# BONSAI: A framework for processing and analysing Electronic Health Records (EHR) data using transformer-based models

Maria Elkjær Montgomery [1*], **Kiril Klein** [1*], **Mikkel Odgaard** [1*], **Stephan Sloth Lorenzen** [1*], **and Zahra Sobhaninia** [1]

1 University of Copenhagen, Denmark ***** These authors contributed equally.

Summary

BONSAI is an end-to-end Python framework for processing and analysing Electronic Health Records (EHR) data. It extends the model described in the CORE-BEHRT paper (Odgaard et al., 2024) and is designed to streamline data preparation, model pre-training, and fine-tuning for patient-level prediction tasks. The framework is built for efficient large-scale processing of EHR data, making it suitable for clinical applications involving substantial volumes of patient records. This enables models for patient outcome prediction and supports scalable clinical research.

The framework accepts EHR data in the MEDS format (Kolo et al., 2024), performs comprehensive preprocessing, and prepares the data for BERT-based modeling, following the structure of the BEHRT model (Li et al., 2020). This includes converting raw EHR data into tokenised inputs by mapping vocabulary to numerical tokens and aligning patient histories into temporally ordered sequences of medical concepts. Each concept is paired with temporal features such as visit timestamps (positions), patient age, and visit-level segment encodings. Additional static features, including date of birth (DOB), gender, and optionally date of death (DOD), are prepended to each sequence. Separator and classification tokens can be optionally included. Numeric values can be binned and added as categorical tokens. Aggregation of similar concepts is supported via regex-based grouping, and the removal of specific concepts is also included via regex.

Modeling is performed using a ModernBERT backbone (Warner et al., 2024) from the Hugging Face library. Model configurations are specified via a YAML config file, with defaults provided. Pre-training uses a masked language modeling (MLM) objective with Cross-Entropy loss. Fine-tuning is performed as a binary classification task using outcome-specific labels derived from the input sequences. Cohort definition and censoring can be done in two ways: either by including all data with post-hoc censoring based on outcome dates and a user-defined window, or via a simulated prospective approach with a fixed cutoff date. The fine-tuning head consists of a BiGRU layer that encodes patient sequences into a single embedding, which is passed to a linear classification layer trained using binary cross-entropy loss and the AdamW optimizer. Optional features include the use of a learning rate scheduler and the ability to freeze pre-trained layers during fine-tuning. Finally, the pipeline also includes an evaluation script that outputs model predictions and, optionally, intermediate model embeddings.

Statement of need

The growing adoption of foundation models in Natural Language Processing (NLP) (Brown et al., 2020; Devlin et al., 2019; Touvron et al., 2023), coupled with the increasing availability of

EHR data, has led to a surge in adapting such models to the clinical domain (Gu & Dao, 2023; Li et al., 2020; Odgaard et al., 2024; Pang et al., 2021, 2024; Rasmy et al., 2021). However, existing frameworks often differ significantly in model architecture, data representations, and preprocessing steps, making comparison and reproducibility challenging. Notably, recent efforts such as the CEHR Benchmark (Pang et al., 2024) have begun to standardise evaluation protocols, underscoring the diversity in approaches across the field.

BONSAI was developed to provide an end-to-end pipeline in a modular setup, enabling flexible experimentation with EHR modeling. Users can easily switch between data representations, sources, normalisation strategies, and fine-tuning heads. Although ModernBERT is the default backbone, the framework supports alternative architectures with minimal configuration changes. It also includes baseline models for comparison and supports deployment on Microsoft Azure, a platform commonly used for working with protected health data, making it practical for working with real-world clinical data.

