
Proceedings Paper:
Ryan Conmy, Philippa Mary, Ozturk, Berk, Habli, Ibrahim et al. (2023) The Impact of Training Data Shortfalls on Safety of AI-based Clinical Decision Support Systems. In: SAFECOMP 2023 (42nd International Conference on Computer Safety, Reliability and Security). International Conference on Computer Safety, Reliability and Security, 20-22 Sep 2023 Lecture Notes in Computer Science. Springer, FRA, pp. 213-226.

https://doi.org/10.1007/978-3-031-40923-3_16

The Impact of Training Data Shortfalls on Safety of AI-based Clinical Decision Support Systems

Philippa Ryan Conmy[1][0000−0003−1307−5207], Berk Ozturk[1], Tom Lawton[2], and Ibrahim Habli[1]

[1] Department of Computer Science, University of York, UK
{philippa.ryan, berk.ozturk, ibrahim.habli}@york.ac.uk
[2] Bradford Royal Infirmary, Bradford Institute for Health Research, BD9 6RJ Bradford, U.K

Abstract. Decision support systems with Artificial intelligence (AI) and specifically Machine Learning (ML) components present many challenges when assuring trust in operational performance, particularly in a safety-critical domain such as healthcare. During operation the Human in/on The Loop (HTL) may need assistance in determining when to trust the ML output and when to override it, particularly to prevent hazardous situations. In this paper, we consider how issues with training data shortfalls can cause varying safety performance in ML. We present a case study using an ML-based clinical decision support system for Type-2 diabetes related co-morbidity prediction (DCP). The DCP ML component is trained using real patient data, but the data was taken from a very large live database gathered over many years, and the records vary in distribution and completeness. Research developing similar clinical predictor systems describe different methods to compensate for training data shortfalls, but concentrate only on fixing the data to maximise the ML performance without considering a system safety perspective. This means the impact of the ML's varying performance is not fully understood at the system level. Further, methods such as data imputation can introduce a further risk of bias which is not addressed. This paper combines the use of ML data shortfall compensation measures with exploratory safety analysis to ensure all means of reducing risk are considered. We demonstrate that together these provide a richer picture allowing more effective identification and mitigation of risks from training data shortfalls.

Keywords: Machine Learning · Training Data · Medical device safety.

1 Introduction

Safety-related decision support systems incorporating Artificial intelligence (AI) and specifically Machine Learning (ML) components are increasingly being developed and deployed [18]. These can have many potential benefits, such as providing faster and richer computational support to complex tasks. However, developing a robust and fit-for-purpose ML algorithm is reliant on good training

data, which reflects the required task. Even with a robust training regime, poor data will influence the performance, and safety of the output from the ML. Given that comprehensive verification of ML across all operating scenarios is typically impossible, these errors may be undetected until it is too late.

During operation the Human In/On The Loop (HTL) working with the system may need assistance in determining when to trust the ML output and when to override it, particularly in cases where there is a safety related outcome. For example, clinical advisory systems typically have a workflow allowing the clinician to override the output, but it may not be clear what the limitations or strengths are of the ML components, making it difficult to trust or ignore certain predictions [22]. This is particularly problematic where there is a difference of opinion between the ML predictor and clinician. Whilst there is research into the impact of different methods to manage training data shortfalls, these concentrate on maximising the ML performance with respect to certain metrics, and do not considering the different risks of varying performance at the system level. Thus the safety impact of data shortfalls is not well understood, nor are all means of reducing risk explored. We argue that taking a systems perspective is necessary for safety critical environments.

In this paper we examine how issues with training datasets, and means to compensate for them, can impact on safety performance. We combine the use of training data shortfall compensation methods and exploratory safety analysis to ensure all means of reducing risk are considered. We apply this combination to a diabetes comorbidity predictor (DCP), implemented using ML, used to support clinical decision making. The DCP is trained using a dataset which contains the real clinical records of the patients taken from the Connected Bradford database [21]. The dataset consists of over 42,000 rows of data for Type-2 diabetic patients from different backgrounds and over 14,000 different types of clinical records (features). Since the dataset records are obtained from different care centres, this causes differences in recorded data. When patients do not attend their visits regularly there can be changes/deficiencies in the recorded laboratory results, which causes the dataset to have a great number of missing values. This makes it critical to conduct systematic safety analysis to prevent and mitigate for misleading outcomes in the manner of patient safety.

