Exploring Fairness in Long-Term Prediction of Type 2 Diabetes Microvascular Complications

Elizabeth Remfry
Queen Mary University of London
e.a.remfry@qmul.ac.uk

Rafael Henkin
Queen Mary University of London
r.henkin@qmul.ac.uk

Zainab Awan
Queen Mary University of London
z.awan@qmul.ac.uk

Rohini Mathur
Queen Mary University of London
r.mathur@qmul.ac.uk

Aakanksha Naik
Allen Institute for AI
aakankshan@allenai.org

Abstract

Existing inequalities are known through out diabetes care which result in poorer health outcomes for ethnic minority groups and those from disadvantaged backgrounds. With the growth of foundation models being deployed to assist with diagnosis and healthcare usage predictions it is essential we understand how these may exacerbate existing biases. We assess the fairness of long-term microvascular complication predictions for individuals living with Type 2 Diabetes. We encoded the entire structured clinical record for each individual as text in order to take advantage of existing knowledge within pretrained clinical language models. Leveraging large-scale EHR data from the UK, we predict the risk of microvascular complications in individuals with Type 2 Diabetes across 6-, 12-, 36- and 60-month prediction windows and assess performance across three fairness metrics; sensitivity, specificity and demographic parity. We find that models demonstrate statistically significant gaps in performance across different protected characteristics such as sex, ethnic group and level of deprivation. These performance gaps were particularly pronounced for ethnic minority groups, and those with missing or unknown ethnicity status.

1 Introduction

Evidence has highlighted that inequalities are present across diabetes care and health outcomes. Type 2 Diabetes (T2DM) is a long-term cardio-metabolic condition that disproportionately impacts ethnic minority groups [3], which experience higher prevalence rates of undiagnosed T2DM compared to White ethnic groups [1] and poorer treatment once diagnosed [10]. These disparities in diagnosis and treatment can lead to worse health outcomes, such as micro- and macro-vascular complications that can result in severe outcomes, such as vision loss, end stage renal disease and amputations [2, 7]. Given these existing inequities, it is crucial that as AI systems become more integrated into healthcare decision making and prediction, that these biases are not proliferated further.

There are well documented examples of how AI models can perpetuate and exacerbate health inequalities [8, 15, 11], and with the wide spread use and proliferation of foundational models, there is risk that we inherit and create biases that could result in inequitable outcomes. Although attention

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

has been paid to biases present in unstructured clinical notes, limited research has explored bias and performance across different socio-demographic groups in models trained on structured real-world electronic healthcare records (EHRs) and how this may change over time.

In this work we assess the impact of biases in a pre-trained model, GatorTron, across various disease prediction tasks in individuals with T2DM. Using large real world EHRs from the UK, we focus on predicting the microvascular complications across 6-, 12-, 36- and 60-month prediction windows, and explore biases across sex, ethnic group and indices of multiple deprivation (IMD).

2 Related work

Considerable research has explored bias in BERT-based models trained on unstructured data. Zhang et al. [18] highlighted linguistic biases as well as differing model performance for genders, ethnicities, language speakers and insurance status on clinical prediction tasks. They found statistically significant performance gaps in sensitivity, specificity and demographic parity across 50 downstream prediction tasks, with models favouring majority groups in gender, language, ethnicity and insurance status.

Jiang et al. [6] assessed the performance of NYUTron on a readmission task stratified by clinical department, age and racial groups and found biases across strata. For racial groups model performance varied, performing best for Chinese patients (AUC 0.85) and worst for Black patients (AUC 0.77), even though Black patients experienced the highest rates of readmission across ethnic groups. Critically they also reported varied performances across clinical departments, performing best in Neurology (AUC 0.90) and worst in Internal Medicine (AUC 0.64). This also wasn't seen to be an effect of sample size, with variations in both performance and number of readmissions. These findings suggest that model performance is poor for specific groups, and this may be amplified by intersecting variables, for example, elderly Black patients seen in Internal Medicine.

Pal et al. [13] compared the performance of various pre-trained BERT-like models on a multi-label classification task identifying smoking and obesity status from unstructured clinical notes. They found that across 5 models on both tasks there was a bias towards males and models performed the worst for those 20 – 40 years. When assessing intersectional bias for the smoking task, the model performed worst for men aged 40 – 60 (micro-F1 0.76) compared to women of the same age (micro-F1 0.92).

These findings paint a picture of systematic bias across a variety of different BERT-like models, healthcare systems and datasets. Our work builds on this to explore biases in sex, ethnic group and IMD in small foundational models fine-tuned on structured EHR data. We particularly focus on the differences in bias over time, exploring both short-term and long-term prediction of microvascular complications in individuals living with T2DM.

