BEDS-Bench: Behavior of EHR-models under Distributional Shift–A Benchmark

Anonymous Author(s)
Affiliation
Address
email

Abstract

Machine learning (ML) has recently demonstrated impressive progress in predictive accuracy across a wide array of tasks. Most ML approaches focus on generalization performance on unseen data that are "similar" to the training data (a.k.a. *In-Distribution*, or IND). However, real world applications and deployments of ML rarely enjoy the comfort of encountering examples that are always IND. In such situations, most ML models commonly display erratic behavior on *Out-of-Distribution* (OOD) examples, such as assigning high confidence to wrong predictions, or vice-versa. Implications of such unusual model behavior are further exacerbated in the healthcare setting, where patient health can potentially be put at risk. It is crucial to study the behavior and robustness properties of models under distributional shift, understand common failure modes, and take mitigation steps before the model is deployed. Having a benchmark that shines light upon these aspects of a model is a first and necessary step in addressing the issue. Recent work and interest in increasing model robustness in OOD settings have focused more on image modality, both in terms of methods as well as benchmarks, while the Electronic Health Record (EHR) modality is still largely under-explored. We aim to bridge this gap by releasing BEDS-Bench, a benchmark for quantifying the behavior of ML models over EHR data under OOD settings. We use two open access, de-identified EHR datasets to construct several OOD data settings to run tests on. The benchmark exercises several clinical prediction tasks, OOD data settings, and measures relevant metrics that characterize crucial aspects of a model's OOD behavior. We evaluate several learning algorithms under BEDS-Bench and find that all of them show poor generalization performance under distributional shift in general. Our results highlight the need and the potential to improve robustness of EHR models under distributional shift, and BEDS-Bench provides one way to measure progress towards that goal.

1 Introduction

Machine Learning models are typically validated on test sets that are similar to the training set. A common assumption in statistical learning theory is that all examples (both train and test) are drawn independently and identically from the same data distribution (IID). Though the IID assumption is a strong one, in practice it is hard to ascertain if it is always being met. When an ML model is deployed in a real world setting, the likelihood of encountering OOD inputs is far higher. In situations when an ML model is presented with OOD inputs, its behavior can be hard to describe theoretically, and tends to be unknown practically (Figure 1). The first step in fixing the behavior of models in OOD settings is to measure and quantify it with benchmarks. Benchmarks and datasets paint a target for the research community to focus and align on, thereby catalyzing the progress of the field [Deng et al., 2009, Dua and Graff, 2017]. They also serve a crucial role as an objective measure of progress

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

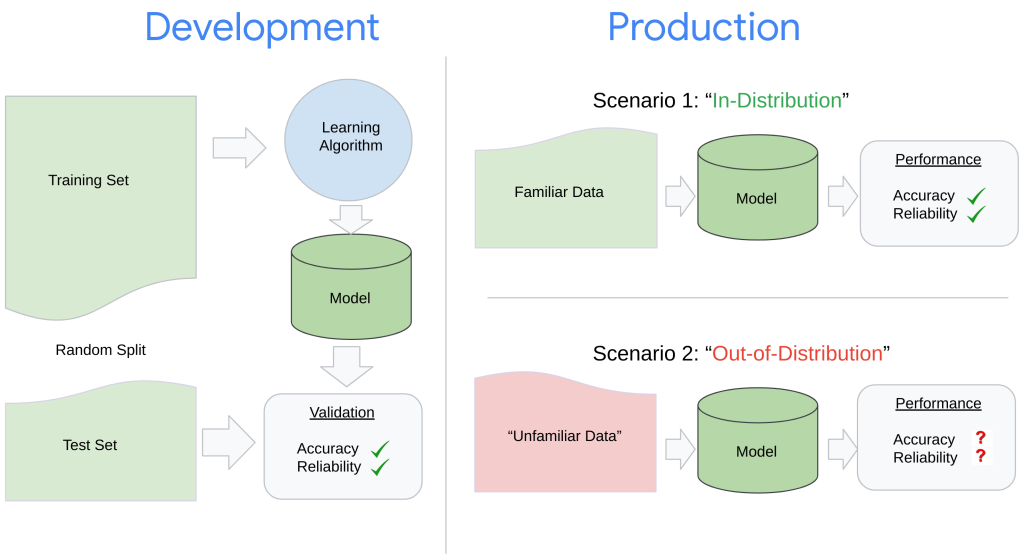

Figure 1: Illustration of a scenario where a model can encounter data that is very dissimilar to the training data distribution.

towards that goal. Yet, there is a lack of good benchmarks for studying the behavior of models on EHR data under OOD settings, which our work attempts to address.

Studying the behavior of EHR models under distributional shift is more than just a purely academic endeavour [Nestor et al., 2019]. There are numerous real world situations where a model may encounter patients who are systematically different from the training data for legitimate reasons. Some specific examples where the train and test distributions may differ include:

- **Changes in the patient population:** The demographics of a patient population may change over time due to gentrification of neighborhoods around a health system, maturing public health policies, global population dynamics etc. Consider for example the rising proportion of females in the Veterans Affairs agency. This may result in models encountering patients from a different distribution than the historical data on which the model was trained.
- **Changes in the practice of medicine:** The COVID-19 pandemic is an example of a dramatic shift in the field of medicine as a whole. It introduced major distributional shifts via changes in the patient population, but also changes in the practice of medicine, the therapies being used and the operational processes of the hospital (e.g. due to resource shortages).
- **Portability of models between health systems:** There is increased sharing of pre-trained EHR models between hospital sites, with vendors offering pre-built models [Tan et al., 2020] and academic consortia such as the Observational Health and Data Sciences Initiative (OHDSI) enabling model portability via common data standards. While this is excellent for broadening the impact of machine leaning and encouraging research reproducibility, it also increases the likelihood of training and deployment datasets being divergent due to differences in both populations and data formats.

When the behavior of an EHR model under distributional shift is unknown, there is a risk that predictions on OOD inputs might be wrong yet highly confident, thereby potentially increasing clinical risk for those patients. This is particularly important as EHR models start to be deployed in real world clinical settings [Sendak et al., 2020].

