Publicly Available Raw Multimodal Datasets (EEG, ECG, PPG, and Imaging) for Cardiovascular Web Applications

1. Executive Summary

This report presents the findings of a comprehensive research effort to identify publicly available, raw, multimodal datasets suitable for web application development, with a specific emphasis on data from wearable sensors (EEG, ECG, PPG) and imaging, particularly in the context of cardiovascular conditions. The investigation reveals a challenging landscape: while numerous medical datasets are publicly accessible, very few simultaneously offer raw EEG, ECG, PPG, *and* linkable imaging data for the same patient cohorts, coupled with licensing terms amenable to web application deployment.

Key findings indicate that the **MIMIC-IV** database, with its various modules (Clinical, Waveform, ECG, and its link to MIMIC-CXR for imaging), stands out as one of the most promising, albeit complex, resources. It offers extensive clinical information, raw ECG and PPG waveforms, diagnostic ECGs, and linkable chest X-ray imaging, primarily from an intensive care setting with a strong cardiovascular component. However, widespread, linkable raw EEG data for these ICU patients is not a standard component. Access requires credentials and adherence to a strict data use agreement.

PhysioNet hosts several relevant datasets. **MC-MED** provides multimodal data from emergency department visits, including ECG and PPG waveforms and links to imaging results, but the availability of raw imaging files and EEG needs careful verification. The "Multimodal Dataset for Investigating Working Memory in Presence of Music" offers ECG, PPG, and fNIRS (a form of neuroimaging) with a permissive license, but its cardiovascular relevance is indirect, and it lacks EEG.

Dedicated imaging datasets like **CMRxRecon2024** (cardiac MRI k-space) and **EchoNet-Dynamic** (echocardiogram videos) offer excellent raw cardiovascular imaging but typically lack concurrently recorded raw EEG, ECG, and PPG from the same subjects. Furthermore, their licenses can be restrictive for commercial web applications.

The primary challenge lies in the intersection of all requirements: the specific combination of raw sensor and imaging modalities for the same individuals, a clear cardiovascular focus, and permissive licensing. EEG data, in particular, is less commonly integrated with comprehensive cardiovascular datasets that also include

ECG, PPG, and traditional cardiovascular imaging.

Recommendations focus on a pragmatic approach: prioritizing datasets like MIMIC-IV that meet most criteria, carefully evaluating licensing terms, and acknowledging potential data gaps (especially for EEG). Strategies for addressing these gaps may involve searching for specialized sub-studies, considering datasets that offer a subset of the required modalities, or being prepared for complex data integration efforts if multiple sources are used. Developers must also consider the significant technical and ethical responsibilities associated with handling raw, sensitive medical data in a web application.

2. Introduction

Purpose of the Report
This report details the outcomes of an in-depth research search for publicly available, raw, multimodal datasets suitable for integration into web applications. The specific criteria for these datasets include the presence of data from wearable sensors—namely electroencephalography (EEG), electrocardiography (ECG), and photoplethysmography (PPG)—alongside imaging data. A particular focus is placed on datasets relevant to cardiovascular conditions.
Importance of Multimodal Data
The integration of diverse data types, such as physiological signals, medical imaging, and clinical information, holds immense scientific and clinical value. Multimodal datasets enable a more holistic understanding of human health and disease progression. In complex areas like cardiovascular medicine, combining information from ECG (heart electrical activity), PPG (blood volume pulse), EEG (brain activity, which can be relevant in conditions affecting autonomic function or in cerebrovascular comorbidities), and various imaging modalities (MRI, CT, echocardiography) can lead to more accurate diagnostic tools, personalized treatment strategies, and a deeper comprehension of pathophysiological mechanisms.1 The development of advanced AI models often relies on such rich, multifaceted data to uncover subtle patterns and correlations that might be missed by unimodal analysis.3
Challenges in Sourcing Suitable Datasets
Sourcing datasets that meet all the specified criteria—comprehensive multimodality for the same subjects, truly "raw" data formats, permissive licensing suitable for web applications, and the specific sensor and imaging combinations—presents significant challenges. Medical data is inherently sensitive, and its collection, curation, and sharing are governed by strict ethical and privacy regulations. While many repositories offer valuable data, they often specialize in either physiological signals or imaging. Finding datasets where EEG, ECG, PPG, and imaging data are all collected from the same individuals, provided in a raw or minimally processed state, and licensed for broad use (including potential web application development) is a non-trivial task. For instance, some rich image databases like the original MedPix® do not allow access to raw data, limiting their utility for training certain AI systems.5

Scope and Structure of the Report

This report aims to navigate these complexities. It begins by defining the key criteria used for dataset selection, providing clarity on terms like "raw data" in different contexts and the importance of data linkage and licensing. It then surveys prominent repositories and individual datasets that show potential for meeting the user's requirements. The most promising candidates are analyzed in greater depth, focusing on their specific modalities, cardiovascular relevance, data accessibility, and licensing terms. The report includes comparative tables to aid in decision-making and concludes with strategic recommendations for dataset integration into a web application, acknowledging the current limitations and potential future directions in the field of multimodal medical data sharing.

3. Key Considerations for Dataset Selection

The evaluation and selection of suitable datasets for a web application integrating raw EEG, ECG, PPG, and imaging data, with a cardiovascular focus, necessitates a clear framework. This section outlines the critical factors considered in this report.

Defining "Raw" Data in Context

The term "raw" data is pivotal but can vary in its interpretation across different modalities. For wearable sensor data (EEG, ECG, PPG), "raw" generally implies access to unprocessed or minimally processed time-series signals. This could mean raw voltage readings directly from the analog-to-digital converters, high sampling rates preserving the original signal fidelity, and minimal application of filters or artifact removal algorithms. This level of rawness provides maximum flexibility for developing and testing novel signal processing pipelines within a web application.

