Towards Comprehensive Benchmarking of Domain Generalization and Adaptation for Time-Series Forecasting in Clinical Multivariate Time-Series Forecasting

Mayra Elwes
Institute for Biomedical Informatics, University of Cologne, Cologne, Germany
mayra.elwes@uk-koeln.de

Abstract

Deep learning models for clinical time-series forecasting face significant challenges in generalizability when deployed across different clinical settings. Various choices in the deep learning pipeline can affect a model's out-of-distribution performance. This abstract introduces a comprehensive benchmarking framework to assess domain generalization and adaptation under varying distribution shifts in multivariate time-series forecasting using real-world data. Preliminary empirical results on the influence of backbone model architecture and a proxy metric, the *robustness slope*, are presented alongside remaining challenges.

1 Introduction

Despite the importance of generalizability in many applied fields, a gap remains between theoretical foundations and practical application for multivariate time-series data. While established statistical learning theories provide generalization bounds and error guarantees, these theoretical limits often prove too abstract and wide. Consequently, recent research has shifted towards empirical experimentation to develop more practical decision guidelines. Domain adaptation research extends this by accounting for the distribution shift between source and target data when estimating generalization bounds [Redko *et al.*, 2022]. In these pipelines, numerous decisions can be taken and combined endlessly. To enable systematic benchmarking, we structured these choices into *degrees of freedom*: data choice, preprocessing, backbone architecture, training strategy, and postprocessing. There is a need to systematically benchmark these variables and their combinations to create practical guidelines as shown in Figure 1.

While a model may be marketed as "generalizable" during development, evidence suggests that such claims often fail to hold up in real-world clinical settings [Chekroud *et al.*, 2024]. My primary objective is to develop a robust domain adaptation pipeline for these "clinic-to-clinic" scenarios, focusing on forecasting blood oxygenation levels (Spo2) and mean arterial blood pressure (ABP Mean) in the Intensive Care Unit. More specifically, this research explores: **RQ1:** How do the various *degrees of freedom* in a machine learning pipeline

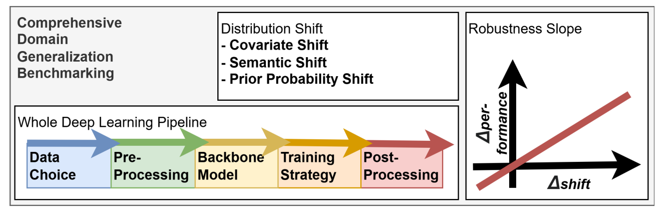

Figure 1: Important aspects in domain generalization benchmarking

interact to influence the generalization of multivariate time-series forecasting models? **RQ2:** What is a suitable proxy metric to benchmark *generalization/adaptation* methods with regard to different magnitudes and types of distribution shift? **RQ3:** How to improve the robustness of domain adaptation methods for clinical multivariate time-series forecasting?

My research extends the state of the art in three key dimensions: task complexity, comprehensive benchmarking, and a proxy metric for *generalizability*. While most domain adaptation research is limited to classification, I focus on multivariate time-series forecasting, addressing the unique temporal and inter-channel dependencies of clinical data. Unlike algorithm-centric benchmarks, my approach considers all *degrees of freedom* and evaluates interactions across the entire pipeline, from source data selection to postprocessing. Importantly, the benchmarking is carried out on clinically grounded, real-world data source-to-target scenarios. I move beyond reporting static "mean performance" by introducing the *robustness slope*. This allows for a continuous assessment of model decay across natural and synthetic distribution shifts, providing a more rigorous empirical validation of theoretical generalization bounds than current benchmarks.

2 Contributions

My contributions consist of methodological frameworks and empirical insights into the robustness of clinical time-series adaptation. I designed and implemented a flexible framework to assess the *degrees of freedom* in generalization and adaptation pipelines. This includes the open-source framework Algorithm2Domain[1], which extends ADATIME [Ragab *et al.*, 2023] with support for multivariate forecasting-specific evaluation metrics. Further, I defined a suite of clin-

[1]https://github.com/BI-K/Algorithm2Domain

ically relevant source-to-target scenarios using the MIMIC-III matched waveform database. To facilitate this, I developed WavePrep[2], an open-source tool for preprocessing high-frequency ICU data into ML-ready datasets[Elwes *et al.*, 2026b].

Preliminary work [Elwes *et al.*, 2026a] identified a proxy metric for generalizability: the *robustness slope*. The *robustness slope* measures the slope of the linear relationship between distribution shift and performance degradation, accounting for shifts across the source train, source test, and target data. Empirical investigations across eight backbone architectures (CNN, (Bi-)LSTM, TCN, LTCN, CfCN, Transformer, GRU, and ResNet18) confirmed the observed linear relation, which has been observed similarly for image-based classification tasks in [Friedman and Chowers, 2025]. The investigations also revealed that target-domain performance is primarily influenced by covariance shift across model backbones. Scenario-averaged performance differed minimally between backbone model architectures for the specific downstream tasks and scenarios. These results still need to be verified by further tests under more extreme shifts. Further experiments addressing other aspects of the *degrees of freedom* will be carried out. Based on these results, a strategy for improving the most robust adaptation pipeline will be developed.