This paper is laid out as follows. In section 2 we describe this real world problem in more detail and describe some related work used to develop our approach. In section 3 we describe the case study, training data issues and safety analysis. We discuss our results and findings in section 4, and finally in section 5 we present our conclusions.

2 ML training data and safety

Increasingly, safety-critical systems with machine learning components are being developed and deployed [8]. Examples include autonomous cars (with or without a safety driver), drones, medical diagnosis systems and agricultural robots. There are many different approaches to machine learning, including supervised,

semi-supervised and unsupervised training methods and models such as neural networks or decision trees. However, a core requirement for each is a set of valid training data which is pre-processed to tailor it for the task and model. For a safety-critical system, poor quality training data can lead to latent faults which can lead to hazardous behaviour. This is illustrated in Figure 1. The top chain of elements indicates the ML training and system integration process. The lower row indicates how the error can propagate throughout the training lifecycle. A training data shortfall can mean the ML doesn't have complete or correct performance with respect to the system requirements. This may not be picked up in verification, as performing complete verification of ML is impossible in all but the most trivial cases. The same issue will affect testing during system integration but it may also be difficult to control the test space (e.g., real world testing of a drone cannot be done in controlled weather conditions), which means a latent failure could lead to a hazard during operation.

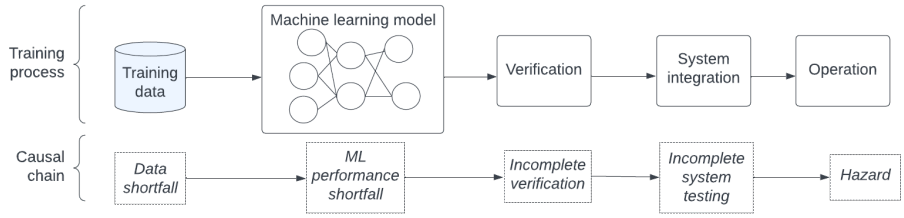

Fig. 1. Causal chain of failure events from training data shortfalls

Consider the following examples. A classifier for an autonomous vehicle object detection system is trained using supervised learning. This uses a labelled set of training data, including images marked with labelled boxes. This data set includes a number of examples of dogs, but even though the training examples are properly labeled and framed, they only show dogs from a side view. The object detector may then fail to detect a dog facing forwards, contributing to a collision. A similar issue contributed to the fatal autonomous vehicle crash in Tempe, Arizona, 2018 [16], where a bicycle was not consistently recognised from the side, and the autonomous driving controller was unable to predict the path of the pedestrian pushing it quickly enough. Another example, would be a decision component determining whether an unmanned drone should return to base if the conditions are unfavourable may be trained using semi-supervised learning. For this, each training sample is marked as safe or unsafe, and contains a series of atmospheric readings on waypoints for the planned path. The ML training process is designed to allow it to look for patterns in the readings/predictions which can be matched to either safe or unsafe. However, the training set has very few samples where the temperature dipped below zero celcius, where icing could be a problem, and each of these samples contained very different sets of other

readings making it hard to generalise. Therefore the ML might incorrectly decide it was safe to continue when in fact there was a severe risk the physical systems of the drone would fail and it would crash e.g., due to ice and low temperature impacting battery power.

When looking for shortfalls we need to consider the source of the data as this may impact on the types encountered. In some situations, the training data can be entirely user generated, such as when simulation software is used. The advantage of this approach is that the data scientist or engineer will have a very high degree of control over the data generated, but it may not be realistic without careful modelling and analysis. However, the opposite approach may be taken, where an off-the-shelf dataset is acquired and curated for the ML training. Real-world sampling will help assure the validity of the data, but the disadvantage is that there are likely to be missing cases or bias to certain situations, or even deliberate data poisoning. Our analysis considers both the normal case for data, where the sample may be valid but overall distribution introduces bias, and the failure case, where the sample may be corrupted in some way.

2.1 Methods for managing data shortfalls

As noted previously there are different types of data shortfalls which may vary depending on the way the training data has been gathered and curated. For example, there are issues of missing data, poorly labelled data, data validity and data distribution [8]. These may be intrinsically linked, for example, if we compensate for missing data using data imputation methods we must ensure the generated data is valid. In this section we examine related literature on data imputation, concentrating on papers where it has been applied to similar prediction problems such as diabetes [7, 11] and Covid-19 [4].

