**Automation Bias in Medicine: The Influence of Automated Diagnoses on Interpreter Accuracy and Uncertainty when Reading Electrocardiograms**

Raymond R. Bond1, Tomas Novotny2, Irena Andrsova2, Lumir Koc2, Martina Sisakova2, Dewar Finlay1, Daniel Guldenring1, James McLaughlin1, Aaron Peace3, Victoria McGilligan4,Stephen J. Leslie5, Hui Wang1, Marek Malik6

1Faculty of Computing and Engineering, Ulster University, Newtownabbey, United Kingdom

2Department of Internal Medicine and Cardiology, University Hospital Brno and Faculty of Medicine of Masaryk University, Brno, Czechia

3Department of Cardiology, Altnagelvin Hospital, Western Health and Social Care Trust

4Northern Ireland Centre for Stratified Medicine, C-TRIC, Ulster University, Derry~Londonderry, United Kingdom

5Cardiac Unit, Ragmore Hospital, Inverness, UK, IV2 3UJ.

6St. Paul’s Cardiac Electrophysiology and Imperial College, London, United Kingdom

**Abstract**

Introduction: Interpretation of the 12-lead Electrocardiogram (ECG) is normally assisted with an automated diagnosis (AD), which can facilitate an ‘automation bias’ where interpreters can be anchored. In this paper, we studied, 1) the effect of an incorrect AD on interpretation accuracy and interpreter confidence (a proxy for uncertainty), and 2) whether confidence and other interpreter features can predict interpretation accuracy using machine learning.

Methods: This study analysed 9000 ECG interpretations from cardiology and non-cardiology fellows (CFs and non-CFs). One third of the ECGs involved no ADs, one third with ADs (half as incorrect) and one third had multiple ADs. Interpretations were scored and interpreter confidence was recorded for each interpretation and subsequently standardised using sigma scaling. Spearman coefficients were used for correlation analysis and C5.0 decision trees were used for predicting interpretation accuracy using basic interpreter features such as confidence, age, experience and designation.

Results:Interpretation accuracies achieved by CFs and non-CFs dropped by 43.20% and 58.95% respectively when an incorrect AD was presented (p<0.001). Overall correlation between scaled confidence and interpretation accuracy was higher amongst CFs. However, correlation between confidence and interpretation accuracy decreased for both groups when an incorrect AD was presented. We found that an incorrect AD disturbs the reliability of interpreter confidence in predicting accuracy. An incorrect AD has a greater effect on the confidence of non-CFs (although this is not statistically significant it is close to the threshold, p=0.065). The best C5.0 decision tree achieved an accuracy rate of 64.67% (p<0.001), however this is only 6.56% greater than the no-information-rate.

Conclusion:Incorrect ADs reduce the interpreter’s diagnostic accuracy indicating an automation bias.Non-CFs tend to agree more with the ADs in comparison to CFs, hence less expert physicians are more effected by automation bias**.** Incorrect ADs reduce the interpreter’s confidence and also reduces the predictive power of confidence for predicting accuracy (even more so for non-CFs). Whilst a statistically significant model was developed, it is difficult to predict interpretation accuracy using machine learning on basic features such as interpreter confidence, age, reader experience and designation.

**Introduction**

Twelve-lead electrocardiogram (ECG) interpretation can be a difficult task due to, 1) the large amount of data being presented, 2) the complexity of the signals, 3) the combination of spatial, vectoral and temporal concepts, 4) the electromechanical link (i.e. associating electrical activity with mechanical activity), 5) the large amount of diagnostic criteria to be referenced, 6) confounding signal noise and artefacts as well as the common variability in ECG acquisition and lead misplacement problems, and finally 6) the lack of competency based training [1-12]. As a result, diagnostic accuracy of ECG interpretation has been widely reported to be as low as 35.90% [11]. In addition, inter-rater and intra-rater reliability has also been low in ECG interpretation [7]. Given these challenges, algorithms have been developed to provide an automated diagnosis (AD) to assist the interpreter. However, the accuracy of these algorithms is sub-optimal, likely because of the difficulty of precise feature extraction from signals comprising of noise and artefacts. Nevertheless, clinicians can be overly influenced by the AD which can be referred to as automation bias [13]. Automation bias exists when humans over rely on automation to complete a task. This phenomenon is similar to other cognitive biases such as anchoring and confirmation bias [14]. Automation bias can introduce two types of errors, 1) commission errors which involve using an incorrect AD, and 2) omission errors, which involve not making a diagnosis or treatment decision due to a lack of an AD. There are many potential reasons for automation bias, including excess workload, complexity of the tasks and limited time and cognitive resources in work intensive environments. Nevertheless, automation bias is related to ‘automation-induced complacency’ since trusting a computer is an option that requires minimal cognitive effort [15]. A related phenomenon is the ‘automation paradox’ which implies clinician deskilling as a result of being increasingly dependent on automation to complete a task. Automation bias could bring about the automation paradox where humans lose their skills, vigilance and a reliable sense of certainty in decision making.