While OOD benchmarks have been extremely impactful in the imaging domain, creating an analagous EHR benchmark is challenging. First, privacy concerns makes it hard to even get access to multiple large EHR datasets. In addition, EHR data is complex, heterogeneous and highly site-specific. This makes it difficult to harmonize multiple EHR datasets in order to perform cross-site experiments to evaluate OOD behaviour. Furthermore, while the benchmark tasks in the imaging domain is typically

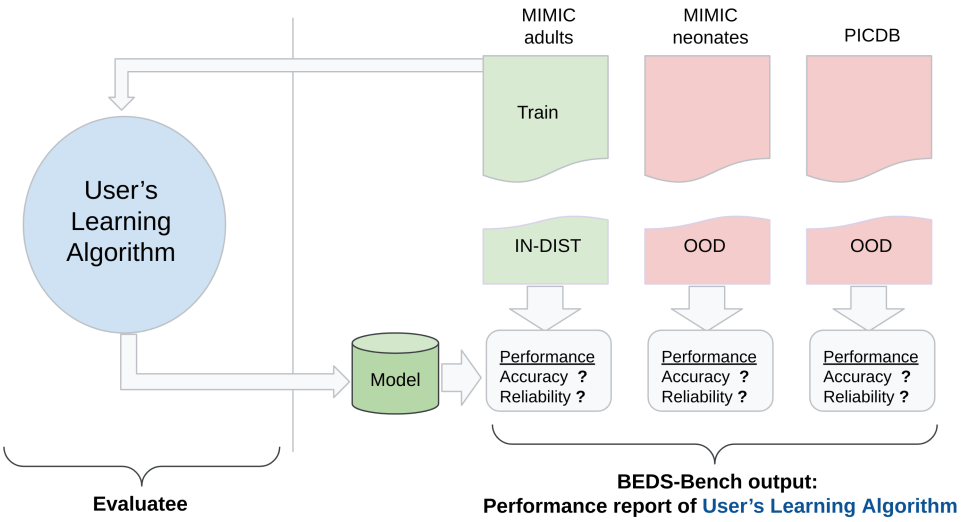

Figure 2: BEDS-Bench setting

a classification problem with readily available labels, EHR tasks are often less straightforward. For example, defining a task involving EHR data necessarily involves nuanced data and temporal considerations, such as deciding a consistent *prediction time* for all examples (e.g. predicting onset of diabetes is meaningful only when the disease is not yet diagnosed), choice of a suitable time window and data sources from which data is extracted for features (broad window makes for more accurate models, but reduces the population who have sufficient data to be applied upon), determining a suitable representation for the extracted sparse and heterogenous data (handling a mixture of real values, categorical values, ordinal values, timestamps, handwritten text, images, missing values etc.), assigning labels (e.g. how to accurately determine which patients actually have diabetes), among other challenges.

To this end, BEDS-Bench is a benchmark created using two open access de-identified EHR datasets. BEDS-Bench simulates OOD settings by creating intentionally dissimilar train and test sets, and measures several metrics around model performance in each of these settings (see Figure 2), for three common downstream classification tasks. The code for pre-processing and model evaluation is open-sourced and we hope that these benchmarks are a useful resource for the EHR community to develop more rigorous methods to characterize OOD behaviour.

2 BEDS-Bench

The BEDS-Bench tool is designed to generate a performance report of a learning algorithm regarding the behavior of models trained by this algorithm under various types of distributional shift in the test data. The general approach taken by BEDS-Bench is to partition data in several ways into *intentionally* dissimilar subsets in order to artificially simulate IND vs OOD settings. Models are trained on the train split of a certain subset for one of the standardized tasks, and tested on the test splits of all the subsets while measuring relevant metrics. The test split corresponding to the subset from which the model was trained is considered IND, while test splits from the other subsets in the partition are considered OOD. Figure 2 describes the workflow in one particular setting. This procedure is repeated by cycling through every subset in every partition to be the IND, while the other subsets in the partition are considered OOD.

In the rest of this section we describe what is an ideal model's behavior under OOD data, the details of the methodology of BEDS-Bench including descriptions of the datasets used, partitions created, tasks for which models are trained for, and metrics measured on the test sets.

2.1 Ideal Model Behavior

Before designing a benchmark, it is crucial to first define what we consider is the ideal behavior of a model. The tests and metrics of the benchmark need to then be chosen to shine light on these aspects of the model and enable objective comparison across multiple algorithms. The following notation of the ideal model behavior informs the design of BEDS-Bench:

- Generalization: When a model is tested on a distribution that is different from the one it was trained on, it is possible, and understandable, for the model performance to drop to some extent. A common generalization metric (in case of classification tasks) is the Area Under the Receiver-Operator Characteristic Curve (AUC) which measures the ability of a model to discriminate between two classes. The drop in generalization performance might likely be larger for test distributions that are "farther" from the train distributions. Yet, *an ideal model should have at least a minimal level of generalization robustness to OOD data, such as not performing worse than random guessing (i.e. maintain AUC $\geq$ 0.5).*
- Calibration: Calibration refers to the property that probabilities output by a model are agree with the observed empirical frequency of events. For example, among all days which had a rain forecast probability of 80%, approximately 8 out of 10 days should observe rain in the long run. Calibration is a property that is orthogonal to discrimination, and hypothetically it is possible to have models with any mix of levels of calibration and discrimination. *An ideal model is not only well-calibrated in its predictions on IND data, but also on OOD data, especially when generalization on OOD data has worsened.*
- Confidence: Closely related to the notion of calibration is *confidence*. Typically confidence of a prediction is measured with metrics such as predictive entropy, or predictive variance. The larger the entropy or variance, the lower the confidence of that prediction. *If an ideal model's OOD generalization performance is lower than IND, then the confidence in the OOD predictions will be lower than in the IND predictions*.

  While we do measure the ability of a model to discriminate OOD vs IND inputs by assigning lower confidence scores to OOD, we also emphasize that this test involves additional nuances that need to be considered before interpreting the results. We discuss this further in Section C.

2.2 Benchmark

The details about the datasets used, data splits defined, tasks evaluated, and metrics that are measured as part of the BEDS-Bench benchmark are described in the Appendix (Section A).

3 Conclusion

In this work, we propose a benchmark, BEDS-Bench, to evaluate an EHR ML model performance under distributional shift of the test data. We evaluate this benchmark on several algorithms, and find that no single algorithm demonstrates satisfactory robustness behavior over a wide range of OOD settings. We also find that no single algorithm works better than another across the board, including algorithms designed with OOD robustness in mind. While prior works have identified that most discriminative models are not reliable in detecting OOD examples reliably on medical tabular data, our work confirm this and in addition find that all the models we tested fare poorly in maintaining calibration under distributional shift of EHR data. This underscores the need for further research into robustness evaluation of EHR models, especially as these models are increasingly deployed in real-world clinical settings.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. OSDI'16, page 265–283, USA, 2016. USENIX Association. ISBN 9781931971331.

Olivier Bodenreider. The unified medical language system (umls): Integrating biomedical terminology, 2004.

Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Michael W. Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M. Dai. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 204–213, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370462. doi: 10.1145/3368555.3384457. URL https://doi.org/10.1145/3368555.3384457.

