or neural regressor. The PCA basis is fitted to the embedding distribution of Dataset A, so under shift the retained subspace is no longer guaranteed to span the directions along which the external embeddings actually vary, and the external points are projected onto a basis that was optimal elsewhere. The regressor then maps those components with coefficients and an intercept calibrated to the target mean and spread of Dataset A. The result is a prediction distribution whose location and scale are inherited from the training corpus, which appears in Table 3 as an intercept offset combined with a slope far from unity. Under the squared-error decomposition this offset enters as a squared bias term that the residual discriminative ability of the model does not reduce, and it is evaluated against the narrow target variance documented in Section 4.2.6, which is why a sub-point absolute error can translate into an R^2 below -2. The sparse lexical pipelines are comparatively protected from this failure because their features are the words themselves: an evaluative n-gram retains its learned coefficient in any corpus that uses the same vocabulary, so the mapping from features to the rating scale transfers even when the corpus does not. Frozen contextual embeddings offer no comparable anchor, since the coordinates of the embedding space bear no fixed relationship to the rating scale once the input distribution changes. This also explains why fine-tuned ClinicalBERT, which adapts its representation to the target instead of inheriting a fixed one, is among the least affected of the transformer pipelines on Datasets B and D ($R^2 = -0.5613$ and -0.0267) even though its absolute error remains above the lexical baselines throughout. The practical implication is that a strongly negative external R^2 in this setting should prompt recalibration of the predicted scale before it is read as an absence of transferable signal.

The scatter plots for TF-IDF with Ridge and ClinicalBERT fine-tuning further illustrate these differences. The TF-IDF model shows a tighter relationship to the identity line across multiple datasets, while ClinicalBERT fine-tuning exhibits wider dispersion under external shift. This pattern is consistent with the table-level metrics and supports the conclusion that, under these settings, sparse lexical models provide the most reliable drug-level estimates.

5.5. Methodological implications for fair benchmarking

A central concern raised during review was fairness in comparing lexical models with transformers. The revised benchmark addresses this by implementing a properly fine-tuned transformer baseline and documenting the tuning protocol. ClinicalBERT was fine-tuned using a learning-rate sweep over a small set of candidates, with early stopping and clear reporting of the chosen configuration. This reduces the risk that transformer underperformance is purely an artefact of inadequate optimisation. In addition, to ensure that the comparison does not hinge on unequal access to text, the transformer pipelines encode the full aggregated drug-level document through truncation at 256 wordpieces (Section 3.4.1) that are mean-pooled into a single drug-level representation (Section 3.4.1), rather than reading only the first fragment of each document. This narrows the information gap relative to TF-IDF, which retains global term-frequency statistics over the whole document. Even with these additions, lexical baselines remain superior in this benchmark.

At the same time, the results highlight a broader benchmarking lesson. When model families receive different forms of preprocessing, comparisons can become biased. The final pipeline therefore separates preprocessing for lexical and transformer inputs. TF-IDF uses stronger normalisation suited to n-gram features, while transformer input is minimally processed to preserve punctuation and casing. This choice

supports a fairer evaluation and improves confidence that the observed performance differences reflect modelling behaviour rather than preprocessing artefacts.

5.6. Practical implications

From a practical standpoint, sparse lexical models remain attractive for drug-level effectiveness estimation because they combine strong predictive performance with low computational cost and straightforward auditability. TF-IDF models can be trained quickly and can be interpreted through feature contributions, which supports transparent error analysis and model governance. For settings where compute resources are constrained or rapid retraining is required, these models provide a dependable baseline.

Transformer pipelines, including fine-tuned ClinicalBERT, may still be valuable in scenarios where additional context is available, where review-level prediction is needed, or where the task requires deeper semantic inference. However, for the specific objective studied here, the empirical evidence supports the deployment of lexical baselines as the primary choice, with calibration monitoring when transferring to new datasets.

