Optimising your training data using model-led iterative confidence-based sample selection

Keiran Tait, Joseph Cronin PhD, Robert Dürichen PhD, Arcturis Data, Building One, Oxford Technology Park, Technology Drive, Kidlington, OX5 1GN, UK

Introduction

Small Language Models (SLMs) have recently shown strong performance in domain-specific NLP tasks while also being more resource-efficient than Large Language Models (LLMs). However, due to dataset variability, they often require finetuning to meet task-specific performance needs. Data annotation is one of the most time-consuming and costly aspects of this process [1], [2], especially in fields like clinical pathology where data availability is limited [3]. While BERT-based models' confidence scores have been used to identify model weaknesses [4], their use in guiding finetuning remains underexplored. We propose **M**odelled **I**terative **C**onfidence based **S**ample **S**election (MICS^2), a human-in-the-loop approach that leverages BioBERT-generated confidence scores [5] to curate training data based on the model's ability to handle specific clinical questions or features. MICS^2 aligns with active learning and uncertainty sampling strategies used in CNN training [6].

Objectives

- Develop MICS2 method leveraging BioBERT's confidence scoring to optimize dataset selection for finetuning.
- Evaluate the accuracy of a MICS2 derived training dataset on an independent test set, comparing it with random sampling and bulk ("en masse") finetuning approaches.
- Analyse computational costs for each strategy and assess the impact of model-led selection on overall efficiency.

Methods

- 9562 anonymised annotated free text pathology reports, written in English, covering 28 clinical features were annotated mean = 350.8 ± 83.4 reports per clinical feature) . A breakdown of each of the covered features can be found in the bar plot in the **Results** section.
- Click here for an animated overview of the **Methods** detailed below.
- For each clinical feature, the initial dataset is split, with 50 reports with a positively identifiable answer (PI) to the question "what is the status of {feature}?" and 50 with an impossible to determine answer (IA) partitioned as an independent test set.
- 5 permutations of the remaining data for each feature are then split into subsets. 1) a permutation test set with 25 each of PI and IA. 2) a validation set with 25 each of PI and IA, and the remaining reserve set with all remaining annotated data.
- The data is run through a two-stage question answering/classification process [7] to extract and classify the BioBERT model's answer. The confidence scores from the question answering are used to then classify the results and identify those the model performed worst on.
- The worst answers from the validation set are added to the training set, the model is trained, the trained model is run over the data, and the cycle repeats. This is repeated until the model reaches an accuracy of 96% on the permutation test set .
- The final training set is then generated by concatenating the 5 permutation sets. A BioBERT model is trained with this set and evaluated against a similar model trained using the en masse approach (all data not used in the independent test set) or the random sampling approach (randomly sampling each feature based on the number of training samples required for the MISC^2 method).
- The evaluation was run on an AWS g6.2xlarge instance, utilising a NVIDIA L4 GPU with 24 GiB of video memory

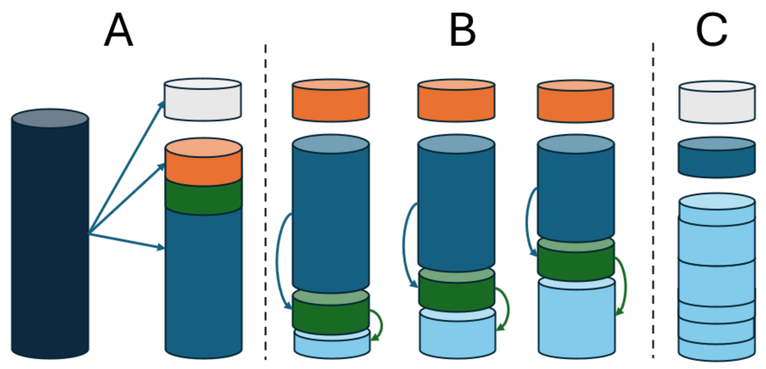


Figure 1 - Diagram showing how the ground truth dataset (dark blue) is partitioned in section A, with the independent test set shown in white, and the data for each permutation in orange (permutation test set), green (initial permutation validation set) and blue (permutation reserve set). Section B shows the flow of reports from the validation to training set, and from reserve to validation set between each iteration. C shows the final construction of the test set from all permutations, the reserve set, and the independent test set.