To address these issues, we used data of our previous publications [10,11] and designed the present study to measure, 1) the effect of an incorrect AD on interpreter accuracy and interpreter confidence (a proxy for uncertainty), and 2) whether supervised machine learning with features such as interpreter confidence can predict diagnostic accuracy of the interpreter. These aspects have not been previously researched.

**Methods**

A total of 9000 ECG interpretations from 30 physicians from University Hospital Brno, (15 cardiology fellows [CFs] and 15 non-CFs) were analysed. The data collection protocol and the types of ECGs and their ground truth diagnoses have been previously described [10-11]. In brief, 100 ECG tracings have been processed three times by each of the fellows, once without any AD, once with one AD of which a half were incorrect, and once with multiple AD choices. The 12-lead ECGs comprised of normal ECGs, acute coronary syndromes and atrial fibrillation to name but a few (refer to [10-11] for more info). The ADs were simulated to control the experiment.

Each ECG interpretation was scored by a consensus of three experienced cardiologists (no additional biomarkers were used). Interpretations were scored as 0 (correct), 1 (almost correct), 2 (incorrect) or 3 (dangerously incorrect). A score of 3 (dangerously incorrect) denotes an interpretation that would result in delayed or inappropriate treatment. This score is used when there is a seriously incorrect classification such as a false negative interpretation that would lead to a lack of treatment in circumstances where it is critical that the patient receives immediate treatment. This score also represents a false positive diagnosis that would lead to unnecessary treatment. In summary, this score represents diagnostic errors that have severe clinical ramifications and dangerous consequences for the patient. Example ECGs and misinterpretations considered ‘dangerously incorrect’ are presented in our previous publication (refer to [10]).

With each interpretation, each subject provided a confidence rating of how correct they believed their interpretation was (where 1=not very confident, 10=very confident). Confidence is measure of uncertainty and plays a role in clinical decision making. For example, a level of uncertainty can determine whether a clinical decision maker requires further tests or additional clinical opinions. Uncertainty is also often implicitly or explicitly used when communicating a diagnosis or treatment options to patients.

Since confidence ratings are a subjective measure, we rescaled confidence ratings using sigma standardisation. This involved transforming each confidence rating to standard units (i.e. the number of standard deviations each individual rating was from each individual mean confidence rating). Spearman coefficient was then used for correlation analysis between confidence rating (both standardised and unstandardized) and interpretation score. We also computed the effect of interpreter confidence when an AD is correct and when an AD is incorrect.

C5.0 rule induction algorithm was used to create a decision tree for supervised machine learning (also known as predictive modelling) to predict whether an ECG interpretation would be correct based on the interpreter’s attributes such as their level of confidence in the interpretation as well as their experience. A decision tree is an algorithm that can be transparently understood by humans since it can be visually depicted as a tree of binary decisions which naturally fork and route the user or in this case a computer to a final decision (also conceptually known as a leaf). Put in technical terms, a decision tree is a series of hierarchal binary decisions (true or false) that eventually lead to an ultimate decision, classification or prediction depending on the intended use of the algorithm. C5.0 is a special algorithm that automatically designs the decision tree by computing the importance of each feature/variable in predicting the outcomes/leafs/decisions that are of interest. C5.0 computes feature/variable importance using a metric called information gain which in turn is used to determine the hierarchy of the tree and where features should be used to split the tree towards a decision. Put differently, a feature/variable that has significant information gain (importance) will be used at the start of the tree to optimize the routing or forking towards a respective decision. Information gain is computed using entropy. Entropy is the amount of disorder in the data (i.e. the number of different classified cases in the dataset before or after a split). Hence, entropy is a measure from 0 to 1 (where 0=homogeneity [all cases are of the same class] and 1=disorder [cases comprise of a large number of heterogeneous classes]). Entropy is defined in equation 1.