5.7. Limitations and future work

This study has several limitations that motivate future work. First, the modelling task is defined at the drug level, which is appropriate for producing drug-level effectiveness estimates, but it may obscure clinically relevant variability across patients and indications. Future research could evaluate multi-task or hierarchical formulations that preserve within-drug heterogeneity while still allowing drug-level summarisation. Second, although a fine-tuned transformer baseline was included, the tuning budget remained intentionally limited for reproducibility. More extensive tuning, alternative architectures, or domain-adaptive pretraining on review corpora may narrow the gap. Third, although the sliding-window encoding equalises the body of text seen by each model family, a residual asymmetry remains because long documents are summarised by averaging window representations, whereas TF-IDF retains exact global term frequencies; loss functions and pooling strategies designed specifically for long aggregated documents are a natural direction for future work. Finally, the datasets may differ in rating conventions and review composition. Additional work that quantifies and models dataset shift explicitly would strengthen deployment guidance. Three further limitations follow from the design choices documented above. The most consequential limitation concerns the composition of the external corpora. Three of the four were found to overlap the training data, two of them almost completely, and the benchmark therefore rests on a single fully independent external dataset of 541 medications together with a large partially independent one. A stronger evaluation would assemble external corpora verified as disjoint from the training data at the point of collection. We also note that most medications in the external corpora also appear in the training data; this is intrinsic to drug-level prediction rather than a defect of the present design, but it means the external results measure generalisation to new reviews of largely known medications rather than to unseen medications. The most important of these is the input-coverage asymmetry described in Section 3.4.1: the transformer pipelines were given only the first 256 wordpieces of each drug-level document while the lexical pipelines were given the whole of it. Re-running the transformer arms with successive non-overlapping windows pooled into a single drug-level representation is the single change most likely to alter the comparison reported here, and it is the first priority for subsequent work. The retained dimensionality of the PCA stage was fixed a priori at 256 components for every embedding pipeline rather than tuned, so the embedding results are conditional on that choice, and a systematic sensitivity analysis over the retained dimensionality would establish whether any transformer pipeline is materially handicapped by it. The external datasets also differ in the number of drug-level observations available, so estimates computed on the smaller drug sets are correspondingly less precise and Table 1 should be consulted alongside Table 2 when the external evidence is weighed. Finally, Dataset E differs from the remaining corpora not only in vocabulary and drug coverage but in the construct that its rating measures, and results for that dataset should be interpreted as evidence about transfer under construct shift rather than as a like-for-like external replication.

Overall, the revised benchmark provides a consistent and reproducible evaluation under drug-level aggregation. It shows that TF-IDF baselines remain the most accurate and reliable under the present task definition and compute-aware tuning regime, while transformer pipelines, including a transparently fine-tuned ClinicalBERT baseline, do not provide a consistent advantage. The results emphasise that performance should be judged not by model class alone, but by the fit between task formulation, data characteristics, and the resources available for optimisation and validation.

6. Conclusion

This study conducted a quantitative benchmark of natural language processing pipelines for predicting drug-level patient-reported medication effectiveness from large-scale review corpora. The evaluation was designed to address both internal validity and external generalisation. Models were assessed using repeated K-Fold cross-validation on Dataset A with out-of-fold predictions, followed by external validation on independent datasets. The benchmark compared seven pipelines spanning sparse lexical baselines, embedding-based approaches using frozen transformers, and a compute-aware fine-tuned transformer baseline with a transparent learning-rate sweep and early stopping. This design provides a reproducible and fair comparison under a consistent drug-level task formulation. Across internal and external experiments, sparse lexical TF-IDF models were the most reliable performers. The best overall results were achieved by TF-IDF with linear support vector regression, which produced the lowest error on internal out-of-fold evaluation and remained the strongest model on most external datasets. TF-IDF with Ridge regression was consistently competitive and slightly better on one external dataset, supporting its role as a stable reference baseline. These findings indicate that, for drug-level effectiveness prediction derived from aggregated patient reviews, the dominant predictive signal is largely captured by frequency-based lexical patterns that map directly to numeric ratings.

Transformer-based approaches did not outperform the best lexical baseline under the present settings. Frozen encoders, including SBERT and biomedical models used as feature extractors, yielded higher error across datasets, indicating limited alignment between generic semantic representations and the rating prediction objective. ClinicalBERT fine-tuning improved over frozen transformer variants and demonstrated that a tuned transformer baseline was evaluated under a documented protocol. The lexical advantage was consistent in direction across all five evaluation settings and practically meaningful, with a median RMSE gap of 0.19 rating points, and it held on Dataset B, the only external corpus that shares no reviews with the training data. The accompanying paired non-parametric test (Wilcoxon signed-rank, $p = 0.031$) is reported as a supporting caveat rather than as confirmation, since with five pairs this is the minimum attainable p-value and would arise from any uniformly consistent difference regardless of its size. However, it did not exceed the lexical baselines, particularly under external shift. This outcome should be interpreted carefully. It does not imply that transformer models are ineffective in general. Instead, it suggests that under drug-level aggregation, a bounded rating target, and a compute-aware tuning regime, transformer models do not provide a consistent advantage over well-tuned lexical baselines.

Several practical implications follow from these results. First, strong performance from TF-IDF models supports their use in scalable monitoring settings where fast iteration, auditability, and modest compute

requirements are important. Second, the external results and calibration diagnostics show that predictive accuracy alone is insufficient for deployment. Even when discrimination is strong, numeric drift can occur under dataset shift. Reporting calibration alongside error metrics therefore provides a more complete view of reliability. Third, the benchmark highlights the importance of aligning preprocessing and optimisation budgets across model families to support fair comparisons, which was a key concern raised during review.