Meire Fortunato, Charles Blundell, and Oriol Vinyals. Bayesian recurrent neural networks, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/gal16.html.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

J Kemp, K Zhang, and A Dai. Mimic-iii - sequenceexamples for tensorflow modeling, 2020.

Balaji Lakshminarayanan, Daniel M. Roy, and Yee Whye Teh. Mondrian forests: Efficient online random forests. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3140–3148, Cambridge, MA, USA, 2014. MIT Press.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness, 2020.

Bret Nestor, Matthew B. A. McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C. Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Alvin Rishi Rajkomar, Eyal Oren, Kai Chen, Andrew Dai, Nissan Hajaj, Mila Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Per Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alex Mossin, Justin Jesada Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel Volchenboum, Kat Chou, Michael Pearson, Srinivasan Madabushi, Nigam Shah, Atul Butte, Michael Howell, Claire Cui, Greg Corrado, and Jeff Dean. Scalable and accurate deep learning for electronic health records. *npj Digital Medicine*, 2018. URL https://www.nature.com/articles/s41746-018-0029-1.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

Mark P. Sendak, Joshua D'Arcy, Sehj Kashyap, Michael Gao, Marshall Nichols, Kristin Corey, William Ratliff, and Suresh Balu. A path for translation of machine learning products into healthcare delivery. *EMJ Innovations*, 2020. doi: https://doi.org/10.33590/emjinnov/19-00172.

Eugene M. Tan, Rahul Kashyap, Ian C. Olson, and John C. OHoro. Validation of a retrospective computing model for mortality risk in the intensive care unit. *Mayo Clinic Proceedings: Innovations, Quality and Outcomes*, 4(5):575 – 582, 2020. ISSN 2542-4548. doi: https://doi.org/10.1016/j.mayocpiqo.2020.09.001. URL http://www.sciencedirect.com/science/article/pii/S2542454820301788.

Dennis Ulmer, Lotta Meijerink, and Giovanni Cinà. Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data, 2020.

Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 222–235, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370462. doi: 10.1145/3368555.3384469. URL https://doi.org/10.1145/3368555.3384469.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, 01 2010. ISSN 1067-5027. doi: 10.1197/jamia.M3378. URL https://doi.org/10.1197/jamia.M3378.

Xian Zeng, Gang Yu, Yang Lu, Linhua Tan, Xiujing Wu, Shanshan Shi, Huilong Duan, Qiang Shu, and Haomin Li. Pic, a paediatric-specific intensive care database. *Scientific data*, 7:14, 2020.

Appendix

A BEDS-Bench

A.1 Datasets

To develop the benchmark, we make use of two open access Intensive Care Unit (ICU) EHR datasets - Medical Information Mart for Intensive Care III (MIMIC-III, or MIMIC) [Johnson et al., 2016], and Paediatric Intensive Care database (PICDB) [Zeng et al., 2020]. Both the datasets are available for download from PhysioNet [Goldberger et al., 2000 (June 13].

The MIMIC dataset has data related to patients who were admitted to the ICU at Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset covers 58,977 ICU stays of 46,520 patients. All the patients were either adults or neonates (newborn babies).

The PICDB dataset was collected at the Children's Hospital of Zhejiang University School of Medicine between 2010 and 2018. It covers a total of 13,450 ICU stays of 12,811 patients who were all minors (newborn up to 18 years of age).

An overview and comparison of the two datasets, including types of data present in each dataset and their encoding formats is presented in Table 5. Both the datasets are represented as relational databases, with a comma separated value (CSV) formatted file per table.

A.2 Harmonization

The way BEDS-Bench works is by creating intentionally dissimilar subsets of data to simulate OOD settings. One natural setting is to consider model behavior when trained on data from MIMIC and tested on PICDB data, and vice versa. Conducting such cross-dataset experiments is quite common, and straight forward with image data. For harmonizing two images datasets, the main considerations are around matching the resolutions, channel count, bits per color etc. which are all quite easily handled. Yet, harmonizing two different sources of EHR data is a lot more involved, with careful considerations required in finding a common set of tables, vocabularies (to codify categorical data), units (to represent continuous data), representation of time, and other semantic reconciliations.

The broad strategy we follow in harmonizing the two datasets is to identify a subset of tables, columns, and rows which can potentially be matched up, and exclude the remaining.

In PICDB the diagnostic codes are coded in the Chinese Edition of the International Classification of Diseases, Tenth Revision (ICD-10CN), whereas MIMIC uses the International Classification of Diseases, Ninth Revision (ICD-9). We perform a one-to-many mapping from ICD-10CN to ICD-9 using the Unified Medical Language System (UMLS) database [Bodenreider, 2004].

For medication codes, we map both the data sources to the RXCUI coding. MIMIC medications are coded in the National Drug Code, which uniquely map to RXCUI. For PICDB we start with the textual descriptions of the medications and run them through the MedEx system to extract the RXCUI codes [Xu et al., 2010].

While the laboratory tests are coded with custom codes in both the datasets, some of the custom codes have an accompanying Logical Observation Identifiers Names and Codes (LOINC) code. We use the LOINC code as the common vocabulary and include only those rows for which the custom code has a corresponding LOINC code.

MIMIC has a very rich representation of vitals and chart events. PICDB on the other hand has a total of nineteen vital and chart event types. We use the event type groupings from the MIMIC-Extract project to map a subset of the MIMIC chart event codes to the corresponding PICDB chart event codes [Wang et al., 2020].

The Inputevents and Outputevents table record the total volumes of different types of fluids that enter and exit the patient during the stay. While MIMIC records both the volumes and the types of the fluids, PICDB only records the volumes (without an associated fluid type). From a medical perspective, while knowing the type of fluid is certainly useful, just knowing the volume of fluids going in and out of the patients is also informative in itself. Thus we exclude the fluid type codes from MIMIC and retain only the volume information for the purposes of harmonization.

Partition	Slice	Criteria
Demographics	MIMIC-adult	DB = "MIMIC" and AGE > 15yr
	PICDB-paed	DB = "PICDB"
	MIMIC-neonate	DB = "MIMIC" and AGE < 1mo
Biological Sex	MIMIC-Female	DB = "MIMIC" and Gender = "F"
	MIMIC-Male	DB = "MIMIC" and Gender $\neq$ "F"
Ageing	MIMIC-lt50	DB = "MIMIC' and 15yrs < AGE $\leq$ 50yrs
	MIMIC-5060	DB = "MIMIC' and 50yrs < AGE $\leq$ 60yrs
	MIMIC-6070	DB = "MIMIC' and 60yrs < AGE $\leq$ 70yrs
	MIMIC-7080	DB = "MIMIC' and 70yrs < AGE $\leq$ 80yrs
	MIMIC-gt80	DB = "MIMIC' and 80yrs < AGE

Table 1: Data partitions and slices to construct IND vs OOD settings

Table 7 in the Appendix summarizes the various code harmonization approaches that were applied.