Equation 1.

Where *S* is a segment of data, *e* is the number of classes and is the percentage of values that are classified into class *i*. Information gain is calculated by subtracting the entropy of the known classified cases before splitting based on a given feature from the entropy in the known classified cases after the split on this given feature. IG is defined in Equation 2.

Equation 2.

Where *F* is a given feature (e.g. experience, confidence or designation), is the entropy in the cases before the split and is the entropy in the cases after the split. The greater the IG, the better that feature or decision is at splitting the data into homogenous classes.

C5.0 was used due to it’s ability to provide a non-linear model (whereas logistic regression, a well known linear model performed poorly). Various decision tree models were developed using a combination of various features including confidence rating, age, clinical experience and designation (CF or non-CF). A model was also developed for CFs only and non-CFs only. The best possible model was built using 80% of the cases with grid optimisation were model accuracies were derived from 10-fold cross validation for each model permutation (each permutation comprised of different model parameters such as winnowing [true/false] and number of trials [up to 20 trials]). The remaining 20% was then used to evaluate the best model.

**Data analysis**

All data analytics were carried out using the R programming language and R Studio (using the caret [16] and C5.0 packages). Wilcoxon tests were used for significance testing (alpha=0.05). Chi-square tests were also used to test the significance of the accuracy achieved by the C5.0 decision tree against the non-information rate (NIR). The NIR is the proportion of the most popular case in the test dataset (can also be called the base rate). For example, if the majority of cases in the test dataset are incorrect interpretations (e.g. 70%) according to the gold standard label, then the NIR would be 70%. The NIR relates to the performance of an algorithm that would predict ‘incorrect’ every time regardless of the case features and would still achieve an accuracy of 70% which can be misleading. This is known as the accuracy paradox.

**Results**

As previously published [10,11] 30 fellows were involved (15 CFs [age 30±2, gender: 3 males and 12 females, months of experience:36±11] and 15 non-CFs [age 28±2, gender: 5 males and 10 females, months of experience:28±13]). Using odds ratios (ORs) derived from logistic regression, our research had previously shown that interpreters are more likely to be correct if the AD is correct (OR=10.87) or if multiple ADs are presented (OR=4.43) [10]. The latter OR being a key result showing that the presentation of multiple ADs maximises the chance of an accurate diagnosis by facilitating differential diagnostics.

A total of 84.86% of CF interpretations were correct when a correct AD was presented but this dropped to 41.66% when an incorrect AD was presented (p<0.001). 86.38% of non-CF interpretations were correct when a correct AD was presented but dropped to 27.43% when an incorrect AD was presented (p<0.001). Prior research [10] has shown that the agreement rates between the human and the machine is high when the AD is correct, and whilst agreement rates do drop when the AD is incorrect, the extent of this drop is not that great (especially for non-CFs). For example, non-CFs will on average (median) agree 42.31% of the time when an incorrect AD is presented, whereas a CF will agree 30.77% of the time when an incorrect AD is presented.

