Past, Present and Future Challenges in Sharing Science: From PhysioNet to Foundation Models

Gari D Clifford[1,2]

[1]Department of Biomedical Informatics, Emory University, USA
[2]Department of Biomedical Engineering, Georgia Institute of Technology, USA

Abstract

Over the last 25 years, the sharing of data and models for research in cardiology has evolved from sneakernet to the internet - from mailing tapes and compact discs of a handful of well-curated recordings of an array of arrhythmias, to the high-speed download of an entire hospital database. Yet, bandwidth and local computation has not kept pace with the rate at which we can strip mine our data archives. Recently, the trend towards the development of large foundational models has required enormous computing power, making large-scale local clusters or centralized cloud instances of data and compute the only viable solution for developing such models. Instead of democratizing access to data and compute, as was the intent of the pioneers in this field, this trend is leading to the balkanization of innovation in the hands of the rich and powerful. Moreover, claims of foundation models in cardiology, physiology, and medical data in general, are largely premature because foundation models must be trained on a broad enough range of data so that they may be applied across a wide range of use cases. To do so requires broad representation, of both individuals and medical conditions. This article examines these trends and provides a discussion of the most promising future directions, together with potential solutions to the concentration of power and lack of diversity in foundation models.

1. Some History

The earliest open access cardiology database, released in 1980, is the MIT-BIH arrhythmia database [1]. Assembled between 1975 and 1979, with supplementary recordings over the next 10 years, the database consisted of 48 30-minute Holter monitor recordings. These were chosen for their rare but clinically important rhythms, QRS morphology or noise/artefacts that may present challenges to ECG analysis software [2]. Importantly, every beat, artifact and event was meticulously labelled. The 600 MB of 109 thousand beats were originally distributed to approximately 100 sites around the world on digital 9-track tape [1]. The database then rapidly became a key resource for researchers and industry alike, with the FDA requiring arrhythmia analysis systems to report performance on the entire database in a standardized manner.

In 1989 the database became available as a CD-ROM, along with seven (later nine) additional ECG databases. By the end of the 1990s, approximately 400 copies of these CD-ROMs had been mailed around the world [2]. (In the late 1990s, I was lucky enough to receive one of these CDs through the mail, with a nice note from George Moody. These gold dust ECGs were fundamental to the success of my doctoral studies, as it was to many other researchers before and since.)

The contemporaneously developed American Heart Association (AHA) ECG database for ventricular arrhythmias has never been made publicly available, so I will not dive into its details and history here. While tremendously influential, it cannot compare to the far-reaching impact of the MIT-BIH DB.

1.1. PhysioNet

In 1999, the Research Resource for Complex Physiologic Signals (commonly known as 'PhysioNet', after the public outreach website), was established with the aim to stimulate current research and new investigations in the study of complex biomedical and physiologic signals [3]. Funded by the auspices of the National Institutes of Health (NIH), it was co-founded by Principal Investigators Ary L. Goldberger at Beth Israel Deaconess Medical Center's/Harvard Medical School's Margret & H. A. Rey Institute for Nonlinear Dynamics in Medicine (Rey-Lab) and Roger G. Mark at MIT's Laboratory for Computational Physiology (LCP). PhysioNet's original and ongoing missions focus on conducting and catalyzing biomedical research and education, in part by offering free access to large collections of physiological and clinical data and related open-source software. Notably, the MIT-BIH DB became easily downloadable by the click of a mouse in just minutes (in a high speed research setting).

