Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022

Matthew A Reyna[1], Yashar Kiarashi[1], Andoni Elola[2], Jorge Oliveira[3], Francesco Renna[4], Annie Gu[1], Erick A Perez Alday[1], Nadi Sadr[1,5], Ashish Sharma[1], Sandra Mattos[6], Miguel T Coimbra[4], Reza Sameni[1], Ali Bahrami Rad[1], Gari D Clifford[1,7]

[1]Department of Biomedical Informatics, Emory University, USA
[2]Department of Electronic Technology, University of the Basque Country UPV/EHU, Spain
[3]REMIT, Universidade Portucalense, Portugal
[4]INESC TEC, Universidade do Porto, Portugal
[5]ResMed, Australia
[6]Unidade de Cardiologia e Medicina Fetal, Real Hospital Português, Brazil
[7]Department of Biomedical Engineering, Georgia Institute of Technology, USA

Abstract

The George B. Moody PhysioNet Challenge 2022 explored the detection of abnormal heart function from phonocardiogram (PCG) recordings.

Although ultrasound imaging is becoming more common for investigating heart defects, the PCG still has the potential to assist with rapid and low-cost screening, and the automated annotation of PCG recordings has the potential to further improve access. Therefore, for this Challenge, we asked participants to design working, open-source algorithms that use PCG recordings to identify heart murmurs and clinical outcomes.

This Challenge makes several innovations. First, we sourced 5272 PCG recordings from 1568 patients in Brazil, providing high-quality data for an underrepresented population. Second, we required the Challenge teams to submit working code for training and running their models, improving the reproducibility and reusability of the algorithms. Third, we devised a cost-based evaluation metric that reflects the costs of screening, treatment, and diagnostic errors, facilitating the development of more clinically relevant algorithms.

A total of 87 teams submitted 779 algorithms during the Challenge. These algorithms represent a diversity of approaches from both academia and industry for detecting abnormal cardiac function from PCG recordings.

1. Introduction

Heart sounds are generated by the vibrations of cardiac valves as they open and close during the cardiac cycle. Pathological cardiovascular structure or function can cause turbulent blood flow that creates audible heart sounds, and cardiac auscultation of these sounds with a stethoscope remains the most common and cost-effective tool for cardiac pre-screening. More recently, digital phonocardiography has emerged as a more sensitive and objective analog of auscultation that can detect inaudible heart sounds, quantify the sounds through physiological waveforms, and remain relatively accessible because expensive equipment and trained professionals are not required for collection [1]. Moreover, while ultrasound imaging is becoming more common for investigating heart defects, phonocardiography can still assist with rapid and low-cost screening [2]. However, experts are still needed to interpret the heart sound recordings, limiting the potential of the phonocardiogram (PCG) in cardiac care. However, the application of algorithmic methods to the PCG physiological waveforms may allow more automated and accessible heart sound analysis and diagnosis.

The 2022 George B. Moody PhysioNet Challenge (formerly the PhysioNet/Computing in Cardiology Challenge) provided an opportunity to address these issues by inviting teams to develop fully automated approaches for detecting abnormal heart function from PCG recordings. We asked teams to identify both heart murmurs and the clinical outcomes from a full diagnostic screening.

2. Methods

2.1. Challenge Data

The 2022 George B. Moody PhysioNet Challenge used the CirCor DigiScope dataset [3]. This dataset contains

 ISSN: 2325-887X DOI: 10.22489/CinC.2022.109

5272 PCG recordings from 1568 primarily pediatrics patients PCG recordings from one or more auscultation locations. It includes demographic data, annotations, and clinical outcomes from full diagnostic screenings.

The dataset was collected during two screening campaigns in the state of Paraiba, Brazil. The study protocol was approved by the 5192-Complexo Hospitalar HUOC/PROCAPE Institutional Review Board, under the request of the Real Hospital Português de Beneficiência em Pernambuco. Details of the dataset can be found in [3, 4].