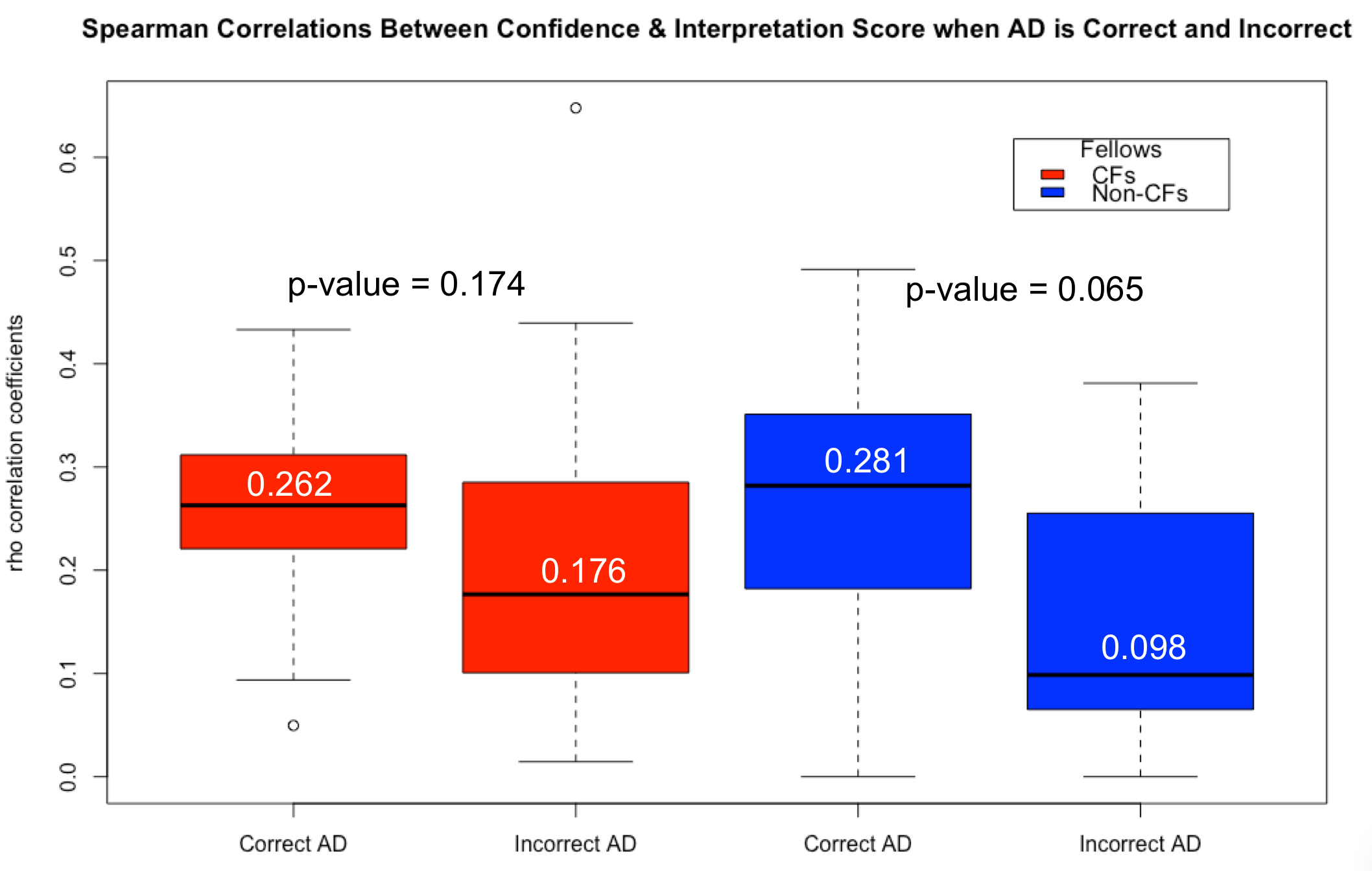
Figure 1 shows the boxplots of confidence ratings and scaled confidence ratings for each interpretation score achieved by CFs and non-CFs. This figure shows that the scaled confidence ratings have an enhanced association with interpretation score in comparison to the non-scaled confidence ratings. The overall correlation between scaled confidence and interpretation score was higher amongst CFs (0.285 vs. 0.221 for non-CFs). For CFs, correlation between confidence and interpretation score decreased from 0.256 to 0.198 when reading ECGs with incorrect ADs. For non-CFs, correlation between confidence and interpretation score decreased from 0.190 to 0.151 when reading ECGs with incorrect ADs.

Confidence ratings amongst both groups dropped by 1 unit when reading ECGs with an incorrect AD (8±4 vs. 7±4 [for non-CFs], p<0.001). This shows that an incorrect AD disturbs the reliability of interpreter confidence in predicting accuracy. Figure 2 shows the boxplots of individual correlations between scaled confidence ratings and interpretation scores and also indicates that an incorrect AD has a greater effect on the confidence of non-CFs (p=0.065). With statistical significance, figure 3 also clearly shows that an incorrect AD effects interpreter confidence when reading an ECG.