A.3 Data Processing

In order to create a supervised learning dataset out of an EHR relational database, certain additional data processing steps are necessary. Each example in the supervised learning dataset corresponds to the data from one hospital admission. First, we exclude all hospital admissions that are shorter than 30 hours, and within those included, we use data up to the first 24 hours since admission. The additional 6 hours "gap" after the first 24 hours of data is common practice to avoid leaking of information of the label into the covariates [Wang et al., 2020]. Further, we only include the first admission of a patient, and exclude admissions after the first discharge, if any. Finally the dataset is randomly divided into train and test splits (80% train, 20% test).

The set of resulting tables after applying the inclusion and exclusion criteria (including both train and test), with their row and column counts is summarized in Table 9 in the Appendix. This is the harmonized dataset using which the various experimental OOD settings are created.

A.4 Data Splits

The benchmark creates three different partitions of the data, each partition having between two to five slices. Within each partition, the slices are completely non-overlapping, and are *characteristically different* to varying degrees depending on the partition. The names and definitions (inclusion criteria) of each of the slices of all the partitions are in Table 1.

The Demographics partition has three slices - MIMIC-adult, MIMIC-neonate, and PICDB-paed. The differences in the slices in this partition are somewhat stark. Not only are the differences between paediatric (especially neonates) and adults particularly pronounced, the MIMIC vs PICDB slices present even more differences, including very distinct populations, health systems and accompanying treatment practices, etc.

The Biological Sex partition separates the MIMIC dataset into Female and Male slices. Both the EHR datasets codify sex as binary and BEDS-Benchfollows the convention. This partition intends to highlight the model behavior under the extreme cases of shift in gender balance.

The Ageing partition slices the adults into different age bands, representing progressively older patients with each band. The age ranges in years used to define the bands are (15-50], (50-60], (60-70], (70-80] and (80,∞).

It may be observed that some of the partitions have slices which are so blatantly dissimilar that sometimes it would be unreasonable to expect a model to ever generalize over to such a distinctly different dataset, or to even consider such generalization goals as clinically relevant. Yet, we argue that these obviously-OOD settings are great examples of scenarios where any reasonably safe model would necessarily need to display some degree of robustness, and hence make for good tests to be included as part of an OOD benchmark suite.

We also note that the distribution of Race in the EHR datasets is quite skewed, with several races having too few examples to be sufficient to form a partition that includes slices for all races. After

considerations of fairness and ethics, we look forward to finding additional EHR datasets that will allow us to construct a more inclusive race based partitioning.

A.5 Supervised Learning Tasks

BEDS-Bench includes three supervised learning tasks to evaluate algorithms on: In-Hospital Mortality (Mort), Remaining Length-of-Stay > 3 days (LoS3+), and Remaining Length-of-Stay > 7 days (LoS7+). All the three tasks are common canonical EHR tasks widely explored in the literature Wang et al. [2020], Rajkomar et al. [2018], and framed as binary classification, with names suggestive of their labels. The Mort task has a label of 1 only if the patient passed away during the hospital stay of that example. Even if the patient passed away soon after discharge or during a follow-up admission, the label remains 0. The LoS3+ (or Los7+) task has a label of 1 only if the patient will end up having at least 3 (or 7) days worth of remaining time in their current stay.

The class balance varies significantly depending on the task and data slice. While mortality of MIMIC-neonate can be as low as 0.5% (fortunately) on the one hand, the three day length of stay for the PICDB-paed slice is as high as 91.2% on the other. The class balances for each of the three tasks on all the slices, along with the number of examples in each slice is listed in the Appendix (Table 8).

A.6 Metrics and Report

The BEDS-Bench evaluates the performance of an algorithm in several test settings as measured by several metrics. For notation, let use denote the number of examples by $n \in \mathbb{N}$, $i \in \mathbb{N}$ as the example index number where $1 \leq i \leq n$, $y_i \in \{0, 1\}$ as the label (correct answer) of the ith example, and $\hat{y}_i \in [0, 1]$ as the predicted probability by a model for the ith example.

- **Task-AUC** - This metric is the Area Under the Receiver Operator Characteristic (ROC) curve (AUROC) measured in the context of the model predicting the downstream task label (Mort, LoS3+, LoS7+).
- **ECE** - Expected Calibration Error. To define ECE, we first divide the probability range [0,1] into K equal non-overlapping intervals, each interval denoted $I_k, k \in [K]$. We also define K corresponding bins $B_k, k \in [K]$ where each bin is the collection of example indices whose predicted probability falls in the interval I_k, i.e. $B_k = \{i : \hat{y}_i \in I_k\}$. With this, the ECE is defined as ECE $= \frac{1}{n} \sum_{k=1}^{K} \left| \sum_{i \in B_k} (y_i - \hat{y}_i) \right|$.
- **OOD-AUC** - This metric measures the ability of a model to assign higher confidence to IND test examples and lower confidence to OOD test examples. This metric requires both IND and OOD test sets for its calculation, whereas the previous two metrics are measured with only one test set (either IND or OOD) at a time. Confidence is typically considered to be the variance or entropy of the predicted Bernoulli distribution in case of a binary classification task. The AUC is measured with the label being set to 1 if the example is OOD (and 0 if IND), and the confidence measure is the score assigned to the example. The OOD-AUC will be high when OOD examples have higher variance or entropy than IND examples. Since both the variance and entropy of a Bernoulli distribution are similarly ordered (with $\hat{y}$=0.5 having the highest variance or entropy, and $\hat{y}$=0 or $\hat{y}$=1 having the lowest variance or entropy), the resulting OOD-AUC metrics with either choice will be the same.

These metrics are measured for each downstream task, on each data slice (IND) and other data slices within the same partition (OOD). The metrics are tabulated by algorithm, presenting the metrics of different algorithms in the same setting side by side.