There are different types of data imputation methods to deal with the missing values, and these methods have been used for different domains. In [23][7] the authors take means of the full set of a particular feature to fill the missing values. However, only taking the average of the entire column and replacing the missing values with the average of the column may lead some bias or misleading outcomes. An issue with both these papers is that they focus purely on the ML performance indicators, and do not consider risk mitigation from a system safety perspective. Maximising the performance may not be required if other risk mitigation measures, such as explainability and transparency [13], are used.

In [5], the authors use kNN Imputation method to deal with the missing values in their dataset. In [1], four different imputation methods (case deletion, mean imputation, median imputation, and kNN Imputation) have been applied to compare these methods. Then, the authors concluded that the kNN Imputation performs better providing a better Mean Squared Error (MSE) value to deal with the missing values. In [24], multiple imputation methods have been compared, and it has been concluded that kNN Imputation has better than mean and median imputation methods. kNN imputation looks for similar cases and nearest neighbours, thus reduces bias from extreme outlying values or overall

distribution. Again, the authors concentrate on ML fitness in isolation of the whole system.

An alternative method is described in [3][4][14] where the authors use the Bag Imputation method to fill the missing values in the dataset. This is a more sophisticated, and computer intensive, nearest neighbour method which uses additional ML to predict missing values, and to avoid overfitting and bias in the dataset [11]. Because we have a large amount of missing values and aim to prevent bias, we have decided to investigate bag imputation as a way to compensate for missing values in our dataset.

2.2 Safety data analysis method

We argue that a system safety perspective is necessary to ensure that the risks associated with data shortfalls are methodically understood. By this we mean considering the impact on the effectiveness of the ML, and then considering how or if this might affect performance at the system level in combination with other information and actors. We also consider other activities during the training process which might reduce the risk. Further, we need to identify and assess the additional risks that using data imputation may introduce.

In Figure 1 we show the ML training and operation lifecycle. There are opportunities to reduce risk at every stage, including controls on how training data is selected, adding specific verification/integration tests for known issues, and how information is presented to the operator, e.g., using explainability, so that they are given a richer picture for individual decisions.

To support the system safety perspective we need an exploratory safety analysis technique which could be effective in identifying types of data shortfall, such as for particular clinical features of importance and how they could propagate. Therefore, we considered bottom-up/inductive analysis safety analysis methods rather than top-down/deductive techniques e.g., Fault Tree Analysis. We argue that by concentrating on the data issues as a starting point we can understand their causal impact more holistically.

Typical inductive safety analysis methods include Failure Modes and Effects Analysis (FMEA)[12], and HAZard and OPerability Studies (Hazops)[15]. Hazops uses particular guidewords, e.g., more, less, early, to provide general categories of failures to engineers performing the analysis. It was originally used in the chemical processing industry but has been successfully used for computer based analysis, both on data flows (such as training data to ML training) and on control flows. On the other hand FMEA is more traditionally applied to physical system safety so we did not consider it further. In [19, 15] the authors have successfully used Hazops to identify safety issues in systems with ML. An alternative version to Hazops is Software Hazard Analysis and Resolution in Design (SHARD) [12] is demonstrated in [9] for a medical decision support system. We note the findings in [15] that SHARD is better suited for scalar data, and given that we are interested in data quantities, using a Hazops type approach may be more meaningful. Therefore, we used Hazops guidewords for our approach.

To summarise, training data shortfalls can lead to latent faults in an ML system which can in turn lead to hazardous behaviour. Whilst there are many methods to manage training data shortfalls, they can themselves introduce further issues such as bias. It may be impossible to train the ML effectively without their use when we are dealing with real-world data. The approach for our case study uses a combination of ML data shortfall compensation methods and exploratory Hazop style system safety analysis to identify and consider means to reduce these risks.

