Exploring multi-site dataset shifts in electronic health records using time series features

Anonymous Author(s)
Affiliation
Address
email

Abstract

Models developed using longitudinal electronic health record (EHR) data can demonstrate inconsistent abilities to generalize to new data at different institutions. Rather than relying only only external validity of performance, we consider how distributional shifts in EHR data can inform multi-site generalizability without the need for task-specific models or annotations. Extending statistical dataset shift detection to time series through feature-based temporal analysis, we compare the EHR data from five different institutions and four different prior patient conditions for patients requiring the administration of an inpatient diuretic. We illustrate which sites exhibit greater variability as well as the EHR measures contributing to the variation, providing valuable insight into downstream deployment.

1 Introduction

There is increasing potential in the ability to harness time series data to improve healthcare. With the adoption of electronic health records (EHRs), many health systems now have comprehensive data representing patient information over time. Indeed, the number of data-driven models developed from EHR-based datasets has expanded significantly [1–5]. Yet despite efforts towards external validation, an extant challenge is ascertaining how such models generalize across myriad sites. There is a well-understood notion that EHR data not only reflect underlying knowledge about health and disease, but also site-specific practices and populations [6]. Consequently, the external validity of models may be inconsistent across different test sites (e.g., a model developed at *Site A* works well at *Site B*, but not *Site C*) [7–14]. Moreover, in uncovering model performance differences between sites, there is a need to provide explanations for these differences – implicitly demanding an understanding of distributional shifts beyond technical performance metrics [15–18]. As such, detection of dataset shift, agnostic of model inference, is a burgeoning issue to provide the necessary insights into model utility within clinical settings. Still, the challenge for models employing EHR datasets – and more broadly time series data – is that deeper dataset comparisons are non-trivial, instead relying on parametric assumptions or a small set of transformations to provide estimates of statistics [19–21]. So while the phenomenon of site variation in performance is acknowledged [6, 7, 13], strategies for identification, understanding, and robust adaptation remain lacking.

Recently, attention has turned to feature-based analysis of time series, in a manner akin to radiomics in imaging [22]. These generalized quantitative metrics describe a range of characteristics related to time series data, providing a simpler approach to assessing distributional shifts in data. Here, we explore a feature-based approach to assess dataset shift in multi-site EHR data, particularly in the context of external validation. The simplicity of the approach offers a framework extensible to different datasets and feature extraction methods, as a way to provide diagnostic support to generalizability. We demonstrate preliminary utility in a multi-institution EHR dataset focused on assessing the dynamics of inpatients given a diuretic. Notably, the use of diuretics is extensive in clinical care, but

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Table 1: Baseline data characteristics across sites. Abbreviations: LOS (length of stay), CAD (coronary artery disease), CKD (chronic kidney disease), CHF (congestive heart failure).

	UCI	UCSD	UCD	UCSF	UCLA
n	19,727	35,060	19,598	29,191	35,246
LOS, mean (SD)	10.1 (12.4)	11.1 (14.3)	13.4 (21.6)	11.7 (16.7)	11.9 (15.4)
Age, mean (SD)	62.6 (16.0)	62.7 (15.5)	62.8 (15.9)	62.4 (16.8)	64.0 (16.8)
Female, n (%)	8,348 (42.3)	14,731 (42.0)	8,071 (41.2)	13,205 (45.2)	15,997 (45.4)
CAD, n (%)	2,093 (10.6)	3,380 (9.6)	1,304 (6.7)	2,482 (8.5)	3,824 (10.8)
CKD, n (%)	2,642 (13.4)	7,467 (21.3)	3,024 (15.4)	4,100 (14.0)	8,340 (23.7)
Diabetes, n (%)	6,530 (33.1)	11,854 (33.8)	5,815 (29.7)	7,385 (25.3)	11,409 (32.4)
CHF, n (%)	6,447 (32.7)	16,224 (46.3)	7,987 (40.8)	9,201 (31.5)	14,598 (41.4)

with different approaches to delivery and monitoring depending on site, underlying condition, and medical domain. We identify notable and inconsistent variations across multi-site temporal EHR data, providing insight into the diversity of the EHR data as a way to inform downstream development and implementation of subsequent models.