The PCG recordings were recorded sequentially using an electronic auscultation device from the aortic, pulmonary, tricuspid, and/or mitral valve. A cardiac physiologist inspected the PCGs by listening to the recordings and visually inspecting the waveforms to identify the presence, absence, or unknown status of murmurs as well as various characteristics of any murmurs, including murmur location, timing, shape, pitch, quality, and grade.

During the data collection sessions, the participants answered a socio-demographic questionnaire and received a clinical examination, including chest radiography, electrocardiogram, and echocardiogram, as appropriate. Patients were either discharged with a normal clinical outcome, or directed for follow-up appointment or treatment with an abnormal clinical outcome.

We publicly released 60% of the recordings as a public training set and sequestered 10% as a hidden validation set and 30% as a hidden test set. These splits attempted to preserve the distributions of the variables and labels. Patients who were represented in the training set were not represented in the validation or test sets. The hidden validation and test sets were used to evaluate the entries of the 2022 Challenge and were not released during the Challenge.

2.2. Challenge Objective

The Challenge was designed to explore the potential for algorithmic pre-screening of abnormal heart function in resource-constrained environments. We asked the Challenge participants to design working, open-source algorithms for identifying heart murmurs and clinical outcomes from PCG recordings.

2.2.1. Challenge Timeline

This year's Challenge was the 23rd George B. Moody PhysioNet Challenge [5]. As with previous years, the Challenge had an unofficial phase and an official phase. The unofficial phase (February 1, 2022 to April 8, 2022) introduced the teams to the Challenge. We publicly shared the Challenge objective, training data, example classifiers, and evaluation metrics and invited the teams to submit their code for evaluation, scoring at most five entries from each team on the hidden validation set. Between the unofficial phase and official phase, we took a hiatus (April 9, 2022 to April 30, 2022) to improve the Challenge. The official phase (May 1, 2022 to August 15, 2022) allowed the teams to refine their approaches for the Challenge. We updated the Challenge objectives, data, example classifiers, and evaluation metric and again invited teams to submit their code for evaluation, scoring at most ten entries from each team on the hidden validation set.

After the end of the official phase, we asked each team to choose a single entry from their team for evaluation on the test set. We only evaluated one entry from each team on the test set to prevent sequential training on the test set. The winners of the Challenge were the teams with the best scores on the test set.

The winners were announced at the end of the Computing in Cardiology (CinC) 2022 conference, where the teams presented and defended their work and published four-page conference proceeding papers describing their work. Only teams that shared their work were eligible for ranking and prizes. We will publicly release the algorithms after the end of the Challenge and the publication of these papers. The full rules and expectations for the Challenge are described in [4].

2.2.2. Challenge Evaluation

To capture the focus of this year's Challenge on algorithmic pre-screening, we developed scoring metrics for each of the two Challenge tasks: detecting heart murmurs and identifying abnormal clinical outcomes from PCGs.

The murmurs are directly identified from the PCGs, but the clinical outcomes used a more comprehensive diagnostic screening, including an echocardiogram as appropriate. However, despite these differences, we asked teams to perform both tasks using only PCGs and routine demographic data so that we could explore the diagnostic potential of algorithmic approaches for interpreting PCGs.

The algorithms for both tasks effectively pre-screen patients for expert referral. If an algorithm inferred abnormal or potentially abnormal cardiac function, then it would refer the patient to a human expert for a confirmatory diagnosis and potential treatment. If the algorithm inferred normal cardiac function, then it would not refer the patient to an expert, and the patient would not receive treatment, even if the patient had abnormal cardiac function that would have been detected by the expert diagnostic screening.

For the murmur detection task, we introduced a weighted accuracy metric that assessed the ability of an algorithm to reproduce the results of a skilled human annotator. For each team and collection of patient recordings, we defined the weighted accuracy metric a_{murmur} as

$$a_{\text{murmur}} = \frac{5m_{PP} + 3m_{UU} + m_{AA}}{5\sum_i m_{iP} + 3\sum_i m_{iU} + \sum_i m_{iA}}, \quad (1)$$

		Expert		
		Present	Unknown	Absent
Model	Present	m_{PP}	m_{PU}	m_{PA}
	Unknown	m_{UP}	m_{UU}	m_{UA}
	Absent	m_{AP}	m_{AU}	m_{AA}

Table 1: Confusion matrix M for murmur detection with murmur present, murmur unknown, and murmur absent classes and the numbers of patients with each combination of expert and model labels.