For imaging data, "raw" can also take several forms. For clinical imaging modalities like Computed Tomography (CT), Magnetic Resonance Imaging (MRI), or X-ray, this usually refers to Digital Imaging and Communications in Medicine (DICOM) files, which contain rich metadata alongside pixel/voxel data. For MRI, an even more fundamental form of raw data is k-space data, which represents the frequency-domain information acquired by the scanner before image reconstruction; datasets like CMRxRecon2024 provide this.6 For ultrasound, raw data might be the direct video feeds from the transducer, as seen in datasets like EchoNet-Dynamic.7

It is important to recognize that "raw" is not an absolute state. Many datasets labeled as "raw" may have undergone some level of preprocessing. Common steps include anonymization (which might involve stripping certain metadata fields to protect patient privacy), resampling to a common frequency (as seen in the BrainBeats sample data, which was resampled to 250 Hz [8]), or initial quality control to remove entirely unusable segments. While the original MedPix® database is a rich source of images and textual data, access to the raw image data is not possible, which limits its use for training multimodal AI systems requiring such input.[5] The ideal "raw" dataset for a web application allows developers to implement their own preprocessing tailored

to their specific algorithms and use cases.

Multimodal Data Linkage (Intra-subject Sensor and Imaging Data)
A fundamental requirement is that the datasets provide EEG, ECG, PPG, and imaging data collected from the same individuals or patient cohorts, with clear mechanisms to link these different data types for each subject. This intra-subject linkage is crucial for any meaningful multimodal analysis or application development. However, this represents a significant challenge in the current public data landscape. Many repositories tend to specialize, for example, PhysioNet is a well-known resource for physiological signals 9, while The Cancer Imaging Archive (TCIA) primarily hosts medical images related to cancer.10 Finding datasets that inherently bridge these specializations, providing all specified modalities for the same participants, is uncommon. Often, researchers must undertake complex data integration efforts themselves, which can be a substantial barrier, especially for initial web application development. Frameworks have been proposed to integrate cardiac images and ECG signals, for instance 1, but these often describe methodologies applied to datasets that may not be publicly available in their entirety or in the required raw format. This report prioritizes datasets where such linkage is explicit and the data is accessible.
Sensor Data Specifics (EEG, ECG, PPG)
For the specified sensor data, certain characteristics are important:

- **EEG:** Number of channels (e.g., full 10-20 system or fewer), sampling rate, type of electrodes, and recording conditions.
- **ECG:** Number of leads (e.g., single-lead, 3-lead, 12-lead diagnostic), sampling rate, and duration of recordings (e.g., short 10-second diagnostic strips or continuous monitoring).
- **PPG:** Source of the PPG signal (e.g., fingertip, wrist), sampling rate, and if available, information about the light wavelengths used, as this can influence signal quality and interpretation.

Imaging Data Modalities
The choice of imaging modality often depends on the specific cardiovascular condition being investigated. Common relevant types include:

- **Cardiac MRI (CMR):** Provides detailed information on cardiac structure, function, perfusion, and tissue characterization.[6]
- **Computed Tomography (CT):** Useful for coronary artery calcium scoring, angiography, and structural assessment.
- **Echocardiography:** Widely used for assessing cardiac function, valvular heart disease, and chamber sizes using ultrasound.[7]
- **Chest X-ray (CXR):** Can provide general information about heart size and shape, and detect pulmonary congestion. Other imaging modalities, such as functional Near-Infrared Spectroscopy (fNIRS), a form of neuroimaging, might be present in some multimodal datasets that also include the target physiological signals.[11] The

relevance of such non-cardiac specific imaging would depend on the web application's scope.

Cardiovascular Relevance
The datasets should ideally have a clear connection to cardiovascular health or disease. This could manifest as:

- Cohorts of patients with specific cardiovascular diagnoses (e.g., myocardial infarction, heart failure, arrhythmias).
- Healthy control subjects who have undergone comprehensive cardiac assessments.
- Longitudinal data that includes information on cardiovascular risk factors or outcomes. It is worth noting that some data sources, like those found on Data.gov, may provide valuable epidemiological or mortality data related to cardiovascular disease [12], but these typically do not contain the raw sensor and imaging data required for the type of web application envisioned.

Public Accessibility and Licensing for Web Applications
For a dataset to be viable for a web application, it must be publicly accessible, and its licensing terms must permit the intended use. This is a critical hurdle, especially if the web application has potential commercial aspects or involves redistribution of data or derived models.
Common license types include:

- **Creative Commons (CC) licenses:** CC0 (public domain dedication), CC BY (attribution required), CC BY-SA (attribution, share-alike). Generally, CC BY is quite permissive for web applications. TCIA, for example, uses CC BY licenses for many of its collections.[10]
- **Open Data Commons (ODC) licenses:** ODC-BY (attribution required), ODbL (Open Database License). PhysioNet often employs licenses like ODC-BY.[13]
- **"Research Use Only" licenses:** These licenses, common for sensitive medical data, often explicitly prohibit commercial use and may restrict redistribution. The EchoNet-Dynamic dataset, for instance, has such a restrictive license.[7]
- **Specific Data Use Agreements (DUAs):** Large clinical databases like MIMIC-IV require users to complete ethics training and sign a DUA that outlines permissible uses and data protection responsibilities.[14] These DUAs typically prohibit attempts to re-identify patients.

The implications for a web application are significant. A restrictive license can render an otherwise excellent dataset unsuitable if the application involves commercialization, hosts the data directly, or distributes models trained on the data in a way that violates the license terms. Clarity on these terms is paramount.

Data Formats and API Availability
The format of the data and the ease of access are practical considerations for web application development.

- **Common formats for physiological signals:** CSV, WFDB (WaveForm DataBase format, common on PhysioNet), EDF (European Data Format for EEG).
- **Common formats for imaging data:** DICOM, NIfTI (Neuroimaging Informatics Technology Initiative),.mat (MATLAB format, used in CMRxRecon2024 [6]). Standardized formats reduce the complexity of data parsing and integration. The availability of Application Programming Interfaces (APIs) for programmatic data download or querying can significantly simplify integration into a web application's backend. For example, TCIA provides a REST API for accessing its data [10], and OpenNeuro datasets can be accessed via DataLad.[16] Even if direct APIs are not available, the presence of well-documented download procedures, accompanying software tools (like the BrainBeats EEGLAB plugin for processing EEG/ECG/PPG [8] or the GitHub repository for CMRxRecon2024 [6]), or data dictionaries can lower the barrier to entry and facilitate use.

4. Survey of Promising Multimodal Datasets and Repositories

This section surveys repositories and specific datasets that hold potential for meeting the user's requirements for raw EEG, ECG, PPG, and imaging data, with a cardiovascular focus. They are categorized based on their primary strengths relative to the query.