3 Case study: ML-based clinical decision support system for Type II diabetes-related co-morbidity prediction

In this section we describe a clinical case study which uses our approach of combining ML data shortfall compensation methods and safety analysis. For this we used training data which contains real clinical patient data from the Connecting Bradford database [21]. The dataset consists of over 42,000 rows for patients with type 2 diabetes mellitus from different backgrounds and over 14,000 different types of clinical records (features). Type 2 diabetes is a lifelong health condition and is the most common type of diabetes in the world. This health condition may cause the level of sugar (glucose) in the blood to become very high, and if not managed properly, it may progress by causing serious comorbidities [2]. When Type-2 diabetes progress, this causes numerous different comorbidities affecting the heart, brain, kidney, and other diseases.

The most frequently recorded disease/condition in our dataset is hypertension. It is known that hypertension is the precursor of the other potential diseases, and having both Type-2 diabetes and hypertension are synergistically dangerous. Hence, this is very critical to make a proper prediction for the risk level of having hypertension. High or low-risk thresholds are calculated using the National Institute for Health Care Excellence (NICE) guidelines used by clinicians [20].

The decision support system is designed to provide a clinician with an independent prediction of whether a patient is at high or low risk of hypertension (e.g., in the next six months), and hence support their decision of whether intervention is required. The clinical workflow is summarised in Figure 2. The DCP will use the most recent patient data set provided as input. It will provide a prediction as to whether the patient is at high or low risk of hypertension, as well as explanation of the prediction. Additionally, the clinician will gather information through discussion with the patient. We discuss provision of contextual information later in the paper, as there are issues of patient confidentiality. In this paper we are specifically considering hypertension which increases the risk of other comorbidities, however the general safety analysis principles discussed can apply to any of the predictions training pathways. Our future work will consider other co-morbidity predictions.

The hazards related to the system are

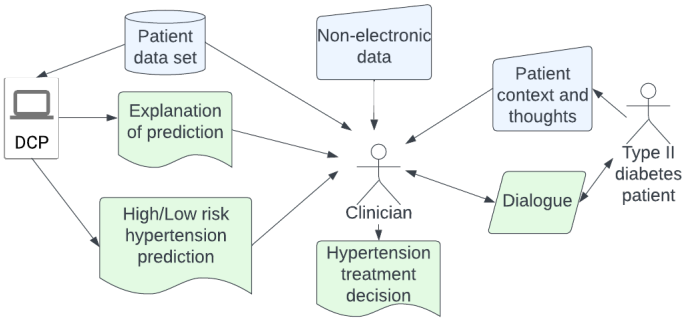

Fig. 2. Clinical workflow using DCP

- **false positive** where a patient is categorised as high risk and given intervention that they do not require, possibly including medication with harmful side effects. For hypertension treatment may range from recommended lifestyle changes to specific medication. Common medications can have a range of minor side-effects (dizziness, headache, cough) to much more severe effects (e.g., angio-oedema). The clinician using the DCP would be making the decision of which medication to administer, and there is no requirement on the DCP to recommend treatment.
- **false negative** where a patient is categorised as low risk and no treatment is provided, leading to the condition not being managed. For hypertension this would mean medication specifically not being provided, potentially putting the patient at risk of severe outcomes such as heart attacks or stroke.

Classical risk analysis expects a combination of severity and likelihood to determine tolerability. Risk severity will depend on the particular comorbidity and potential outcomes of the false positive and false negative clinical decisions. In the case of hypertension, there is a potentially catastrophic outcome of heart attack and death if it is not treated. Calculating the likelihood of an incorrect prediction will require an understanding of the ML's performance for that particular comorbidity, but we note that there may be certain groups or individual patients where the predictions are less or more reliable. This may be due to weaknesses in the training data used or issues with specific information about an individual patient. An additional consideration is that the clinical decision may be influenced by other predictions from the DCP. As it is infeasible to assess likelihood of incorrect prediction for each individual patient, or to calculate overall likelihood with accuracy, we instead consider means to reduce the risk as far as possible at each stage of development and use.

3.1 The DCP training data

As noted, the training dataset consists of over 42,000 rows with 14,000 variables (features). A row represents a visit of the Type-2 diabetic patient and each

feature represents the observations or test results gained during the visits. Some of the patients have been attending for many years so have many rows in the database, whereas newer patients only have a few rows. We need to consider whether too many samples from the same patient would introduce bias.

Data shortfalls will impact on training effectiveness, but not all of the features will impact safety any may be irrelevant or of low importance. Given the large number of features in the database (14,000) it is infeasible to perform safety analysis for each of them. Further, using ML across the 14,000 would have a very high calculation cost, and may not be meaningful. Hence we need to reduce this set to be be more meaningful.