2 Related Work

There has been extensive work on comparing time series using statistical tests [19–21]. Currently, there is a stronger emphasis on the extraction of a diverse array of informative features from time series, whether through interpretable characterizing algorithms [23–25] or deep representation learning [26–30]. The primary objective of these approaches has been to improve predictive or forecasting capabilities of different models. Makredly, these features can also be used to provide insight into cross-site time series dataset shifts. Work on detecting dataset shift is also extensive, but focuses on static distributions with minimal consideration for different temporal characteristics [15–18]. Existing approaches to assess EHR dataset shifts focus on variation in coding practices and phenotypic variations rather than collected observations (e.g., laboratory measures, vital signs) over time [31–34]. There is also focus on inter- and intra-site changes in disease characterization over time due to the evolution of technology, quality of observations, and clinical understanding, requiring the detection and mitigation of such temporal shifts [35–37]. We aim to demonstrate initial utility of bridging the developments of feature-based time series analysis and dataset shift characterization to inform the generalizability of models learned from EHR data.

3 Methods

We obtained data from the University of California Health Data Warehouse (UCHDW) provided by the Center for Data-driven Insights and Innovation at UC Health (CDI2). This data warehouse collects data from five academic medical sites: UC Irvine (UCI), UC Davis (UCD), UC San Diego (UCSD), UC San Fransico (UCSF), and UC Los Angeles (UCLA). From this data repository, we identified 138,822 patients across four different prior conditions (coronary artery disease (CAD), chronic kidney disease (CKD), diabetes, congestive heart failure) with inpatient encounters requiring the administration of a diuretic between 2019-2024 (Table 1). For these patients, we gather the hourly data of 17 longitudinal lab (blood urea nitrogen, creatinine, estimated glomerular filtration rate, glucose, magnesium, platelet count, potassium, sodium) and vital measurements (diastolic blood pressure, systolic blood pressure, mean arterial pressure, oxygen flow rate, heart rate, respiration rate, oxygen saturation, temperature, weight) for their first 96 hours of inpatient admission. Missing data is imputed using the last observation carried forward.

From the time series profiles of multiple users and measures, we extract 22 CAnonical Time-series CHaracteristic (catch22) features for each measure [25]. Over all pairs of the five different UC sites, we compare each feature using Dunn's test with Holm correction for multiple testing [38]. As each measure is decomposed into multiple (correlated) features, the significance of each *measure* is obtained via an omnibus permutation test approach [39, 40]. Concretely, the number of significantly different *features* for each measure is compared against the estimated count under the null hypothesis through multiple permutations of the outcome variable (i.e., site) to obtain a single p-value for

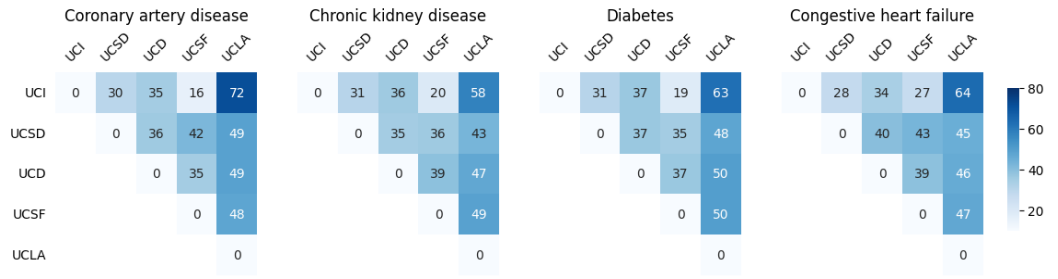

Figure 1: Heatmaps illustrating the number of significantly different time series features between sites for each prior condition.