3 Methods

3.1 Data and study population

This study uses the Clinical Practice Research Datalink (CPRD), real-world anonymised patient data on 19 million patients from across the UK [16]. We analysed EHRs from CPRD AURUM and included all individuals ≥ 18, diagnosed with at least one long-term condition, permanently registered to any General Practice in London between 01/01/2010 and 01/01/2020.

A diagnosis of T2DM, retinopathy, neuropathy or nephropathy were identified using validated phenotype definitions and we used the first occurring diagnosis date [4]. Individuals with microvascular complications prior to a diagnosis of T2DM were excluded. Our dataset included 140,186 individuals diagnosed with T2DM, 19,954 with nephropathy, 31,091 with retinopathy and 8,135 with neuropathy (Table 1).

Study entry was defined as the first EHR event, up until the event before the prediction window, or to the last recorded event for those without complications. For example, for the 6-month prediction window for retinopathy, we kept all data until 6 months prior to the first diagnosis of retinopathy. See Appendix (Figure 4) for an example.

Table 1: Characteristics of individuals with T2DM complications

Characteristic		**Retinopathy**	**Nephropathy**	**Neuropathy**
Total		31,091	19,954	8,135
Sex	Male	17,481	10,806	4,853
	Female	13,610	9,148	3,282
Ethnic group	White	9,415	8,527	3,983
	Asian or Asian British	8,941	5,123	1,696
	Black or Black British	6,555	4,471	1,792
	Any Other Ethnic Group	2,033	1,029	403
	Mixed	414	212	89
	Missing	3,067	362	113
	Unknown	666	230	59
IMD	1 (least deprived)	1,700	1,231	451
	2	3,061	2,042	697
	3	5,525	3,536	1,427
	4	11,026	6,641	2,693
	5 (most deprived)	9,779	6,504	2,867

3.2 EHR pre-processing

We utilised data on diagnoses, symptoms, demographics, referrals, hospitalisations, procedures and medications. Each event in CPRD is associated with a clinical code and textual descriptor, for example the ICD10 code *E11.9* is associated with *type 2 diabetes mellitus without complications*. We concatenated the textual descriptor for every event in a patient's EHR chronologically to generate textual sentences for each patient.

Within CPRD, sex is recorded as a binary variable (Male/Female). IMD, a well used measure of deprivation, is calculated based on an individual's address and grouped into quintiles. For ethnic group, we aggregated into 6 categories: White, Black or Black British, Asian or Asian British, Any Other Ethnic Group, Mixed, Unknown. For any individuals without an ethnicity code we generated a Missing category, as missingness can be informative [5].

3.3 Model architecture

We utilised a pre-trained clinical language model, GatorTron-base [17] to encode the tokenized EHR sequences, truncated or padded to 512 tokens, the maximum length for standard encoder-only models. We calculate the median length and interquartile range (IQR) for each dataset (Appendix A).

We fine-tuned a total of 12 models, one for each microvascular complication and risk prediction window. The models consisted of a fine-tuned encoder with a single linear output layer for the classification task. We split our data 80/10/10 into training, test and validation, downsampled the train datasets and used weighted cross entropy due to class imbalance. We report on recall, F1, Area Under the Receiver Operating Characteristic (AUROC) and area under the precision recall curve (AUPRC) calculated on the held-out test set.

Models were fine-tuned using a learning rate of 2e-5, on the entire dataset with early stopping. Losses were monitored for overfitting. For more information on pre-processing and architecture see Appendix A.

3.4 Evaluation of model fairness

We evaluated the classifiers performance gaps on sensitivity, specificity and demographic parity. We used bootstrapping of 1000 samples from the test set to establish 95% confidence intervals (CI) for each gap. We use three definitions of fairness: sensitivity, specificity and demographic parity. Sensitivity, also known as recall or the true positive rate, is a ratio of correctly identified positive samples over the total number of positive samples. A higher value [0, 1] indicates better prediction of the positive class. Sensitivity is an important metric in clinical diagnostic tools as it is preferable to capture as many true cases of a disease as possible whilst minimising false negatives (18). Sensitivity is expressed as:

$$P(\hat{Y} = 1|Y = 1) = P(\hat{Y} = 1|Y = 1, Z = z), \forall z \in Z$$

Specificity is the ratio of correctly identified negative samples over the total number of all negative samples. A higher specificity value [0, 1] represents accurately predicting the negative class. Specificity denoted as:

$$P(\hat{Y} = 0|Y = 0) = P(\hat{Y} = 0|Y = 0, Z = z), \forall z \in Z$$

Demographic parity is a commonly used fairness metric, also known as statistical parity or group fairness, that refers to equal positive prediction rates across groups regardless of the true outcome. In a clinical setting, this means that for a prediction model of nephropathy in individuals with T2DM, the proportion of men and women identified is equal regardless of whether the true rates of disease differ between groups. Demographic parity is expressed as:

$$P(\hat{Y} = y) = P(\hat{Y} = y|Z = z), \forall z \in Z$$

4 Results

We assess the performance of the pre-trained model on 3 downstream tasks over 4 prediction windows. We benchmark the performance across each of these settings (Table 2).

Table 2: Model Performance metrics for different prediction windows

Time Period		**Recall**	**F1**	**AUROC**	**AUPRC**
6-months	Nephropathy	0.81	0.42	0.42	0.52
	Retinopathy	0.83	0.54	0.83	0.66
	Neuropathy	0.72	0.22	0.77	0.30
12-months	Nephropathy	0.82	0.42	0.81	0.53
	Retinopathy	0.83	0.55	0.83	0.68
	Neuropathy	0.74	0.21	0.77	0.29
36-months	Nephropathy	0.85	0.44	0.56	0.63
	Retinopathy	0.87	0.56	0.88	0.75
	Neuropathy	0.73	0.23	0.80	0.43
60-months	Nephropathy	0.84	0.47	0.89	0.70
	Retinopathy	0.89	0.58	0.91	0.81
	Neuropathy	0.76	0.26	0.85	0.55

Across all three tasks models performed better at the longest prediction window of 60-months, compared to a 6-month window and for tasks with lower levels of class imbalance. The model for retinopathy, the largest class, performed best across all metrics with an AUPRC score of 0.81 at 60-month prediction window, as compared to neuropathy with an AUPRC of 0.55.

4.1 Variation in performance gaps for ethnic groups

We visualise the sensitivity, specificity and parity gap for ethnic group over different prediction windows across Retinopathy (Figure 1), Nephropathy (Figure 2) and Neuropathy (Figure 3), which highlights a series of significant gaps in performance over time.

A positive bar indicates that the model performs better for that ethnic group at that specific time point compared to the reference group (White), a negative bar indicates poorer performance for that ethnic group. The 95% CI is included to aid in interpretation of statistical significance. Where the 95% CI crosses zero, the gap is not statistical significant. For ease we also report the number of significant gaps, over total number of gaps for each socio-demographic area.

For the prediction of retinopathy, gaps in sensitivity were significant 8 out of 24 times, almost all of which favoured ethnic minority groups. Specificity was poorer for ethnic minority groups (11/24), particularly Asian or Asian British and those with Missing ethnicity. Demographic parity was better