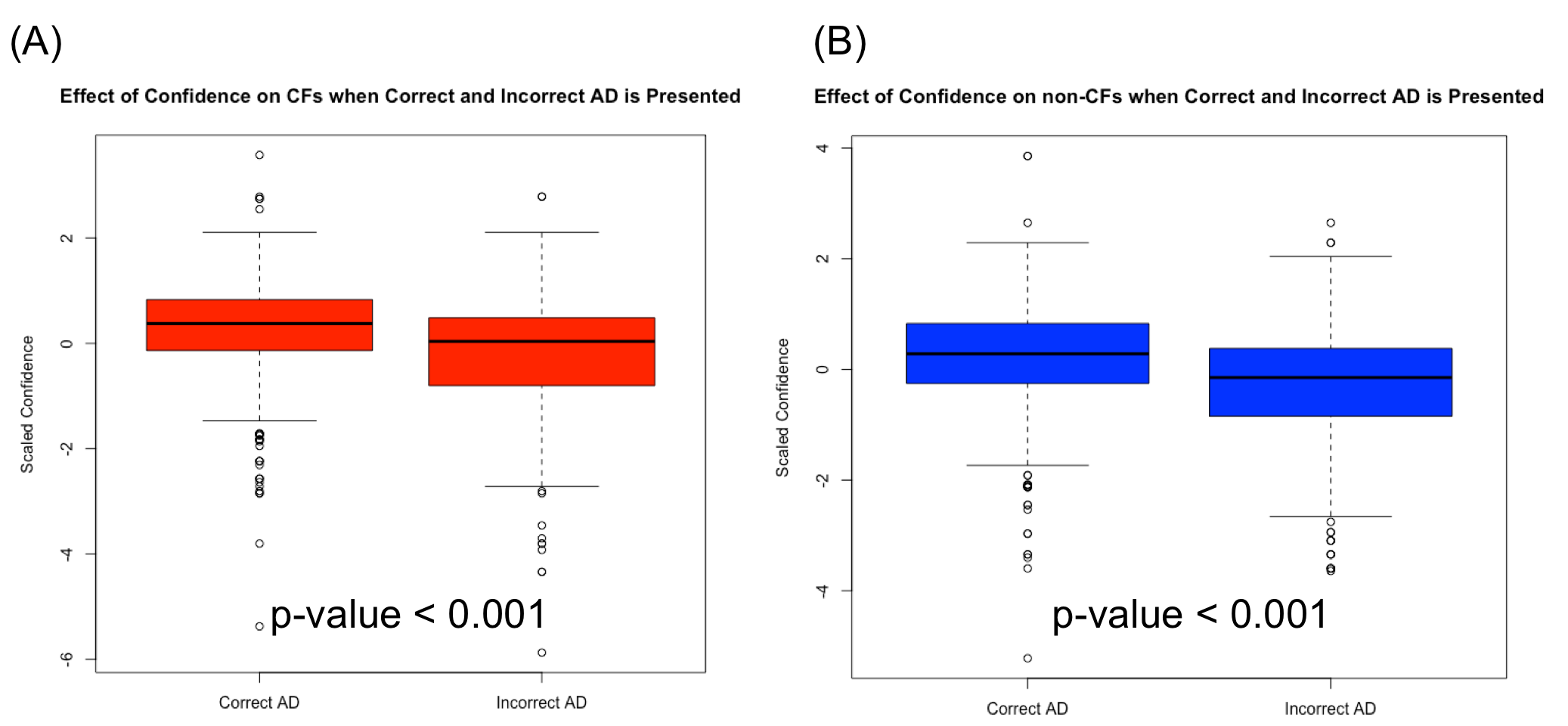
Tables 1 shows the results achieved by C5.0 decision trees for predicting interpreter accuracy (correct/incorrect interpretation). Adding age, experience, and whether the reader was a CF did not substantially improve the predictive power of the model. The best model achieved a statistically significant accuracy rate of 64.67% (p<0.001), however this is only 6.56% greater than the NIR. Table 2 also indicates that a tailored model for predicting performance of CFs only and a tailored model for predicting performance of non-CFs also achieve poor accuracy scores. Nevertheless, the decision tree out performs logistic regression (best logit model accuracy= 55.60%, p=0.539).