The framework is intended for machine learning researchers and clinical data scientists who are working with or interested in EHR data and wish to explore, benchmark, or develop transformer-based models in a scalable and reproducible manner. While BONSAI builds on the earlier CORE-BEHRT framework (Odgaard et al., 2024), it introduces improvements in modularity and configurability, as well as support for more data sources (such as numerical input). In contrast to frameworks like Med-BERT (Rasmy et al., 2021) or BEHRT (Li et al., 2020), which have more rigid assumptions about data preprocessing and modeling setup, BONSAI offers a general-purpose, flexible design adaptable to a wide range of EHR tasks.
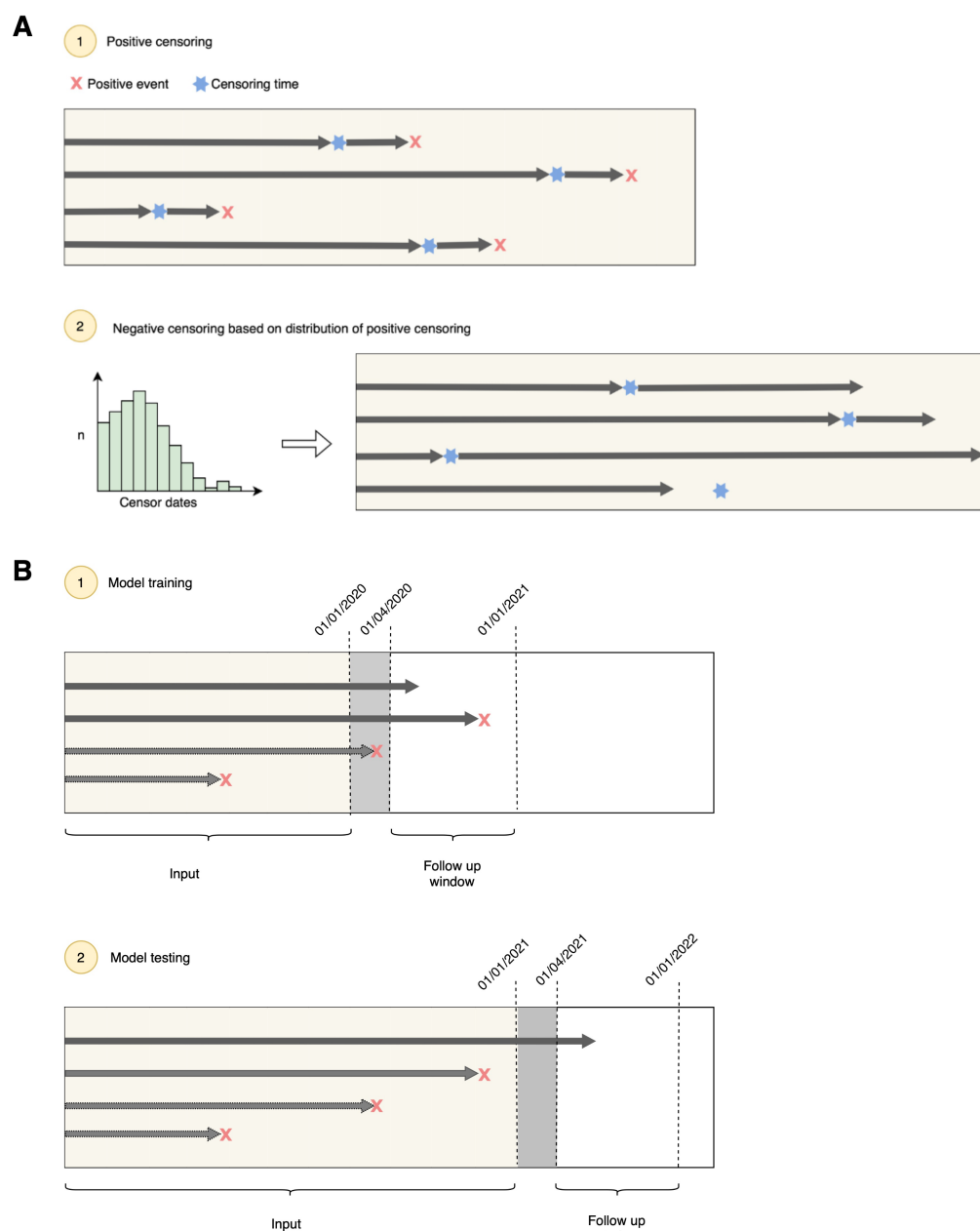
Figures

Figure 1 depicts the overall pipeline of BONSAI.
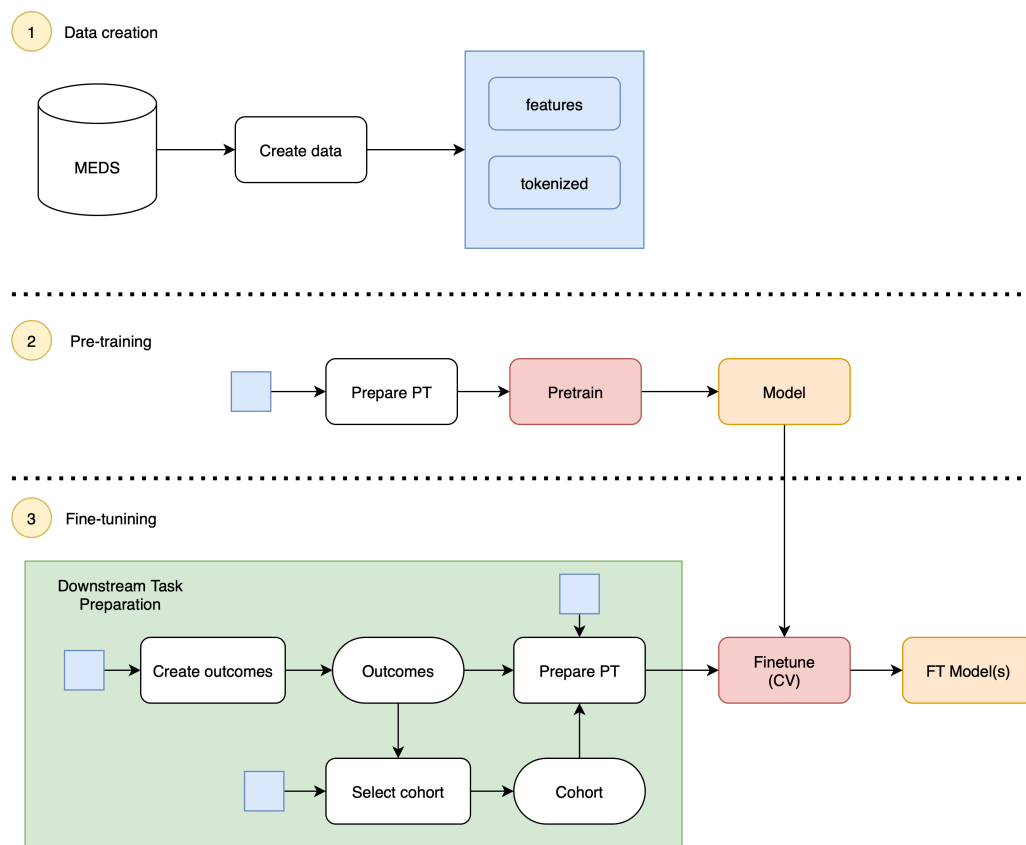
Figure 1: The BONSAI pipeline.

