Agentic Construction of EPAR-Structured Toxicity Dossiers: Split Evaluation and a Rapamycin Case Study

Anonymous Author(s)
Affiliation
Address
email

Abstract

We build an AI-driven system that compiles medicine safety dossiers in the structure of the European Public Assessment Report (EPAR)—the format regulators use to document benefits and risks. We focus on the "Undesirable effects" section (§4.8) and use sirolimus (rapamycin) as a case study. To judge quality, we compare the system's side-effect list to the official patient leaflet by matching each side-effect term one-to-one and checking whether the attached study citation is correct. On sirolimus, our approach covers more of the leaflet's terms than a strong web-enabled baseline while keeping good citation accuracy. We release tables, figures, and a simple recipe so others can repeat the evaluation on their own agents.

1 Introduction

Why this matters. Clinicians rely on the prescriber label (*Summary of Product Characteristics, SmPC*), patients on the *Patient Information Leaflet (PIL)*, and the public on the *European Public Assessment Report (EPAR)* for transparent benefit–risk information. These sources evolve, and synthesizing them manually is slow and error-prone.

One-sentence idea. Use an AI system to gather, cross-check, and organize safety information directly into an EPAR-like dossier, with emphasis on §4.8 (side effects).

What is new. (i) *Structured from the start:* the agent writes into a regulator-style layout so outputs are easy to navigate and *comparable across drugs*. (ii) *Testable against ground truth:* we judge the side-effects section against the official leaflet's preferred terms. (iii) *One claim = one side effect:* we split sentences so each claim contains exactly one side-effect term for unambiguous scoring. (iv) *Source-aware:* we verify whether attached study citations are correct.

Glossary (plain language)

EPAR: European Public Assessment Report (regulator's public assessment).
SmPC: Summary of Product Characteristics (prescriber label).
Side effects ("Undesirable effects"): unwanted effects while taking a medicine.
Preferred term: standardized side-effect name (e.g., "stomatitis").
SOC: System Organ Class (body-system category, e.g., *Blood and Lymphatic System Disorders*).

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

2 System and Pipeline (Progress Report)

EPAR structure as blueprint. The EPAR/SmPC schema defines how safety information is organized across quality, non-clinical toxicology, and clinical safety [1]. Writing directly into this schema makes outputs navigable and comparable across drugs.

Dossier content. Scope covers mechanism-of-action (safety-relevant), preclinical toxicology, clinical side effects, warnings/contraindications, and risk management.

Data acquisition. Programmatic access to EMA (EPAR/SmPC) [1], FDA labeling and FAERS [2], plus public repositories; OCR is used if needed.

Multi-source integration. We integrate SIDER, TOXRIC, and FAERS alongside peer-reviewed literature [2–4]. Terminology is harmonized and mapped to SOCs.

Automated updates. Scheduled refresh, normalization and merge; a simple website reflects updates automatically.

Benchmarking vs. leaflets. We benchmark §4.8 coverage against official leaflets with versioning.

3 Methods

Ground truth. Leaflet §4.8 preferred terms with their SOC and frequency bands serve as ground truth. Leaflet preferred terms align with MedDRA Preferred Terms (PT) [5].

Producing the dossier sections (§4.1 and §4.8). *§4.1 Therapeutic Indications* is derived from pivotal trials and consensus guidance; we cross-check scope and population against SmPC/EPAR wording for consistency [1]. *§4.8 Undesirable Effects* is built by (i) splitting generated text so each row contains one side-effect term; (ii) mapping to MedDRA PT and SOC [5]; (iii) verifying citations (DOI/PMID); and (iv) quality checks for duplicates, wrong SOCs, and unverifiable sources. Auxiliary evidence bases include SIDER, FAERS, and TOXRIC [2–4].

Split matching and metrics. A claim is counted as covered if it contains the leaflet preferred term (case-insensitive) within the correct SOC. We report: (i) term coverage, (ii) SOC macro coverage (macro-average of per-SOC recall), (iii) frequency-weighted coverage (weights from leaflet frequency bands), and (iv) citation accuracy on sourced rows. 95% CIs for proportions use Wilson/Newcombe intervals [6]; macro/weighted coverage CIs use a nonparametric bootstrap [7].