Since the record types differ according to the different sites, or if patients were not able to attend their appointments regularly, we have a large amount of missing values in our dataset (typically over more than half for each feature). We need to compensate for this during training, using data imputation, in order to train the ML. It is of note that missing data may itself be significant (see section 4) however understanding the varying reasons for missing data, which could be clinically significant or simply due to different reporting practice across multiple clinics, would be difficult to infer without guidance, and lead to more uncertainty in the quality of outputs of the DCP.

An additional problem we cannot compensate for is that there may be groups of patients which are completely missing, e.g. from certain age groups or backgrounds. Also, we cannot compensate for validity issues, as whilst we can run some simple sanity checks e.g. for negative values for BMI, the issue of plausible but wrong data remains. Training data issues are illustrated in Figure 3.

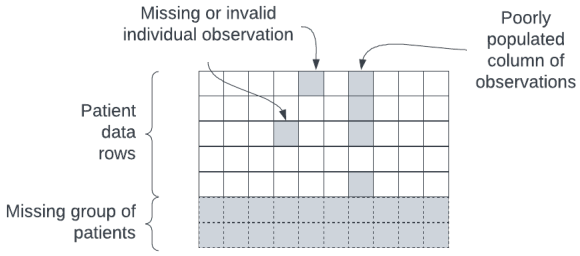

Fig. 3. Training data issues

The systematic data pre-processing techniques applied for our study are shown in Figure 4. First, a data frame has been prepared from the stored data. We have determined the most 20 frequent which are related to Type 2 diabetes FOIs and use these as a sub-set.

All the patients have been filtered by Type-2 diabetes, and duplicated or mistyped records have been deleted. When it has been ensured that we have unique records for each patient, the records are checked for the missing values. To

fill all the missing values, we have used bag imputation method (see section 2.1) incorporated in the R-Studio suite, as it reduces the risk of bias, and overfitting by predicting the missing values using ML. After dealing with all the missing values, we have normalized the dataset to fit all the values between 0 and 1 and to prevent from bias caused by the variation of the features. After finishing all the filtering and the necessary data pre-processing steps, the training dataset has been trained by the ML model.

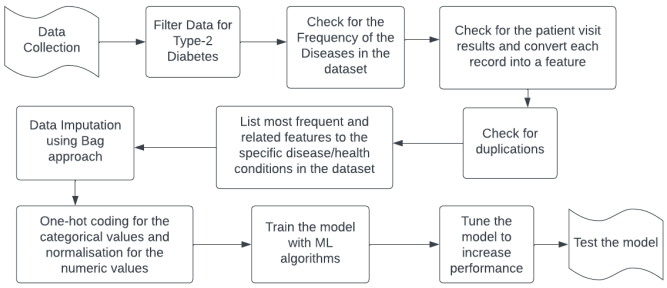

Fig. 4. Work-flow of the ML-based Type-2 Diabetes Progression Prediction

After training the ML model, the feature importance of each variable has been calculated. Figure 5 shows us each variable's weighted importance levels to predict the output. This provides us some level of explainability of the model and also helps us to have a better understanding of the reasons behind of the model's predictions. Further, it allows us to focus the exploratory safety analysis on the FOIs. In order to ensure the validity at this stage, these FOIs were reviewed to confirm that they are plausible. Body Mass Index (BMI) is considered a good predictor, and was our highest FOI, so we have concentrated on that for this paper. Some of the other FOIs may be caused by hypertension, rather than being predictive. Note that the FOIs were gathered using an ensemble of different ML methods (including neural networks and random forests [17]), and future work is looking at comparing these individually.

3.2 Hazop Analysis

In this section we present an extract of the exploratory safety analysis of the FOIs as identified in in the previous section and shown in Figure 5. We have used a hazops style analysis (see 2.2) to consider how shortfalls in the training dataset could lead to hazards on output if not mitigated for each FOI, or groups of FOIs. The "flow" was interpreted as the flow of data into the ML training process. We used standard Hazop guidewords as inspiration for possible issues. We note that additional guidewords may be needed to capture unusual data shortfalls, although we did not identify any in this analysis. Some examples of how we interpret the guidewords are as follows.