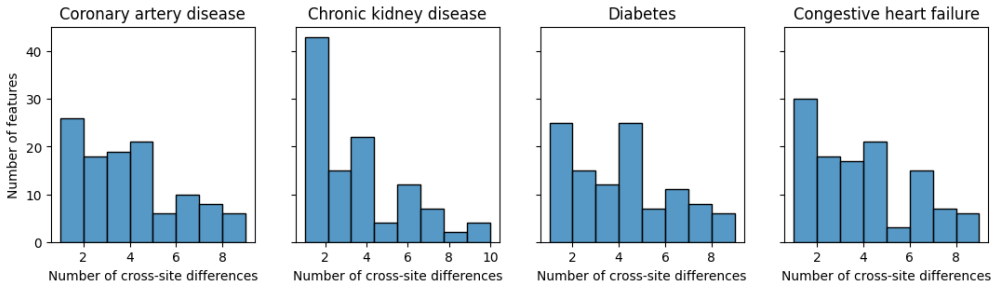

Figure 2: Histograms illustrating frequency of a features difference across all cross-site comparisons.

significance of one time series measure, regardless of the decomposition of the measure into multiple features. To mitigate Type 1 error due to differences in population size, we calculate the effect size using Glass's rank-biserial correlation and identify significant features with non-negligible effect [41, 42]. While we used catch22 features as proof-of-concept in this study, the permutation test approach is extensible to compare cross-site differences in temporal EHR data, regardless of the number of features extracted and approach to extract the features. This process is outlined in A.1.

We consolidate the results into visualizations that illustrate cross-site variations. Specifically, we highlight not only which sites exhibit greater differences, but the constituent time series variables contributing to these differences. Using one site as a reference, we identify variables that differ between the reference site and others. We collect the sets of differing variables and identify if they are unique to a particular site comparison, or if they differ across multiple sites. To this end, we use UpSet plots to illustrate, for different site-to-site comparisons, which measures differ across one site, few sites, or across all sites [43].

4 Results

Figure 1 demonstrates the emergence of variation across sites and prior condition, suggesting differences in the extent of dataset drift between sites. Across all prior conditions, the top row of each heatmap illustrates the number of time series features that differ between patients from UCI and the other sites; there are notably more differences between the patients from UCI vs. UCLA and fewer differences between patients from UCI vs. UCSF especially for those with prior CAD. Overall, the UCLA site exhibits a higher rate of variation towards all other sites. Figure 2 shows the distribution of time series feature differences across the multiple site comparisons. The right-skew indicates that differences are more heterogeneous across the site-comparisons. These histograms demonstrate that the variability of a longitudinal EHR measure is not necessarily consistent across different sites. While some may be consistently different, reflecting institutional variation in population or reporting, others are different with select sites, indicating the potential for selective translation of models.

Figure 3 focuses on UCI as the reference institution and illustrates common vs. unique differences relative to other sites. After applying permutation testing to identify measures differing according to

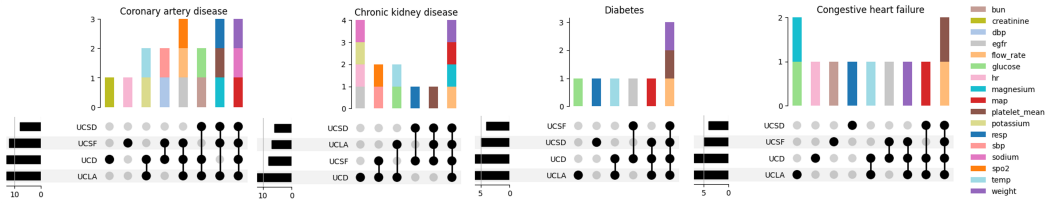

Figure 3: UpSet plots illustrating which measures differ across sets of sites. UCI is used as the reference. Abbreviations: BUN (blood urea nitrogen), DBP (diastolic blood pressure), eGFR (estimated glomerular filtration rate), HR (heart rate), MAP (mean arterial pressure), Resp (respiration rate), SBP (systolic blood pressure), SpO_2 (blood oxygen saturation), Temp (temperature).

the comparisons across the multiple extracted features, we analyze the sets of measures across the site comparisons with respect to the reference. Particularly, we observe if a significantly different variable only exhibits variation across one site comparison or across multiple sites. For example, for patients with a history of CAD, patients at UCI differ the most from patients from UCLA but no lab measures emerge as uniquely different between these two sites. Rather, most differences are shared across multiple sites. In contrast, for patients with a history of CKD, we observe a bimodality such that a large number of differences between UCI and UCD are attributable to unique differences in laboratory and vital measures (sodium, potassium, heart rate, and eGFR), while another set of differences are shared across all sites (weight, mean arterial pressure, magnesium, and oxygen flow rate). For patients with prior diabetes, differences are generally shared across sites, with variations in weight, platelets, and oxygen flow rate. Lastly, for patients with a history of CHF, there is a relatively uniform distribution of differences, indicating higher and more inconsistent cross-site variation.

