New artificial intelligence prediction model using serial prothrombin time international normalized ratio measurements in atrial fibrillation patients on vitamin K antagonists: GARFIELD-AF

Shinichi Goto¹, Shinya Goto²*, Karen S. Pieper³, Jean-Pierre Bassand³,⁴, Alan John Camm⁵, David A. Fitzmaurice⁶, Samuel Z. Goldhaber⁷, Sylvia Haas⁸, Alexander Parkhomenko⁹, Ali Oto¹⁰, Frank Misselwitz¹¹, Alexander G.G. Turpie¹², Freek W.A. Verheugt¹³, Keith A.A. Fox¹⁴, Bernard J. Gersh¹⁵, and Ajay K. Kakkar³,¹⁶; for the GARFIELD-AF Investigators

¹Department of Cardiology, Keio University School of Medicine, Shinanomachi 35, Shinjuku 160-8582, Tokyo, Japan; ²Department of Medicine (Cardiology), Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan; ³Department of Clinical Research, Thrombosis Research Institute, Emmanuel Kaye Building, Manresa Road, Chelsea, London SW3 6LR, UK; ⁴Department of Cardiology, University of Besançon, Boulevard Fleming, 25000 Besançon, France; ⁵Cardiology Clinical Academic Group, Molecular & Clinical Sciences Institute, St. George’s University of London, Cranmer Terrace, Tooting, London, UK; ⁶Department of Cardio-respiratory Primary Care, Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK; ⁷Department of Medicine, Harvard Medical School, Brigham and Women’s Hospital, 75 Francis Street, Boston, MA 02115, USA; ⁸Formerly Klinikum rechts der Isar, Technical University of Munich, Normannstr. 34a, Munich 80333, Germany; ⁹National Scientific Center, Strazhesko Institute of Cardiology, 5 Narodnogo Opolcheniya Street, Kiev 03680, Ukraine; ¹⁰Department of Cardiology, Memorial Ankara Hospital, Sihhiye, 06100, Ankara, Turkey; ¹¹Therapeutic areas Thrombosis & Hematology, Bayer AG, Müllestraße 178, 13353 Berlin, Germany; ¹²Department of Medicine, McMaster University, 237 Barton St E Hamilton, Ontario L8L 2X2, Canada; ¹³Department of Cardiology, Onze Lieve Vrouwe Gasthuis (OLVG), Oosterpark 9, NL-1091 AC Amsterdam, Netherlands; ¹⁴Edinburgh Centre for Cardiovascular Science, University of Edinburgh, Queen’s Medical Research Institute, 47 Little France Crescent, Edinburgh EH16 4TJ, UK; ¹⁵Department of Cardiovascular Medicine, Mayo Clinic College of Medicine, Rochester, MN, USA; and ¹⁶Department of Surgery, University College London, Gower St, Bloomsbury, London WC1E 6BT, UK

Received 23 October 2019; revised 14 November 2019; editorial decision 28 November 2019; accepted 5 December 2019

Aims

Most clinical risk stratification models are based on measurement at a single time-point rather than serial measurements. Artificial intelligence (AI) is able to predict one-dimensional outcomes from multi-dimensional datasets. Using data from Global Anticoagulant Registry in the Field (GARFIELD)-AF registry, a new AI model was developed for predicting clinical outcomes in atrial fibrillation (AF) patients up to 1 year based on sequential measures of prothrombin time international normalized ratio (PT-INR) within 30 days of enrolment.

Methods and results

Patients with newly diagnosed AF who were treated with vitamin K antagonists (VKAs) and had at least three measurements of PT-INR taken over the first 30 days after prescription were analysed. The AI model was constructed with multilayer neural network including long short-term memory and one-dimensional convolution layers. The neural network was trained using PT-INR measurements within days 0–30 after starting treatment and clinical outcomes over days 31–365 in a derivation cohort (cohorts 1–3; n = 3185). Accuracy of the AI model at predicting major bleed, stroke/systemic embolism (SE), and death was assessed in a validation cohort (cohorts 4–5; n = 1523). The model’s c-statistic for predicting major bleed, stroke/SE, and all-cause death was 0.75, 0.70, and 0.61, respectively.