Category A: Datasets with Strong Integrated Physiological Signals and Clinical/Imaging Data

- **MIMIC-IV (Medical Information Mart for Intensive Care IV)**
  - **Description:** MIMIC-IV is a large, comprehensive, and freely accessible de-identified health database sourced from patients admitted to intensive care units (ICUs) at the Beth Israel Deaconess Medical Center (BIDMC).[14] It is designed with a modular structure, allowing researchers to link data across different components.[14]
  - **Relevant Modalities:**
    - **Clinical Data:** Extensive structured and semi-structured data including patient demographics, diagnoses (ICD codes), procedures, medication administrations, laboratory measurements, and de-identified free-text clinical notes (e.g., discharge summaries, radiology reports).[14] This provides rich contextual information, highly relevant for cardiovascular research.
    - **MIMIC-IV Waveform Database:** This crucial module contains

high-resolution physiological signals recorded from bedside monitors. Signals include ECG, PPG (often from pulse oximeters), continuous invasive blood pressure, and respiration, among others.[17] Data is typically available in the WFDB format. An initial release contained 200 records, with a larger release of around 10,000 records planned.[17]

- **MIMIC-IV-ECG Module:** This module provides approximately 800,000 diagnostic 12-lead ECGs, each 10 seconds in length and sampled at 500 Hz, from nearly 160,000 unique patients.[18] These ECGs can be linked to the broader MIMIC-IV clinical database and, where available, to cardiologist reports.[4]
- **Imaging Link (MIMIC-CXR):** While technically a separate dataset, MIMIC-CXR contains over 370,000 chest X-rays (CXRs) in DICOM format, associated with patients from the MIMIC-IV clinical database.[5] Many patients in MIMIC-IV also have corresponding CXRs in MIMIC-CXR, enabling linkage between clinical data, physiological signals, and imaging. The CXR Foundation model, for instance, was trained using MIMIC-CXR among other datasets.[19]

- **Raw Data:** Yes. Waveform data (ECG, PPG) is provided in raw or near-raw format (WFDB). Diagnostic ECGs are available as raw signals. Clinical data is structured but represents source information. Chest X-rays from MIMIC-CXR are in DICOM format.
- **Cardiovascular Focus:** Extensive. Given its origin in ICUs, a significant proportion of patients have primary cardiovascular conditions or develop cardiovascular complications. Standard ICU monitoring heavily involves cardiovascular parameters.
- **Access & Licensing:** Access to MIMIC-IV (including its waveform and ECG modules) requires credentialing. Users must complete recognized human subjects research training (e.g., CITI Program) and sign a data use agreement (DUA). The DUA emphasizes data privacy and strictly prohibits any attempts to re-identify patients.[14] While suitable for research, any web application utilizing MIMIC-IV data must ensure rigorous compliance with these terms, particularly concerning data security, de-identification of any displayed results, and user access controls.
- **Considerations:** MIMIC-IV is a powerful resource for ECG, PPG, and linkable CXR imaging, all within a rich clinical cardiovascular context. However, raw EEG data is not a standard, universally available component for all patients within MIMIC-IV. While some ICU patients may have EEG monitoring, its systematic availability and linkage within the main MIMIC-IV release for cardiovascular cohorts would need specific investigation, possibly through

associated research projects or specialized data requests if feasible. This potential gap in EEG data is a key consideration if all four modalities (EEG, ECG, PPG, Imaging) are strict requirements for every subject in the target dataset.

- PhysioNet Ecosystem
  PhysioNet is a repository managed by the MIT Laboratory for Computational Physiology, offering free access to a wide range of physiological signals and related data, often accompanied by open-source software for their analysis.9
  - **MC-MED (Multimodal Clinical Monitoring in the Emergency Department)**
    - **Description:** MC-MED is a comprehensive, multimodal, de-identified dataset from 118,385 adult emergency department (ED) visits to monitored beds at the Stanford adult ED between September 2020 and September 2022.[20]
    - **Relevant Modalities:** The dataset includes continuously monitored vital signs, physiological waveforms (ECG, PPG, and respiration), patient demographics, medical histories, home medications, orders placed and medications administered during the visit, laboratory and imaging results (including free-text radiology reports), diagnoses, visit disposition, and length of stay.[20]
    - **Raw Data:** Continuously monitored waveforms (ECG, PPG, respiration) are included. The term "imaging results" and "free-text radiology reports" suggests that while information about imaging studies is present, direct access to the raw imaging files (e.g., DICOMs) within this specific dataset release needs confirmation from the dataset's documentation on PhysioNet.
    - **Cardiovascular Focus:** High. EDs are frontline facilities for managing acute cardiovascular events (e.g., myocardial infarction, arrhythmias, heart failure exacerbations), making this dataset highly relevant. The dataset also uniquely covers the period during and after the peak of the COVID-19 pandemic.[20]
    - **Access & Licensing:** As a PhysioNet dataset, access typically involves agreeing to a data use agreement. The specific license for MC-MED would be detailed on its PhysioNet project page and would likely require credentialing, similar to MIMIC.
    - **Considerations:** MC-MED explicitly offers multimodal data, including the desired ECG and PPG waveforms, and links to imaging information within an ED context. The primary uncertainty is whether "imaging results" encompasses access to raw image files or is limited to reports. Additionally, EEG data is not mentioned as a component of MC-MED.[20]

- **"A Multimodal Dataset for Investigating Working Memory in Presence of Music"**
  - **Description:** This dataset was created to investigate the effects of music on working memory and cognitive arousal, recording a variety of multimodal physiological signals and behavioral data.[11]
  - **Relevant Modalities:** The dataset includes skin conductance (SC), electrocardiogram (ECG), skin surface temperature (SKT), respiration, photoplethysmography (PPG), functional near-infrared spectroscopy (fNIRS), electromyogram (EMG), de-identified facial expression scores, and behavioral data (task responses, reaction times).[11] fNIRS is a non-invasive optical neuroimaging technique that measures changes in hemoglobin concentrations in the brain.
  - **Raw Data:** Yes, the physiological signals are provided in raw format. For example, data recorded via Biopac systems have a sampling frequency of 2 kHz and are available in CSV format.[11] fNIRS data would also be in its raw or minimally processed form.
  - **Cardiovascular Focus:** Low, as the primary focus is on cognitive neuroscience. However, the inclusion of raw ECG and PPG signals makes it partially relevant if these specific signals are of interest, irrespective of a direct cardiovascular disease context.
  - **Access & Licensing:** The dataset is available on PhysioNet.[11] The license is the Open Data Commons Attribution License v1.0 (ODC-BY 1.0).[13] This license is generally permissive and allows for commercial use, modification, and distribution, provided attribution is given. This is favorable for web application development.
  - **Considerations:** This dataset provides raw ECG, PPG, and a form of imaging (fNIRS), along with a permissive license. However, it does not include EEG. The imaging modality (fNIRS) is neuroimaging, not traditional cardiovascular imaging (like cardiac MRI or echocardiography). The sample size is noted as small in one of the descriptions [11], which might limit its utility for training robust deep learning models. Its cardiovascular relevance is indirect, stemming only from the presence of ECG/PPG signals.