Year	Name of Database	# of Recordings	Sampling Frequency	Av. Length	# of Leads	Geographic Location
1980	MIT DB	48	360 Hz	30 minutes	2	NE USA
1985	AHA DB (*)	80	250 Hz	30 minutes	2	Midwest USA
1996	STAFF III DB (**)	104	1000 Hz	1-9 minutes	12	Appalachia, USA
1997	QTB DB	100	360 Hz	15 minutes	2	USA + Western EU
2000	ESC DB	90	250 Hz	2 hours	2	Western Europe
2004	PTB DB	549	1000 Hz	10-100 seconds	15	Germany
2008	THEW	3,000	180-1000 Hz	24 hours	3-12	US, Europe, and South America
2018	CPSC	$>$13,000	500 Hz	6-144 seconds	12	China
2020	SaMi-Trop	$\approx$ 2,000	400 Hz	10 seconds	12	Brazil
2020	GA ECG DB	$\approx$ 21,000	500 Hz	5-10 seconds	12	SE USA
2020	UMich DB	20,000	250-500 Hz	10 seconds	12	NE USA
2020	CU DB	45,152	500 Hz	10-100 seconds	12	SW US and China
2020	CODE 15%	$\approx$ 200,000	400 Hz	10 seconds	12	Brazil
2022	PTB-XL DB	18,869	500 Hz	10-100 seconds	12	Germany
2023	SPH DB	$>$ 24,666	200 Hz	24-hr	12	USA
2023	MIMIC-IV	800,000	500 Hz	10 seconds	12	USA
2023	H-E ECG DB	$>$10 million	500 Hz	10 seconds	12	USA

Table 1: Key open ECG databases in chronological order. * Not public. ** Not available on PhysioNet until 2017

PhysioNet was remarkable, not just because it offered an ever-growing collection of high quality labelled data, but also high quality documentation, and high quality software for analyzing and evaluating many aspects of the ECG (and related biosignals).

Since 1999, many databases have begun to appear, increasing in size over time. Table 1 summarizes the key open access ECG databases in chronological order, with a comparison of number of individuals, recording length, number of leads and the geographic origin of their subjects. It's unsurprising that many of these became available through the PhysioNet Resource.

Of course, as databases increased in size, our ability to label every beat and event has ebbed away. With the latest database of over 10 million ECGs, it is clearly impractical and cost-prohibitive to label the data using experts. Instead, we are increasingly reliant on the weaker labels of algorithms. While this might be concerning, creating a biased set of labels that limit performance to the last generation of algorithms, modern machine learning has provided ways to mitigate these issues, and leverage the information in the broad range of data, as I explore in section 2.

1.2. The PhysioNet Challenges

With the inception of PhysioNet came the PhysioNet Challenges [4]. In cooperation with the Computing in Cardiology conference, PhysioNet hosted an annual series of biomedical 'Challenges', focusing research on unsolved clinical or engineering problems in clinical and basic science. In 2022 the Challenges were renamed 'the George B. Moody PhysioNet Challenges' in honor of George's key role in inventing and running the Challenges over a 15 year period.

The Challenges have been critical to spurring the development of novel databases, ranging from novel handheld ECGs focused on identifying atrial fibrillation to multimodal waveform databases for predicting sepsis, or joint ECG image-signal databases for digitizing paper ECGs.

In addition to the rich new data these Challenges generate and the focus on a particular unsolved problem, they also introduce novel metrics for evaluating the solutions [5], and of course, a range of approaches to the problem. It is not the 'winner' that matters, as the top 5 or so algorithms are usually highly similar in performance, but rather the range of methods used to solve the problem that matter. The result is an evolution in databases that was started by PhysioNet, where the code and data are two parts of a whole. They show others how to use the data, and they discover errors in the data (and metrics) as they attempt to solve the problem over a 6+ month period. The PhysioNet Challenge has been particularly innovative in ensuring that the models built on the data are fully reproducible, even down to requiring them to be fully retrainable.

2. Recent Developments using 'AI'

The confluence of high performance computing, deep learning and large ECG databases (hundreds of thousands to millions of recordings) has led to a gold rush in the field, with endless claims of 'cardiologist-level' diagnostics abilities, and prescient predictive abilities. There are too many to cite, and singling out any single article would be an injustice, since there is no single work that seems to justify the hyperbolic claims. Even more recently, the successes of large language models (like ChatGPT) has spawned a trend for foundation models in health data.

There is a natural gold rush and groups are branding large models as foundation models, so I want to be clear about the definitions here. IBM defines them as:
'*Models that are trained on a broad set of unlabeled data that can be used for different tasks, with minimal fine-*

tuning ... [and] using self-supervised learning and transfer learning, the model can apply information it's learnt about one situation to another.' [6]

Weak foundation model: *A machine learning or deep learning model that is trained on broad data such that it can be applied across a wide range of use cases, synonymous with the term large 'AI model'* [7].