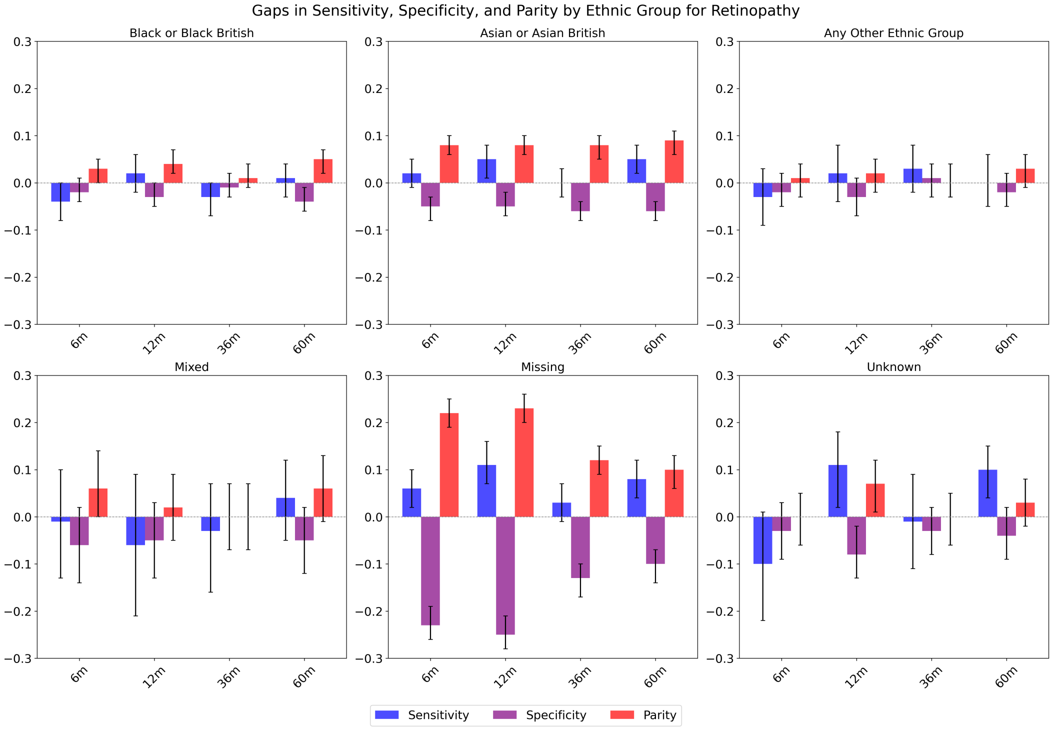

Figure 1: Performance gaps for Retinopathy

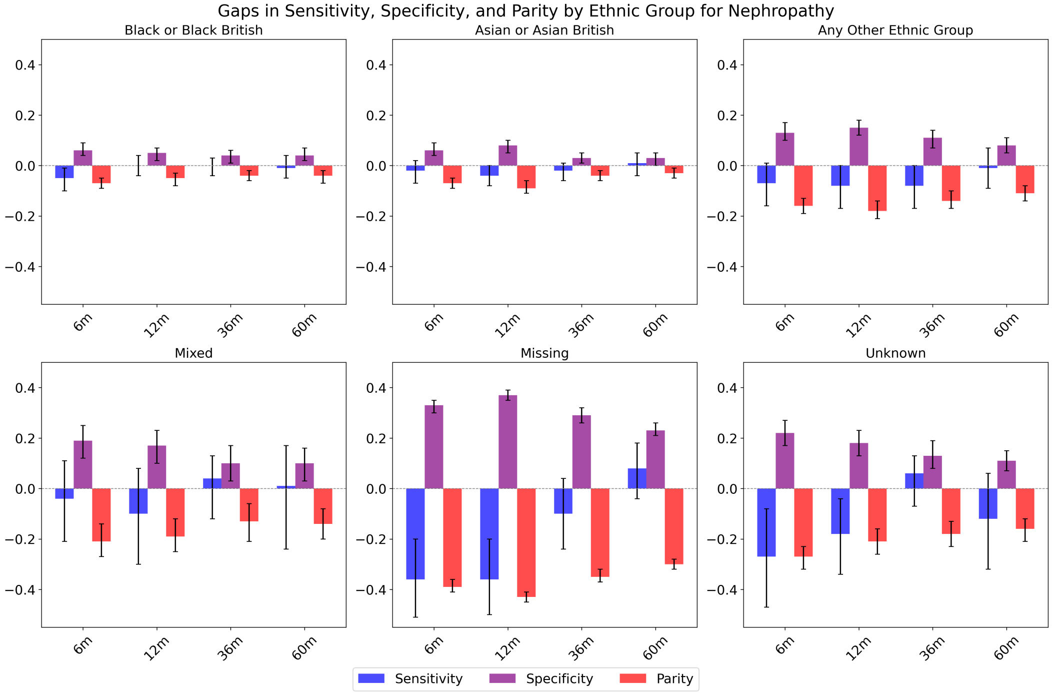

Figure 2: Performance gaps for Nephropathy

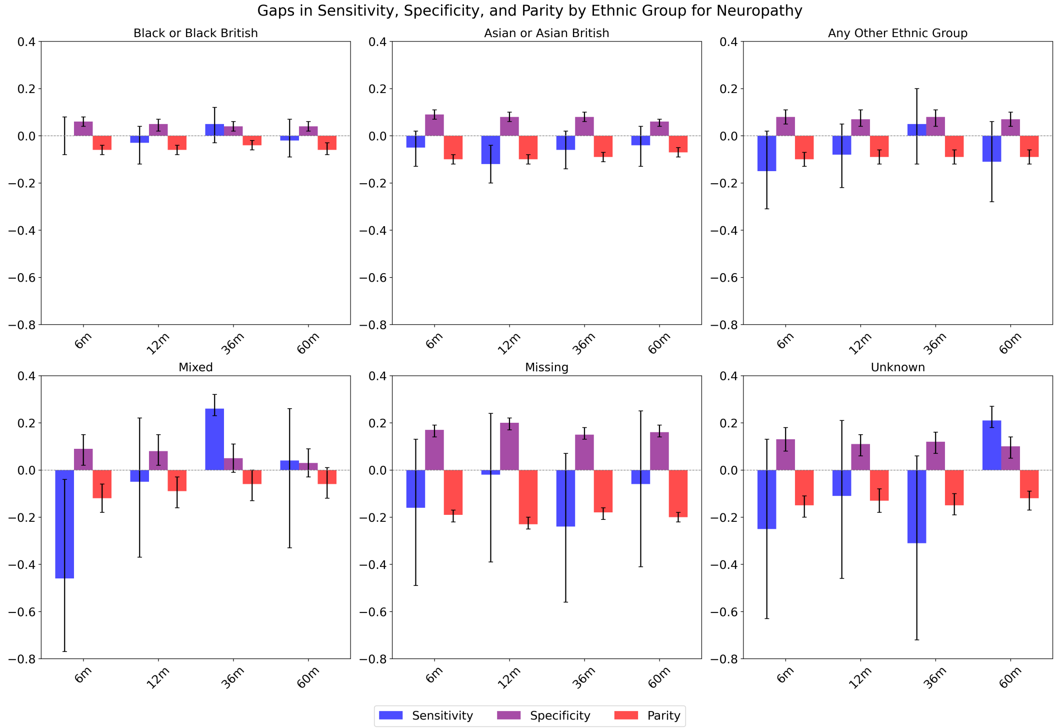

Figure 3: Performance gaps for Neuropathy

for most ethnic minority groups (12/24), which means that the model is predicting disease presence at a higher rate compared to the White ethnic groups.

For nephropathy prediction, there were fewer significant gaps in sensitivity (7/24), all of which favoured the majority reference group. Whilst specificity was better for all ethnic minority groups across all time windows (24/24), meaning that the models are better at correctly identifying negative cases in ethnic minority groups as compared to the White reference group. The gap in demographic parity was significant in all cases (24/24) and favoured the reference group.

Finally for neuropathy prediction, fewer gaps were significant (4/24) and were equally spread between favouring ethnic minority groups or the reference group. Specificity favoured ethnic minority groups in almost every case (23/24) whilst demographic parity was worse for ethnic minority groups in comparison to the reference group (23/24).

Overall model performance varied across each prediction task and highlights the importance of investigating model performance over a variety of metrics. Although the gaps in performance were not always significant for sensitivity, they were for specificity and demographic parity. The biggest variation and significant differences could be seen in the category with Missing ethnicity, which experienced poorer sensitivity and parity across nephropathy and neuropathy compared to the reference group. In practice this could result in an underdiagnosis of those with a Missing ethnicity for nephropathy and neuropathy, given that more true positive cases are missed.

4.2 Ethnic group bias may decrease over time