* Corresponding author. Tel: +81-463-93-1121, Fax: +81-463-93-6679, Email: sgoto3@mac.com

© The Author(s) 2019. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Conclusions

Using serial PT-INR values collected within 1 month after starting VKA, the new AI model performed better than time in therapeutic range at predicting clinical outcomes occurring up to 12 months thereafter. Serial PT-INR values contain important information that can be analysed by computer to help predict adverse clinical outcomes.

Keywords

Atrial fibrillation • Artificial intelligence • Machine learning

Introduction

In chronic diseases such as atrial fibrillation (AF) risk stratification using prediction models is useful for clinical decision-making. Several models predict clinical events such as stroke and bleeding. The CHA₂DS₂-VASc and HAS-BLED scores are widely used to select suitable AF patients for oral anticoagulation (OAC). However, some of the variables in these scoring systems are not consistently related to outcomes. Novel machine learning technology has facilitated the development of more accurate models such as the Global Anticoagulant Registry in the Field (GARFIELD)-AF risk model. However, these models incorporate data obtained at a single time-point, baseline. Although computers can process multi-dimensional data such as changes of variables over time, few models have used these inputs to predict future clinical events.

Vitamin K antagonists (VKAs) continue to be prescribed for the prevention of stroke in patients with AF, despite the more recent introduction of non-VKA oral anticoagulants (NOACs). The VKAs are the only recommended choice of OAC for AF patients with haemodynamically overt mitral stenosis and mechanical heart valve. Clinicians adjust the dose of VKA based on an individual patient’s prothrombin time international normalized ratio (PT-INR) at each visit. Time in therapeutic range (TTR) is widely used to standardize the effects of VKA therapy over periods beyond 6 months. Various bleeding risk scores feature a TTR component to enhance accuracy, and TTR has predictive power for thrombotic and bleeding events. However, information on serial changes in PT-INR during early-phase VKA therapy, which may reflect many occult clinical characteristics of patients such as genotype, concomitant medications, and lifestyle, were not included in these TTR-based models.

Advances in artificial intelligence (AI) using recurrent neural networks allow the identification and translation of multi-dimensional data including time-series data directly into meaningful models. Herein, we describe a new AI model for predicting clinical outcomes over 31–365 days after patient enrolment. The model evaluates serially measured PT-INR within the first 30 days of treatment only without other clinical parameters, using data from the largest multinational prospective registry in AF, GARFIELD-AF. The predictive accuracy of the AI model was compared with that of TTR. The working hypothesis was to test whether serially measured PT-INR in early phase can provide information to predict future clinical events.

Methods

Design

The AI model was derived from prospective GARFIELD-AF data gathered in adults with newly diagnosed AF. Three independent AI models were developed with the same composite of neural network structure with multi-dimensional patient-level PT-INR values obtained within the first 30 days after starting treatment. The model tabulated the clinical events of major bleed, ischaemic stroke/systemic embolism (SE), and death occurring within days 31–365.

Registry population

The GARFIELD-AF is an ongoing, international, prospective registry of newly diagnosed patients with AF at risk of stroke. The study design, baseline characteristics, and main results have been published. Eligible patients were adults aged >18 years who had been diagnosed with non-valvular AF within the previous 6 weeks and had at least one risk factor for stroke as judged by the investigator. Risk factors were not pre-specified in the protocol. Any use of antithrombotic agents was shared decision between clinicians and patients only. Patients with a transient reversible cause of AF and those for whom follow-up was not envisaged were excluded. The present analysis was conducted in patients enrolled in GARFIELD-AF cohorts 1–5 between March 2010 and August 2016. Data were extracted from the study database in November 2017.

Study population

Patients who received anticoagulation therapy with VKA and had three or more PT-INR measurements within the first 30 days after enrolment were included in the model. Patients were excluded if they had experienced any outcome events such as serious bleeding or stroke or died within the first 30 days. In this analysis, day of first visit was set as day 0. Patients from cohorts 1–3 (recruited between March 2010 and October 2014) were included in the derivation cohort whereas those in cohort 4–5 (recruited March 2014 to August 2016) comprised the validation cohort. This study design was considered stringent because each GARFIELD-AF cohort exhibited substantial differences in terms of participating countries, use of anticoagulants, and outcomes.