5 Conclusion

Through a feature-based comparison of temporal EHR data, we elucidated variability across sites to provide insight into the generalizability of models developed on site-specific data. Findings from this exploration are key sources of variability between patients at the different sites (e.g., UCLA and UCI, especially for patients with a history of CAD). This result highlights how extra care may be advised when developing EHR-based models when developing models on combined data and the impact of the variability on learning useful patterns. Conversely, there may be more confidence in the ability of models developed using data from some sites to generalize to each other (e.g., UCI and UCSF). These considerations can also be made in the context of features used to develop models. For example, labs like potassium and sodium exhibit unique differences between patients with CKD from UCI and UCD, potentially reflecting different monitoring frequency and diuretic aggressiveness; this may be more important and require site calibration in certain clinical applications than others, such as predictive models of electrolyte derangement. In comparison, shared differences in weight for patients with a history of diabetes between UCI and all other sites suggest more general population differences requiring careful consideration of generalizability. Importantly, these differences should be identified early-on, providing useful insight into downstream decisions.

Our approach combines current advances in time series feature extraction and basic statistical techniques to analyze dataset drift. As such, there is no dependency on annotations to train and evaluate predictive models, nor on the models themselves to assess changes in cross-site performance, enabling a way to perform dataset shift evaluation early on and irrespective of model development. There are several limitations and future steps we will further explore. We have not disentangled whether differences are due to underlying patient characteristics or site-specific data collection practices, which is an ongoing goal in the understanding of generalizability of models for healthcare. We also aim to confirm the impact of decisions made due to early exploration of shift on downstream deployment [44]. Lastly, the features we extracted, while intentionally simple for this preliminary work, are not the only possibility. Deep features from recurrent, Transformer, or foundation models can provide more expressive representations of time series at the expense of interpretability – but also improve consideration of additional complexities such as cross-series correlations.

References

[1] Amitava Banerjee, Ashkan Dashtban, Suliang Chen, Laura Pasea, Johan H. Thygesen, Ghazaleh Fatemifar, Benoit Tyl, Tomasz Dyszynski, Folkert W. Asselbergs, Lars H. Lund, Tom Lumbers, Spiros Denaxas, and Harry Hemingway. Identifying subtypes of heart failure from three electronic health record sources with machine learning: an external, prognostic, and genetic validation study. *The Lancet Digital Health*, 5(6):e370–e379, June 2023. ISSN 2589-7500. doi: 10.1016/S2589-7500(23)00065-1.

[2] Nathan Brajer, Brian Cozzi, Michael Gao, Marshall Nichols, Mike Revoir, Suresh Balu, Joseph Futoma, Jonathan Bae, Noppon Setji, Adrian Hernandez, and Mark Sendak. Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission. *JAMA Network Open*, 3(2):e1920733, February 2020. doi: 10.1001/jamanetworkopen.2019.20733.

[3] Sayon Dutta, Dustin S. McEvoy, Lisette N. Dunham, Ronelle Stevens, David Rubins, Gearoid M. McMahon, and Lipika Samal. External Validation of a Commercial Acute Kidney Injury Predictive Model. *NEJM AI*, 1(3), February 2024. doi: 10.1056/AIoa2300099.

[4] Ariel Avraham Hasidim, Matthew Adam Klein, Itamar Ben Shitrit, Sharon Einav, Karny Ilan, and Lior Fuchs. Toward the standardization of big datasets of urine output for AKI analysis: a multicenter validation study. *Scientific Reports*, 15(1):20009, June 2025. doi: 10.1038/s41598-025-95535-4.

[5] Davina Zamanzadeh, Jeffrey Feng, Panayiotis Petousis, Arvind Vepa, Majid Sarrafzadeh, S. Ananth Karumanchi, Alex A. T. Bui, and Ira Kurtz. Data-driven prediction of continuous renal replacement therapy survival. *Nature Communications*, 15(1):5440, June 2024. doi: 10.1038/s41467-024-49763-3.