Results

- The model trained on the compounded MICS^2 dataset was evaluated against each of the independent test sets, achieving an average accuracy of 98.46% [95% CI = 0.5%] and required an average of 76.43 (±43.59) training examples per clinical feature.
- In comparison, an en masse approach that took all annotated data that was not partitioned into the independent test sets and used that for finetuning (242.63 ± 89.79 training examples per feature) resulted in a model with 95.46% average accuracy [95% CI = 1.6%]. This took an average of 13.15 (± 19.25) iterations per permutation.
- The poorest performance was associated with a random sampling approach. This took the same number of samples as was used in MICS^2 (76.43 ±43.59 samples per feature) but randomly sampled them from the remaining data not used in the test sets. This led to an Accuracy of 92.29%, [95% CI = 1.91%].
- MICS^2 took significantly longer to run (~32 hrs and 34 mins for all clinical features) compared to the en masse (~20 mins for all features) or random sampling (~29 mins for all features) approaches.
- MICS^2 s significantly longer run time is due to the iterative nature of the data sampling, and the requirement to retrain the model for each iteration which increases in length with each iteration due to the increasing finetuning set size.

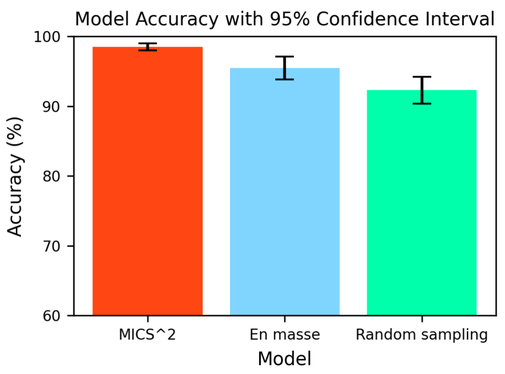
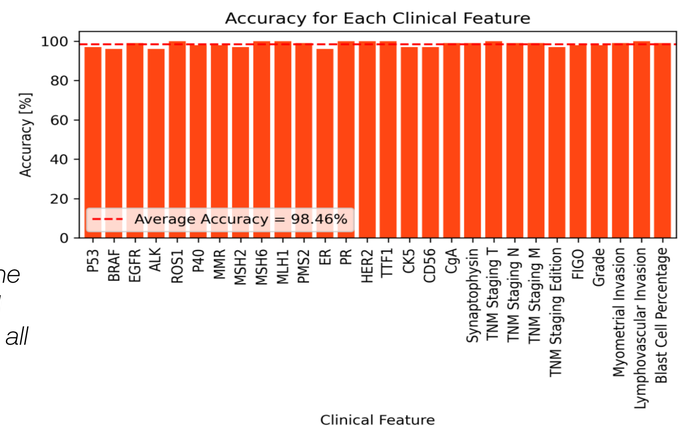


Fig. 3 (right) – Accuracy of MICS2 against the independent test set for each of the clinical features, with the average accuracy across all features displayed with the dashed line

Fig. 2 (left) – Accuracy of MICS^2 (left, orange) compared with training a model with all remaining data (middle, blue), and randomly sampling the remaining data based off the number of training samples used in the MICS^2 training set (right, green)



Conclusions

- MICS2 produced by the BioBERT model to aid dataset curation helps to generate models with higher accuracy despite smaller training sets.
- Model performance is not directly linked to larger finetuning dataset sizes, as demonstrated by the poorer performance shown by the en masse approach.
- Equally, model performance does not peak with a certain dataset size as demonstrated by the random sampling approach using the same finetuning set size as MICS2
- Using MICS2 is significantly more time consuming and therefore more costly to run but allows for curation of smaller finetuning sets without compromising model performance, which is ideal when data is limited.