Category B: Datasets with Strong Imaging (especially Cardiovascular) and Potential Sensor Links

- **CMRxRecon2024**
  - **Description:** This is currently the largest and most protocol-diverse publicly available k-space dataset specifically for cardiac MRI reconstruction. It

includes data from 330 healthy volunteers and covers multiple commonly used clinical protocols, modalities (cine, T1 and T2 mapping, tagging, phase-contrast/2D flow, black-blood imaging), and anatomic views.[6] It was released as part of a challenge to advance cardiac MRI reconstruction techniques.[22]
    - **Relevant Modalities:** Cardiac MRI (raw k-space data and reconstructed images).
    - **Raw Data:** Yes, the dataset provides raw multicoil k-space data, typically in.mat (MATLAB) format, which is a very fundamental form of MRI data before image formation.[6]
    - **Cardiovascular Focus:** Explicitly and exclusively cardiac.
    - **Access & Licensing:** The dataset can be downloaded from the Synapse repository. It is openly accessible for educational and research purposes. However, commercial use of the dataset *itself* is prohibited. The use of the dataset for developing, testing, or refining software or algorithms for academic research is not restricted.[6] The associated arXiv pre-print mentions an "arXiv.org perpetual non-exclusive license".[22]
    - **Considerations:** CMRxRecon2024 offers state-of-the-art raw cardiac MRI data, invaluable for research in cardiac imaging and AI. However, there is no mention of concurrently recorded EEG, ECG, or PPG signals for these 330 subjects within this dataset. The licensing term "commercial use of the dataset itself is prohibited" needs careful interpretation for a web application that might be commercial. If the application uses models trained on this data, or displays results derived from it, legal counsel might be needed to clarify if this constitutes prohibited commercial use of the dataset.
- **EchoNet-Dynamic**
    - **Description:** A large dataset consisting of 10,030 labeled echocardiogram videos from unique patients, accompanied by human expert annotations such as ejection fraction, left ventricular volumes at end-systole and end-diastole, and tracings of the left ventricle.[7]
    - **Relevant Modalities:** Echocardiogram videos (cardiac ultrasound).
    - **Raw Data:** The video data itself (e.g., in.avi or similar format) is provided, representing the raw output of the ultrasound examination.
    - **Cardiovascular Focus:** Explicitly and exclusively cardiac.
    - **Access & Licensing:** Access requires registration via the Stanford AIMI Center.[7] The dataset is governed by the "Stanford University School of Medicine EchoNet-Dynamic Dataset Research Use Agreement." This agreement strictly limits use to personal, non-commercial research purposes only. Any commercial use, sale, or other monetization is explicitly prohibited.[7]

Redistribution and creation of derivative works are also heavily restricted.
    - **Considerations:** EchoNet-Dynamic is a valuable resource for cardiac ultrasound video analysis due to its size and expert annotations. However, it does not include EEG, ECG, or PPG data for these subjects. More critically for web application development, its license is highly restrictive and generally unsuitable for any application with potential commercial aspects or broader distribution.
- **The Cancer Imaging Archive (TCIA)**
    - **Description:** TCIA is a large, publicly accessible archive that de-identifies and hosts medical images, primarily related to cancer research. Data is organized into "collections," often grouped by disease, image modality (MRI, CT, digital histopathology, etc.), or research focus. The primary file format is DICOM. Supporting data, such as patient outcomes, treatment details, genomics, and expert analyses, are also provided when available.[10]
    - **Relevant Modalities:** A wide variety of imaging modalities, including CT, MRI, PET, and others.
    - **Raw Data:** Yes, imaging data is typically provided in DICOM format.
    - **Cardiovascular Focus:** Not the primary focus. However, some collections might contain imaging relevant to cardiovascular structures (e.g., chest CTs for lung cancer screening will also image the heart and major vessels) or involve patients with cardiovascular comorbidities. A targeted search within TCIA's collections would be necessary to identify such datasets.
    - **Access & Licensing:** Most TCIA datasets are publicly downloadable. The majority are available under Creative Commons Attribution 3.0 Unported (CC BY 3.0) or 4.0 International (CC BY 4.0) licenses, which permit commercial use, modification, and distribution with appropriate attribution.[10] Some collections may have different licenses or require registration, so individual collection details should always be checked.
    - **Considerations:** TCIA offers a vast repository of imaging data with generally permissive licenses, which is attractive for web application development. The main challenge is the lack of integrated, raw EEG, ECG, or PPG signals for the same subjects within these imaging collections. The cardiovascular relevance would depend on finding specific collections that happen to align with this interest.

Category C: Other Repositories and Potential (but less direct-fit) Sources

- **OpenNeuro**
    - **Description:** A free and open platform for validating and sharing neuroimaging data, compliant with the Brain Imaging Data Structure (BIDS)

