Classification of Cardiovascular Conduction Disorder Using 12-lead Electrocardiograms

Amr El-Bokl, Patrick Johnson, Colin Ornelas, Yaxuan Zhang

High-level problem description

Sudden cardiac death is the unexpected demise of an otherwise healthy person in a short period of time; namely within 1 hour of symptom onset. The most common cause of such rapid deaths is arrhythmias; a fatal disturbance in the electrical conduction of the heart causing inefficient filling and ejection and subsequent hypoxia of major organs and death (Srinivasan et al.). Given the rapid timeline from symptom onset to demise, arrhythmias are extremely difficult to diagnose and treat.

In the majority of cases, there are undiagnosed underlying cardiac conditions that predispose people to fatal arrhythmias. An example of this is hypertrophic cardiomyopathy, the most common underlying cause of sudden cardiac death in teenage athletes. Hypertrophic cardiomyopathy is a condition of myocardial fiber disarray causing subclinical inefficiency in contraction and ejection. To compensate for the disarray, the cardiac muscle hypertrophies (augments in thickness) beyond its own ability to provide blood supply to itself, eventually causing subendocardial ischemia and scarring, which provide niduses for fatal ventricular arrhythmia. Another example is myocarditis, an inflammatory condition of the heart that causes disturbance of myocardial ion channel function and, in its fulminant subtype, can cause fatal arrhythmias. Other examples include long QT syndrome, left anomalous coronary artery, Brugada syndrome, arrhythmogenic right ventricular cardiomyopathy, and myocardial infarction.

12 lead electrocardiograms (ECGs) are the gold standard for diagnosis and identification of fatal arrhythmias. Moreover, most of the underlying conditions listed above have identifiable, recognizable patterns on ECGs prior to the onset of symptoms, which provides a window of opportunity for screening and sudden cardiac death prevention.

Though ECGs have such potential for screening and sudden cardiac death prevention, they are currently not recommended as a universal screening tool for the general population. This is because epidemiological studies have uncovered a low yield-to-cost ratio, especially in healthy young teenagers and adults. Our group argues that this lack of yield is likely due to our dependence on human interpretation for large volume screening. It is much more cost-effective to automate a large volume screening program utilizing a machine learning algorithm. Moreover, we argue that an algorithm may be able to detect features on ECG much more accurately and completely than the human eye and potentially make new connections between changes in electrocardiograms and pathology. Thus through this project, we take the first step by designing algorithms for the detection of conduction disturbances.

Technical ML problem

Electrocardiograms present a feature-rich dataset that lends itself well to many different machine learning approaches. Each ECG consists of 12 leads sampled at 500 hertz for 10 seconds, which ultimately results in a feature space of p=60000 for one given test. Although feature rich, it can be difficult to break down all this data into relevant information that can be used for disease classification.

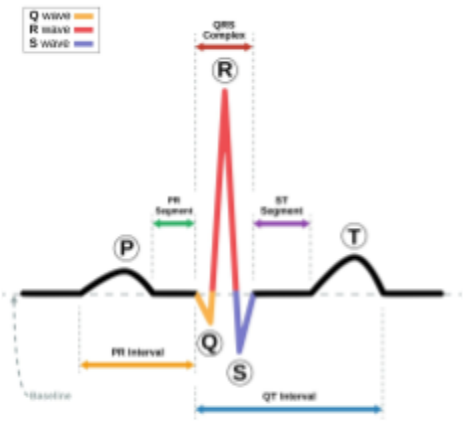

Figure 1: ECG Beat: https://en.wikipedia.org/wiki/Pan%E2%80%93Tompkins_algorithm

One common approach is to use a Convolutional Neural Network (CNN) (Zachi et al.). CNNs work especially well for ECGs for several different reasons. Each heartbeat consists of a P, Q, R, S, and T wave. These features are critical in diagnosing disease classes, namely, their order, morphology, frequency, and amplitude. Add to this potential complex relationships waves may have with one another to reflect an underlying physiological process. A CNN is a way to implicitly extract this information without a large amount of preprocessing and contextual knowledge. A CNN accomplishes this through repeating layers of convolutions and max poolings, followed by a fully connected layer with an output layer and activation function.