Acknowledgments and References

This work uses anonymised data collected by NHS Trusts as part of routine care. We believe that the safe, transparent, and ethical use of anonymised patient data is vital to improve health and care for everyone, and we would like to thank Oxford University Hospitals NHS Foundation Trust and the Thames Valley and Surrey Secure Data Environment for their contribution. This research was supported by the National Institute for Health and Care Research (NIHR) Oxford Biomedical Research Centre. e views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

1. Q. Ding, D. Ding, Y. Wang, C. Guan, and B. Ding, 'Unraveling the landscape of large language models: a systematic review and future perspectives', Journal of Electronic Business & Digital Economics, vol. 3, no. 1, pp. 3–19, Dec. 2023, doi: 10.1108/JEBDE-08-2023-0015.
2. S. Ren, B. Tomlinson, R. W. Black, and A. W. Torrance, 'Reconciling the contrasting narratives on the environmental impact of large language models', Sci Rep, vol. 14, no. 1, p. 26310, Nov. 2024, doi: 10.1038/s41598-024-76682-6.
3. F. Bai, A. Ritter, and W. Xu, 'Pre-train or Annotate? Domain Adaptation with a Constrained Budget', May 13, 2022, arXiv: arXiv:2109.04711. doi: 10.48550/arXiv.2109.04711.
4. K. Shen and M. Kejriwal, 'Quantifying confidence shifts in a BERT-based question answering system evaluated on perturbed instances', PLOS ONE, vol. 18, no. 12, p. e0295925, Dec. 2023, doi: 10.1371/journal.pone.0295925.
5. J. Lee et al., 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.
6. K. De Angeli et al., 'Deep active learning for classifying cancer pathology reports', BMC Bioinformatics, vol. 22, no. 1, p. 113, Mar. 2021, doi: 10.1186/s12859-021-04047-1.
7. Tait, K., Cronin, J., Wiper, O., Wallis, J., Davies, J., & Dürichen, R. ArcTEX—a novel clinical data enrichment pipeline to support real-world evidence oncology studies. Frontiers in Digital Health, 7. May 2025. https://doi.org/10.3389/fdgth.2025.1561358

Introduction

Small Language Models (SLMs) have recently shown strong performance in domain-specific NLP tasks while also being more resource-efficient than Large Language Models (LLMs). However, due to dataset variability, they often require finetuning to meet task-specific performance needs. Data annotation is one of the most time-consuming and costly aspects of this process [1], [2], especially in fields like clinical pathology where data availability is limited [3]. While BERT-based models' confidence scores have been used to identify model weaknesses [4], their use in guiding finetuning remains underexplored. We propose Modelled Iterative Confidence based Sample Selection (MICS^2), a human-in-the-loop approach that leverages BioBERT-generated confidence scores [5] to curate training data based on the model's ability to handle specific clinical questions or features. MICS^2 aligns with active learning and uncertainty sampling strategies used in CNN training [6].

Objectives

- Develop MICS2 method leveraging BioBERT’s confidence scoring to optimize dataset selection for finetuning.
- Evaluate the accuracy of a MICS2 derived training dataset on an independent test set, comparing it with random sampling and bulk ("en masse") finetuning approaches.
- Analyse computational costs for each strategy and assess the impact of model-led selection on overall efficiency.

Methods

- 9562 anonymised annotated free text pathology reports, written in English, covering 28 clinical features were annotated mean = 350.8 ± 83.4 reports per clinical feature) . A breakdown of each of the covered features can be found in the bar plot in the Results section.
- Click here for an animated overview of the Methods detailed below.
- For each clinical feature, the initial dataset is split, with 50 reports with a positively identifiable answer (PI) to the question “what is the status of {feature}?” and 50 with an impossible to determine answer (IA) partitioned as an independent test set.
- 5 permutations of the remaining data for each feature are then split into subsets. 1) a permutation test set with 25 each of PI and IA. 2) a validation set with 25 each of PI and IA, and the remaining reserve set with all remaining annotated data.
- The data is run through a two-stage question answering/classification process [7] to extract and classify the BioBERT model’s answer. The confidence scores from the question answering are used to then classify the results and identify those the model performed worst on.
- The worst answers from the validation set are added to the training set, the model is trained, the trained model is run over the data, and the cycle repeats. This is repeated until the model reaches an accuracy of 96% on the permutation test set .
- The final training set is then generated by concatenating the 5 permutation sets. A BioBERT model is trained with this set, and evaluated against a similar model trained using the en masse approach (all data not used in the independent test set) or the random sampling approach (randomly sampling each feature based on the number of training samples required for the MISC^2 method).
- The evaluation was run on an AWS g6.2xlarge instance, utilising a NVIDIA L4 GPU with 24 GiB of video memory

