# 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 imaging ultrasound 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 provides several innovations. First, we sourced 5272 PCG recordings from 1568 patients in Brazil, providing high-quality data for a diverse population. Second, we required the Challenge teams to submit 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, allowing us to facilitate the development of more clinically relevant algorithms.*

*A total of 89 teams submitted 780 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. Turbulent blood flow from pathological cardiovascular structure or function can create audible heart sounds. Cardiac auscultation 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 and quantify the richness of the heart sounds through physiological waveforms while remaining relatively accessible [1]. While ultrasound imaging is becoming more common for investigating heart defects, phonocariography still has the potential to assist with rapid and low-cost screening [2]. The caveat i that experts are needed to interpret the heart sound recordings to detect murmurs and identify different pathologies, limiting the potential of the phonocardiogram (PCG) as part of cardiac care. However, the application of algorithmic methods to the physiological waveforms from the PCG poses the potential for automated 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 consists

of 5272 PCG recordings from 1568 primarily pediatrics patients with one or more PCG recordings from several auscultation locations. It also includes demographic data, annotations, of the recordings, and clinical outcomes from a full diagnostic screening.

The dataset was collected during two screening campaigns of primarily pediatric patients 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 using an electronic auscultation device from as many as four prominent auscultation locations on the body: aortic valve, pulmonary valve, tricuspid valve, and/or mitral valve. The PCGs were recorded by the same operator sequentially (not simultaneously) from different locations on the patient's body. A cardiac physiologist then inspected the PCGs by listening to the audio recordings and by visually inspecting the waveforms to identify the presence, absense, or unknown status of murmurs and various characteristics of any murmurs, including murmur location, timing, shape, pitch, quality, and grade.

During the data collection sessions, the participants also answered a socio-demographic questionnaire and received a clinical examination and cardiac investigations, including chest radiography, electrocardiogram, and echocardiogram as appropriate. Patients were either discharged, directed for a follow-up appointment, or referred to cardiac catheterization or heart surgery as appropriate. The clinical outcome annotations indicate if the clinical outcome as diagnosed by the medical expert was normal or abnormal.

We publicly released 60% of the recordings as the training set and sequestered the remaining 10% as the validation set and 30% as the test set. These splits were approximately matched to preserve the distributions of the variables and labels, and 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 will only be released after the end of 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. For each patient encounter, each algorithm interprets the PCG recordings and demographic data for the patient.

### 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 infers 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 infers normal

| | | Expert | | |
|---|---|---|---|---|
| | | Present | Unknown | Absent |
| Model | Present | $m_{\text{PP}}$ | $m_{\text{PU}}$ | $m_{\text{PA}}$ |
| | Unknown | $m_{\text{UP}}$ | $m_{\text{UU}}$ | $m_{\text{UA}}$ |
| | Absent | $m_{\text{AP}}$ | $m_{\text{AU}}$ | $m_{\text{AA}}$ |

Table 1: Confusion matrix $M$ for murmur detection with three classes: murmur present, murmur unknown, and murmur absent. The entries are the numbers of patients with each combination of expert and model outputs.

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. We defined

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

where Table 1 is a confusion matrix $M = [m_{ij}]$ for the murmur present, murmur unknown, and murmur absent classes. The coefficients in (1) emphasize patients with murmurs or potential murmurs because to reflect the 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. We defined

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

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;

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

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 $n_{\text{patients}}$ is the total number of patients.

We 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.

| | | Expert | |
|---|---|---|---|
| | | Abnormal | Normal |
| Model | Abnormal | $n_{\text{TP}}$ | $n_{\text{FP}}$ |
| | Normal | $n_{\text{FN}}$ | $n_{\text{TN}}$ |

Table 2: Confusion matrix $N$ for clinical outcome detection with two classes: clinical outcome abnormal and clinical outcome normal. The entries are the numbers of patients with each combination of expert combination and model output.

## 3. Challenge Results

A total of 89 teams submitted 780 algorithms during the course of the Challenge. We will share an analysis of the Challenge results in an updated version of this manuscript after the Challenge concludes.

## 4. Discussion

We will share a discussion of the Challenge in an updated version of this manuscript after the Challenge concludes.

## 5. Conclusions

We will share conclusions about the Challenge in an updated version of this manuscript after the Challenge concludes.

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. 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 receives financial support from the Spanish Min-

isterio de Ciencia, Innovacion y Universidades through grant RTI2018-101475-BI00, jointly with the Fondo Europeo de Desarrollo Regional (FEDER), and by the Basque Government through grant IT1229-19. 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.

## References

[1] Vermarien H. Phonocardiography. John Wiley & Sons, Ltd. ISBN 9780471732877, 2006; .

[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 perinodic 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;1–1.

[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.

Address for correspondence:

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