There is a visual trend towards a decreasing bias over time across all three metrics. This is more pronounced in specific settings, such as the sensitivity and parity gap for individuals with Missing ethnicity data in the nephropathy task. This may be due to a variety of factors, which includes a selection bias. In order to contribute data to the 60-month prediction model an individual is required to have at least 60-months of data, whilst those contributing to 6-month prediction models are only required to have at least 6 months. Due to this inherent selection bias due to the set up of the study, there may be differences in individuals that are included at each time point. For example, at 6 months there are 8135 individuals with neuropathy, whilst for the 60-month prediction task there were 7164 individuals (Table 3).

Table 3: Number of cases by disease and prediction window

Condition	6-months	12-months	36-months	60-months
Neuropathy	8,135	8,031	7,646	7,164
Nephropathy	19,954	19,773	19,055	17,999
Retinopathy	31,091	30,505	28,381	26,034

4.3 Biases less prominent across other demographics

For both sex and IMD there were considerably fewer statistical differences. We report the total number of statistically significant gaps across all three tasks (retinopathy, neuropathy and nephropathy) and all four prediction windows (6-, 12, 36- and 60-months) together. A score of 12 would demonstrated a significant gap over all models and time periods for each metric. We also report of the total number of gaps that favour the reference group in brackets.

Of the statistically significant gaps in performance, all gaps favoured the reference groups (males, and IMD quartile 1) for sensitivity and demographic parity, but favoured all other groups for specificity. In medical terms this is less desirable, as a higher specificity but lower sensitivity can result in fewer false alarms but also fewer actual cases being identified.