[6] Thomas A. Lasko, Eric V. Strobl, and William W. Stead. Why do probabilistic clinical models fail to transport between sites. *npj Digital Medicine*, 7(1):1–8, March 2024. doi: 10.1038/s41746-024-01037-4.

[7] Mylene W. M. Yao, Elizabeth T. Nguyen, Matthew G. Retzloff, L. April Gago, John E. Nichols, John F. Payne, Barry A. Ripps, Michael Opsahl, Jeremy Groll, Ronald Beesley, Gregory Neal, Jaye Adams, Lorie Nowak, Trevor Swanson, and Xiaocong Chen. Machine learning center-specific models show improved IVF live birth predictions over US national registry-based model. *Nature Communications*, 16(1):3661, April 2025. doi: 10.1038/s41467-025-58744-z.

[8] Seung Wook Lee, Hyung-Chul Lee, Jungyo Suh, Kyung Hyun Lee, Heonyi Lee, Suryang Seo, Tae Kyong Kim, Sang-Wook Lee, and Yi-Jun Kim. Multi-center validation of machine learning model for preoperative prediction of postoperative mortality. *npj Digital Medicine*, 5(1):91, July 2022. doi: 10.1038/s41746-022-00625-6.

[9] Fatemeh Amrollahi, Supreeth P. Shashikumar, Andre L. Holder, and Shamim Nemati. Leveraging clinical data across healthcare institutions for continual learning of predictive risk models. *Scientific Reports*, 12(1):8380, May 2022. doi: 10.1038/s41598-022-12497-7.

[10] Jethro C. C. Kwong, Adree Khondker, Eric Meng, Nicholas Taylor, Cynthia Kuk, Nathan Perlis, Girish S. Kulkarni, Robert J. Hamilton, Neil E. Fleshner, Antonio Finelli, Theodorus H. van der Kwast, Amna Ali, Munir Jamal, Frank Papanikolaou, Thomas Short, John R. Srigley, Valentin Colinet, Alexandre Peltier, Romain Diamand, Yolene Lefebvre, Qusay Mandoorah, Rafael Sanchez-Salas, Petr Macek, Xavier Cathelineau, Martin Eklund, Alistair E. W. Johnson, Andrew Feifer, and Alexandre R. Zlotta. Development, multi-institutional external validation, and algorithmic audit of an artificial intelligence-based Side-specific Extra-Prostatic Extension Risk Assessment tool (SEPERA) for patients undergoing radical prostatectomy: a retrospective cohort study. *The Lancet Digital Health*, 5(7):e435–e445, July 2023. doi: 10.1016/S2589-7500(23)00067-5.

[11] Hojjat Salehinejad, Anne M. Meehan, Parvez A. Rahman, Marcia A. Core, Bijan J. Borah, and Pedro J. Caraballo. Novel machine learning model to improve performance of an early warning system in hospitalized patients: a retrospective multisite cross-validation study. *eClinicalMedicine*, 66, December 2023. doi: 10.1016/j.eclinm.2023.102312.

[12] Colin G. Walsh, Michael A. Ripperger, Yirui Hu, Yi-han Sheu, Hyunjoon Lee, Drew Wilimitis, Amanda B. Zheutlin, Daniel Rocha, Karmel W. Choi, Victor M. Castro, H. Lester Kirchner, Christopher F. Chabris, Lea K. Davis, and Jordan W. Smoller. Development and multi-site external validation of a generalizable risk prediction model for bipolar disorder. *Translational Psychiatry*, 14(1):58, January 2024. doi: 10.1038/s41398-023-02720-y.

[13] Jenny Yang, Andrew A. S. Soltan, and David A. Clifton. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. *npj Digital Medicine*, 5(1):69, June 2022. doi: 10.1038/s41746-022-00614-9.

[14] Vallijah Subasri, Amrit Krishnan, Ali Kore, Azra Dhalla, Deval Pandya, Bo Wang, David Malkin, Fahad Razak, Amol A. Verma, Anna Goldenberg, and Elham Dolatabadi. Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models. *JAMA Network Open*, 8(6):e2513685, June 2025. doi: 10.1001/jamanetworkopen.2025.13685.