standard. It primarily hosts MRI, PET, MEG, EEG, and iEEG data.[16]
    - **Relevant Modalities:** Strong focus on EEG and various neuroimaging modalities (MRI, MEG). Some datasets might include ECG or PPG if recorded as auxiliary physiological measures during neuroimaging experiments.
    - **Raw Data:** Yes, the BIDS standard encourages the sharing of raw or minimally processed data in an organized structure. For example, dataset ds004306 ("EEG Semantic Imagination and Perception Dataset") provides raw EEG data.[27]
    - **Cardiovascular Focus:** Generally low, as the platform is primarily oriented towards neuroscience research.
    - **Access & Licensing:** Open access. Licenses vary by dataset but often include permissive options like CC0 (public domain dedication) or CC BY.
    - **Considerations:** OpenNeuro is an excellent resource for raw EEG data. However, it is less likely to feature datasets that combine EEG with comprehensive cardiovascular imaging (like cardiac MRI or echo) *and* all three specified sensor types (EEG, ECG, PPG) in a cardiovascular context.
- **BrainBeats Toolbox and Sample Data**
    - **Description:** BrainBeats is an open-source EEGLAB plugin designed for the joint analysis of EEG and cardiovascular signals (ECG/PPG). It offers protocols for assessing heartbeat-evoked potentials, feature-based analysis, and removing cardiac artifacts from EEG.[8] The toolbox comes with a sample dataset for tutorial purposes.
    - **Relevant Modalities (Sample Data):** The sample dataset contains 63 EEG channels, one ECG channel, and one PPG channel, all time-synchronized. This data was resampled from 1000 Hz to 250 Hz and corresponds to subject sub-032 (resting state, eyes open) from an open-source dataset available on the NEMAR (NeuroElectroMagnetic Data Archive and Tools Resource) platform.[8]
    - **Raw Data (Sample Data):** The sample data itself is resampled. The key would be to access the *original, full raw dataset* on the NEMAR platform from which this sample was derived.
    - **Cardiovascular Focus:** Yes, the toolbox and its intended use focus on brain-heart interactions.
    - **Access & Licensing (Toolbox):** The BrainBeats toolbox is open-source. The license of the underlying full dataset on NEMAR would need to be checked independently.
    - **Considerations:** The BrainBeats sample data demonstrates the availability of synchronized EEG, ECG, and PPG. If the full source dataset on NEMAR is accessible, contains genuinely raw (not just resampled) data for a larger

cohort, and potentially includes or can be linked to imaging data (which is not mentioned for the sample), it could be a candidate. The utility hinges on the characteristics of this full NEMAR dataset.

- **Hugging Face Datasets**
  - **Description:** Hugging Face is a rapidly growing platform that hosts a vast collection of datasets, models, and tools, primarily for AI and machine learning. It includes an increasing number of medical and multimodal datasets.[29]
  - **Relevant Modalities:** Highly varied. Examples include RadGenome/PMC-VQA (radiology image-caption pairs for visual question answering) [29], the GMAI-VL-5.5M dataset (image-text pairs from various medical modalities for general medical AI) [3], CXR Foundation (pre-trained model and embeddings for chest X-ray analysis, using images from MIMIC-CXR and private datasets) [19], and MedMCQA (medical multiple-choice question answering).[30]
  - **Raw Data:** Less common for the specific type of raw sensor and imaging data sought by the user. Many datasets are processed, feature-level, or consist of image-text pairs rather than comprehensive raw signal and pixel/voxel data collections suitable for broad, from-scratch model development.
  - **Cardiovascular Focus:** Possible in some specific datasets but not a general theme across the platform's medical offerings.
  - **Access & Licensing:** Varies significantly by dataset. Some are fully open, while others, like CXR Foundation, require users to review and agree to specific terms of use.[19]
  - **Considerations:** Hugging Face is a dynamic platform worth monitoring for new multimodal medical datasets. However, at present, it appears less focused on hosting the kind of large-scale, raw, integrated sensor (EEG, ECG, PPG) and imaging datasets required for this specific query. The user would need to filter extensively.
- **MedPix / MedPix 2.0**
  - **Description:** MedPix® is a free, open-access online database of medical images, teaching cases, and clinical topics, managed by the National Library of Medicine (NLM). It integrates images with textual metadata, including case scenarios, findings, diagnoses, and discussions.[5] MedPix 2.0 is an initiative to create a structured MongoDB version of this dataset to improve its utility for AI applications.[5]
  - **Relevant Modalities:** A wide variety of medical images (CT, MRI, X-ray, etc.) accompanied by rich textual clinical information.
  - **Raw Data:** A significant limitation of the original MedPix® is that "it is not possible to access to the raw data" in a way that is easily downloadable in

bulk for AI model training.[5] While MedPix 2.0 aims to make the data more structured and accessible (e.g., via JSON documents in MongoDB), it remains unclear if this translates to easy bulk download of raw pixel/voxel data for all images.

- **Cardiovascular Focus:** The database includes cardiovascular cases among its many topics and organ systems.
- **Access & Licensing:** Free and open-access.
- **Considerations:** Despite its rich content of images and clinical descriptions, the historical difficulty in accessing bulk raw image data and the absence of any mention of associated EEG, ECG, or PPG signals make MedPix less suitable for the current requirements.

Table 1: Overview of Key Multimodal Dataset Repositories

Repository Name	Primary Data Types	Availability of EEG	Availability of ECG	Availability of PPG	Availability of Imaging	Cardiovascular Focus	General Access Notes & Licensing Type	Key Reference(s)
MIMIC-IV (including Waveform, ECG, CXR link)	Clinical EHR, Physiological Waveforms, Diagnostic ECGs, Chest X-rays	Limited/Uncertain	Yes	Yes	Yes (CXR, potentially others via reports)	High	Credentialed access, Data Use Agreement (DUA), restricts re-identification.	[14]
PhysioNet (General)	Physiological Signals,	Varies by dataset	Varies by dataset	Varies by dataset	Varies by dataset	Varies	Often open, licenses vary	[9]

	Clinical Data, some Imaging						(e.g., ODC-BY, specific DUAs for some datasets like MC-MED).	
↳ MC-MED (on PhysioNet)	ED Clinical Data, ECG/PPG/Resp Waveforms, Imaging Results	No (not mentioned)	Yes	Yes	Yes (Reports, raw files TBC)	High	Likely credentialed via PhysioNet, specific DUA.	[20]
↳ "Working Memory & Music" (on PhysioNet)	ECG, PPG, fNIRS, other physiological signals	No	Yes	Yes	Yes (fNIRS - neuroimaging)	Low/Indirect	Open Data Commons Attribution License v1.0 (ODC-BY 1.0) - permissive.	[11]
CMRx Recon 2024 (via Synapse)	Cardiac MRI (k-space)	No (not mentioned)	No (not mentioned)	No (not mentioned)	Yes (Cardiac MRI)	High (Healthy Vols)	Open for research/education; commercial	[6]

							use of dataset itself prohibited.	
EchoNet-Dynamic (via Stanford AIMI)	Echocardiogram Videos	No (not mentioned)	No (not mentioned)	No (not mentioned)	Yes (Echocardiography)	High	Registration required; "Research Use Only," commercial use strictly prohibited.	[7]
The Cancer Imaging Archive (TCIA)	Medical Images (CT, MRI, PET, etc.), some clinical data	No (generally)	No (generally)	No (generally)	Yes (Various)	Partial/ Incidental	Public download; mostly CC-BY 3.0/4.0 (permissive), some collections may vary.	[10]
Open Neuro	Neuroimaging (MRI, PET, MEG), EEG, iEEG	Yes (Primary)	Varies (Auxiliary)	Varies (Auxiliary)	Yes (Neuroimaging)	Low	Open access; BIDS format; licenses vary (e.g., CC0, CC-BY).	[16]

