データの説明

データ番号 （コモンズセンターで記入します）	DPPSCdbp_2025-06
タイトル	闘病記医療イベント要約コーパスに対する汎用臨床医学情報アノテーション
作成者	矢田竣太郎（筑波大学図書館情報メディア系）
問い合わせ先	矢田竣太郎　yada@slis.tsukuba.ac.jp
概要（データの内容や作成方法）	100 名の乳がん患者による闘病記ブログの医療イベント要約（TobyokiSummary）に対し、PRISM アノテーションガイドラインに基づく汎用臨床医学情報アノテーションを付与したコーパスである。13 種類のエンティティ（病変、臓器、検査、投薬、時間表現など）と 10 種類の関係（時間関係、部位関係など）が付与されている。アノテーションは LLM（大規模言語モデル）を用いて自動的に行い、そのためのツール prism-annotator を開発・公開した。
更新履歴（版・ファイル名・年月日）	v1.0 / tobyoki-prism-dist.zip / 2026 年 3 月
データの形式	XML（個別文書ファイルおよび統合コーパスファイル）、JSON（エンティティ・関係情報）、HTML（閲覧用ビューア）
データのサイズ	約 11MB（ZIP 圧縮時約 1.5MB）。文書数：100 件、エンティティ総数：14,797 個、関係総数：7,035 個
利用上の注意（メタデータの利用など）	本データは闘病記ブログから抽出した医療イベント要約（https://github.com/sociocom/TobyokiSummary）に対するアノテーションであり、ブログ原文テキストは含まれない。アノテーションは LLM により自動生成されたものであり、人手による厳密な検証は行っていない。
関連報告書・論文等	矢田, 菅原, 木﨑, 西山 (2025). 闘病記ブログの感情要因イベントを LLM で抽出する試み. 第 20 回言語処理若手シンポジウム (YANS 2025). Shimizu, Hisada, Uno, Yada, Wakamiya, & Aramaki (2025). Exploring LLM Annotation for Adaptation of Clinical Information Extraction Models under Data-sharing Restrictions. Findings of ACL 2025. Yada et al. (2020). Towards a Versatile Medical-Annotation Guideline. LREC 2020.
備考	アノテーションツール：https://github.com/sociocom/prism-annotator（PyPI: prism-annotator）