Another common approach is to extract features directly from the waveforms. For example, it is a common approach to look at heart rate variability (HRV). In this approach, the relative time between peaks is found, and statistics are taken. However, it is often difficult to extract these features and their relationships with one another. It also requires a good contextual understanding of heart functionality and potential problems. Traditionally, however, this is the approach that has been used by statisticians for over half a century prior to advancements available through machine learning (Rafie et al.).

This project will take into account both of these approaches and try to baseline against some well-known CNNs used for ECG disease classification.

Description of data and preprocessing

Due to the inability to obtain ECG data from the University of Minnesota study, the group pivoted to an open-source ECG dataset, called PTB-XL. PTB-XL is a part of PhysioNet, a freely available medical research data managed by the MIT Laboratory for Computational Physiology. It contains 12-lead ECG data for 21799 patients in both 100hz and 500hz frequency, as well as diagnostic labels and demographic information for each patient. The data has already been statistically summarized by its owners, and it contains a sample script that extracts ECGs into 12x5000 arrays. The diagnostic labels are divided into the following superclasses shown in Table 1:

Superclass	Description
NORM	Normal ECG
MI	Myocardial Infarction
STTC	ST/T Change
CD	Conduction Disturbance
HYP	Hypertrophy

Table 1: PTB-XL Superclasses

Some patients have more than one superclass designated, which adds nuance to any multi-class classification effort. Since myocarditis labels were not present (original project goal), the group decided to focus on Conduction Disturbance (CD), and Non-Conduction Disturbance (Non-CD), which will contain all other labels. CD was chosen because of the large sample size in the dataset and the broad nature of disease symptoms. Since myocarditis is an inflammatory condition that often presents with conduction disturbances such as premature contractions, repolarization abnormalities, or tachyarrhythmias, we felt that transitioning to a model for the detection of conduction disturbances would be comparable. Furthermore, predicting conduction abnormalities would have wider applications, namely the prevention of sudden cardiac death.

ECG signals can be inherently noisy and are highly susceptible to baseline wandering and power line interference (PLI). To address these issues, we experimented with multiple filtering approaches. In our final implementation, we landed on an IIR notch filter to address wandering and PLI, complimented by a bandpass (butter) filter to remove noise (Chavan et al.). Figure 2 shows a sample ECG pre and post-processing. During the early phases of the analysis, additional techniques, such as peak alignment, were tested.

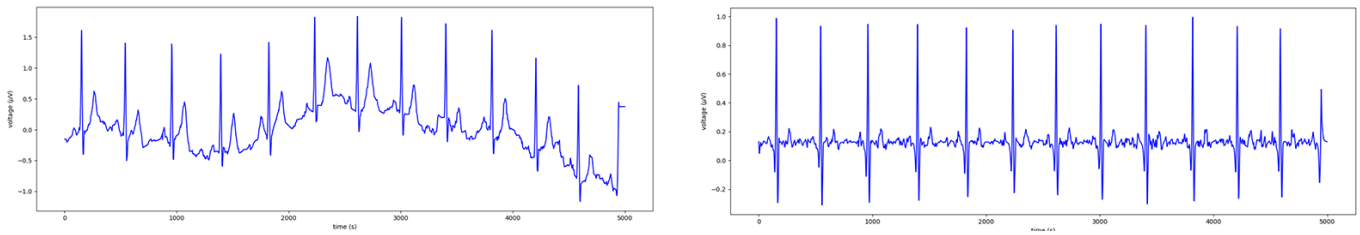

Figure 2: Pre and Post Filtered ECG Signal

Applied Algorithms

Approach 1: Feature Extraction and AutoML

Feature Extraction Method 1: Autoencoder: In a preliminary analysis, an autoencoder was generated to extract relevant features from 12-lead ECG data. For this task, ECGs with a sampling frequency of 100Hz were used. The encoder was constructed using four fully connected layers, each with a ReLu activation function. The input to the encoder was [32, 12000], where 32 represents the batch size, and 12000 represents a flattened ECG signal (original 12 leads). The autoencoder outputs 64 latent features. Similarly, the decoder is constructed using four fully connected layers using a ReLu activation function and one final sigmoid activation function. The decoder takes the output from the encoder, and attempts to recreate the original flattened ECG. During training, mean squared error (MSE) loss was used as an objective function, and Adam optimizer was used. The model was trained over 25 epochs, and cumulative loss was observed. After training, the encoder was used to generate extracted ECG features to be used for further modeling.