BrainBeats Sample (source: NEMAR)	EEG, ECG, PPG (sample data)	Yes (Sample)	Yes (Sample)	Yes (Sample)	No (not mentioned for sample)	Medium (Brain-Heart)	Toolbox open-source; full source dataset on NEMAR needs license check.	[8]
Hugging Face Datasets	Diverse (Image-Text, Embeddings, Tabular, etc.)	Varies	Varies	Varies	Varies	Varies	Access and licenses vary greatly by dataset; some require accepting terms.	[3]
MedPix / MedPix 2.0	Medical Images, Textual Clinical Data	No (not mentioned)	No (not mentioned)	No (not mentioned)	Yes	Yes	Open access; raw data accessibility for bulk download has been a limitation.	[5]

This table provides a high-level comparison, highlighting that datasets excelling in one area (e.g., OpenNeuro for EEG, CMRxRecon2024 for cardiac MRI) often lack the other required modalities for the same subjects. MIMIC-IV and potentially MC-MED appear to be the most integrated but come with access conditions and potential gaps

(notably EEG for MIMIC-IV).

5. In-Depth Analysis of Top Candidate Datasets for Cardiovascular Web Applications

Based on the initial survey, datasets that offer the most comprehensive combination of physiological signals (especially ECG and PPG), linkable imaging (even if not all types), and a strong cardiovascular context, along with at least a pathway to raw data, are prioritized for deeper analysis. MIMIC-IV (with its associated modules) emerges as a primary candidate, despite the EEG gap. MC-MED is also promising but requires clarification on raw imaging access.

MIMIC-IV (Medical Information Mart for Intensive Care IV) and Associated Modules

- **Comprehensive Modality Profile:**
  - **EEG:** EEG data is not a standard, systematically collected and released component across the entirety of the MIMIC-IV patient cohort. While some ICU patients undergo EEG monitoring for specific clinical indications (e.g., seizure detection, altered mental status), these recordings are not part of the core MIMIC-IV Waveform Database or MIMIC-IV-ECG releases in a widespread, easily linkable manner for general cardiovascular cohorts. Finding substantial raw EEG data linked to cardiovascular patients within MIMIC-IV would likely require identifying specific research sub-studies that might have collected and (potentially) shared such data, or if such data exists within the hospital's raw archives but has not yet been processed for public release. This represents a significant gap if raw EEG is a hard requirement for all subjects.
  - **ECG:**
    - **MIMIC-IV Waveform Database:** Provides continuous ECG waveforms from bedside monitors, typically 2 to 5 leads, sampled at rates like 125 Hz or 250 Hz. These are available in WFDB format and represent the raw or minimally processed signals.[17] Duration can span hours to days.
    - **MIMIC-IV-ECG Module:** Contains approximately 800,000 standard 12-lead diagnostic ECGs. These are typically 10-second recordings, sampled at 500 Hz, stored in a DICOM-like structure with access to the raw signal data.[18] These are linkable to patient admissions and clinical data.
  - **PPG:** The MIMIC-IV Waveform Database includes continuous PPG signals, usually derived from pulse oximeters, with sampling rates similar to the bedside ECGs (e.g., 125 Hz). These are also in WFDB format.[17]

    - **Imaging:**
        - **MIMIC-CXR:** Provides over 370,000 chest X-ray images in DICOM format for a large subset of MIMIC-IV patients.[5] These images can be linked to the clinical data and, by extension, to the physiological signals of the same patient.
        - **Other Imaging:** MIMIC-IV clinical notes (e.g., radiology reports) contain textual descriptions of other imaging studies performed (CT, MRI, ultrasound, etc.). While the raw image files for these are not directly part of the MIMIC-IV release (beyond CXR), the reports provide valuable clinical context about imaging findings. Accessing these other raw imaging modalities would require separate initiatives or collaborations.
- Cardiovascular Relevance and Cohort Characteristics:
  MIMIC-IV's data is sourced from ICU patients, a population with a high prevalence of cardiovascular diseases, including myocardial infarction, heart failure, arrhythmias, shock, and post-cardiac surgery recovery. Cardiovascular monitoring is standard in the ICU. The database includes detailed diagnostic codes (ICD-9/10), procedure codes, medication administrations (including many cardiovascular drugs), and laboratory tests relevant to cardiac function (e.g., troponins, BNP). Patient demographics are available, though dates are shifted for de-identification, and ages over 89 are capped. There isn't a "healthy control" group in the traditional sense, as all patients are ICU admissions.
- **Raw Data: Veracity and Accessibility:**
    - **Waveforms (ECG/PPG):** Considered raw or minimally processed, provided in WFDB format. Downloadable from PhysioNet after credentialed access.
    - **Diagnostic ECGs:** Raw 12-lead signal data is accessible. Downloadable from PhysioNet after credentialing.
    - **CXR Images:** DICOM format, considered raw. Downloadable from PhysioNet after credentialing for MIMIC-CXR.
    - **Data Organization:** MIMIC-IV is modular. Data is organized into tables (e.g., patients, admissions, labevents, chartevents). Waveforms and ECGs are stored in specific directory structures linked by subject_id and hadm_id (hospital admission ID) or study_id (for ECGs).
    - **Size:** MIMIC-IV is very large (terabytes, especially with waveforms and imaging). This has significant implications for local storage and processing infrastructure for a web application.
- Licensing Terms and Implications for Web Application Use:
  Access to MIMIC-IV is governed by a Data Use Agreement (DUA) with PhysioNet. Users must complete human subjects research ethics training (e.g., CITI).14 Key terms include:

    - Prohibition of attempts to re-identify patients.
    - Requirement to protect data confidentiality.
    - Citation of the database in publications. The DUA does not explicitly prohibit commercial use of *models trained* on MIMIC-IV, but directly hosting or redistributing substantial portions of the raw MIMIC-IV data via a public web application would likely be problematic and require careful legal review. The focus is on enabling research. Any web application must ensure that its use of MIMIC-IV data (or models derived from it) strictly adheres to the DUA, particularly regarding patient privacy and data security. Displaying individualized predictions or data would need extreme caution to prevent re-identification.
- **Data Quality, Annotation, and Supporting Resources:**
    - MIMIC-IV is a real-world clinical dataset, so it contains noise, missing data, and variability inherent in clinical practice.
    - Annotations include clinical labels (diagnoses, procedures), machine-generated measurements from monitors, and cardiologist reports for some ECGs.
    - Extensive documentation is available on the MIMIC-IV website, including data dictionaries, tutorials for data extraction and analysis (e.g., for waveforms [17]), and example SQL queries. A large research community uses MIMIC, leading to many shared tools and publications.