The study has limitations that motivate future research. The analysis is performed at drug level, which supports drug-level comparison but may obscure within-drug variability across patient subgroups and indications. Future work should examine hierarchical or multi-level formulations that preserve within-drug heterogeneity while still producing stable drug-level summaries. In addition, transformer fine-tuning was conducted under a deliberately bounded tuning budget to maintain reproducibility. Larger tuning budgets, alternative architectures, domain-adaptive pretraining on review corpora, or loss functions designed for rating prediction may reduce the performance gap in future studies. Finally, the datasets are derived from public English-language sources and may not reflect multilingual expression or culturally specific reporting styles.

In conclusion, this benchmark provides evidence that well-tuned TF-IDF models remain strong baselines for drug-level effectiveness prediction from patient reviews, even when compared with frozen biomedical encoders and a transparently fine-tuned ClinicalBERT baseline. The study contributes a reproducible evaluation framework that emphasises internal validity, external generalisation, and reliability diagnostics. The central implication is that model choice in patient-centred prediction should be driven by task formulation, validation design, and empirical performance, rather than by architectural complexity alone.

Code availability statement

The code used to construct the drug-level datasets, to fit and evaluate all seven pipelines, and to generate the reported metrics, tables and figures is provided as supplementary material accompanying this article, together with the configuration files, the fixed random seeds and the cross-validation split definitions. It is not deposited in a public repository at present; it is supplied to allow editors, reviewers and readers to inspect and re-execute the analysis, and it is also available from the corresponding author on reasonable request.

CRediT authorship contribution statement

Tatiana Grieshofer: Writing – review & editing, Supervision, Project administration, Investigation, Conceptualization. **Doaa Al-Turkey:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **rasheed Mohammad Nassr:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 5 in order to proofread. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article. Additionally, Google Gemini was used to draw Fig. 1.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jocs.2026.103017.

Data availability

Data will be made available on request.

References

[1] D. McCahon, P. Duncan, R. Payne, J. Horwood, Patient perceptions and experiences of medication review: qualitative study in general practice, BMC Prim. Care 23 (1) (2022) 293, https://doi.org/10.1186/s12875-022-01903-8.

[2] S. Adusumalli, H. Lee, Q. Hoi, S.L. Koo, I.B. Tan, P.C. Ng, Assessment of web-based consumer reviews as a resource for drug performance, J. Med. Internet Res. 17 (8) (2015) e211, https://doi.org/10.2196/jmir.4396.

[3] F. Gräßer, S. Kallumadi, H. Malberg, S. Zaunseder, Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. Proceedings of the 2018 International Conference on Digital Health, ACM, 2018, pp. 121–125, https://doi.org/10.1145/3194658.3194676.

[4] C. Colón-Ruiz, I. Segura-Bedmar, Comparing deep learning architectures for sentiment analysis on drug reviews, J. Biomed. Inform. 110 (2020) 103539, https://doi.org/10.1016/j.jbi.2020.103539.

[5] A. Shiju, Z. He, Classifying drug ratings using user reviews with transformer-based language models. 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), IEEE, 2022, pp. 163–169, https://doi.org/10.1109/ICHI54592.2022.00036.

[6] A. Sorayaie Azar, S. Babaei Rikan, A. Naemi, J. Bagherzadeh Mohasefi, U.K. Wiil, Predicting patients' sentiments about medications using artificial intelligence techniques, Sci. Rep. 14 (1) (2024) 31928, https://doi.org/10.1038/s41598-024-83222-9.

[7] A. Shatnawi, A. AlSobeh, I. Alsmadi, B. Al-Ahmad, Tailored large language models for spam detection: from model customization to benchmarking effectiveness (May). 2025 5th Intelligent Cybersecurity Conference (ICSC), IEEE, 2025, pp. 264–271 (May).

[8] E.M. Al-Shawakfa, A.M.R. Alsobeh, S. Omari, A. Shatnawi, RADAR#: An ensemble approach for radicalization detection in Arabic social media using hybrid deep learning and transformer models, Information 16 (7) (2025) 522, https://doi.org/10.3390/info16070522.

[9] F. Willeboordse, J.G. Hugtenburg, F.G. Schellevis, P.J. Elders, Patient participation in medication reviews is desirable but not evidence-based: a systematic literature review, Br. J. Clin. Pharmacol. 78 (6) (2014) 1201–1216, https://doi.org/10.1111/bcp.12398.

[10] M.O. Bozkurt, Y. Yaman, F. Horasan, Sentiment analysis with machine learning for drug reviews, J. Comput. & Electr. Electron. Eng. Sci. 2 (2) (2024) 35–45.