Baseline. A general-purpose, web-enabled LLM prompted to "search and compile side effects of a specific substance (e.g., sirolimus)" without EPAR schema enforcement, SOC mapping, split normalization, or DOI verification.

Contributions

EPAR-aligned safety dossiers: an AI system that writes directly into a regulator-style structure.
Transparent scoring: split-based evaluation where each claim maps to exactly one side-effect term.
Source checking: study citations are attached and assessed for correctness.
Reproducible artifacts: data tables and a simple recipe to re-run the evaluation on other agents and drugs.

4 Results

Overview. We report coverage and citation results for §4.8 on sirolimus. Fig. 1 compares term, SOC macro, and frequency-weighted coverage; Fig. 2 adds 95% CIs and shows citation accuracy (system, sourced rows).

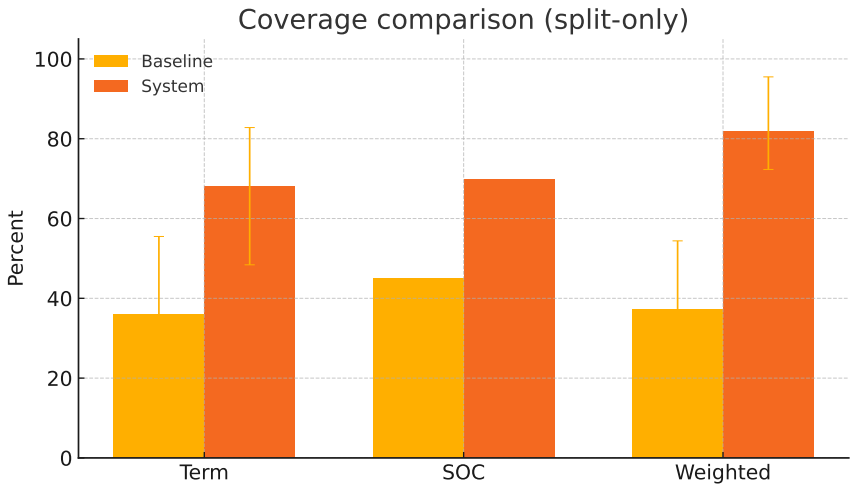

Figure 1: Term, SOC macro, and weighted coverage (§4.8, split-only).

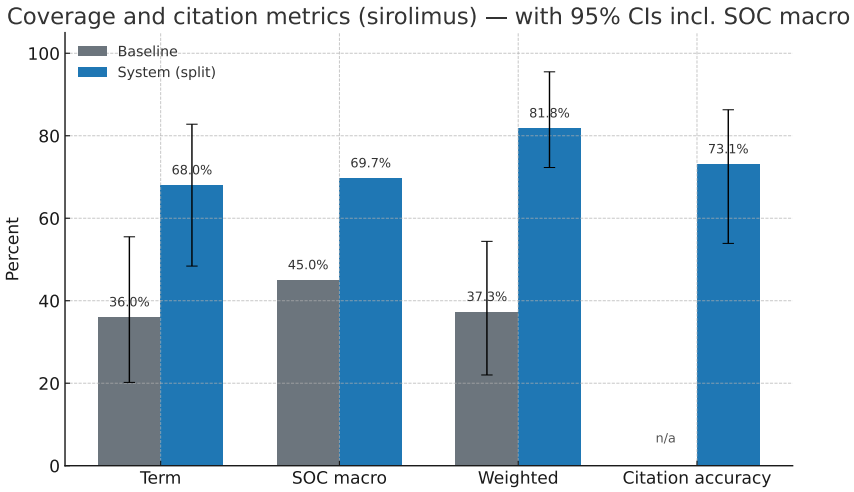

Figure 2: Coverage and citation metrics (sirolimus) with 95% confidence intervals for term, SOC macro (bootstrap), and weighted coverage; citation accuracy shown for the system (sourced rows).

Metrics at a glance. Term coverage: **68.0%** (95% CI: 48.4–82.8) vs. baseline **36.0%** (20.2–55.5); SOC macro coverage: **69.7%** vs. baseline **45.0%**; frequency-weighted coverage: **81.8%** (72.3–95.5) vs. baseline **37.3%** (22.0–54.4); citation accuracy (system, sourced rows): **73.1%** (53.9–86.3). Per-SOC distributions are in Fig. 3 and Fig. 4.