[15] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport, February 2019. arXiv:1812.04597 [stat].

[16] Zachary C. Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and Correcting for Label Shift with Black Box Predictors, July 2018. arXiv:1802.03916 [cs].

[17] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[18] Felipe Maia Polo, Rafael Izbicki, Evanildo Gomes Lacerda, Juan Pablo Ibieta-Jimenez, and Renato Vicente. A unified framework for dataset shift diagnostics. *Information Sciences*, 649: 119612, November 2023. doi: 10.1016/j.ins.2023.119612.

[19] E.A. Maharaj. Cluster of Time Series. *Journal of Classification*, 17(2):297–314, July 2000. doi: 10.1007/s003570000023.

[20] Elizabeth Ann Maharaj, Paula Brito, and Paulo Teles. A test to compare interval time series. *International Journal of Approximate Reasoning*, 133:17–29, June 2021. doi: 10.1016/j.ijar.2021.02.008.

[21] I. V. Basawa, L. Billard, and R. Srinivasan. Large-sample tests of homogeneity for time series models. *Biometrika*, 71(1):203–206, April 1984. doi: 10.1093/biomet/71.1.203.

[22] Hannah Horng, Apurva Singh, Bardia Yousefi, Eric A. Cohen, Babak Haghighi, Sharyn Katz, Peter B. Noël, Russell T. Shinohara, and Despina Kontos. Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. *Scientific Reports*, 12(1):4493, March 2022. doi: 10.1038/s41598-022-08412-9.

[23] Trent Henderson and Ben D. Fulcher. An Empirical Evaluation of Time-Series Feature Sets, October 2021. arXiv:2110.10914 [cs].

[24] Gašper Petelin, Gjorgjina Cenikj, and Tome Eftimov. Towards understanding the importance of time-series features in automated algorithm performance prediction. *Expert Systems with Applications*, 213:119023, March 2023. doi: 10.1016/j.eswa.2022.119023.

[25] Carl H. Lubba, Sarab S. Sethi, Philip Knaute, Simon R. Schultz, Ben D. Fulcher, and Nick S. Jones. catch22: CAnonical Time-series CHaracteristics, January 2019. arXiv:1901.10200 [cs].

[26] Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism, February 2017. arXiv:1608.05745 [cs].

[27] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1): 6085, April 2018. doi: 10.1038/s41598-018-24271-9.

[28] Vijay Ekambaram, Arindam Jati, Nam Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. TSMixer: Lightweight MLP-Mixer Model for Multivariate Time Series Forecasting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 459–469, August 2023. doi: 10.1145/3580305.3599533. arXiv:2306.09364 [cs].

[29] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. Patient Subtyping via Time-Aware LSTM Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 65–74, Halifax NS Canada, August 2017. ACM. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3097997.

[30] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the Language of Time Series, March 2024. arXiv:2403.07815 [cs].

[31] Hoda Abbasizanjani, Fatemeh Torabi, Stuart Bedston, Thomas Bolton, Gareth Davies, Spiros Denaxas, Rowena Griffiths, Laura Herbert, Sam Hollings, Spencer Keene, Kamlesh Khunti, Emily Lowthian, Jane Lyons, Mehrdad A. Mizani, John Nolan, Cathie Sudlow, Venexia Walker, William Whiteley, Angela Wood, and Ashley Akbari. Harmonising electronic health records for reproducible research: challenges, solutions and recommendations from a UK-wide COVID-19 research collaboration. *BMC Medical Informatics and Decision Making*, 23:8, January 2023. doi: 10.1186/s12911-022-02093-0.

[32] Arian Aminoleslami, Geoffrey M. Anderson, and Davide Chicco. EHRs Data Harmonization Platform, an easy-to-use shiny app based on recodeflow for harmonizing and deriving clinical features, November 2024. arXiv:2411.10342 [cs] version: 1.

[33] Carlos Sáez, Alba Gutiérrez-Sacristán, Isaac Kohane, Juan M García-Gómez, and Paul Avillach. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. *GigaScience*, 9(8):giaa079, July 2020. doi: 10.1093/gigascience/giaa079.