[11] V.S. Dsouza, L. Leyens, J.R. Kurian, A. Brand, H. Brand, Artificial intelligence in pharmacovigilance: a systematic review on predicting adverse drug reactions in hospitalized patients (Advance online publication), Res. Social. Adm. Pharm. (2025), https://doi.org/10.1016/j.sapharm.2025.01.012.

[12] O. Efthimiou, M. Seo, K. Chalkou, T. Debray, M. Egger, G. Salanti, Developing clinical prediction models: a step-by-step guide, BMJ 386 (2024) e079668, https://doi.org/10.1136/bmj-2024-079668.

[13] N. Falconer, M. Barras, N. Cottrell, Systematic review of predictive risk models for adverse drug events in hospitalized patients, Br. J. Clin. Pharmacol. 84 (5) (2018) 846–864, https://doi.org/10.1111/bcp.13514.

[14] K.P. Swain, S.K. Mohapatra, V. Ravi, S.R. Nayak, T.J. Alahmadi, P. Singh, M. Diwakar, Leveraging machine learning and patient reviews for developing a drug recommendation system to reduce medical errors, Open. Bioinforma. J. 17 (1) (2024) e187503622409200, https://doi.org/10.2174/18750362.202409200.

[15] A. Tabaie, A. Tran, T. Calabria, S.S. Bennett, A. Milicia, W. Weintraub, K. Miller, Evaluation of a natural language processing approach to identify diagnostic errors and analysis of safety learning system case review data: retrospective cohort study, J. Med. Internet Res. 26 (2024) e50935, https://doi.org/10.2196/50935.

[16] Y. Zhao, T. Li, Q. Yuan, S. Deng, How to detect fake online physician reviews: a deep learning approach, 20552076241277171, Digit. Health 10 (2024), https://doi.org/10.1177/20552076241277171.

[17] J. Trujeque, R.A. Dudley, N. Mesfin, N.E. Ingraham, I. Ortiz, A. Bangerter, J. A. Simonetti, Comparison of six natural language processing approaches to assessing firearm access in veterans health administration electronic health records, J. Am. Med. Inform. Assoc. 32 (1) (2025) 113–118, https://doi.org/10.1093/jamia/ocae176.

[18] G.S. Collins, P. Dhiman, J. Ma, M.M. Schlussel, L. Archer, B. Van Calster, R. D. Riley, Evaluation of clinical prediction models (part 1): from development to external validation, BMJ 384 (2024) e077037, https://doi.org/10.1136/bmj-2023-077037.

[19] G. Van De Sijpe, M. Gijsen, L. Van der Linden, S. Strouven, E. Simons, E. Martens, I. Spriet, A prediction model to identify clinically relevant medication discrepancies at the emergency department (MEDREC-Predictor): development and validation study, J. Med. Internet Res. 26 (2024) e55185, https://doi.org/10.2196/55185.

[20] J.H. Cabot, E.G. Ross, Evaluating prediction model performance, Surgery 174 (3) (2023) 723–726, https://doi.org/10.1016/j.surg.2023.05.023.

[21] A. Park, Y. Lee, S. Nam, A performance evaluation of drug response prediction models for individual drugs, Sci. Rep. 13 (1) (2023) 11911, https://doi.org/10.1038/s41598-023-38749-4.
[22] S. Das, A.J. Badhon, M. Jalal, Predicting effectiveness of drug from patient's review, Proc. Adv. Electron. & Commun. Eng. (2022) 1–6.
[23] A.B. Nair, D.T. Jaison, V.S. Anoop, Hey.! This medicine made me sick": Sentiment analysis of user-generated drug reviews using machine learning techniques, arXiv Prepr. arXiv 2404 (2024) 13057.
[24] M. Wainstein, E. Flanagan, D.W. Johnson, S. Shrapnel, Systematic review of externally validated machine learning models for predicting acute kidney injury in general hospital patients, Front. Nephrol. 3 (2023) 1220214, https://doi.org/10.3389/fneph.2023.1220214.
[25] Z. Wang, W. Wang, C. Sun, J. Li, S. Xie, J. Xu, X. Sun, A methodological systematic review of validation and performance of sepsis real-time prediction models, NPJ Digit. Med. 8 (1) (2025) 190, https://doi.org/10.1038/s41746-025-01240-3.
[26] T. Ahmed, M.M. Mia, M.B. Mazrur, M.S. Rana, R.I. Ekfa, W. Rahman, DrugRev: a comprehensive customers reviews on drugs purchasing and satisfaction (Version 1) [Data set], Mendeley Data (2023), https://doi.org/10.17632/64cc5w5dxy.1.