B Experiments

We evaluate seven algorithms on BEDS-Bench and analyze their performance: Logistic Regression (LogReg), Gaussian Process (GP) [Rasmussen and Williams, 2005], Random Forest (RF)[Breiman, 2001], Mondrian Forest (MF)[Lakshminarayanan et al., 2014], Multi Layer Perceptron (MLP), Bayesian Recurrent Neural Network (BRNN) [Fortunato et al., 2019] with the same setup as [Dusenberry et al., 2020], and Spectral-normalized Neural Gaussian Process (SNGP) [Liu et al., 2020]. We use both the Scikit-Learn [Pedregosa et al., 2011] and Tensorflow [Abadi et al., 2016] software

In Hospital Mortality - AUC (larger values are better)								
Train	Test	LogReg	GP	RF	MF	MLP	BRNN	SNGP
Adult	Adult	0.706	0.703	0.736	0.650	0.691	**0.784**	0.710
	Neonate	0.935	**0.946**	0.650	0.122	0.836	0.394	0.878
	Paediatric	0.459	**0.685**	0.567	0.590	0.307	0.381	0.528
Neonate	Neonate	0.790	**0.970**	0.759	0.890	0.777	0.762	0.956
	Adult	0.562	**0.594**	0.540	0.529	0.440	0.383	0.459
	Paediatric	0.555	0.610	0.577	0.438	0.398	0.602	**0.615**
Paediatric	Paediatric	0.814	0.813	**0.834**	0.747	0.796	0.787	0.799
	Adult	0.488	0.480	**0.500**	0.492	0.495	0.426	0.492
	Neonate	0.893	0.905	0.944	0.702	**0.950**	0.539	0.692
Male	Male	0.765	0.775	0.814	0.716	0.773	**0.823**	0.764
	Female	0.781	0.788	**0.823**	0.718	0.771	0.809	0.751
Female	Female	0.779	0.788	**0.816**	0.727	0.765	0.797	0.766
	Male	0.762	0.763	0.804	0.726	0.754	**0.812**	0.775
Age 15-50yr	Age 15-50yr	0.767	0.783	**0.784**	0.741	0.767	0.775	0.717
	Age 50-60yr	0.757	0.761	0.790	0.747	0.748	**0.793**	0.732
	Age 60-70yr	0.711	0.702	0.712	0.677	0.691	**0.744**	0.668
	Age 70-80yr	0.661	0.640	**0.706**	0.667	0.631	0.696	0.649
	Age 80+yr	0.611	0.591	0.623	0.609	0.614	**0.667**	0.600

Table 2: Performance of various algorithms as measured with AUC against the In-Hospital Mortality task. The IID numbers are in dark yellow. When the OOD performance drops to random guessing or worse, the value is colored Red. If the OOD performance happens to be significantly better than the IND performance, those values are colored Green. Within each row, the best performing algorithm's value is in bold.

frameworks for the experiments, depending on the specific algorithm. Six of these models use a fixed length representation and one of them (BRNN) uses sequential embedding representation. The summary of models evaluated in this work is presented in Table 6 in the Appendix.

The fixed length representation is calculated as an array of binary indicators for each of the possible codes that might appear in the training data. Age is represented in years, and volumes (inputevents and outputevents) are aggregated over the 24-hour period. The inputs are standardized by column.

The sequential embedding representation creates an embedding for each categorical data and maintains the temporal ordering of all the codes. The data format of the generated representation matches that of [Kemp et al., 2020].

In all our experiments, within each partition, we randomly subsample (without replacement) the training examples to the size of the smallest training slice. This keeps all the training sets of equal size and makes is easy to compare metrics across different training slices.

The results of one of the tasks (In-Hospital Mortality) are presented in Tables 2 (for AUC), 3 (for ECE) and 4 (for OOD). The metrics are described in Section A.6. Each row corresponds to a specific combination of train and test set, and each column (starting from the third column) corresponds to a learning algorithm. In the AUC table (Table 2) and the ECE table (Table 3), the IND rows have values colored in gray. The red colored values are those where an improvement in performance is desired, and the green colored values are those where the performance is, somewhat unexpectedly, better than expected. Within each row, the best performing algorithm's value is set in bold. We do not color code the OOD table (Table 4) since interpreting those numbers is a little more nuanced and should only be viewed in conjunction with the corresponding values in the other two tables, due to reasons described in the following section.

The remaining results for the other tasks are reported in the Appendix.

C Discussion

In the results from our experiments, we broadly observe that among the algorithms we tested no algorithm dominates another in its performance across the board. We also observe that every model

In Hospital Mortality - ECE (smaller values are better)								
Train	Test	LogReg	GP	RF	MF	MLP	BRNN	SNGP
Adult	Adult	0.201	0.221	0.195	0.203	**0.175**	0.196	0.216
	Neonate	0.123	0.129	0.203	0.236	**0.101**	0.204	0.140
	Paediatric	**0.114**	0.243	0.192	0.152	0.117	0.177	0.152
Neonate	Neonate	0.010	0.013	**0.008**	**0.008**	0.030	0.014	0.327
	Adult	0.370	0.438	0.239	0.168	**0.127**	0.129	0.387
	Paediatric	**0.061**	0.292	0.112	0.070	**0.061**	0.064	0.362
Paediatric	Paediatric	0.102	0.113	**0.098**	0.102	0.101	0.107	0.142
	Adult	0.175	0.232	0.235	0.170	**0.129**	0.135	0.273
	Neonate	0.051	0.037	0.041	0.043	**0.008**	0.089	0.091
Male	Male	0.177	0.188	0.162	0.170	0.162	**0.155**	0.166
	Female	0.178	0.186	0.160	0.169	0.162	**0.157**	0.167
Female	Female	0.179	0.184	**0.162**	0.168	0.164	0.169	0.179
	Male	0.179	0.186	**0.164**	0.170	**0.164**	0.167	0.179
Age 15-50yr	Age 15-50yr	0.122	0.148	0.119	0.119	0.139	0.129	**0.115**
	Age 50-60yr	0.143	0.174	0.185	0.140	0.162	**0.137**	0.141
	Age 60-70yr	0.164	0.201	0.201	**0.156**	0.186	**0.156**	0.164
	Age 70-80yr	0.190	0.227	0.220	0.180	0.212	**0.176**	0.191
	Age 80+yr	0.255	0.315	0.264	0.228	0.265	**0.225**	0.290

Table 3: Performance of various algorithms as measured with ECE against the In-Hospital Mortality task. The IID numbers are in Gray. When the OOD performance is significantly worse (or better), the value is colored Red (or Green). Within each row, the best performing algorithm's value is in bold.