Methods

- The final training set is then generated by concatenating the 5 permutation sets. A BioBERT model is trained with this set and evaluated against a similar model trained using the en masse approach (all data not used in the independent test set) or the random sampling approach (randomly sampling each feature based on the number of training samples required for the MISC[2] method).
- The evaluation was run on an AWS g6.2xlarge instance, utilising a NVIDIA L4 GPU with 24 GiB of video memory

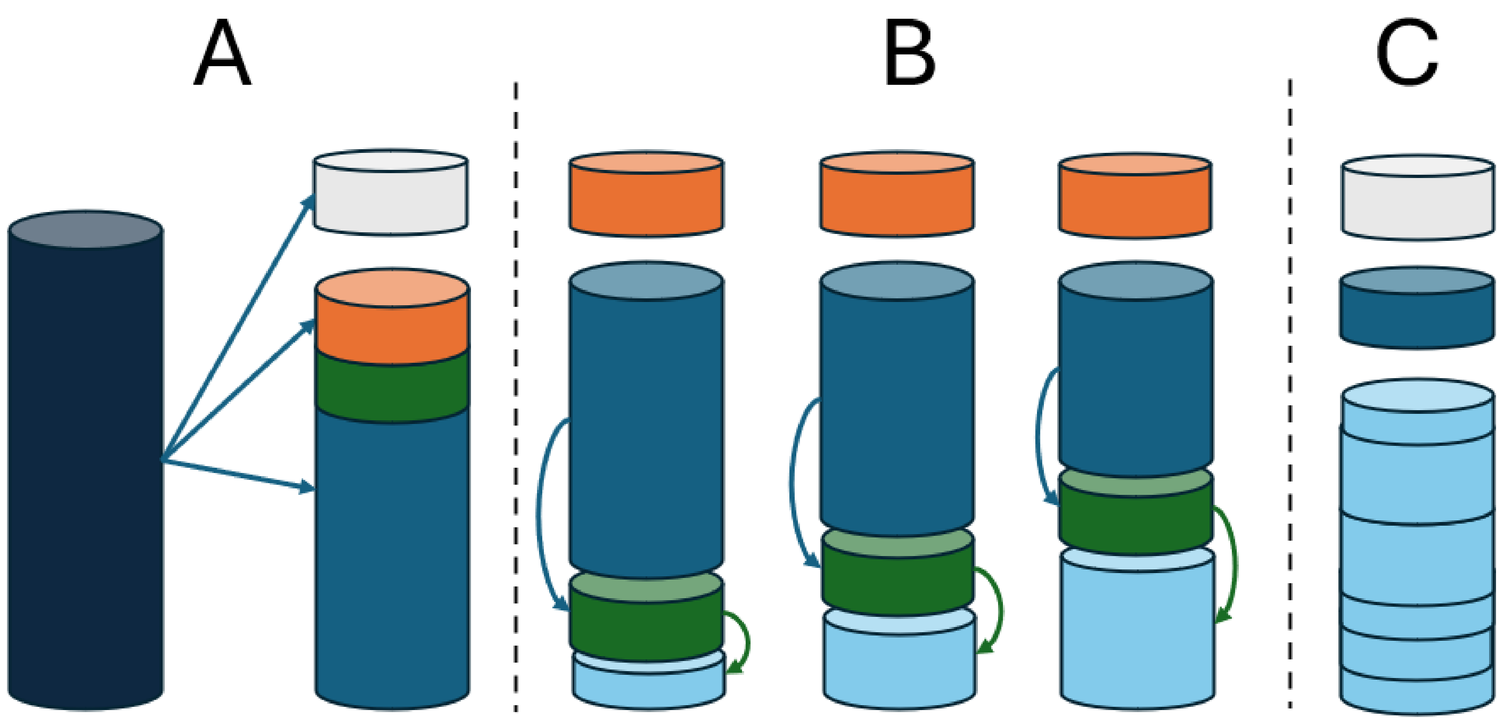


Figure 1 - Diagram showing how the ground truth dataset (dark blue) is partitioned in section A, with the independent test set shown in white, and the data for each permutation in orange (permutation test set), green (initial permutation validation set) and blue (permutation reserve set). Section B shows the flow of reports from the validation to training set, and from reserve to validation set between each iteration. C shows the final construction of the test set from all permutations, the reserve set, and the independent test set.