**Figure 1. (A) and (B) show scaled and unscaled confidence ratings for each interpretation score achieved for all subjects, (C) and (D) show scaled and unscaled confidence ratings for each interpretation score achieved by CFs [red boxplots], and (E) and (F) show scaled and unscaled confidence ratings for each interpretation score achieved by non-CFs [blue boxplots].**



**Figure 2. Boxplots showing the spread of individual spearman correlations between scaled confidence ratings and interpretation score for both groups when the AD is correct and incorrect (red boxplots=CFs, blue boxplots=non-CFs).**



**Figure 3. (A) shows the difference in scaled confidence ratings amongst CFs when the AD is correct and incorrect, and (B) shows the difference in scaled confidence ratings amongst non-CFs when the AD is correct and incorrect.**

**Table 1. Performance metrics achieved by C5.0 decision trees for predicting interpretation accuracy (correct/not correct) for all subjects.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Confidence** | **Scaled-Confidence** | **Confidence / Scaled-Confidence** | **Scaled Confidence / Age / Experience / Group** | **Scaled Confidence / Age / Experience / Group /**  **Type\*** |
| **Accuracy**  **(%, 95% CI)** | 60.22  (57.92, 62.49) | 61.22  (58.93, 63.48) | 61.44  (59.15, 63.70) | 64.06  (61.79, 66.28) | 64.67  (62.41, 66.88) |
| **Kappa** | 0.125 | 0.198 | 0.186 | 0.238 | 0.272 |
| **Sensitivity (%)** | 28.78 | 51.06 | 44.30 | 45.76 | 56.76 |
| **Specificity (%)** | 82.89 | 68.55 | 73.80 | 77.25 | 70.36 |
| **Pos Pred Value (%)** | 54.80 | 53.92 | 54.93 | 59.18 | 57.99 |
| **Neg Pred Value (%)** | 61.75 | 66.02 | 64.77 | 66.39 | 69.30 |
| **P-value, i.e. Accuracy > No-Information Rate (58.11%)** | *p*=0.036 | *p*=0.004 | *p*=0.002 | <0.001 | <0.001 |
| *\* Type is a nominal variable indicating 1) ‘ECG with an AD’, 2) ‘ECG with multiple ADs’ or 3) ‘ECG without any AD’.* | | | | | |

**Table 2. Performance metrics achieved by C5.0 decision trees that are tailored for each group (CFs and non-CFs) for predicting interpretation accuracy (correct/not correct) for all subjects.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **CF:**  **Scaled-Confidence** | **Non- CF:**  **Scaled-Confidence** | **CF:**  **Scaled Confidence / Age / Experience** | **Non- CF:**  **Scaled Confidence / Age / Experience** | **CF:**  **Scaled Confidence / Age / Experience /**  **Type** | **Non- CF:**  **Scaled Confidence / Age / Experience /**  **Type** |
| **Accuracy in % (95% CI)** | 66.01  (62.84, 69.08) | 58.64  (55.32, 61.91) | 65.57  (62.39, 68.65) | 59.21  (55.89, 62.47) | 68.09  (64.96, 71.10) | 60.56  (57.26, 63.80) |
| **Kappa** | 21.94 | 16.02 | 24.65 | 17.93 | 26.07 | 21.35 |
| **Sensitivity (%)** | 38.69 | 44.74 | 49.11 | 54.07 | 39.58 | 63.88 |
| **Specificity (%)** | 81.87 | 71.09 | 75.13 | 63.81 | 84.63 | 57.60 |
| **Pos Pred Value (%)** | 55.32 | 58.07 | 53.40 | 57.22 | 59.91 | 57.42 |
| **Neg Pred Value (%)** | 69.71 | 58.97 | 71.78 | 60.82 | 70.71 | 64.05 |
| **P-value, i.e. Accuracy > No-Information Rate \*** | 0.046 | <0.001 | 0.079 | <0.001 | <0.001 | <0.001 |
| *\* CF No-Information Rate = 63.28%, Non-CF No-Information Rate = 52.77%* | | | | | | |