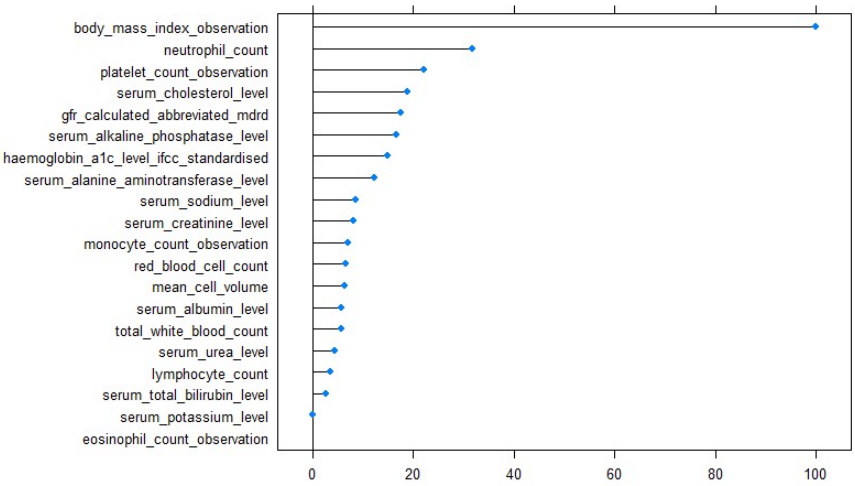

Fig. 5. Feature Importance Levels

- More - indicates a bias in the data, e.g., over representation of particular patient group in the dataset
- No or Not - FOI or set of FOIs are missing
- Less - fewer examples of FOI than are desirable for good performance are present
- Early/Before - indicates that a FOI may be present but out of date with respect to the co-morbidity presenting itself
- Late/After - indicates that a FOI is present, but the overall patient data sample might be late in progression of the co-morbidity and hence not a useful predictor
- Part of - indicates missing data which needs to be compensated for
- Reverse - opposite diagnosis provided (i.e., False positive/negative)
- Instead - indicates the wrong FOI being used
- As well as - no interpretation

In Table 1 we show an extract of our analysis considering BMI as FOI, as this is the most critical. We list the guideword, identified deviation in the training data set, possible causes, effect on system safety and means to mitigate the deviation. Note that this analysis focuses on training data only, it may be useful to do a similar analysis for problems on the data used for an actual prediction as part of the overall safety assurance case, e.g. when there is no FOI in the patient record. The analysis has uncovered a number of risk mitigation measures which could be used, where practical, to reduce the risk of a latent failure caused by shortfalls in the training data leading to an incorrect prediction. These include technical approaches to data imputation and data sampling, but also manual data review, and explanations provided to the clinician. We have also included the discussion that the clinician would have with the patient as a

mitigation (Figure 2) to emphasise that the ML decision is not used in isolation of other independent data sources. From the analysis we have a much richer understanding of risks and their mitigations.

Table 1. Extract of Hazop analysis of BMI FOI in Training Data

Guideword	Deviation	Cause	Effect	Mitigation
No or not	Samples for ethnic group not included in training data (TD)	No/limited patients of ethnic group were patients	ML not trained or verified adequately for ethnic group with higher genetic risk of hypertension	Manual review of DB by expert, show clinician prototypical examples, patient discussion
Part of	Partially missing BMI in TD samples	BMI not consistently recorded	ML performance biased based on the data imputation method used, leads to poor performance for high or low BMI patients	Use bag imputation for TD records to reduce bias, recommend collection of BMI for future TD, show clinician prototype examples, patient discussion
More	Over representation in TD of high BMI patients	Most patients examined had high BMI	Prediction biased towards patients with high BMI, meaning patients with low BMI have less accurate predictions	Manual review of DB by expert, training samples picked across all ranges, show clinician prototype examples, patient discussion
More	Over representation in TD of certain ethnic group	Over diagnosis by trained ML for patients of other ethnic groups	TD dominated by ethnic group with genetic disposition to hyper tension	Manual review of DB by expert, show clinician prototype examples, patient discussion
Early/ Before and More	BMI data is out of date and training patients have changed BMI by time of diagnosis	DB not kept up to date, TD sampled from wrong part of patient history	ML underestimates likelihood of hypertension	TD selected from samples near to hypertension diagnosis, manual review of DB by expert, patient discussion
Instead	BMI value no longer highest FOI for some FOI distribution	Performance outlier from ML	Wrong prediction for hypertension	Show clinician FOI from training and for each prediction at point of use, patient discussion

4 Discussion

In the previous section we presented a case study combining system safety analysis with ML data shortfall compensation measures. In this section we discuss the findings in more depth.