Results

- The model trained on the compounded MICS^2 dataset was evaluated against each of the independent test sets, achieving an average accuracy of 98.46% [95% CI = 0.5%] and required an average of 76.43 (±43.59) training examples per clinical feature.
- In comparison, an en masse approach that took all annotated data that was not partitioned into the independent test sets and used that for finetuning (242.63 ± 89.79 training examples per feature) resulted in a model with 95.46% average accuracy [95% CI = 1.6%]. This took an average of 13.15 (± 19.25) iterations per permutation.
- The poorest performance was associated with a random sampling approach. This took the same number of samples as was used in MICS^2 (76.43 ±43.59 samples per feature) but randomly sampled them from the remaining data not used in the test sets. This led to an Accuracy of 92.29%, [95% CI = 1.91%].
- MICS^2 took significantly longer to run (~32 hrs and 34 mins for all clinical features) compared to the en masse (~20 mins for all features) or random sampling (~29 mins for all features) approaches.
- MICS^2 s significantly longer run time is due to the iterative nature of the data sampling, and the requirement to retrain the model for each iteration which increases in length with each iteration due to the increasing finetuning set size.

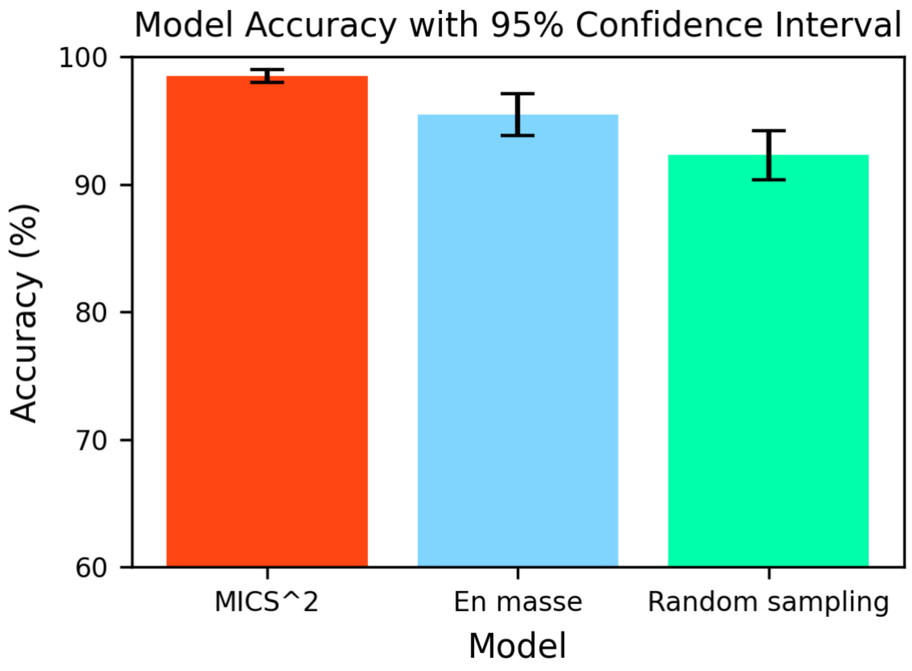
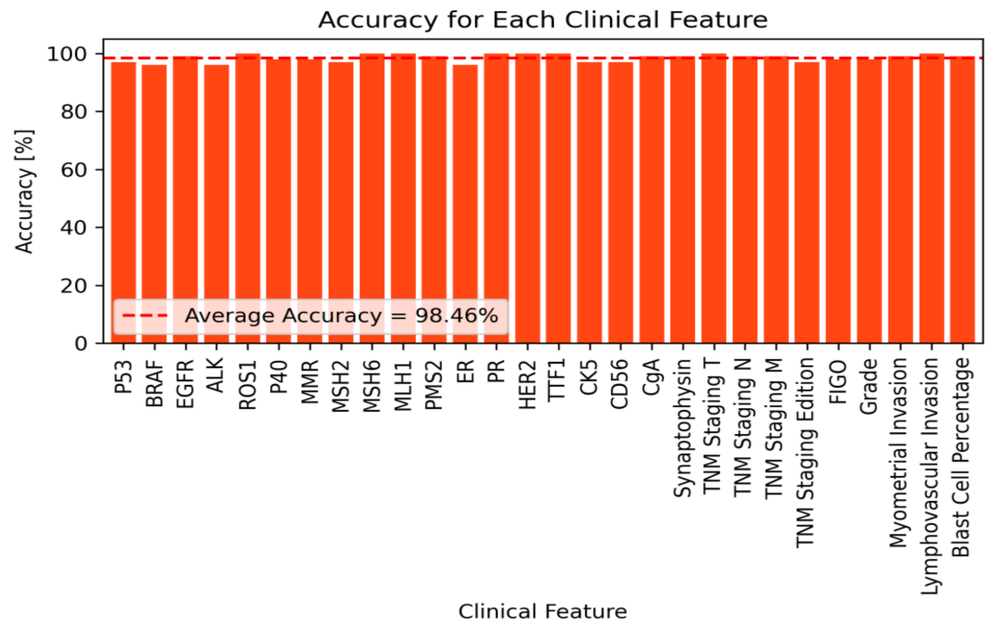