		Expert	
		Abnormal	Normal
Model	Abnormal	n_{TP}	n_{FP}
	Normal	n_{FN}	n_{TN}

Table 2: Confusion matrix N for clinical outcome identification with clinical outcome abnormal and clinical outcome normal classes and the numbers of patients with each combination of expert and model labels.

where Table 1 is a confusion matrix $M = [m_{ij}]$ for the murmur present, murmur unknown, and murmur absent classes. The coefficients in (1) emphasized patients with murmurs or potential murmurs to reflect a preference for false alarms over missed treatment.

For the clinical outcome identification task, we introduced a cost-based scoring metric that reflected the cost of human diagnostic screening as well as the costs of timely, delayed, and missed treatments. For each team and collection of patient recordings, we defined the total cost metric $c_{\text{outcome}}^{\text{total}}$ as

$$\begin{aligned} c_{\text{outcome}}^{\text{total}} = {} & f_{\text{algorithm}}(n_{\text{patients}}) \\ & + f_{\text{expert}}(n_{\text{TP}} + n_{\text{FP}},\ n_{\text{patients}}) \\ & + f_{\text{treatment}}(n_{\text{TP}}) \\ & + f_{\text{error}}(n_{\text{FN}}), \end{aligned} \tag{2}$$

where $f_{\text{algorithm}}(s) = 10s$, $f_{\text{treatment}}(s) = 10000s$, and $f_{\text{error}}(s) = 50000s$ are the costs of algorithmic pre-screening, treatment, and missed or late treatment, respectively, for s individuals;

$$f_{\text{expert}}(s,t) = 25t + 397s - 1718\frac{s^2}{t} + 11296\frac{s^4}{t^3} \tag{3}$$

is the cost of expert screening for s individuals out of a cohort of t individuals; Table 2 is a confusion matrix $N = [n_{ij}]$ for the clinical outcome abnormal and normal classes; and n_{patients} is the total number of patients.

To compare costs for databases with different numbers of patients, e.g., the training, validation, and test databases, we defined the mean per-patient cost of diagnosis and treatment with algorithmic pre-screening as

$$c_{\text{outcome}} = \frac{c_{\text{outcome}}^{\text{total}}}{n_{\text{patients}}}. \tag{4}$$

We motivated and described both metrics in detail in [4]. The team with the highest weighted accuracy metric won the murmur detection task, and the team with the lowest cost-based scoring metric won the clinical outcome identification task.

3. Challenge Results

A total of 87 teams submitted 779 algorithms during the course of the Challenge, including 81 teams with 167 successful entries during the unofficial phase and 63 teams with 306 successful entries during the official phase. After the end of the official phase, we attempted to score one entry from each team with a successful official phase entry on the hidden test set; 40 teams had a successful entry for the murmur detection task on the test set for the murmur detection task and met the other Challenge criteria for rankings, while 39 teams had a successful entry for the clinical outcome identification task on the test set and met the other Challenge criteria for rankings.

Table 3 summarizes the highest-ranked teams for murmur detection task, and Table 4 summarizes the highest-ranked teams for clinical outcome identification task. All of these teams met the Challenge requirements, which are described in [4, 6]. Team summaries, additional scores, and other required and encouraged criteria, such as robust training code, are available on [6].

Rank	Team Name	Score
1	HearHeart	0.780
2	CUED_Acoustics	0.776
2	HearTech+	0.776
4	PathToMyHeart	0.771
5	CAU_UMN	0.767

Table 3: Five teams with the highest weighted accuracy metric scores (1) on the test set for the murmur detection task; higher scores are better, and only ranked teams are shown.