Feature Extraction Method 2: Features based on domain knowledge: In a secondary approach, feature extraction was performed using well-known and defined ECG and signal features. Multiple approaches and packages were used during this process, including ecgdetectors, neurokit2, wfdb, and scipy. The extracted features include data points such as R-R intervals, QRS complex locations, heart rate, heart rate variability, wavelet coefficients, and ST-segment elevation/depression. Furthermore, statistical measures of these values were calculated (e.g., mean, skew, and standard deviation). This process generated a high volume of features, which were then reduced based on a missingness threshold (50%) and a low variance threshold (0.003, after normalization).

AutoML: After feature extraction using both methods, AutoML was applied to each feature set to get performance metrics across a broad range of model types (e.g., Gradient Boosting Classifier, AdaBoost Classifier, Random Forest, SVM, Naive Bayes, and LDA) using PyCaret. Models were generated on training data (80%) using stratified K-fold cross-validation. Final metrics were reported on the final 20% test data. For the second approach (well-known features), a feature importance plot was generated for the gradient-boosting classifier.

Approach 2: CNN

For this analysis, 20% of ECGs were held out as a test set (n=4359). Due to the highly imbalanced nature of the remaining 17439 records (13554 Non-CD, 3885 CD), 4500 Non-CD records were randomly selected and included in the model building, while the rest were discarded. In disease classification, highly imbalanced datasets can lead to a number of problems, such as majority class bias, poor generalization, and misleading evaluation metrics. This left an overall population of 8385 records. These records were then split using 87.5/12.5 training-validation split. The final architecture of the CNN is shown in the Appendix. The model was fit using a binary cross-entropy loss function, and Adam was the optimizer. The best optimal performance was found using 3 convolutional layers, followed by Max Pooling and Dropout. The kernel size of the first convolutional layers started small, but it was found that larger layers produced better results. This is likely due to relevant features spanning longer intervals and multiple ECG leads. Max Pooling was used to maintain the most relevant features while keeping the dimensionality in check. Larger pooling sizes in the first convolutional layer helped keep the model to reasonable dimensionality. A dropout of 0.6 was decided because of initial problems with overfitting. The overall architecture was derived from a basic MNist classification problem and modified. The parameters were optimized primarily by guess and check, checking the training and validation accuracy through each epoch. The model was trained for 30 epochs, which was the point the validation loss no longer decreased.

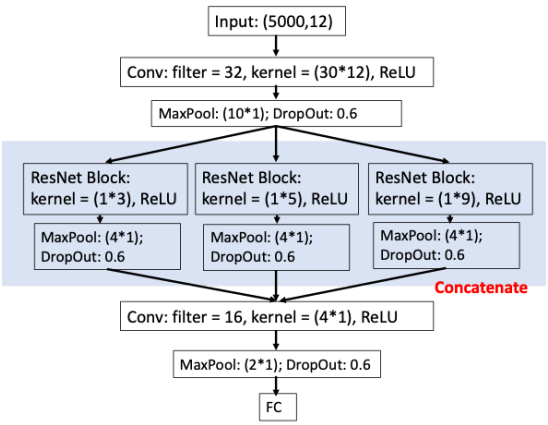

Figure 3: Multi-Scale ResNet CNN Architecture

Approach 3: Multi-Scale ResNet CNN

Based on the CNN model, the multi-scale ResNet was used for better information extraction. Similar to the CNN, the model starts with a convolutional layer with a large kernel size to capture relationships across leads. After that, the three residual blocks with kernel sizes (1,3), (1,5), and

(1,9) ran twice separately, followed by Max Pooling and DropOut. The multi-scale ResNet was used for two reasons: 1) The added residual blocks can skip some convolutional connections to ensure some important features carry through to the final layers. 2) Different kernel sizes were used to capture features at different scales. The outputs of the three residual blocks were concatenated together as the input for the third convolutional layer with a smaller kernel size and lastly a fully connected layer with a sigmoid function for prediction. The model was trained 30 epochs, and the ResNet showed a satisfying result in the test set. The model architecture is summarized in Figure 3.