In Hospital Mortality - OOD								
Train	Test	LogReg	GP	RF	MF	MLP	BRNN	SNGP
Adult	Neonate	0.644	0.507	0.752	**0.793**	0.678	0.694	0.620
	Paediatric	0.229	**0.784**	0.684	0.544	0.456	0.536	0.309
Neonate	Adult	0.576	0.999	0.914	0.976	0.000	0.020	**1.000**
	Paediatric	0.162	0.966	0.853	0.950	0.047	0.327	**0.989**
Paediatric	Adult	0.242	0.746	**0.799**	0.701	0.061	0.160	0.795
	Neonate	0.613	0.417	0.500	0.604	0.120	**0.756**	0.536
Male	Female	0.496	**0.498**	0.492	0.493	0.494	0.494	0.496
Female	Male	0.500	0.498	0.505	0.505	0.500	0.506	**0.510**
Age 15-50yr	Age 50-60yr	0.545	0.566	**0.782**	0.511	0.535	0.490	0.558
	Age 60-70yr	0.563	0.619	**0.774**	0.493	0.568	0.488	0.571
	Age 70-80yr	0.603	0.656	**0.780**	0.518	0.587	0.496	0.618
	Age 80+yr	0.713	0.765	**0.778**	0.537	0.656	0.488	0.726

Table 4: OOD detection performance of various algorithms on the In-Hospital Mortality task. There are no rows corresponding to IND performance since OOD detection is not well defined in this case. Within each row, the best performing algorithm's value is in bold.

does particularly poorly in staying calibrated in OOD settings, with the exception of testing on Neonates.

Among the various algorithms tested in our experiments, a few of them are specifically designed with robustness to OOD inputs in mind. The SNGP, MF, and GP algorithms in particular are known to have stronger uncertainty estimation properties, even under distributional shift, relative to other algorithms. The SNGP algorithm in our experiment is essentially an MLP with additional GP layer as the final layer and spectral normalization in the fully connected layers. The SNGP algorithm is designed to be distance preserving at each of the individual layers in the deep model, with the hypothesis that it helps prevent collapsing of OOD and IND inputs at the final layer. What we observe is that, while in some cases these algorithms do perform better in OOD settings sometimes with respect to AUC score, the overall performance, especially with the ECE score has scope for improvement. Our hypothesis, and hope, is that as these algorithms get increasingly tested against EHR data modality, as they have been so far against images, improvements to the methods would result in increased robustness and better performance in such OOD settings.

It is also interesting to note that the difference between Male and Female distributions seem to not matter much for the downstream tasks, and the models are able to generalize over just fine with respect to both AUC and ECE. Indeed, this is also reflected in the fact that the OOD performance between these two subgroups is very close to 0.5. This example highlights the nuances in interpreting the OOD table, especially in isolation. While other works [Ulmer et al., 2020] highlight the fact that the models failing to distinguish Male vs Female distributions with appropriate levels predictive uncertainty and consider it as a failure mode, a more careful analysis shows that this might not be a failure in itself. When models are able to perfectly generalize from Male to Female and vice-versa with respect to both AUC and ECE metrics, the expectation for them to assign lower confidence (higher uncertainty) to the OOD examples is really undue, and even incorrect. Having a high OOD-AUC (i.e. assigning lower confidence to OOD examples) becomes a desiderata only when the model fails to generalize well to those examples. Our view is that while OOD detection performance can be interesting in some situations, the AUC and ECE metrics carry most of the story in terms of the model's OOD behavior.

A related consideration is about how to measure *confidence* from a model's prediction that is in turn used to detect OOD examples. Among the class of models which involve a set of predictions for each input, such as ensembles [Lakshminarayanan et al., 2017] or MC dropout [Gal and Ghahramani, 2016], confidence is typically measured with the standard deviation (or mutual information [Ulmer et al., 2020]) using the set of predictions. A large standard deviation among the predictions represent higher uncertainty and vice versa.

For the second class of models, including many we have tested in this work, which output just one prediction per input, confidence is typically measured as the entropy of the predicted Bernoulli distribution (or Categorical for multi-class classification). A Bernoulli distribution with mean parameter $p = 0.5$ has the highest entropy in its family, and therefore represents a prediction with least confidence, while those predictions with mean parameter close to 0 or 1 have lower entropy and hence represent predictions with high confidence.

Among this second class of models, an alternate way of describing confidence would be to consider the model's discrimination ability, and inspect whether a given prediction is close to the threshold of maximum discrimination, or away from it (closer being "less confident"). Here the threshold of maximum discrimination refers to the threshold value that maximizes the mean (either arithmetic or geometric) of the sensitivity and specificity of the model.

The question now is, is a low confidence prediction a Bernoulli distribution with high entropy, or is it a Bernoulli distribution whose mean is close to the threshold of maximum discrimination? This distinction is typically a moot point when there is a perfect class balance between the positive and negative classes (i.e. the marginal probability $p(y) = 0.5$). However when the classes are not well balanced, the two interpretations of confidence start to diverge, as illustrated in Figure 3. The left plot is from a situation where the OOD examples are closer to the threshold of maximum discrimination. The right plot is from a different situation where the OOD examples have higher entropy overall (note the scale of the X-axes in both the plots differ), and thus achieves high OOD-AUC score. Yet, one might also reasonably interpret the right plot as the OOD examples being assigned a "strongly positive" score relative to the two IND classes, and hence the model is in fact being "more confident" on the OOD examples and therefore ought to have a low OOD-AUC score under an appropriately

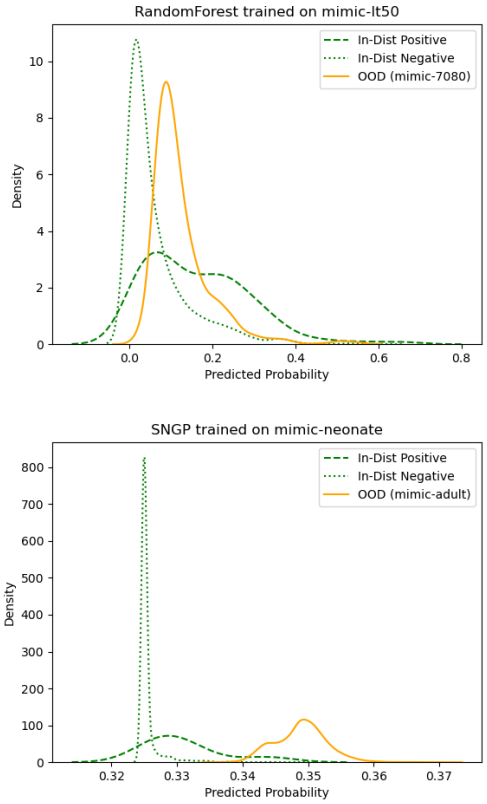

Figure 3: Two scenarios (models) plotting the In-Distribution postives, In-Distribution negatives, and OOD histograms (with smoothing) under class imbalance. The negative class is the majority and positives are few. In the left plot, the OOD class is "in-between" the positive and negative class. In the right plot, the OOD class is "closer to 0.5" than both the positives and negatives. Which of these two models is assigning "lower confidence" to the OOD examples?

chosen measure of confidence. These observations, in our opinion, makes the problem of OOD-detection itself a little less well defined under class imbalance situations for the class of models that use predictive entropy as a measure of confidence, in addition to it not being a very useful metric in isolation.