**Discussion**

Given the plethora of research in the area of automated diagnostics and AI in medicine, it is important to consider how we can mitigate or control for automation bias. Usually, automation bias is mitigated using routine trainingand by emphasizing the clinician’s accountability. Emphasizing the clinician’s accountability could be used to attenuate the influence computer systems have on the clinician’s final decision by disseminating that its not the computer that is responsible but the human. Providing transparency of an AD can also mitigate automation bias by making clinicians aware of the machine’s rationale and reasoning for its decision. We often require transparency and rationale from human decision making hence it is important to have transparency for an AD, which is a key principle in ethical intelligent systems [17]. To follow these transparency principles, it would be ideal if every ECG machine provided an optional print out of the rules that were used for any given AD. In addition, given we often ask how uncertain or certain a physician is in their diagnostics, perhaps a machine should also be obliged to provide an indicator of its uncertainty. Additionally, such an ECG algorithmic uncertainty index for ADs could help mitigate automation bias allowing readers to calibrate their bias based on how confident the machine is. Other researchers have used a similar algorithmic confidence index in aviation to control for automation bias in pilots [18]. Pilots were previously known to commit errors due to an over reliance on automated assistance. This particular study [18] used machine confidence indexes for what they call ‘trust calibration’. Hence, it is perhaps possible to take a similar approach in ECG interpretation. A potential ECG algorithmic uncertainty index could be a composite score comprising of a number of features including signal-to-noise ratio, likely presence of artefacts and lead misplacement patterns, the percentage of diagnostic criteria met and the prevalence of disease being suggested. However, more importantly this uncertainty index should include the algorithm’s accuracy for the disease being suggestion (also known as class accuracy, which can be much different to the algorithm’s overall accuracy).

Conceivably, automation bias could be mitigated using interaction design since ECG interpretation in the future is likely to be carried out using touch-screen tablet PCs. This would allow the ECG leads to be presented over a number of screens augmented with prompts, questions, supplement visualisations [19]. A number of ADs could be presented in the final screen to avoid the bias. A decision support tool of this nature has already be developed and evaluated [5-6]. Touch screens are likely to be the immediate modality given their popularity, utility, cost-effectiveness and growing use in industry, education and also healthcare. However, the distant future could involve wearable glasses with augmented reality to visualize medial signals or indeed using holographic projections.

Limitations: This work is limited to the small number of subjects (n=30) however the number of interpretations (300 per subject, 9000 in total) provides a reliable representation of diagnostic accuracy. Moreover, other than months of experience and designation, we had no other measurement of expertise for baseline competence for both groups, CF and non-CFs. In addition, the study is limited by being a single center study in one country and in one hospital and may not generalize to other population groups. Moreover, a key limitation is that the subjects were not informed as to what clinical criteria they should use when interpreting the ECG, hence different subjects may have used different diagnostic criteria which could have had an effect on the results achieved.

**Conclusion**

Overall, ADs positively influence decision making. However,incorrect ADs significantly reduce the reader’s diagnostic accuracy indicating an automation bias.Providing multiple ADs increase the chance of being correct suggesting that multiple ADs facilitate a differential diagnosis and mitigate automation bias. Non-CFs tend to agree more with the ADs than CFs, hence they are effected more by automation bias**.** This indicates that less expert personnel are more susceptible to automation bias.There is a poor (but statistically significant) correlation between interpreter confidence and interpretation score. Expert confidence ratings are more predictive of accuracy in comparison to non-expert rating. Incorrect ADs reduce the interpreter’s confidence and also reduces the predictive power of confidence for predicting accuracy (even more so for non-CFs). Whilst a statistically significant model was developed, it is difficult to predict interpretation accuracy using machine learning on basic features such as interpreter confidence, age, reader experience and designation. Future work involves measuring the effect of an uncertainty index that would be affiliated with an AD. We will also compute specific commission and omission errors. We also hope to measure ‘decision switching’ by recording human interpretations before and after the exposure of ADs. In addition, future work may involve studying automation bias amongst specific cases, for example amongst an AF or STEMI population. Also, automation bias could be tested amongst different types of physicians from various disciplines including general practice, however there is a strong hypothesis that the less expert a physician is the more susceptible they are to automation bias and hence naïve to computer-aided diagnoses.

**Acknowledgements**

R Bond, A Peace, V McGilligan and S Leslie are supported by the European Union’s INTERREG VA Programme, managed by the Special EU Programmes Body (SEUPB). Project: Centre for Personalised Medicine – Clinical Decision Making and Patient Safety.

The research was supported in part by Ministry of Health, Czech Republic - conceptual development of research organization (FNBr, 65269705)

**References**

[1] Bond RR, Finlay DD, Nugent CD, Moore G, Guldenring D. A simulation tool for visualizing and studying the effects of electrode misplacement on the 12-lead electrocardiogram. Journal of electrocardiology. 2011 Jul 1;44(4):439-44.

[2] Bond RR, Finlay DD, Nugent CD, Breen C, Guldenring D, Daly MJ. The effects of electrode misplacement on clinicians’ interpretation of the standard 12-lead electrocardiogram. European journal of internal medicine. 2012 Oct 1;23(7):610-5.