4. Discussion

Due to difficulty of assessing clinical outcomes from phonocardiograms (PCGs) alone, we expected that the Challenge algorithms would generally perform worse on the clinical outcome identification task than on the murmur detection task. Indeed, the algorithms performed worse across a variety of evaluation metrics on the clinical outcome identification task, e.g., the mean accuracy of the

Rank	Team Name	Score
1	CUED_Acoustics	11144
2	prna	11403
2	Melbourne_Kangas	11735
4	CeZIS	11916
5	CAU_UMN	11933

Table 4: Five teams with the lowest cost scores (4) on the test set for the clinical outcome identification task; lower scores are better, and only ranked teams are shown.

ranked algorithms on the test set dropped from 0.72 for the murmur detection task to 0.54 for the clinical outcome identification task, even though the former task had more classes.

However, despite the differences between tasks, we saw that algorithms had similar rankings on both tasks (Spearman's $\rho = 0.59$), suggesting that features and approaches that allowed algorithms to detect murmurs were also informative for identifying abnormal outcomes.

Unlike recent Challenges, the Challenge training, validation, and test sets were sourced from the same database. However, algorithm performance for the ranked algorithms still decreased from the public training set to the hidden test set by a mean of 15% across both tasks and a variety of metrics[1]; the cost metric increased by a mean of 49% across all ranked teams and a mean of 38% for the top 5 ranked teams for the clinical outcome identification task, demonstrating potential overtraining even for algorithms better able generalize to the unseen data.

5. Conclusions

This year's Challenge explored the potential for algorithmic pre-screening of abnormal heart function in resource-constrained environments. We asked the Challenge participants to design working, open-source algorithms for identifying heart murmurs and clinical outcomes from phonocardiogram (PCG) recordings; the first task was more tractable than the second task, but successful algorithms made progress for both tasks. By reducing human screening of patients with normal cardiac function, algorithms can lower healthcare costs and increase the accessibility of cardiac screening and care for patients with abnormal cardiac function in low-resourced environments.

Acknowledgements

This research is supported by the National Institute of General Medical Sciences (NIGMS) and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under NIH grant numbers 2R01GM104987-09 and R01EB030362 respectively, the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378, as well as the Gordon and Betty Moore Foundation and MathWorks under unrestricted gifts. GC has financial interests in Alivecor, LifeBell AI and Mindchild Medical. GC also holds a board position in LifeBell AI and Mindchild Medical. AE is supported by the Spanish Ministerio de Ciencia, Innovación y Universidades under Grant RTI2018-101475-BI00, jointly with the Fondo Europeo de Desarrollo Regional (FEDER), by the Basque Government under Grant IT1717-22 and by the University of the Basque Country (UPV/EHU) under Grant COLAB20/01. The work of FR and MC is financed by National Funds through the Fundação para a Ciência e a Tecnologia within project UIDB/50014/2020. None of the aforementioned entities influenced the design of the Challenge or provided data for the Challenge. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the above entities.

[1] I.e., area under the receiver-operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), F-measure, accuracy, and the weighted accuracy scoring metric.

References

[1] Vermarien H. Phonocardiography. In Webster JG (ed.), Encyclopedia of Medical Devices and Instrumentation, volume 5, 2nd edition. John Wiley & Sons, Ltd, 2006; 278–290.

[2] Viviers PL, Kirby JAH, Viljoen JT, Derman W. The diagnostic utility of computer-assisted auscultation for the early detection of cardiac murmurs of structural origin in the periodic health evaluation. Sports Health 2017;9(4):341–345.

[3] Oliveira JH, Renna F, Costa P, Nogueira D, Oliveira C, Ferreira C, et al. The CirCor DigiScope dataset: From murmur detection to murmur classification. IEEE Journal of Biomedical and Health Informatics 2021;26(6):2524–2535.

[4] Reyna MA, Kiarashi Y, Elola A, Oliveira J, Renna F, Gu A, et al. Heart murmur detection from phonocardiogram recordings: The George B. Moody PhysioNet Challenge 2022. medRxiv 2022;URL https://doi.org/10.1101/2022.08.11.22278688.

[5] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[6] Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022. https://physionetchallenges.org/2022/. Accessed: 2022-09-30.

Address for correspondence:

Matthew A Reyna; DBMI, 101 Woodruff Circle, 4th Floor East, Atlanta, GA 30322; matthew.a.reyna@emory.edu