Table 4: Comparison of sensitivity, specificity, and parity by sex and IMD

Category	Sensitivity	Specificity	Demographic Parity
Sex (Male vs Female)	3 (3)	8 (0)	6 (6)
IMD 1 vs 2	1 (1)	-	-
IMD 1 vs 3	-	7 (0)	5 (5)
IMD 1 vs 4	2 (2)	3 (0)	4 (4)
IMD 1 vs 5	1 (1)	4 (0)	4 (4)

5 Discussion and future work

We assessed the fairness of a pre-trained language model in a series of microvascular complication prediction tasks over different prediction windows. These models demonstrated differences in performance across ethnic groups, sex and IMD across a variety of metrics.

These performances may reflect known biases in the data. For example, research shows that although ethnic minority groups experience higher rates of diabetes complications [12], they are not always diagnosed at the same rate as White ethnic groups. The models in this study may perform better for the majority group as these are the trends captured within the available data.

Additionally, there are issues with messy EHR data. Performance varied across the 7 ethnic groups, each group contains other granular ethnicity categories which are collapsed for a larger sample size. Future work should look at more granular ethnicity categories to explore within group differences. The Missing ethnicity group typically experience the poorest performance. Research has found that those with missing ethnicity data are generally younger, male and living with fewer co-morbidities [9, 14] which suggests that this group may be a relatively healthy group that does not interact with the healthcare service regularly, thus reducing the possibility to capture ethnicity data. It is common in research to exclude this group from modelling, but this work highlights the need to understand how model performance varies under real-world conditions where missing ethnicity data can be common.

A limitation of this work is that we do not engage in understanding where the biases emerge from, whether clinical practices, data quality, class imbalance or other sources, nor do we attempt to account or correct for the biases in the pipeline. In future work, to get a deeper understanding of bias we will consider a counterfactual evaluation, whereby all data remains the same whilst we alter one or more sensitive attributes, such as ethnic group or sex, and then compare model performance.

Future work will explore the temporal aspects to bias, particularly to understand the potential reduction in bias over time. This could be in the form of explainability, to understand the features that drive prediction at 6 months versus 60 months as well as analysing changes in the cohort over longer prediction windows to assess any systematic differences in these cohorts. Additionally it is important to understand how inequalities intersect, and a particular focus should be on understanding and mitigating any intersectional biases.

Acknowledgments and Disclosure of Funding

ER is funded by the Wellcome Trust Health Data in Practice (HDiP) Programme (218584/Z/19/Z). RH is funded by the AI for Multiple Long Term Conditions (AIM) Programme (NIHR203982). RM is supported by Barts Charity (MGU0504).

This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. http://doi.org/10.5281/zenodo.438045

References

[1] Risk factors for pre-diabetes and undiagnosed type 2 diabetes in England - Office for National Statistics. URL https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/bulletins/riskfactorsforprediabetesandundiagnosedtype2diabetesinengland/2013to2019.

[2] Jack R. W. Brownrigg, Cian O. Hughes, David Burleigh, Alan Karthikesalingam, Benjamin O. Patterson, Peter J. Holt, Matthew M. Thompson, Simon de Lusignan, Kausik K. Ray, and Robert J. Hinchliffe. Microvascular disease and risk of cardiovascular events among individuals with type 2 diabetes: a population-level cohort study. *The Lancet Diabetes & Endocrinology*, 4 (7):588–597, July 2016. ISSN 2213-8587, 2213-8595. doi: 10.1016/S2213-8587(16)30057-2. URL https://www.thelancet.com/journals/landia/article/PIIS2213-8587(16)30057-2/abstract. Publisher: Elsevier.

[3] T. M. E. Davis. Ethnic diversity in Type 2 diabetes. *Diabetic Medicine*, 25(s2):52–56, 2008. ISSN 1464-5491. doi: 10.1111/j.1464-5491.2008.02499.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1464-5491.2008.02499.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1464-5491.2008.02499.x.

[4] Fabiola Eto. MULTIPLY-Initiative, August 2023. URL https://github.com/Fabiola-Eto/MULTIPLY-Initiative. original-date: 2020-11-25T17:37:13Z.

[5] Rolf H. H. Groenwold. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and Prognostic Research*, 4(1):8, July 2020. ISSN 2397-7523. doi: 10.1186/s41512-020-00077-0. URL https://doi.org/10.1186/s41512-020-00077-0.

[6] Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, Madeline Miceli, Nora C. Kim, Cordelia Orillac, Zane Schnurman, Christopher Livia, Hannah Weiss, David Kurland, Sean Neifert, Yosef Dastagirzada, Douglas Kondziolka, Alexander T. M. Cheung, Grace Yang, Ming Cao, Mona Flores, Anthony B. Costa, Yindalon Aphinyanaphongs, Kyunghyun Cho, and Eric Karl Oermann. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06160-y. URL https://www.nature.com/articles/s41586-023-06160-y. Publisher: Nature Publishing Group.