The figures below depict the censoring scheme for the data preprocessing, where Figure 2 A shows the censoring scheme for the post-hoc setup, and Figure 2 B shows the censoring scheme in a prospective setup.

Figure 2: Censoring in the post-hoc setup (A) and simulated prospective setup (B).

Software dependencies

The codebase is primarily built on PyTorch v2.5.1 for deep learning (Paszke et al., 2019), with transformer architectures implemented using the Hugging Face Transformers library (v4.48.0 and above) (*Transformers*, n.d.). Additional support for Azure workflows is provided via azureml-mlflow (*Mlflow.azureml*, n.d.).

Other key dependencies include:

- numpy (2.1.3)
- pandas (2.2.3)
- pyarrow (18.0.0)

- python_dateutil (==2.9.0.post0)
- PyYAML (6.0.2)
- scikit-learn (1.5.2)
- setuptools (==75.6.0)
- tqdm (==4.67.1)
- xgboost (3.0.2)

All dependencies are listed in the requirements.txt file. Installation instructions and environment setup are detailed in the main README.md.

Acknowledgements

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv Preprint arXiv:2312.00752*. https://doi.org/10.48550/arXiv.2312.00752

Kolo, A., Pang, C., Choi, E., Steinberg, E., Jeong, H., Gallifant, J., Fries, J. A., Chiang, J. N., Oh, J., Xu, J., & others. (2024). *MEDS decentralized, extensible validation (MEDS-DEV) benchmark: Establishing reproducibility and comparability in ML for health*. https://openreview.net/pdf?id=DExp3tRRel

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., & Salimi-Khorshidi, G. (2020). BEHRT: Transformer for electronic health records. *Scientific Reports*, *10*(1), 7155. https://doi.org/10.1038/s41598-020-62922-y

Mlflow.azureml. (n.d.). https://mlflow.org/docs/1.22.0/python_api/mlflow.azureml.html.

Odgaard, M., Klein, K. V., Thysen, S. M., Jimenez-Solem, E., Sillesen, M., & Nielsen, M. (2024). Core-behrt: A carefully optimized and rigorously evaluated behrt. *arXiv Preprint arXiv:2404.15201*. https://doi.org/10.48550/arXiv.2404.15201

Pang, C., Jiang, X., Kalluri, K. S., Spotnitz, M., Chen, R., Perotte, A., & Natarajan, K. (2021). CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. *Machine Learning for Health*, 239–260. https://doi.org/10.48550/arXiv.2111.08585

Pang, C., Jiang, X., Pavinkurve, N. P., Kalluri, K. S., Minto, E. L., Patterson, J., Zhang, L., Hripcsak, G., Gürsoy, G., Elhadad, N., & others. (2024). CEHR-GPT: Generating electronic health records with chronological patient timelines. *arXiv Preprint arXiv:2402.04400*. https://doi.org/10.48550/arXiv.2402.04400

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An

---

imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. https://doi.org/10.48550/arXiv.1912.01703

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, *4*(1), 86. https://doi.org/10.1038/s41746-021-00455-y

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., & others. (2023). Llama: Open and efficient foundation language models. *arXiv Preprint arXiv:2302.13971*. https://doi.org/10.48550/arXiv.2302.13971

Transformers. (n.d.). https://huggingface.co/docs/transformers/en/index.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., & others. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv Preprint arXiv:2412.13663*. https://doi.org/10.48550/arXiv.2412.13663