Strong foundation models: *A 'general-purpose AI' or 'GPAI' system. These are capable of a range of general tasks (such as text synthesis, image manipulation and audio generation. Notable examples are OpenAI's GPT-3 and GPT-4, foundation models that underpin the conversational chat agent ChatGPT* [8].

(Here I use 'weak' and 'strong' not to mean 'bad' or 'good', but rather 'relaxed' or 'stringent' definitions, as in weak and strong stationarity.) In general, I don't think anyone is even close to a strong foundation model in cardiology. But even with the weak model, we are lacking a key additional condition for a model to be thought of as a foundation model.

There have been a few attempts to use clinical data (particularly time series data) to make a 'foundation model' but nothing is particularly convincing just yet. To qualify for even a weak foundation model, it has to substantially outperform training on the available domain data, using a broad range of data, and, leveraging unsupervised pretraining and transfer learning.

2.1. Foundation and Empire - the New Colonialism?

Claims of foundation models in cardiology, physiology, and medical data in general, are largely premature because foundation models must be trained on a broad enough range of data so that they may be applied across a wide range of use cases. To do so requires a broad representation of both individuals and medical conditions. We might argue that the very latest datasets [9, 10] draw on a diverse populations in the USA and Brazil, but there are many social and structural factors that mean that Africa, Asia, and Latin America are still vastly under-represented in our databases. In addition, the researchers working on these data are primarily drawn from Europe and North America, entrenching the neocolonialism of the AI revolution.

Another key issue we face in developing Foundation models is the computing power required. Large-scale local clusters or centralized cloud instances of data and compute have become the only viable solution for developing large models. Instead of democratizing access to data, as was the intent of the pioneers in this field (such as Mark, Moody and Goldberger), this trend is leading to the balkanization of innovation in the hands of the rich and powerful. Moreover, even if access is made free (as in 'free beer', in addition to 'open'), the access becomes meaningless without the finances to run large-scale computing on the data.

2.2. The Carbon Footprint of Success

The elephant in the room with foundation models is their carbon footprint. Training GPT-3 generated 552 ton of CO_2 (tCO_e) and ChatGPT has more than 13 million users per day with 5 questions each, so 65 million queries would generate 30 tCO_e per day or 0.5 gCO_{2e} per query [11]. GTP4 cost 10 to 100 times as much to train ($100 million), and although training has become more efficient, the carbon footprint is likely to still be an order of magnitude larger [12].

We cannot close Pandora's box, and perhaps we don't want to, but we do need to be more thoughtful about how and when we train large models on our vast quantities of data, how and when the models are used, and how data are collected. The developments in distributed learning and edge computing hold enormous promise for energy efficiency, privacy preservation, and the empowerment of healthcare workers in the most remote regions, where AI has the potential to have the most impact.

2.3. What about Interpretability?

There's an argument against complex and deep models - that we'll never understand how the classification or predictions are made. I think this is probably overblown. I am not sure if there's a single person who can explain how any commercial cardiology algorithm works. They have been modified countless times by so many individuals, that there's no full understanding left. The only code I can really explain is the QRS detector and neural network I wrote from scratch in C during my PhD. But that was just me working on that, and I had to think about every operation and memory allocation. Every library I used from 'Numerical Recipes in C', I spent time to understand how they worked, line by line. After that, I mostly wrote in higher level languages using many inbuilt languages, and rarely wrote code alone, and mostly in higher level languages. These days, I'm horrified at how little thought goes into the use of Python libraries with default settings. I am finally getting old!

3. The Future of ECG Analysis

It's obvious that the trend for large and foundation models isn't going to stop any time soon, and I strongly believe we'll have a useful generalizable model in short order. But there will be some false starts, and we will need to call them out, and force their creators/owners to fix them.