Results

Model Performance: Models were evaluated using AUC (Area under the ROC curve), and final values were reported on the test set. AUC was used as the primary evaluation metric here, in part, because metrics more susceptible to class imbalance, such as error rate, would have shown overly optimistic results. Summaries of AUC are described in Table 2, in descending order of performance. For brevity, only a few select models are highlighted from the AutoML approach. The top performing algorithm was the ResNet CNN, with an AUC of 0.885. As a general rule of thumb, an AUC of anything >.8 overall represents a 'good' or 'very good' fit. For this model, a confusion matrix is shown in Figure 4, with an overall accuracy of 85.7%

Model	AUC
Multi-Scale ResNet CNN	0.885
CNN	0.873
Gradient Boosting - Known Features	0.764
Random Forest - Known Features	0.764
Ada Boost Classifier - Autoencoder	0.526

Table 2: AUC Results for several models

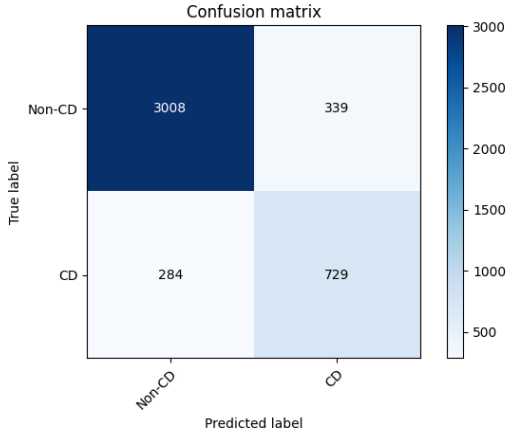

Figure 4: Multi-Scale ResNet Confusion Matrix

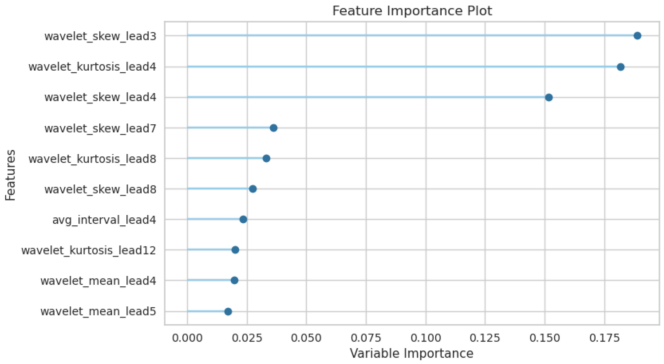

Figure 5: Gradient Boost Model Feature Importance

Model Interpretability: The top 10 most important predictors, ranked based on variable importance, for the gradient boosting classifier are shown in Figure 5.

External Validation: For most clinical models, it is generally not sufficient to only perform internal validation. In other words, once a model is developed using data from one or more hospitals, it must be then evaluated on an external cohort that was not part of the original model creation. To mimic this process, we validated the best-performing model (ResNet CNN) on a cohort of patients from Mayo Clinic (with IRB approval). The Electronic Health Records (EHR) were mined for a sample of recently performed ECGs. Labels were generated using International Classification of Disease (ICD) codes. The same preprocessing algorithms were applied to the model, then predictions were generated. The overall AUC was 0.851. Notably, while the code used to generate this validation is provided (with the exception of database connection info and database schema names), the data for this portion of the analysis can not be shared.

Discussion

In this work, we have applied a number of machine learning techniques to an open-source ECG dataset. Models started with relatively simple approaches, and increased in complexity until a final model was selected. First, a fairly naive approach of using an autoencoder to extract features from ECGs was applied. Based on plots generated from data that passed from the encoder to the decoder, it was clear only the main peaks of an ECG were being identified, and most of the nuances and complexities of an ECG waveform were lost. This technique yielded poor results. Therefore this warranted manual feature extraction based on domain-level expertise. One advantage to this approach is that the results are easy to interpret (e.g., how heart rate variability is associated with CD), which is frequently desired by both clinicians and patients. Despite this advantage, the results were not good enough to justify its use, which led to the more complex model architectures of a CNN and a Multi-Scale CNN.