[3] Bond RR, Finlay DD, McLaughlin J, Guldenring D, Cairns A, Kennedy A, Deans R, Waldo AL, Peace A. Human factors analysis of the CardioQuick Patch®: A novel engineering solution to the problem of electrode misplacement during 12-lead electrocardiogram acquisition. Journal of electrocardiology. 2016 Nov 1;49(6):911-8.

[4] Finlay DD, Nugent CD, Nelwan SP, Bond RR, Donnelly MP, Guldenring D. Effects of electrode placement errors in the EASI-derived 12-lead electrocardiogram. Journal of electrocardiology. 2010 Nov 1;43(6):606-11.

[5] Cairns AW, Bond RR, Finlay DD, Guldenring D, Badilini F, Libretti G, Peace AJ, Leslie SJ. A decision support system and rule-based algorithm to augment the human interpretation of the 12-lead electrocardiogram. Journal of electrocardiology. 2017 Nov 1;50(6):781-6.

[6] Cairns AW, Bond RR, Finlay DD, Breen C, Guldenring D, Gaffney R, Gallagher AG, Peace AJ, Henn P. A computer-human interaction model to improve the diagnostic accuracy and clinical decision-making during 12-lead electrocardiogram interpretation. Journal of biomedical informatics. 2016 Dec 1;64:93-107.

[7] Bond RR, Zhu T, Finlay DD, Drew B, Kligfield PD, Guldenring D, Breen C, Gallagher AG, Daly MJ, Clifford GD. Assessing computerized eye tracking technology for gaining insight into expert interpretation of the 12-lead electrocardiogram: an objective quantitative approach. Journal of electrocardiology. 2014 Nov 1;47(6):895-906.

[8] Peace A, Ramsewak A, Cairns A, Finlay D, Guldenring D, Clifford G, Bond R. Using computerised interactive response technology to assess electrocardiographers and for aggregating diagnoses. Journal of electrocardiology. 2015 Nov 1;48(6):995-9.

[9] Canning A, Bond RR, Finlay DD, Guldenring D, Aleong G, Peace A. Variable diagnostic accuracy in reading ECGs in a nurse-led primary PCI pathway. Journal of Electrocardiology. 2018 Jan 1;51(1):e8-9.

[10] Novotny T, Bond R, Andrsova I, Koc L, Sisakova M, Finlay D, Guldenring D, Spinar J, Malik M. The role of computerized diagnostic proposals in the interpretation of the 12-lead electrocardiogram by cardiology and non-cardiology fellows. International journal of medical informatics. 2017 May 1;101:85-92.

[11] Novotny T, Bond RR, Andrsova I, Koc L, Sisakova M, Finlay DD, Guldenring D, Spinar J, Malik M. Data analysis of diagnostic accuracies in 12-lead electrocardiogram interpretation by junior medical fellows. Journal of electrocardiology. 2015 Nov 1;48(6):988-94.

[12] Kennedy A, Finlay DD, Guldenring D, McLaughlin J. ECG recording sites for improving signal-to-noise ratio during atrial depolarisation. InComputing in Cardiology Conference (CinC), 2014 2014 Sep 7 (pp. 581-584). IEEE.

[13] Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association. 2011 Jun 16;19(1):121-7.

[14] Kahneman D. Thinking, fast and slow. Macmillan; 2011 Oct 25.

[15] Parasuraman R, Molloy R, Singh IL. Performance consequences of automation-induced'complacency'. The International Journal of Aviation Psychology. 1993 Jan 1;3(1):1-23.

[16] Kuhn M. Caret package. Journal of statistical software. 2008 Feb 29;28(5):1-26.

[17] Chatila R, Firth-Butterflied K, Havens JC, Karachalios K. The IEEE global initiative for ethical considerations in artificial intelligence and autonomous systems [standards]. IEEE Robotics & Automation Magazine. 2017 Mar;24(1):110-110.

[18] McGuirl JM, Sarter NB. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Human factors. 2006 Dec;48(4):656-65.

[19] Bond RR, Finlay DD, Nugent CD, Moore G, Guldenring D. Methods for presenting and visualising electrocardiographic data: From temporal signals to spatial imaging. Journal of electrocardiology. 2013 May 1;46(3):182-96.