Hopefully the usual interplay between industry and open academia will enable this. Industry is already beginning to

rush in, with bold and brash startups promising a lot. This is the roller coaster society has built for itself. Whether we like it or want it, it's hard to avoid. I, for one, am up for the ride. I fell in love with the promise of neural networks in the mid 1990s, and I smacked my head on small databases, slow processors and tiny memory footprints, reluctantly learning feature engineering, and kernel tricks while embracing the beauty of electrical engineering. I could taste the promise, but was 25 years away.

However, we're still not 'there', and we have an imperative to collect more diverse datasets. Significant holes in public data that reflect the diversity of humanity still exist, with sparse sampling of Asian and Latin American populations, and virtually no representation from Africa. The Global Majority rarely appears in our algorithms.

Just as importantly, we need to build more diverse teams and promote Global Majority researchers, and governments need to ensure that we don't enact a new age of colonialism. I.e., we need to ensure that the Global South controls their data for commercialization, so that they directly and substantially benefit from potential profits.

Finally, we need to be more thoughtful about how and when we train large models on our vast quantities of data, how and when the models are used, and how data are collected. The developments in distributed learning and edge computing hold enormous promise for energy efficiency, privacy preservation, and the empowerment of healthcare workers in the most remote regions, where AI has the potential to have the most impact.

I'll end with a counterpoint. Multimodal databases enable multimodal foundation models, which are becoming more and more powerful. We will rapidly move beyond text plus ECG, to add ultrasound, and other imaging, social determinants of health, and many other medical or meta data. Eventually this will include non-medical data such as emotional expressions during daily living, location, movement, and social interactions. These will take enormous efforts, and the costs may turn out to be prohibitive, until the next revolution in algorithmic efficiency, of course.

Perhaps the most exciting thing to see these days is the reinvention of signal processing in the form of data-driven machine learning. It may be a cliché for an old person to bemoan that the new generation are just rediscovering all the things we already know. However, there is a justifiable concern that the lack of understanding of sampling theory, aliasing, and transformations into other domains (such as Fourier) will produce inefficient and even dangerous algorithms. But this is the journey of discovery that every new generation must take, and with each generation, we make a new future. Hopefully the great researchers will connect it with the past, though. As a great 20th Century philosopher once said, 'in this bright future you can't forget your past' [13] ... and we shouldn't.

Acknowledgements

This research is supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB: R01EB030362). The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or PhysioNet.

References

[1] Moody G, Mark R. The MIT-BIH Arrhythmia Database on CD-ROM and software for use with it. In [Proceedings Computers in Cardiology, volume 17. 1990; 185–188.

[2] Moody G, Mark R. The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine 2001;20(3):45–50.

[3] 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.

[4] Clifford GD, Reyna MA. The George B. Moody PhysioNet Challenges. Online. URL https://physionetchallenges.org. Accessed: 2024/08/21.

[5] Reyna MA, Nsoesie EO, Clifford GD. Rethinking algorithm performance metrics for artificial intelligence in diagnostic medicine. JAMA ;328:329–330. Publisher: American Medical Association.

[6] Murphy M. What are foundation models? Online, 09 May 2022. URL https://research.ibm.com/blog/what-are-foundation-models. Accessed: 2024/08/21.

[7] CMA. AI Foundation Models: Initial Report, 18 Sept. 2023. URL https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf. Competition and Markets Authority, UK Government. Accessed: 2024/08/21.

[8] Jones E. Foundation models: An explainer. Online. URL https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/. Ada Lovelace Institute Accessed: 2024/08/21.

[9] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. Nature Communications 2020;11:1760.

[10] Moura Junior V, Reyna M, Hong S, Gupta A, Ghanta M, Sameni R, et al. Harvard-emory ecg database, 2023.

[11] de Vries A. The growing energy footprint of artificial intelligence. Joule October 2023;7(10):2191–2194.

[12] Patterson D, Gonzalez J, Le Q, Liang C, Munguia LM, Rothchild D, et al. Carbon Emissions and Large Neural Network Training, 2021. URL https://arxiv.org/abs/2104.10350.

[13] Marley B. Bob Marley - Natural Mystic. Hal Leonard Corporation, 1996. ISBN 0793551528. First edition.

Address for correspondence:

101 Woodruff Circle, Atlanta, GA gari@physionet.org