Table 1: Summary metrics for §4.8 (split-only). Brackets show 95% CIs where applicable.

Metric	Baseline	System
Term coverage (%)	36.0 [20.2, 55.5]	**68.0** [48.4, 82.8]
SOC macro coverage (%)	45.0	**69.7**
Weighted coverage (%)	37.3 [22.0, 54.4]	**81.8** [72.3, 95.5]
Citation accuracy (%)	—	**73.1** [53.9, 86.3]

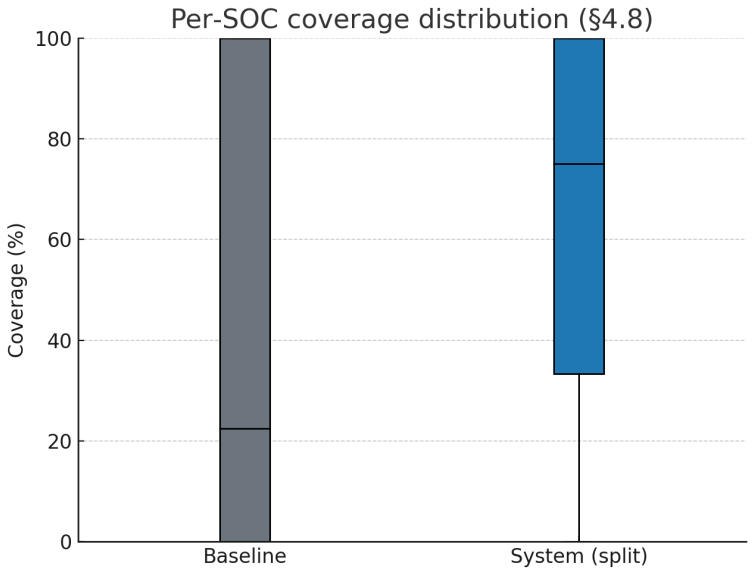

Figure 3: Per-SOC coverage distribution (§4.8). Boxes show median and IQR; whiskers indicate spread; points are outliers.

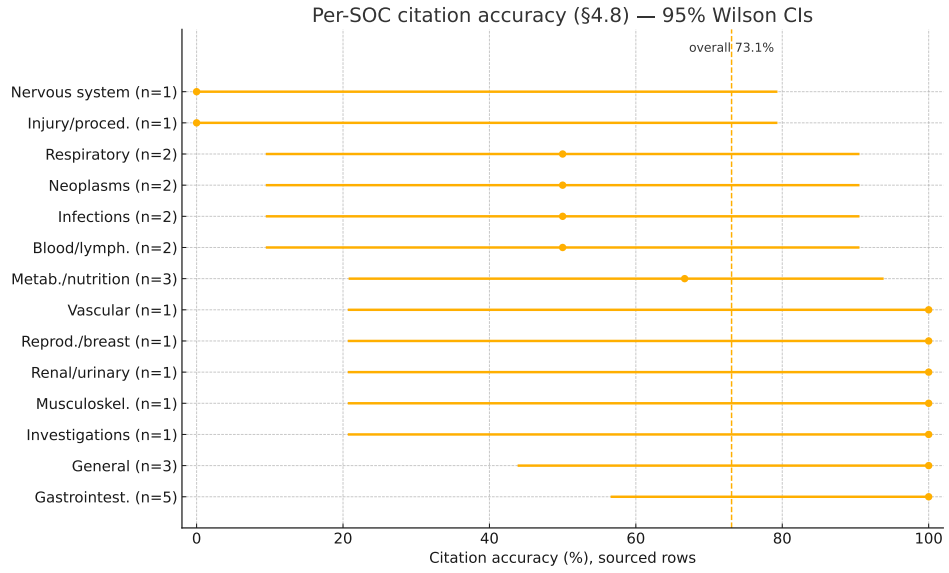

Figure 4: Per-SOC citation accuracy for the system (§4.8, sourced rows only). Points show accuracy; horizontal lines are 95% Wilson confidence intervals [6]; the dashed line indicates overall accuracy.

Table 2: Per-SOC coverage (leaflet §4.8). Percent of leaflet terms covered; Δ is System minus Baseline.