Follow-up

Collection of follow-up data occurred at 4 monthly intervals based on medical records and, sometimes, telephone interviews up to 24 months. The incidence of ischaemic stroke, transient ischaemic attack (TIA), SE, acute coronary syndrome, hospitalization, death (cardiovascular and non-cardiovascular), chronic heart failure (CHF: occurrence or worsening), and bleeding (severity and location) was documented. An audit and quality control programme was applied, and data were examined for completeness and accuracy by the co-ordinating centre (TRI, London, UK). By design, 20% of all electronic case report forms in the GARFIELD-AF registry were monitored against source documentation at sites over the 8 years of recruitment and follow-up.

Outcomes

Outcome measures used in this analysis were major bleeding, stroke/SE, and all-cause death occurring between days 31 and 365. Major bleed was classified by investigators according to International Society on Thrombosis and Haemostasis definition. Stroke/SE was defined as a combined Endpoint of ischaemic stroke, SE, and TIA.
Artificial intelligence model

The structure of neural networks for the AI model is shown in Figure 1A. To deal with serial data on raw PT-INR measurements, the AI model was constructed by stacking multiple layers of special neurons that can deal with time-dependent data, namely one-dimensional convolution layer and long short-term memory (LSTM) layer. The LSTM layer transfers rectified data to each neighboring neuron. This structure allows the layer to learn time-dependent data in sequential order.

The neural network model was trained independently for each outcome event. For training, PT-INR measurement patterns for each individual patient were converted to a 30 dimensional PT-INR vector as shown in Figure 1B. All PT-INR measurements obtained within the first 30 days were input to the model. The measured PT-INR value was inserted into the n-th dimension of the 30 dimensional vector, where n is the number of days after starting VKA. Un-measured data-points were filled with 0. Each vector for patients was labelled with the occurrence of outcome (0 for no event and 1 for event for all three outcome measures) within days 31–365. The neural networks were trained with the multi-dimensional data-set of the PT-INR vector and outcome label as shown in Figure 1C.

Model training

The process of model training is shown in Figure 2. The training was performed using only patient data from the derivation cohort. The derivation cohort was further split into training (70%) and testing (30%) datasets. The training was performed for 500 epochs and each training epoch included a mini-batch of 455 patients randomly selected from the training dataset. Conceptually, the performance of the model is designed to improve by training with longer epochs. However, this approach can also result in overfitting. To avoid this pitfall and select the model with best performance, the model was evaluated using the testing dataset at the end of each epoch. The final model was that which performed best with the testing dataset. The performance was measured by calculating the c-statistics of the prediction model for all the data in testing dataset. No data from validation cohort were used for training.

Model validation

The derived models were validated by inputting the 30 day PT-INR vector and obtaining prediction scores for each outcome. Predicted outcomes were compared with the actual clinical course for each individual patient in the validation cohort. Receiver operating characteristic (ROC) curves were drawn to evaluate the predictive value of the model. The threshold to achieve overall best accuracy for the model was determined and the model’s sensitivity and specificity calculated at that threshold. To test the ability of the model to discriminate between high- and low-risk patients for each event, three sets of Kaplan–Meier plots were drawn for event rates stratified as high and low risk with the threshold.
Table 1

Baseline characteristics are displayed in Table 1. Ninety-eight patients were excluded (92 with an outcome event within the first 30 days and were included in the an novo AF patients treated with VKA, 4806 had at least three PT-INR measurements during initial 30 days were slightly less likely to have paroxysmal AF and CHF than those with fewer than three INR measurements. No difference in patients’ sex, age, body mass index, and left ventricular ejection fraction. Patients with at least three PT-INR measurements during initial 30 days were slightly less likely to have paroxysmal AF and CHF than those with fewer than three INR measurements. No difference in baseline characteristics was noted between derivation and validation cohorts.