It was infeasible to review all the potential features in the training data as there were over 14,000 of these. This meant that performing safety analysis prior on the data prior to pre-processing was not possible. Instead, it was necessary to reduce the set to 20 FOIs. The initial ML training (using data imputation to manage missing values) was performed to prioritise features and focus the safety analysis. However, it may be the case that training the ML using a much larger set of features would uncover a link or pattern of causes of hypertension which had not been considered previously. This is an avenue for further research.

When undertaking the safety analysis (Table 1) we suggested a number of risk mitigation methods which require further thought. One method for reducing the risks is a manual review of the patient database, for example to look for missing ethnic groups of patients or ensure up to date records have been kept. In practice this may be difficult to do effectively given the size of the database and some automation would be needed.

Another operational mitigation is to show the clinician similar patients from the TD to the one which the predictor has been applied to (i.e., prototypical examples as described in [10]). This would allow the clinician to review similar cases, their progression, and provides context to a particular prediction. However, the raw training data cannot be presented to the clinician for reasons of patient confidentiality and would need to be anonymised or obfuscated in some way. An avenue for further research is to consider whether using methods such as k-anonymise [6] would reduce the effectiveness such that this isn't a useful mitigation/explainability method for DCP.

Finally, it was noted by our clinical expert that missing data can be an important indicator of an underlying problem, for example if the patient was ill with another condition they may not have attended the clinic. In our clinical workflow (Figure 2) we see this can potentially be considered via discussion with the patient. Another consideration is the progression of the comorbidity in the patient and whether this can improve predictive performance. Both cases may require a different and more complex ML model and training regime.

5 Conclusions

In this paper we have demonstrated that using a combination of ML data shortfall compensation measures, and exploratory safety analysis provides an effective method for the identification and mitigation of risks from training data shortfalls for a DCP. This takes a whole system perspective on risk identification and mitigation that is not found in similar literature in the area.

We have identified a number of avenues for further work including applying this methodology to an expanded predictor with multiple comorbidities (e.g.,

for brain diseases). Another is to review performance of different ML models with respect to bias from the different data imputation methods whilst balancing optimal performance against risk mitigations. Additional issues raised by the case study included balancing patient confidentialy with explainability, and wider contextual issues such as the clinical importance of missing data in the training data.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EP/W011239/1) and the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York.

References

1. Acuna, E., Rodriguez, C.: The treatment of missing values and its effect on classifier accuracy. In: Classification, Clustering, and Data Mining Applications: International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004. pp. 639–647. Springer (2004)
2. Alonso-Morán, E., Orueta, J., Esteban, J., Axpe, J., González, M.l., Toro, N., Loiola, P., Sonia, G., Nuño-Solinís, R.: The prevalence of diabetes-related complications and multimorbidity in the population with type 2 diabetes mellitus in the basque country. BMC public health **14**, 1059 (10 2014). https://doi.org/10.1186/1471-2458-14-1059
3. Bourdon, C., Lelijveld, N., Thompson, D., Dalvi, P.S., Gonzales, G.B., Wang, D., Alipour, M., Wine, E., Chimwezi, E., Wells, J.C., et al.: Metabolomics in plasma of Malawian children 7 years after surviving severe acute malnutrition:"ChroSAM" a cohort study. EBioMedicine **45**, 464–472 (2019)
4. Churpek, M.M., Gupta, S., Spicer, A.B., Parker, W.F., Fahrenbach, J., Brenner, S.K., Leaf, D.E.: Hospital-level variation in death for critically ill patients with COVID-19. American journal of respiratory and critical care medicine **204**(4), 403–411 (2021)
5. Driss, K., Boulila, W., Batool, A., Ahmad, J.: A Novel Approach for Classifying Diabetes' Patients Based on Imputation and Machine Learning. In: 2020 International Conference on UK-China Emerging Technologies (UCET). pp. 1–4 (2020). https://doi.org/10.1109/UCET51115.2020.9205378
6. Emam, K.E., Dankar, F.K.: Protecting Privacy Using k-Anonymity. Journal of the American Medical Informatics Association **15** (2008)
7. Hasan, M.K., Alam, M.A., Das, D., Hossain, E., Hasan, M.: Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access **8**, 76516–76531 (2020). https://doi.org/10.1109/ACCESS.2020.2989857
8. Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., Habli, I.: Guidance on the assurance of machine learning in autonomous systems (AMLAS). arXiv preprint arXiv:2102.01564 (2021)
9. Jia, Y., Lawton, T., Burden, J., McDermid, J., Habli, I.: Safety-driven design of machine learning for sepsis treatment. Journal of Biomedical Informatics **117**, 103762 (2021). https://doi.org/https://doi.org/10.1016/j.jbi.2021.103762, https://www.sciencedirect.com/science/article/pii/S1532046421000915