MC-MED (Multimodal Clinical Monitoring in the Emergency Department)

- **Comprehensive Modality Profile:**
    - **EEG:** Not mentioned as a component of the MC-MED dataset.[20] This is a gap for the user's full requirements.
    - **ECG:** Continuously monitored ECG waveforms are included.[20] Details on the number of leads, sampling rates, and format would be available on the PhysioNet project page for MC-MED.
    - **PPG:** Continuously monitored PPG waveforms are included.[20] Sampling rate and format details would be on the PhysioNet project page.
    - **Imaging:** The dataset includes "laboratory and imaging results" and "free-text radiology reports".[20] This confirms the availability of information *about* imaging studies. However, it is crucial to determine if this also includes direct access to the raw imaging files (e.g., DICOMs of CT scans, X-rays, or ultrasounds performed in the ED) or if it's limited to the textual reports. If raw images are not directly part of the MC-MED release, this would be a significant limitation for applications requiring image processing.
- Cardiovascular Relevance and Cohort Characteristics:

High cardiovascular relevance, as the data comes from adult ED visits.20 EDs frequently manage patients with acute cardiovascular conditions like chest pain, arrhythmias, heart failure, and stroke. The dataset covers a diverse ED population from September 2020 to September 2022, including the COVID-19 pandemic period. Patient demographics, medical histories, and home medications are included.

- **Raw Data: Veracity and Accessibility:**
  - **Waveforms (ECG/PPG/Respiration):** Described as "continuously monitored vital signs and physiologic waveforms," suggesting raw or near-raw data.[20] Access would be via PhysioNet, likely requiring credentialing.
  - **Imaging Data:** As noted, the critical point is whether "imaging results" includes the raw images themselves. If so, formats (likely DICOM) and accessibility need to be confirmed.
  - **Data Organization:** Details would be provided in the dataset documentation on PhysioNet.
- Licensing Terms and Implications for Web Application Use:
  MC-MED is hosted on PhysioNet. Access would require agreeing to PhysioNet's terms and likely a specific DUA for MC-MED, similar to MIMIC-IV. The license terms would need careful review for implications regarding web application use, especially concerning commercialization, data hosting, and redistribution. Given its nature as sensitive patient data, restrictions on re-identification and data security would be paramount.
- **Data Quality, Annotation, and Supporting Resources:**
  - As a real-world ED dataset, it will reflect clinical realities, including potential noise and data variability.
  - Annotations include clinical data (diagnoses, orders, medications, lab results, visit outcomes).[20]
  - Supporting documentation, data dictionaries, and potentially example usage code would be expected on the PhysioNet project page for MC-MED.

Table 2: Detailed Comparison of Top Candidate Datasets

Feature	MIMIC-IV (with Waveforms, ECG, CXR link)	MC-MED (Multimodal Clinical Monitoring in the ED)
Specific EEG details	Not a standard component; would require finding	Not mentioned as a component.[20] Major gap.

	specialized sub-studies or data not yet released. Major gap.	
Specific ECG details	Continuous (2-5 leads, ~125Hz, WFDB) from Waveform DB.[17] Diagnostic 12-lead (10s, 500Hz, DICOM-like) from ECG Module.[18]	Continuous ECG waveforms included.[20] Specifics (leads, Hz, format) TBD from PhysioNet page.
Specific PPG details	Continuous PPG from pulse oximeters (~125Hz, WFDB) from Waveform DB.[17]	Continuous PPG waveforms included.[20] Specifics (Hz, format) TBD from PhysioNet page.
Specific Imaging details	MIMIC-CXR: Chest X-rays (DICOM).[5] Other imaging via radiology reports (text).	"Imaging results" and "free-text radiology reports".[20] Availability of raw image files (e.g., DICOM) needs confirmation.
Cardiovascular Conditions Covered	Extensive: MI, heart failure, arrhythmias, shock, post-cardiac surgery, etc. (ICU population).[14]	High relevance: Acute cardiovascular events common in ED (chest pain, arrhythmias, etc.).[20]
Raw Data Access & Formats	Waveforms (WFDB), ECGs (DICOM-like signals), CXRs (DICOM). Clinical data (structured tables). Via PhysioNet (credentialed).	Waveforms (format TBD). Imaging (raw files TBD, reports are text). Clinical data (structured). Via PhysioNet (likely credentialed).
Licensing & Key Terms	PhysioNet DUA: Requires ethics training, prohibits re-identification, emphasizes data privacy. Commercial use of trained models generally okay, but raw data redistribution problematic.[14]	Likely PhysioNet DUA: Similar restrictions to MIMIC-IV expected. Terms need careful review for web app implications.
Web App Suitability	**Medium to High (with**	**Medium (pending imaging**

	caveats): Strong for ECG/PPG/CXR + clinical data. EEG gap. Licensing requires careful handling for web app. Large scale.	**clarification):** Promising for ECG/PPG + ED clinical context. EEG gap. Raw imaging access TBC. Licensing needs review.
Pros	Very large, rich clinical context, well-documented, strong research community, raw ECG/PPG/CXR.	Modern ED data (incl. COVID era), multimodal by design (ECG, PPG, clinical, imaging info).
Cons	No standard EEG. DUA compliance for web app. Data size. Complexity.	No EEG. Raw imaging file access unconfirmed. Potentially restrictive DUA.
Key Reference(s)	5	20

This detailed comparison underscores that while MIMIC-IV offers a more established and comprehensive package for ECG, PPG, and linked CXR imaging within a rich clinical context, it has a clear gap regarding EEG. MC-MED is newer and potentially very valuable, especially for its ED focus, but key details regarding raw imaging access and its specific DUA terms need to be ascertained from its PhysioNet page. Both datasets would require careful planning for web application integration due to data size, complexity, and licensing/privacy obligations.