Predictive value of artificial intelligence model

The ROC curve compiled for the validation cohort (Figure 4A) revealed that the AI model had a statistically higher predictive value compared with TTR with c-statistics for major bleeding and all-cause death 0.75 and 0.61, respectively (both \( P = 0.01 \) vs. TTR). A similar trend albeit nonsignificant was observed for stroke, with a c-statistic 0.70 (\( P = 0.08 \) vs. TTR). Forest plots of 95% CI for AUC of ROC curves (Figure 4B) show that the AI model performed better than random; c-statistics for major bleed, stroke, and all-cause death were 0.75 (95% CI, 0.62–0.87), 0.70 (95% CI, 0.56–0.83), and 0.61 (95% CI, 0.54–0.67), respectively, whereas TTR was not significantly different compared with random (for same outcomes: 0.47 [95% CI, 0.32–0.61], 0.47 [95% CI, 0.31–0.64], and 0.48 [95% CI, 0.42–0.54], respectively). Table 2 shows the accuracies, sensitivities, and specificities for the best thresholds derived from the ROC curve for major bleed, stroke, and all-cause death. The model showed good predictive accuracy for major bleeding with a sensitivity 0.79 and specificity 0.78. These results were similar for the training dataset (Supplementary material online, Table 1 and Figure 1A, B).

Survival analysis

Kaplan–Meier plots stratified by risk determined from the AI model are shown in Figure 5. The threshold of prediction score was calculated for each event to achieve the best accuracy according to the ROC curve. The best thresholds for major bleed, stroke/SE, and all-cause death were 0.27, 0.44, and 0.49, respectively. Note that these output values from our model are arbitrary numbers related to risk of future events but not actual probabilities. Patients who had a model output higher than or equal to threshold were classified as high risk and the remainder were low risk. Among 1523 patients in the validation cohort, 354 were classified high risk for major bleeding, 738 for stroke, and 560 for death. High-risk patients had higher cumulative event rates (major bleed, stroke/SE, and all-cause death) compared with low-risk patients.

The same analysis was performed for the derivation dataset. No threshold calculations were performed for the derivation dataset and the same thresholds obtained from the validation dataset were used for this analysis. The results were similar, supporting the robustness of the threshold (Supplementary material online, Figure 2).