SOC	Baseline (%)	System (%)	Δ (pp)
Blood/lymph.	100.0	100.0	0.0
Gastrointest.	20.0	100.0	80.0
Investigations	100.0	100.0	0.0
Metab./nutrition	100.0	100.0	0.0
Skin/subcut. tissue	0.0	100.0	100.0
Vascular	0.0	100.0	100.0
General	25.0	50.0	25.0
Musculoskel.	50.0	50.0	0.0
Infections	0.0	33.3	33.3
Nervous system	0.0	33.3	33.3

5 Discussion

Baseline characterization. Without EPAR schema enforcement, SOC mapping, split normalization, and DOI verification, the baseline yields less structured, less reproducible outputs. Our agent's structure and checks likely drive the gains in coverage and the more stable sourcing.

Strengths and limitations. EPAR-aligned ground truth and split matching reduce undercounting of multi-term sentences; frequency weighting emphasizes clinically important effects. Limits include exact-term matching (synonyms are not credited), indirect treatment of frequency, sourcing measured only on cited rows, and potential lag between labels and databases.

Error examples. *False negative (FN):* "mouth sores" not matched to PT *stomatitis* (fixed by synonym map). *False positive (FP):* generic "infection" mapped to wrong SOC (fixed by enforcing leaflet SOC).

Ablation note. Dropping the split-by-term rule makes matching ambiguous and reduces term coverage in dry runs; structured splitting is a simple but impactful control.

6 Case Study: Rapamycin Dossier (EPAR-Aligned)

6.1 4.1 Therapeutic Indications

Sirolimus (rapamycin) is indicated for prophylaxis of organ rejection in adult kidney-transplant recipients, typically in combination with a calcineurin inhibitor (e.g., cyclosporine) and corticosteroids [8, 9]. In practice, sirolimus is often started early post-transplant with cyclosporine and steroids; once stable graft function is achieved, tapering or withdrawal of the calcineurin inhibitor can be considered, with sirolimus maintenance thereafter [10]. For sporadic lymphangioleiomyomatosis (LAM), sirolimus stabilizes lung function and improves outcomes in adults with abnormal or declining function [11]; contemporary guidance recommends sirolimus as first-line therapy for such patients [12].

6.2 4.8 Undesirable Effects

Frequent effects include changes in blood counts (thrombocytopenia, anemia, leukopenia) and raised blood lipids (hyperlipidemia, hypertriglyceridemia), often requiring monitoring and lipid-lowering therapy [13, 14]. High blood sugar and new-onset diabetes have been observed in transplant recipients on sirolimus [14, 15]. Dose-related mouth ulcers (stomatitis), skin eruptions, and gastrointestinal symptoms such as nausea, diarrhea, constipation, and abdominal pain are common [13, 14]. Hypertension and peripheral edema also occur and may require treatment or dose change [13, 14]. Some effects correlate with higher trough levels and improve with dose reduction; severe or refractory events—such as intractable stomatitis, non-infectious pneumonitis, or proteinuria—can require interruption or discontinuation [13–15]. As an immunosuppressant, sirolimus increases susceptibility to infections; long-term immunosuppression is associated with higher malignancy risk, especially

lymphomas and skin cancers [13]. In kidney-transplant recipients with prior cutaneous squamous-cell carcinoma, conversion to sirolimus reduced new skin cancers albeit with higher adverse-event and discontinuation rates [16]. Sirolimus can impair wound healing and promote lymphatic complications (delayed wound healing, lymphocele). Non-infectious interstitial pneumonitis is uncommon but important and typically resolves after dose reduction or withdrawal [14]. Although not directly nephrotoxic, sirolimus has been linked to de novo or worsening proteinuria in some patients, sometimes progressing to nephrotic-range proteinuria or focal segmental glomerulosclerosis, prompting discontinuation [15]. Reproductive effects: in men, impaired spermatogenesis and reduced paternity rates have been reported [17].

Reproducibility Statement

We release the exact inputs and derived outputs used for all results. Ground truth is the leaflet §4.8 preferred-term list with SOC and frequency bands. System and baseline predictions are provided as split-by-term claim tables with SOC mapping and citation flags. Matching is deterministic and auditable; synonyms are not credited. We compute four metrics (term, SOC macro, frequency-weighted coverage, and citation accuracy) and report 95% CIs (Wilson for proportions; bootstrap for macro/weighted). A single script rebuilds all tables and figures from three CSV files.