[7] Parvin Akter Khanam, Sayama Hoque, Tanjima Begum, Samira Humaira Habib, and Zafar Ahmed Latif. Microvascular complications and their associated risk factors in type 2 diabetes mellitus. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 11: S577–S581, December 2017. ISSN 1871-4021. doi: 10.1016/j.dsx.2017.04.007. URL https://www.sciencedirect.com/science/article/pii/S1871402117300747.

[8] David Leslie, Anjali Mazumder, Aidan Peppin, Maria K. Wolters, and Alexa Hagerty. Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? *BMJ*, 372:n304, March 2021. ISSN 1756-1833. doi: 10.1136/bmj.n304. URL `https://www.bmj.com/content/372/bmj.n304`. Publisher: British Medical Journal Publishing Group Section: Analysis.

[9] R. Mathur, L. Palla, R.E. Farmer, N. Chaturvedi, and L. Smeeth. Ethnic differences in the severity and clinical management of type 2 diabetes at time of diagnosis: A cohort study in the UK Clinical Practice Research Datalink. *Diabetes Research and Clinical Practice*, 160:108006, February 2020. ISSN 0168-8227. doi: 10.1016/j.diabres.2020.108006. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7042884/`.

[10] Rohini Mathur, Ruth E. Farmer, Sophie V. Eastwood, Nish Chaturvedi, Ian Douglas, and Liam Smeeth. Ethnic disparities in initiation and intensification of diabetes treatment in adults with type 2 diabetes in the UK, 1990–2017: A cohort study. *PLOS Medicine*, 17(5):e1003106, May 2020. ISSN 1549-1676. doi: 10.1371/journal.pmed.1003106. URL `https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003106`. Publisher: Public Library of Science.

[11] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aax2342. URL `https://www.science.org/doi/10.1126/science.aax2342`.

[12] Chandra Y. Osborn, Mary de Groot, and Julie A. Wagner. Racial and Ethnic Disparities in Diabetes Complications in the Northeastern United States: The Role of Socioeconomic Status. *Journal of the National Medical Association*, 105(1):51–58, March 2013. ISSN 0027-9684. doi: 10.1016/S0027-9684(15)30085-7. URL `https://www.sciencedirect.com/science/article/pii/S0027968415300857`.

[13] Ridam Pal, Hardik Garg, Shashwat Patel, and Tavpritesh Sethi. Bias Amplification in Intersectional Subpopulations for Clinical Phenotyping by Large Language Models, March 2023. URL `https://www.medrxiv.org/content/10.1101/2023.03.22.23287585v1`. Pages: 2023.03.22.23287585.

[14] Marta Pineda-Moncusí, Freya Allery, Antonella Delmestri, Thomas Bolton, John Nolan, Johan H. Thygesen, Alex Handy, Amitava Banerjee, Spiros Denaxas, Christopher Tomlinson, Alastair K. Denniston, Cathie Sudlow, Ashley Akbari, Angela Wood, Gary S. Collins, Irene Petersen, Laura C. Coates, Kamlesh Khunti, Daniel Prieto-sAlhambra, and Sara Khalid. Ethnicity data resource in population-wide health records: completeness, coverage and granularity of diversity. *Scientific Data*, 11(1):221, February 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-02958-1. URL `https://www.nature.com/articles/s41597-024-02958-1`. Publisher: Nature Publishing Group.

[15] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, December 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01595-0. URL `https://www.nature.com/articles/s41591-021-01595-0`. Publisher: Nature Publishing Group.

[16] Achim Wolf, Daniel Dedman, Jennifer Campbell, Helen Booth, Darren Lunn, Jennifer Chapman, and Puja Myles. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *International Journal of Epidemiology*, 48(6):1740–1740g, December 2019. ISSN 0300-5771, 1464-3685. doi: 10.1093/ije/dyz034. URL `https://academic.oup.com/ije/article/48/6/1740/5374844`.

[17] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. A large language model for electronic health records. *npj Digital Medicine*, 5(1):1–9, December 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00742-2. URL `https://www.nature.com/articles/s41746-022-00742-2`. Publisher: Nature Publishing Group.

[18] Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120, Toronto Ontario Canada, April 2020. ACM. ISBN 978-1-4503-7046-2. doi: 10.1145/3368555.3384448. URL `https://dl.acm.org/doi/10.1145/3368555.3384448`.