	MIMIC	PICDB
Center	Beth Israel Deaconess Medical Center	Children's Hospital of Zhejiang University School of Medicine
Duration	2001 - 2012	2010 - 2018
Patient Count	46,520	12,881
Age range	0-1mo, 16-89yrs (obfuscated at 90)	0-18yrs
Encounters	58,977	13,450
Diagnostics	ICD-9 (6986 unique)	ICD-10CN (1122 unique)
Medication	NDC (4211 unique)	NCCD (657 unique)
Chart Events	330,712,484 Events (6646 types)	2,278,979 Events (19 types)
Lab Events	27,854,056 Events (727 unique, 137 LOINC)	10,094,118 Events (822 unique, 118 LOINC)
Microbiology	631,727 Events (94 unique)	183,870 Events (43 unique)
Procedures	240,096 ICD-9 (3833 unique) 573,147 CPT (2019 unique)	N/A
Notes	2,083,180 Notes (15 types)	N/A

Table 5: Summary and comparison of the MIMIC-III and PICDB datasets.

Algorithm	Representation	Category
Logistic Regression (LogReg)	Fixed-Length	Flat
Gaussian Process Classifier (GP)	Fixed-Length	Flat
Random Forest (RF)	Fixed-Length	Tree-based
Mondrian Forest (MF)	Fixed-Length	Tree-based
Multi Layer Perceptron (MLP)	Fixed-Length	Deep Learning
Bayesian RNN (BRNN)	Sequential, Embeddings	Deep Learning
SNGP + MLP (SNGP)	Fixed-Length	Deep Learning

Table 6: Algorithms evaluated with BEDS-Bench.

	MIMIC	PICDB	Harmonized To	Mapping Sources
Diagnostics	ICD-9	ICD-10CN	ICD-9	UMLS
Prescriptions	National Drug Code (NDC)	Text description	RXCUI	UMLS MedEx
Lab Tests	LOINC+custom	LOINC+custom	LOINC only	N/A
Vitals / Charts	Custom	Custom	MIMIC $\rightarrow$ PICDB	MIMIC-Extract
Input/Output Events	Custom	N/A	Volumes only (ignore type)	N/A

Table 7: Details of harmonizing MIMIC-III and PICDB by category

	N	Mort	LoS3+	LoS7+
MIMIC-adult	36,909	13.3%	78.1%	42.9%
PICDB-paed	12,293	6.0%	91.2%	72.4%
MIMIC-neonate	7,651	0.5%	55.2%	29.8%
MIMIC-male	25,004	10.5%	74.0%	40.6%
MIMIC-female	19,556	10.7%	74.2%	40.5%
MIMIC-lt50	7,795	7.2%	70.4%	38.5%
MIMIC-5060	6,405	9.7%	77.6%	41.6%
MIMIC-6070	7,587	11.9%	81.9%	45.1%
MIMIC-7080	7,673	14.6%	82.3%	47.1%
MIMIC-gt80	7,449	19.9%	78.7%	42.4%

Table 8: Sizes of data slices and their class balances for the three tasks

Table	Rows	Columns
PATIENTS	59,401	5
ADMISSIONS	72,425	6
CHARTEVENTS	45,577,975	7
INPUTEVENTS	16,068,433	5
OUTPUTEVENTS	4,300,561	5
LABEVENTS	34,306,923	7
PRESCRIPTIONS	5,320,452	7
DIAGNOSES_ICD	664,116	4

Table 9: Summary of the resulting dataset after harmonizing and pre-processing the combined MIMIC-III and PICDB datasets.

Length of Stay 3+ days - AUC								
Train	Test	LogReg	GP	RF	MF	MLP	BRNN	SNGP
Adult	Adult	0.681	0.685	0.743	0.626	0.684	**0.756**	0.679
	Neonate	0.511	0.305	0.233	0.332	0.214	0.196	**0.583**
	Paediatric	0.366	**0.649**	0.514	0.478	0.364	0.398	0.582
Neonate	Neonate	0.888	0.883	0.876	0.837	0.884	**0.910**	0.887
	Adult	0.551	0.576	0.365	**0.699**	0.526	0.476	0.435
	Paediatric	0.375	**0.588**	0.393	0.449	0.306	0.440	0.569
Paediatric	Paediatric	0.766	0.790	**0.801**	0.740	0.789	0.767	0.767
	Adult	0.401	0.410	0.377	**0.584**	0.444	0.495	0.419
	Neonate	0.680	0.522	0.463	0.372	0.622	**0.847**	0.711
Male	Male	0.718	0.760	0.795	0.709	0.747	**0.800**	0.739
	Female	0.729	0.768	**0.803**	0.692	0.756	0.791	0.750
Female	Female	0.733	0.772	**0.803**	0.713	0.741	0.780	0.750
	Male	0.725	0.772	**0.798**	0.719	0.748	0.792	0.754
Age 15-50yr	Age 15-50yr	0.722	0.731	**0.757**	0.669	0.702	0.753	0.726
	Age 50-60yr	0.699	0.701	**0.772**	0.681	0.718	0.758	0.710
	Age 60-70yr	0.670	0.698	**0.752**	0.666	0.670	0.728	0.681
	Age 70-80yr	0.640	0.654	**0.716**	0.630	0.641	0.680	0.656
	Age 80+yr	0.558	0.575	**0.713**	0.624	0.623	0.672	0.589

Length of Stay 3+ days - ECE								
Train	Test	LogReg	GP	RF	MF	MLP	BRNN	SNGP
Adult	Adult	0.330	0.331	0.290	0.318	0.323	**0.285**	0.312
	Neonate	0.468	0.476	0.521	0.477	0.484	0.584	**0.459**
	Paediatric	0.230	0.302	0.263	0.221	0.367	0.448	**0.202**
Neonate	Neonate	0.318	0.284	0.262	0.266	0.263	**0.250**	0.289
	Adult	0.578	0.456	0.386	0.495	**0.213**	0.784	0.322
	Paediatric	0.415	0.399	0.515	0.367	0.267	0.511	**0.220**
Paediatric	Paediatric	0.150	0.158	0.148	**0.144**	0.179	0.149	0.160
	Adult	0.272	0.332	0.331	0.272	0.244	**0.217**	0.275
	Neonate	0.447	0.448	0.452	0.464	0.446	**0.417**	0.447
Male	Male	0.332	0.336	0.290	0.318	0.343	**0.283**	0.319
	Female	0.329	0.335	0.288	0.319	0.340	**0.284**	0.315
Female	Female	0.330	0.336	**0.290**	0.318	0.330	0.300	0.324
	Male	0.332	0.337	**0.292**	0.316	0.330	0.297	0.323
Age 15-50yr	Age 15-50yr	0.374	0.382	0.348	0.391	0.379	**0.346**	0.358
	Age 50-60yr	0.335	0.351	0.330	0.354	0.342	**0.293**	0.311
	Age 60-70yr	0.310	0.323	0.312	0.341	0.328	**0.260**	0.286
	Age 70-80yr	0.302	0.313	0.316	0.340	0.327	**0.244**	0.273
	Age 80+yr	0.300	0.319	0.348	0.368	0.347	**0.264**	0.298