Responsible AI Statement

The system is intended for drafting and triage under expert supervision and must not replace clinical judgment or regulatory review. We report limitations and risks and adhere to community ethics guidelines.

AI Contribution Disclosure (Checklist)

Hypothesis/scoping: agent-led with human oversight.
Data handling (leaflet parsing): agent-led with human oversight.
Method design (split, metrics): agent-led.
Experiments: agent-led.
Writing: agent-led with human oversight.
Visualization: agent-led.
Human role: oversight and final editing.

References

[1] European Medicines Agency. European public assessment reports (epar): structure and content. Available at https://www.ema.europa.eu/en/medicines/european-public-assessment-reports, 2025. Accessed 16 September 2025.

[2] U.S. Food and Drug Administration. FDA adverse event reporting system (faers). Available at https://www.fda.gov/drugs/surveillance/fda-adverse-event-reporting-system-faers, 2025. Accessed 16 September 2025.

[3] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079, 2016.

[4] Yonghyun Kim and et al. TOXRIC: a comprehensive toxicology knowledgebase. *Nucleic Acids Research*, 51(D1):D1502–D1510, 2023.

[5] E. G. Brown, L. Wood, and S. Wood. The medical dictionary for regulatory activities (meddra). *Drug Safety*, 20(2):109–117, 1999.

[6] Robert G. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8):857–872, 1998.

[7] Bradley Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7(1): 1–26, 1979.

[8] Barry D. Kahan and Rapamune US Study Group. Efficacy of sirolimus compared with azathioprine for reduction of acute renal allograft rejection: a randomised multicentre study. *The Lancet*, 356(9225):194–202, 2000. doi: 10.1016/S0140-6736(00)02480-6.

[9] A. S. MacDonald and Rapamune Global Study Group. A worldwide, phase III, randomized, controlled, safety and efficacy study of a sirolimus/cyclosporine regimen for prevention of acute rejection in recipients of primary mismatched renal allografts. *Transplantation*, 71(2): 271–280, 2001. doi: 10.1097/00007890-200101270-00019.

[10] R. Oberbauer and Rapamune Maintenance Regimen Study Group. Long-term improvement in renal function with sirolimus after early cyclosporine withdrawal in renal transplant recipients: 2-year results. *Transplantation*, 76(2):364–370, 2003. doi: 10.1097/01.TP.0000074360.62032.39.

[11] Francis X. McCormack and et al. Efficacy and safety of sirolimus in lymphangioleiomyomatosis. *New England Journal of Medicine*, 364(17):1595–1606, 2011. doi: 10.1056/NEJMoa1100391.

[12] Francis X. McCormack, Nishant Gupta, Gregory R. Finlay, and et al. Official ats/jrs clinical practice guidelines: Lymphangioleiomyomatosis. *American Journal of Respiratory and Critical Care Medicine*, 194(6):748–761, 2016. doi: 10.1164/rccm.201607-1384ST.

[13] Bruce Kaplan, Yasir Qazi, and Jason R. Wellen. Strategies for the management of adverse events associated with mTOR inhibitors. *Transplantation Reviews*, 28(3):126–133, 2014. doi: 10.1016/j.trre.2014.03.002.

[14] Laurent S. Nguyen, Mélanie Vautier, Yann Allenbach, and et al. Sirolimus and everolimus in solid organ transplantation: safety profile and synthesis of adverse events. *Drug Safety*, 42(7): 813–825, 2019. doi: 10.1007/s40264-019-00806-8.

[15] Owen Johnston and et al. Sirolimus is associated with new-onset proteinuria in kidney transplant recipients. *Journal of the American Society of Nephrology*, 19(7):1411–1418, 2008. doi: 10.1681/ASN.2007111202.

[16] Sébastien Euvrard and et al. Sirolimus and secondary skin cancers in kidney transplantation: randomized trial. *New England Journal of Medicine*, 367(4):329–339, 2012. doi: 10.1056/NEJMoa1204166.

[17] J. Zuber, D. Anglicheau, C. Elie, L. Bererhi, M.-O. Timsit, and et al. Sirolimus may reduce fertility in male renal transplant recipients. *American Journal of Transplantation*, 8(7):1471–1479, 2008. doi: 10.1111/j.1600-6143.2008.02267.x.