Discussion

We created a new method to convert early time-series measurements of PT-INR to a long-term prediction model. The novel AI model, constructed with neural networks including one-dimensional convolution and LSTM, garnered useful information from the raw PT-INR values, and measurement dates over 30 days after VKA initiation and converted this to predict major bleeding events over the next 11 months. Although TTR is widely used to standardize VKA therapy over the long term, its accuracy for predicting future thrombotic/bleeding events is low. A previous report showed that GARFIELD-AF patients with 1 year TTR <65% had worse outcomes.
than those with greater values.\textsuperscript{36} Within 30 days, TTR has low predictive power because early-phase PT-INR values vary greatly due to a number of influencing factors including genetics,\textsuperscript{31,32} choice of commercial thromboplastin and coagulometer device,\textsuperscript{37–39} and patients’ lifestyles.\textsuperscript{40} With the use of AI, we show here the presence of important information in raw PT-INR patterns over first 30 days that can support patient management.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>≥3 PT-INRs (N = 4708)</th>
<th>0–2 PT-INRs (N = 9630)</th>
<th>≥3 PT-INRs subgroup Derivation (N = 3185)</th>
<th>Validation (N = 1523)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>2085 (44.3)</td>
<td>4330 (45.0)</td>
<td>1420 (44.6)</td>
<td>665 (43.7)</td>
</tr>
<tr>
<td>Male</td>
<td>2623 (55.7)</td>
<td>5300 (55.0)</td>
<td>1765 (55.4)</td>
<td>858 (56.3)</td>
</tr>
<tr>
<td>Age at dx, years</td>
<td>72.1 (9.9)</td>
<td>70.0 (10.7)</td>
<td>72.2 (9.7)</td>
<td>72.0 (10.2)</td>
</tr>
<tr>
<td>BMI, kg/m(^2)</td>
<td>28.7 (5.9)</td>
<td>28.1 (5.7)</td>
<td>28.6 (5.7)</td>
<td>29.0 (6.1)</td>
</tr>
<tr>
<td>LVEF, %</td>
<td>53.7 (12.9)</td>
<td>55.7 (12.7)</td>
<td>53.2 (13.2)</td>
<td>54.7 (12.2)</td>
</tr>
<tr>
<td>Type of AF, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>2409 (51.2)</td>
<td>4087 (42.4)</td>
<td>1706 (53.6)</td>
<td>703 (46.2)</td>
</tr>
<tr>
<td>Paroxysmal</td>
<td>798 (16.9)</td>
<td>2207 (22.9)</td>
<td>567 (17.8)</td>
<td>231 (15.2)</td>
</tr>
<tr>
<td>Permanent</td>
<td>877 (18.6)</td>
<td>1514 (15.7)</td>
<td>487 (15.3)</td>
<td>390 (25.6)</td>
</tr>
<tr>
<td>Persistent</td>
<td>624 (13.3)</td>
<td>1822 (18.9)</td>
<td>425 (13.3)</td>
<td>199 (13.1)</td>
</tr>
<tr>
<td>CHF, n (%)</td>
<td>721 (15.3)</td>
<td>2149 (22.3)</td>
<td>466 (14.6)</td>
<td>255 (16.7)</td>
</tr>
<tr>
<td>CAD, n (%)</td>
<td>878 (18.6)</td>
<td>1896 (19.7)</td>
<td>511 (16.0)</td>
<td>367 (24.1)</td>
</tr>
<tr>
<td>ACS</td>
<td>461 (9.8)</td>
<td>872 (9.1)</td>
<td>292 (9.2)</td>
<td>169 (11.1)</td>
</tr>
<tr>
<td>CHA(_2)DS(_2)-VASc</td>
<td>3.4 (1.5)</td>
<td>3.3 (1.5)</td>
<td>34 (1.5)</td>
<td>33 (1.4)</td>
</tr>
<tr>
<td>HAS-BLED</td>
<td>1.4 (0.9)</td>
<td>1.4 (0.9)</td>
<td>15 (0.9)</td>
<td>14 (0.9)</td>
</tr>
</tbody>
</table>

Values are mean (SD) unless specified otherwise.
ACS, acute coronary syndrome; AF, atrial fibrillation; BMI, body mass index; CAD, coronary artery disease; CHF, congestive heart failure; LVEF, left ventricular ejection fraction.

Overall, output values from our AI model are related to risk of future events but not their probability. Therefore, calibration of the model with typical Hosmer-Lemeshow goodness-of-fit (GOF) test is not feasible.

**Study limitations**

Several limitations of this analysis should be noted. First, validation of the AI models was performed using datasets derived from the GARFIELD-AF registry. External validations of the AI model were not conducted. Thus, validity of this model beyond GARFIELD-AF patients is unknown. On the other hand, large dissimilarities between cohorts 1–3 and 4 and 5 were noted, suggesting that our model is sufficiently robust to apply in daily clinical practice. Prothrombin time international normalized ratio within 30 days may be influenced by concomitant dosing with parenteral anticoagulants. However, our model attempted to account for all influencing factors beyond the effects of VKA. We hope that other researchers will test our model’s performance in external datasets.

Second, the AI model was trained only with PT-INR data and did not include other information such as sex, age, biomarkers, concomitant drugs, or other serially measured values. Although consecutive patient data were analysed, unrecognized confounders may exist. Many other known risk factors for adverse outcome events were not considered in our models.
Third, by selecting only patients with >3 PT-INR measurements within 30 days, two-thirds of the entire cohort were excluded, which could introduce selection bias. Furthermore, patients do not necessarily remain stable after day 31 and our model cannot capture changes at time-points later than day 31. Future studies will examine the impact of time periods beyond 30 days in relation to AI risk prediction.

Fourth, although our results suggest the presence of crucial information within the PT-INR measurement pattern to predict patients’ clinical course, the nature of that information is unknown. It might be present in the target PT-INR value, PT-INR fluctuations, PT-INR measurement frequency, or elsewhere.