The work performed here adds to the rapidly growing field of auto-interpreted ECGs using machine learning. With an AUC over 0.88, this model could serve as a meaningful screening tool in a clinical setting. Indeed, this model performance exceeds many tools used in cardiovascular medicine today. For reference, the CHA_2DVASc Score (screening tool or stroke risk) and commonly used peptide tests for heart failure have an AUC of 0.6-0.7, and outperforms other well-cited ECG prediction tools (Wu et al. and Bhlla et al.) .

One of the trickiest aspects of working in a clinical domain is the huge variation of patient populations, treatment types, electronic health record systems, and types of care provided. Harmonizing data across institutions can be challenging, if not impossible, and oftentimes models that work for one hospital system do not perform elsewhere. While we did see a small drop in AUC during our external validation, the result was still high enough to remain valuable.

Next steps/future work

Our models were limited by the broad target of conduction disturbance. We hypothesize that one way to hone in our model and improve our predictions is to narrow down our targets. This can be done either by separating the broad category of arrhythmia or the mechanism. So our next step would be to separate bradycardias from tachycardias, acquired from congenital, or ischemia mediated versus non-ischemic arrhythmias. Moreover, we believe that model performance could be built upon in a few ways, including (1) adding additional demographic and clinical features (e.g., age and sex) into the models, (2) gaining access to higher-performing compute environments that would allow for shorter training times and more hyperparameter tuning, and (3) adding more model explainability to aid in interpretation, including things like saliency maps. Finally, it would be of interest to try and generate a single lead model. The first lead of a standard 12-lead ECG is nearly identical to that of a single-lead ECG recorded on wearable devices (e.g., a smartwatch). A successful model that could predict conduction disorders on a wearable device would greatly improve its ability to serve as a screening tool for asymptomatic patients.

Feedback encorporation

Overall the group found the feedback to be helpful and incorporated it into two main facets. One question that was received multiple times was why we were interested in applying a CNN. As a result, we used this feedback and took a step back to evaluate other possible model types and architectures, as well as ensure we could justify its use. We used these other models as a benchmark to compare the performance of the CNN. Based on feedback, the group also realized that a single CNN would not be an appropriate scope for a project of 4. Based on this, the group decided to explore multiple methods for disease classification, such as manual feature extraction/classification, implicit feature extraction/classification, as well as multiple CNN architectures.

Appendix:

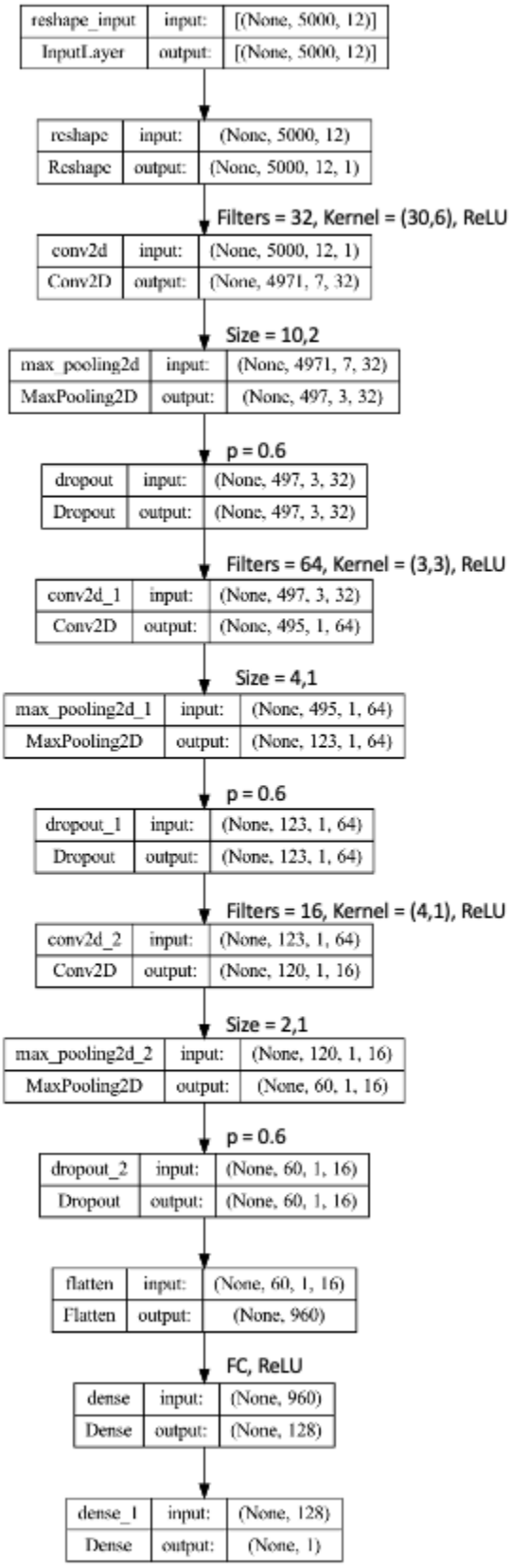

Figure 6: CNN Architecture

Sources:

- PTB-XL
  - Wagner, P., Strodthoff, N., Bousseljot, R., Samek, W., & Schaeffter, T. (2022). PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). *PhysioNet*. https://doi.org/10.13026/kfzx-aw45.
- PhysioNet
  - Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. E215–e220.
- Autoencoder image
  - https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798
- Multi-Scale ResNet CNN
  - Liu, R., Wang, F., Yang, B., & Qin, S. J. (2019). Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions. *IEEE Transactions on Industrial Informatics*, *16*(6), 3797-3806.
- Srinivasan, N. T., & Schilling, R. J. (2018). Sudden Cardiac Death and Arrhythmias. In Arrhythmia & Electrophysiology Review (Vol. 7, Issue 2, p. 111). Radcliffe Group Ltd. https://doi.org/10.15420/aer.2018:15:2
- Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., Pellikka, P. A., Enriquez-Sarano, M., Noseworthy, P. A., Munger, T. M., Asirvatham, S. J., Scott, C. G., Carter, R. E., & Friedman, P. A. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. In Nature Medicine (Vol. 25, Issue 1, pp. 70–74). Springer Science and Business Media LLC. https://doi.org/10.1038/s41591-018-0240-2
- Rafie, N., Kashou, A. H., & Noseworthy, P. A. (2021). ECG Interpretation: Clinical Relevance, Challenges, and Advances. In Hearts (Vol. 2, Issue 4, pp. 505–513). MDPI AG. https://doi.org/10.3390/hearts2040039
- Chavan, Mahesh & Agarwala, Ra & Uplane, Mahadev. (2008). Suppression Of Baseline Wander And Power Line Interference in ECG Using Digital IIR Filter. International Journal of Circuits, Systems and Signal Processing. 2.
- Wu, J.-T., Wang, S.-L., Chu, Y.-J., Long, D.-Y., Dong, J.-Z., Fan, X.-W., Yang, H.-T., Duan, H.-Y., Yan, L.-J., & Qian, P. (2017). CHADS₂ and CHA₂DS₂-VASc Scores Predict the Risk of Ischemic Stroke Outcome in Patients with Interatrial Block without Atrial Fibrillation. In Journal of Atherosclerosis and Thrombosis (Vol. 24, Issue 2, pp. 176–184). Japan Atherosclerosis Society. https://doi.org/10.5551/jat.34900
- BHALLA, V., ISAKSON, S., BHALLA, M., LIN, J., CLOPTON, P., GARDETTO, N., & MAISEL, A. (2005). Diagnostic ability of B-type natriuretic peptide and impedance cardiography: Testing to identify left ventricular dysfunction in hypertensive patients. In American Journal of Hypertension (Vol. 18, Issue 2, pp. 73–81). Oxford University Press (OUP). https://doi.org/10.1016/j.amjhyper.2004.11.044