Fig. 3 (right) – Accuracy of MICS2 against the independent test set for each of the clinical features, with the average accuracy across all features displayed with the dashed line

Fig. 2 (left) – Accuracy of MICS2 (left, orange) compared with training a model with all remaining data (middle, blue), and randomly sampling the remaining data based off the number of training samples used in the MICS2 training set (right, green)



Conclusion

- MICS2 produced by the BioBERT model to aid dataset curation helps to generate models with higher accuracy despite smaller training sets.
- Model performance is not directly linked to larger finetuning dataset sizes, as demonstrated by the poorer performance shown by the en masse approach.
- Equally, model performance does not peak with a certain dataset size as demonstrated by the random sampling approach using the same finetuning set size as MICS2
- Using MICS2 is significantly more time consuming and therefore more costly to run but allows for curation of smaller finetuning sets without compromising model performance, which is ideal when data is limited.

Acknowledgments and References

This work uses anonymised data collected by NHS Trusts as part of routine care. We believe that the safe, transparent, and ethical use of anonymised patient data is vital to improve health and care for everyone, and we would like to thank Oxford University Hospitals NHS Foundation Trust and the Thames Valley and Surrey Secure Data Environment for their contribution. This research was supported by the National Institute for Health and Care Research (NIHR) Oxford Biomedical Research Centre. e views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

1. Q. Ding, D. Ding, Y. Wang, C. Guan, and B. Ding, ‘Unraveling the landscape of large language models: a systematic review and future perspectives’, Journal of Electronic Business & Digital Economics, vol. 3, no. 1, pp. 3–19, Dec. 2023, doi: 10.1108/JEBDE-08-2023-0015.
2. S. Ren, B. Tomlinson, R. W. Black, and A. W. Torrance, ‘Reconciling the contrasting narratives on the environmental impact of large language models’, Sci Rep, vol. 14, no. 1, p. 26310, Nov. 2024, doi: 10.1038/s41598-024-76682-6.
3. F. Bai, A. Ritter, and W. Xu, ‘Pre-train or Annotate? Domain Adaptation with a Constrained Budget’, May 13, 2022, arXiv: arXiv:2109.04711. doi: 10.48550/arXiv.2109.04711.
4. K. Shen and M. Kejriwal, ‘Quantifying confidence shifts in a BERT-based question answering system evaluated on perturbed instances’, PLOS ONE, vol. 18, no. 12, p. e0295925, Dec. 2023, doi: 10.1371/journal.pone.0295925.
5. J. Lee et al., ‘BioBERT: a pre-trained biomedical language representation model for biomedical text mining’, Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.
6. K. De Angeli et al., ‘Deep active learning for classifying cancer pathology reports’, BMC Bioinformatics, vol. 22, no. 1, p. 113, Mar. 2021, doi: 10.1186/s12859-021-04047-1.
7. Tait, K., Cronin, J., Wiper, O., Wallis, J., Davies, J., & Dürichen, R. ArcTEX—a novel clinical data enrichment pipeline to support real-world evidence oncology studies. Frontiers in Digital Health, 7. May 2025. https://doi.org/10.3389/fdgth.2025.1561358