3 Analysis Method

I evaluate current state-of-the-art methods on two primary downstream tasks: SpO2 and ABP Mean forecasting. The benchmark utilizes eight clinically derived source-to-target scenarios. To ensure a fair and rigorous comparison, each evaluated method undergoes hyperparameter tuning. Sensitivity lies in the chosen distribution shift metric; currently, I am reviewing the literature for metrics suitable for assessing covariate, semantic, and prior probability shifts in multivariate time-series prediction tasks, as well as for combining different types of distribution shift.

Currently, the primary evaluation metric is the *robustness slope*. While initial results suggest a linear relationship between distribution shift and performance decay, I am currently expanding this validation via synthetic stress testing. Using a controllable shift-generation method inspired by [Friedman and Chowers, 2025] adapted for multivariate time-series forecasting, I generate synthetic time-series data to simulate extreme distribution shifts. This allows me to test the limits of the observed linearity and identify potential ”breaking points” where the proxy metric may become nonlinear. To move beyond public datasets, ongoing efforts in accordance with the German Health Data Utilization Act are collecting relevant data from local patient cohorts across two units of the University Hospital Cologne, thereby enhancing the test scenarios with natural ”clinic-to-clinic” scenarios. This local validation serves as the final ”gold standard” for testing the clinical applicability of the adaptation strategies.

4 Discussion

The uncovered *robustness slope* provides a way to contextualize and report the performance of a generalization/adaptation method in a comparable manner across application domains and downstream tasks, where standard benchmarking datasets are no longer available. ”AI as Medical Device” regulations require strict post-market monitoring and robustness [Cuocolo *et al.*, 2025]; this benchmarking approach provides a path to certification for a certain safe range of distribution shifts. Refinements of the strategy for quantifying distribution shift may be used to improve the design of loss metrics in domain adaptation methods.

[2]https://github.com/BI-K/WavePrep

References

[Chekroud *et al.*, 2024] Adam M. Chekroud, Matt Hawrilenko, Hieronimus Loho, Julia Bondar, Ralitza Gueorguieva, Alkomiet Hasan, Joseph Kambeitz, Philip R. Corlett, Nikolaos Koutsouleris, Harlan M. Krumholz, John H. Krystal, and Martin Paulus. Illusory generalizability of clinical prediction models. *Science*, 383(6679):164–167, January 2024.

[Cuocolo *et al.*, 2025] Renato Cuocolo, Diana Bernardini, Daniel Pinto Dos Santos, Michail E. Klontzas, Tugba Akinci D’Antonoli, Luís Curvo Semedo, Robin Decoster, Merel Huisman, Elmar Kotter, Luis Martí-Bonmatí, Costin Minoiu, Emanuele Neri, Konstantin Nikolaou, Maija Radzina, Evis Sala, Susan C. Shelmerdine, Laurens Topff, Michelle C. Williams, and European Society of Radiology (ESR). AI medical device post-market surveillance regulations: consensus recommendations by the European Society of Radiology. *Insights into Imaging*, 16(1):275, December 2025.

[Elwes *et al.*, 2026a] Mayra Elwes, Jonas Alfitian, Karen Hornung, Bhanu Koppolu, Oya Beyan, and Ekaterina Kutafina. Distribution Shift Analysis in Generalizable Modelling: Intensive Care Time-Series Data - Preprint. In *Studies in Health Technology and Informatics*. IOS Press, May 2026.

[Elwes *et al.*, 2026b] Mayra Elwes, Jonas Alfitian, Karen Hornung, Bhanu Koppolu, Oya Beyan, and Ekaterina Kutafina. WavePrep - Create Machine-Learning-Ready Datasets from MIMIC Waveforms- Preprint. Innsbruck University Press, 2026.

[Friedman and Chowers, 2025] Roy Friedman and Rhea S Chowers. Control+Shift: Generating Controllable Distribution Shifts. In Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi, editors, *Computer Vision – ECCV 2024 Workshops*, volume 15642, pages 321–334. Springer Nature Switzerland, Cham, 2025. Series Title: Lecture Notes in Computer Science.

[Ragab *et al.*, 2023] Mohamed Ragab, Emadeldeen Eldele, Wee Ling Tan, Chuan-Sheng Foo, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. ADATIME: A Benchmarking Suite for Domain Adaptation on Time Series Data. *ACM Trans. Knowl. Discov. Data*, 17(8):106:1–106:18, May 2023.

[Redko *et al.*, 2022] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees, July 2022. arXiv:2004.11829 [cs].