Fifth, the c-statistics, sensitivity, and specificity of our models are far from perfect. Further studies to improve predictive accuracy possibly by adding other clinical characteristics and measurements are necessary.

Sixth, statistical significance was not achieved in either the derivation or validation cohort in comparison with TTR for prediction of all-cause death and stroke. This could be explained by low numbers of events limiting statistical power. Moreover, even though the number of deaths observed was not low, they could have been caused by factors not

---

**Table 2** Best predictive accuracies and corresponding sensitivities and specificities (95% CIs) for validation cohort

<table>
<thead>
<tr>
<th>AI</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major bleed</td>
<td>0.78 (0.40–0.92)</td>
<td>0.79 (0.50–1.00)</td>
<td>0.78 (0.39–0.93)</td>
</tr>
<tr>
<td>Stroke</td>
<td>0.53 (0.24–0.98)</td>
<td>0.85 (0.31–1.00)</td>
<td>0.53 (0.23–0.99)</td>
</tr>
<tr>
<td>All-cause death</td>
<td>0.64 (0.51–0.69)</td>
<td>0.63 (0.50–0.76)</td>
<td>0.65 (0.50–0.70)</td>
</tr>
</tbody>
</table>

AI, artificial intelligence.
related to anticoagulation. Validation of the model in larger cohorts with higher numbers of events may demonstrate better predictive power for death and stroke. Despite these limitations, our results suggest that AI can capture important information to predict future outcomes from early-phase PT-INR measurements.

**Conclusions**

In AF patients treated with VKA, we developed new AI models to predict all-cause death, stroke, and major bleeding events occurring between months 2 and 12. The models’ predictive accuracy was greatest for major bleeding, followed by all-cause mortality and stroke/SE. Our results imply that AI can capture important information to predict future outcomes from early-phase PT-INR measurements.

**Supplementary material**

Supplementary material is available at *European Heart Journal – Cardiovascular Pharmacotherapy* online.

**Acknowledgments**

Alex Kahney of Thrombosis Research Institute, London, UK provided editorial support.

**Funding**

The GARFIELD-AF registry is an independent academic research initiative sponsored by the Thrombosis Research Institute (London, UK) and supported by an unrestricted research grant from Bayer AG (Berlin, Germany). The funding source had no involvement in the data collection, data analysis, or data interpretation.

**Conflict of interest**

S.G. (1st author) has no financial competing interest to disclose. S.G. (2nd author) acknowledges financial support from MEXT/JSPS KAKENHI 17K19669, partly by 18H01726 and 19H03661, and from Bristol-Myers Squibb for independent research support project (33999603). S.G. (2nd author) acknowledges grant support from Vehicle Racing Commemorative Foundation and Nakatani Foundation for Advancement of Measuring Technologies in Biomedical Engineering. S.G. (2nd author) received research funding from Sanofi, Pfizer, and Ono, and a modest personal fee from Bayer. S.G. (2nd author) is an associate editor for *Circulation, Journal of Biochemistry, and Archives of Medical Science* and section editor for *Thrombosis and Haemostasis*. K.S.P. has no financial competing interest to disclose. J.P.B. reports personal fees from Thrombosis Research Institute, during the conduct of the study. A.J.C. has received institutional grants and personal fees from Bayer, Boehringer Ingelheim, BMS/Pfizer, Daiichi-Sankyo, and M.S.C. A.P. has received consulting fees and honoraria from Bayer HealthCare, Boehringer Ingelheim, BMS/Pfizer, Daiichi-Sankyo, and Portola. S.Z.G. has received personal fees from Bayer, Daiichi-Sankyo, and Portola. K.A.A.F. has received grants from Bayer/Janssen and AstraZeneca and consultation fees from Bayer/Janssen, Sanofi/Regeneron, and Verseon. B.J.G. is a consultant for Janssen Pharmaceuticals. A.K.K. has received research support from Bayer AG and personal fees from Bayer AG, Boehringer-Ingelheim Pharma, Daiichi-Sankyo Europe, Pfizer, Janssen Pharma, Sanofi SA, and Verseon.