[34] John J. Stephen, Padraig Carolan, Amy E. Krefman, Sanaz Sedaghat, Maxwell Mansolf, Norrina B. Allen, and Denise M. Scholtens. psHarmonize: Facilitating reproducible large-scale pre-statistical data harmonization and documentation in R. *Patterns*, 5(8):101003, August 2024. doi: 10.1016/j.patter.2024.101003.

[35] Lin Lawrence Guo, Ethan Steinberg, Scott Lanyon Fleming, Jose Posada, Joshua Lemmon, Stephen R. Pfohl, Nigam Shah, Jason Fries, and Lillian Sung. EHR foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports*, 13(1):3767, March 2023. doi: 10.1038/s41598-023-30820-8.

[36] Lin Lawrence Guo, Stephen R. Pfohl, Jason Fries, Jose Posada, Scott Lanyon Fleming, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Systematic Review of Approaches to Preserve Machine Learning Performance in the Presence of Temporal Dataset Shift in Clinical Medicine. *Applied Clinical Informatics*, 12(4):808–815, August 2021. doi: 10.1055/s-0041-1735184.

[37] Seungyeon Lee, Changchang Yin, and Ping Zhang. Stable clinical risk prediction against distribution shift in electronic health records. *Patterns*, 4(9):100828, September 2023. doi: 10.1016/j.patter.2023.100828.

[38] Olive Jean Dunn. Multiple Comparisons Using Rank Sums. *Technometrics*, 6(3):241–252, 1964. doi: 10.2307/1266041.

[39] Thomas E. Nichols and Andrew P. Holmes. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1):1–25, October 2001. doi: 10.1002/hbm.1058.

[40] Douglas M. Potter. Omnibus permutation tests of the association of an ensemble of genetic markers with disease in case-control studies. *Genetic Epidemiology*, 30(5):438–446, 2006. doi: 10.1002/gepi.20155.

[41] Kane Meissel and Esther S. Yao. Using Cliff's Delta as a Non-Parametric Effect Size Measure: An Accessible Web App and R Tutorial. *Practical Assessment, Research, and Evaluation*, 29 (1), January 2024. doi: 10.7275/pare.1977.

[42] Ewa Tomczak and Maciej Tomczak. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *TRENDS in Sport Sciences*, 21(1), 2014.

[43] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. UpSet: Visualization of Intersecting Sets. *IEEE transactions on visualization and computer graphics*, 20(12):1983–1992, December 2014. doi: 10.1109/TVCG.2014.2346248.

[44] Haoran Zhang, Harvineet Singh, Marzyeh Ghassemi, and Shalmali Joshi. "Why did the Model Fail?": Attributing Model Performance Changes to Distribution Shifts. In *Proceedings of the 40th International Conference on Machine Learning*, pages 41550–41578. PMLR, July 2023.

A Supplementary Material

A.1 Outline of procedure

Input
Data: EHR time series across $\mathbb{B}$ sites for a total of N patients and M measures over t timepoints, such that $\mathbb{D} = (X_1, X_2, \ldots, X_N)$, where $X_i = (m_{i,1}, m_{i,2}, \ldots, m_{i,M}), m_{i,j} \in \mathbb{R}^t$
Feature extractor: $f : \mathbb{R}^t \to \mathbb{R}^z$.

Steps
Feature extraction
initialize $\mathbb{D}^* = []$
for patient X_i in $\mathbb{D}$ **do**
  initialize $Z_i = []$
  for measure $m_{i,j}$ in X_i **do**
    $z_{ij} = f(m_{ij})$
    append $Z_i \leftarrow Z_i + z_{ij}$ ▷ append
  end for
  append $\mathbb{D}^* \leftarrow \mathbb{D}^* + Z_i$ ▷ append
end for ▷ $\mathbb{D} \in \mathbb{R}^{N,M,t}$ to $\mathbb{D} \in \mathbb{R}^{N,(M \times z)}$
Multi-site comparison
for measure j in M **do**
  for feature k in z **do**
    p (per feature) $\leftarrow$ Dunn's test (Holm's correction) for feature k of measure j over $\mathbb{B}$ sites
    $eff \leftarrow$ Glass's biserial rank correlation for feature k of measure j over $\mathbb{B}$ sites
  end for
  p (per measure) $\leftarrow$ permutation test of number of significant features over 1000 iterations
end for