10. Jia, Y., McDermid, J., Lawton, T., Habli, I.: The Role of Explainability in Assuring Safety of Machine Learning in Healthcare. IEEE Transactions on Emerging Topics in Computing **10**(4), 1746–1760 (2022). https://doi.org/10.1109/TETC.2022.3171314
11. Luo, F., Qian, H., Wang, D., Guo, X., Sun, Y., Lee, E.S., Teong, H.H., Lai, R.T.R., Miao, C.: Missing Value Imputation for Diabetes Prediction. In: 2022 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2022)
12. McDermid, J.A., Nicholson, M., Pumfrey, D.J., Fenelon, P.: Experience with the application of HAZOP to computer-based systems. In: IEEE proceedings of the 10th Conference on Computer Assurance Systems Integrity, Software Safety and Process Security. pp. 37–48 (1997)
13. McDermid, J.A., Jia, Y., Porter, Z., Habli, I.: Artificial intelligence explainability: the technical and ethical dimensions. Philosophical Transactions of the Royal Society A **379**(2207), 20200363 (2021)
14. Modabbernia, A., Janiri, D., Doucet, G.E., Reichenberg, A., Frangou, S.: Multivariate patterns of brain-behavior-environment associations in the adolescent brain and cognitive development study. Biological psychiatry **89**(5), 510–520 (2021)
15. Molloy, J.J., McDermid, J.A.: Safety Assessment for Autonomous Systems' Perception Capabilities. ArXiv **abs/2208.08237** (2022)
16. National Transportation Safety Board: Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018, NTSB/HAR-19/03 (2019), https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf
17. Ozturk, B., Lawton, T., Smith, S., Habli, I.: Predicting Progression of Type 2 Diabetes using Primary Care Data with the Help of Machine Learning. In: Medical Informatics Europe 2023 (2023)
18. Picardi, C., Hawkins, R., Paterson, C., Habli, I.: A pattern for arguing the assurance of machine learning in medical diagnosis systems. In: Computer Safety, Reliability, and Security: 38th International Conference, SAFECOMP 2019, Turku, Finland, September 11–13, 2019, Proceedings 38. pp. 165–179. Springer (2019)
19. Qi, Y., Conmy, P.R., Huang, W., Zhao, X., Huang, X.: A Hierarchical HAZOP-Like Safety Analysis for Learning-Enabled Systems. In: AISafety 2022 (2022)
20. Ritchie, L.D., Campbell, N.C., Murchie, P.: New NICE guidelines for hypertension (2011)
21. Sohal, K., Mason, D., Birkinshaw, J., West, J., McEachan, R.R.C., Elshehaly, M., Cooper, D., Shore, R., McCooe, M., Lawton, T., Mon-Williams, M., Sheldon, T., Bates, C., Wood, M., Wright, J.: Connected Bradford: a Whole System Data Linkage Accelerator. Wellcome open research **7**, 26 (2022). https://doi.org/10.12688/wellcomeopenres.17526.2, https://europepmc.org/articles/PMC9682213
22. Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I., Reynolds, N.: Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health & Care Informatics **26**(1), e100081 (2019)
23. Wei, S., Zhao, X., Miao, C.: A comprehensive exploration to the machine learning techniques for diabetes identification (2018). https://doi.org/10.1109/WF-IoT.2018.8355130
24. Zainuri, N.A., Jemain, A.A., Muda, N.: A comparison of various imputation methods for missing values in air quality data. Sains Malaysiana **44**(3), 449–456 (2015)