Length of Stay 3+ days - OOD								
Train	Test	LogReg	GP	RF	MF	MLP	BRNN	SNGP
Adult	Neonate	0.560	0.313	0.542	0.246	0.310	**0.827**	0.511
	Paediatric	0.185	0.600	0.542	0.345	0.719	**0.833**	0.221
Neonate	Adult	0.268	**0.926**	0.728	0.901	0.002	0.015	0.356
	Paediatric	0.530	0.839	**0.883**	0.837	0.401	0.800	0.275
Paediatric	Adult	0.366	**0.828**	0.788	0.704	0.190	0.066	0.371
	Neonate	0.444	0.272	0.441	**0.539**	0.177	0.298	0.441
Male	Female	**0.501**	0.500	0.497	0.491	0.499	**0.501**	0.495
Female	Male	0.495	0.496	**0.502**	0.499	0.499	**0.502**	0.500
Age 15-50yr	Age 50-60yr	0.422	0.434	**0.514**	0.475	0.469	0.395	0.458
	Age 60-70yr	0.361	0.366	**0.477**	0.451	0.435	0.313	0.419
	Age 70-80yr	0.345	0.345	**0.490**	0.458	0.444	0.261	0.430
	Age 80+yr	0.255	0.278	**0.546**	0.504	0.437	0.237	0.392

Length of Stay 7+ days - AUC								
Train	Test	LogReg	GP	RF	MF	MLP	BRNN	SNGP
Adult	Adult	0.653	0.658	**0.723**	0.620	0.647	0.715	0.639
	Neonate	0.554	0.487	**0.664**	0.655	0.326	0.205	0.381
	Paediatric	0.539	0.461	0.444	**0.583**	0.535	0.560	0.457
Neonate	Neonate	0.913	0.915	0.898	0.873	0.916	**0.920**	0.912
	Adult	0.485	0.570	0.502	0.567	**0.586**	0.482	0.538
	Paediatric	0.507	0.493	0.401	0.445	0.451	**0.529**	0.503
Paediatric	Paediatric	0.721	**0.745**	0.731	0.695	0.713	0.719	0.717
	Adult	0.445	0.413	0.504	**0.563**	0.434	0.502	0.442
	Neonate	0.815	0.644	0.485	0.292	0.468	0.766	**0.825**
Male	Male	0.698	0.732	**0.763**	0.674	0.714	0.749	0.699
	Female	0.698	0.733	**0.763**	0.679	0.720	0.757	0.698
Female	Female	0.697	0.731	**0.762**	0.691	0.717	0.742	0.707
	Male	0.698	0.733	**0.765**	0.690	0.719	0.738	0.703
Age 15-50yr	Age 15-50yr	0.701	0.700	**0.753**	0.665	0.708	0.731	0.709
	Age 50-60yr	0.678	0.683	**0.712**	0.637	0.662	0.704	0.674
	Age 60-70yr	0.650	0.672	**0.743**	0.634	0.665	0.690	0.678
	Age 70-80yr	0.621	0.651	**0.694**	0.628	0.614	0.677	0.644
	Age 80+yr	0.550	0.542	**0.688**	0.583	0.572	0.651	0.556

Length of Stay 7+ days - ECE								
Train	Test	LogReg	GP	RF	MF	MLP	BRNN	SNGP
Adult	Adult	0.474	0.458	**0.420**	0.459	0.455	**0.420**	0.479
	Neonate	0.473	0.502	0.468	**0.455**	0.533	0.517	0.506
	Paediatric	0.545	**0.511**	0.514	0.530	0.534	0.559	0.515
Neonate	Neonate	0.250	0.215	0.210	0.212	**0.197**	0.201	0.212
	Adult	0.447	0.492	0.497	0.465	**0.427**	0.442	0.511
	Paediatric	0.609	0.517	0.569	0.553	0.463	0.640	**0.415**
Paediatric	Paediatric	0.368	0.356	**0.348**	**0.348**	0.354	0.361	0.357
	Adult	0.541	**0.528**	0.532	0.529	0.569	0.555	0.541
	Neonate	**0.604**	0.651	0.624	0.656	0.636	0.609	0.621
Male	Male	0.451	0.424	**0.380**	0.419	0.413	0.388	0.430
	Female	0.449	0.421	**0.379**	0.415	0.409	0.384	0.429
Female	Female	0.449	0.421	**0.378**	0.409	0.407	0.385	0.428
	Male	0.451	0.423	**0.380**	0.413	0.410	0.390	0.428
Age 15-50yr	Age 15-50yr	0.451	0.427	**0.388**	0.431	0.422	0.411	0.429
	Age 50-60yr	0.462	0.440	**0.422**	0.448	0.439	0.428	0.447
	Age 60-70yr	0.466	0.449	**0.412**	0.449	0.439	0.435	0.448
	Age 70-80yr	0.479	0.461	**0.436**	0.461	0.463	0.446	0.461
	Age 80+yr	0.478	0.491	**0.431**	0.462	0.466	0.459	0.476

Length of Stay 7+ days - OOD								
Train	Test	LogReg	GP	RF	MF	MLP	BRNN	SNGP
Adult	Neonate	0.655	0.814	0.766	0.575	0.637	0.439	**0.850**
	Paediatric	0.350	**0.860**	0.785	0.615	0.562	0.542	0.714
Neonate	Adult	0.054	**0.957**	0.950	0.865	0.157	0.151	0.821
	Paediatric	0.607	**0.966**	0.876	0.906	0.280	0.608	0.630
Paediatric	Adult	0.476	**0.743**	0.476	0.524	0.372	0.053	0.547
	Neonate	0.243	0.080	**0.391**	0.309	0.275	0.214	0.303
Male	Female	0.492	0.496	0.496	0.494	0.500	**0.502**	0.497
Female	Male	0.499	0.507	0.504	**0.512**	0.505	0.504	0.496
Age 15-50yr	Age 50-60yr	0.501	0.546	**0.574**	0.489	0.499	0.531	0.531
	Age 60-70yr	0.507	**0.571**	0.567	0.496	0.493	0.533	0.549
	Age 70-80yr	0.516	**0.607**	0.591	0.503	0.520	0.581	0.562
	Age 80+yr	0.603	0.565	0.583	0.511	0.547	0.600	**0.616**