A Appendix

A.1 Pre-processing

Patients were only included if they were eligible for data linkage to Hospital Episode Statistics (HES) and Office for National Statistics (ONS) registries. This ensured that only patients with primary and secondary care were included. Data was pre-processed to remove duplicate events (identical rows), impossible events (dates of events that occur before birth or after deregistration), events with missing dates, or missing clinical code (events without a textual descriptor). Due to data quality issues only records between 1985 and 2020 were included [16]. Only individuals with at least 3 unique events were included.

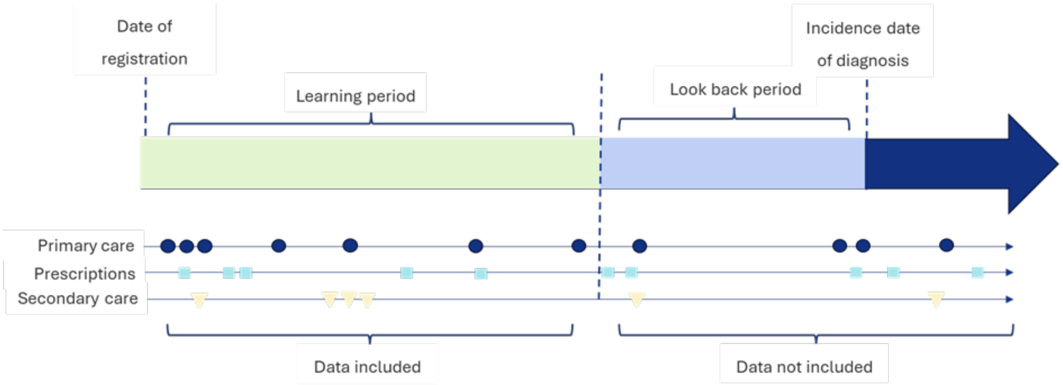

Figure 4: Data window

A.2 Model architecture

Gatortron-base is a small foundational model with 345M parameters. It was trained on 82B words of de-identified clinical notes, 6.1B words from PubMed, 2.5B words from WikiText and 0.5B words of de-identified clinical notes from MIMIC-III.

For all prediction tasks and prediction windows, input was first tokenized and special token [CLS] added. The tokenized sequences, special tokens and positional embeddings were fed into the pretrained encoder-only model. The final hidden state of the [CLS] token was used as input to the fully connected layer. A sigmoid activation function was applied to logits to produce independent probabilities for each label.

We searched for a learning rate that gave the lowest F1 score (1e-3, 2e-5, 3e-5, 4e-5, 5e-5) and fine-tuned on the entire dataset for 48000 steps with early stopping. Models were fine-tuned on one NVidia A100 GPU.

A.3 Average token length

We also provide the median token length of patient's EHR sequences for each disease and prediction window. We also calculate the interquartile range, displayed as the 25th and 75th percentile. The vast majority of EHR sequences are truncated due to the standard maximum length of 512 although this decreased gradually the longer predicton window lengths.

Table 5: Median [IQR] and percentage truncated by disease and prediction window

Condition	Prediction Window	Median [IQR]	% Truncated
Nephropathy	6-months	2773 [1059, 5877]	86.31%
	12-months	2698 [1006, 5741]	85.31%
	36-months	2367 [806, 5333]	81.48%
	60-months	2118 [657, 4995]	78.29%
Neuropathy	6-months	3080 [1176, 6550]	87.87%
	12-months	3026 [1151, 6486]	87.42%
	36-months	2858 [1041, 6225]	85.66%
	60-months	2720 [946, 6051]	83.99%
Retinopathy	6-months	2374 [912, 5004]	84.40%
	12-months	2268 [840, 4864]	82.94%
	36-months	1921 [628, 4397]	77.98%
	60-months	1670 [470, 4111]	73.83%

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in this paper are clearly layed out in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations can be found in discussion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

Justification: N/A

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Additional information is provided in the supplementary materials. However, the dataset is private, researchers need to apply to use the data, going through a strict application, as well as paying for access which limits reproducibility. This is one of the many challenges of EHR data. We will release all the code on acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: It is not possible to provide open access to the data, as described above. We will release code on acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: These are layed out in the methods and supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Where appropriate we have provided confidence intervals and we outline how these were caluculated.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the appendix we include details on the compute utilised.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines: The paper conforms to the guidelines.

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The findings are discussed in relation to impact in the clinical setting.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: This paper uses pre-trained models which are widely available for use.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA] .

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: This project was reviewed by CPRD and the work is approved under licence.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.