6. Strategic Recommendations for Dataset Integration

Integrating raw, multimodal medical data (EEG, ECG, PPG, and imaging) into a web application is a complex undertaking. Based on the preceding analysis, the following strategic recommendations are offered.

Prioritized List of Datasets
Given the stringent requirements, no single publicly available dataset perfectly matches all criteria (raw EEG, ECG, PPG, and cardiovascular imaging for the same subjects, with a permissive license for web applications). However, a pragmatic approach would be:

1. **MIMIC-IV (with Waveform Database, ECG Module, and MIMIC-CXR linkage):**
    - **Rationale:** Offers the most comprehensive combination of raw ECG, PPG, linkable raw imaging (Chest X-rays), and extensive clinical data with a strong cardiovascular focus from a large ICU cohort.[5] It is well-documented and widely used in research.
    - **Caveats:** The primary gap is the lack of systematically available raw EEG data for this cohort. The Data Use Agreement (DUA) requires careful adherence,

especially concerning patient privacy and data security in a web application context. The sheer size and complexity of the data also pose technical challenges.

2. **MC-MED (Multimodal Clinical Monitoring in the Emergency Department) (pending clarifications):**
    - **Rationale:** Specifically designed as a multimodal dataset from an ED setting, including raw ECG and PPG waveforms and links to imaging information.[20] Its focus on ED patients provides relevance for acute cardiovascular conditions.
    - **Caveats:** Also lacks EEG data. Crucially, confirmation is needed regarding whether "imaging results" includes direct access to raw imaging files or only textual reports. The specific DUA terms must be thoroughly reviewed for web application suitability.
3. **Specialized Datasets for Specific Modalities (if augmentation is considered):**
    - **OpenNeuro datasets (e.g., ds004306 [27]):** For raw EEG data, if this modality is critical and cannot be found within MIMIC-IV or MC-MED for relevant cohorts. Licenses are generally permissive.[16]
    - **CMRxRecon2024 [6] or EchoNet-Dynamic [7]:** For high-quality raw cardiac MRI or echocardiography videos, respectively, if the primary datasets (MIMIC-IV, MC-MED) lack sufficient depth or quality in these specific imaging types for the application's needs. However, these datasets do not include the required sensor data, and their licenses (especially EchoNet-Dynamic's "research use only" [7] and CMRxRecon2024's prohibition on commercial use of the dataset itself [6]) may conflict with web application goals.

Addressing Data Gaps

It is highly probable that a single dataset will not fulfill all requirements simultaneously. The combination of raw EEG with robust cardiovascular imaging (like cardiac MRI or echo) and other sensors (ECG, PPG) for the same large cohort, all under a permissive license, is exceptionally rare. Strategies to address this include:

- **Prioritize and Augment:** Select a primary dataset that meets the majority of critical requirements (e.g., MIMIC-IV for ECG, PPG, CXR, and cardiovascular context). Then, if EEG is indispensable for a specific feature of the web application, explore ways to:
    - Search for specialized sub-studies or ancillary projects related to the primary dataset that might have collected the missing modality (e.g., EEG studies on MIMIC patients).
    - Consider if a smaller, separate dataset with the missing modality (e.g., an OpenNeuro EEG dataset) could be used for a distinct component of the application, acknowledging the challenge of not having all data from the same

subjects.

- **Relax Constraints:** Evaluate if any constraints can be moderately relaxed without compromising the core goals of the web application. For example:
    - Is EEG absolutely essential for all aspects, or could the application initially focus on ECG, PPG, and imaging, with EEG planned as a future enhancement if suitable data becomes available?
    - Can "raw" be redefined to "minimally processed" if it significantly broadens the pool of available datasets? For instance, if some standard denoising has been applied but the fundamental signal characteristics are preserved.
    - Is a direct cardiovascular disease cohort essential, or could data from healthy individuals with comprehensive cardiovascular assessments (like in CMRxRecon2024, though it lacks sensors) be used for certain baseline modeling tasks?

Data Fusion Considerations

If the strategy involves using multiple datasets to cover all required modalities (e.g., one dataset for strong EEG, another for ECG/PPG + Imaging), the complexities of data fusion must be acknowledged. If data comes from different cohorts, direct per-subject multimodal analysis is impossible. If attempts are made to link subjects across different datasets (a highly challenging and often infeasible task with de-identified public data), issues of robust subject matching, temporal alignment of recordings, and harmonization of features and data formats become paramount. This level of data integration is a significant research and engineering undertaking in itself.

Technical Considerations for Web Applications

Deploying a web application that utilizes large volumes of raw medical data involves several technical considerations:

- **Data Storage:** Raw physiological waveforms and medical images, especially DICOMs or k-space data, can consume terabytes of storage. Scalable and cost-effective storage solutions (e.g., cloud storage like AWS S3, Google Cloud Storage) will be necessary.
- **Processing Pipelines:** Efficient pipelines will be needed to ingest, preprocess, analyze, and serve the raw data or derived results. This may involve specialized libraries for signal processing (e.g., SciPy, WFDB Software Package) and image analysis (e.g., Pydicom, SimpleITK, MONAI).
- **Ethical and Privacy Implications:** Beyond licensing, handling sensitive medical data in a web application carries profound ethical and privacy responsibilities.
    - If dealing with data that could potentially be linked to US patients (even if de-identified in the source dataset), considerations around HIPAA (Health Insurance Portability and Accountability Act) compliance for the application's infrastructure and data handling practices may be relevant, especially if any

protected health information (PHI) is inadvertently generated or handled.
    - For data involving European subjects, GDPR (General Data Protection Regulation) principles regarding data minimization, purpose limitation, and user rights must be respected.
    - Robust security measures to prevent data breaches are essential. Anonymization and de-identification techniques should be rigorously applied to any data presented to end-users to minimize re-identification risks.
- **Computational Resources:** Training complex AI models on multimodal data or performing on-the-fly analysis for a web application will require significant computational resources (CPUs, GPUs, memory).

Long-term Data Management and Updates
Some repositories, like MIMIC, release updated versions of their datasets over time. The web application's design should consider how to manage these updates. This might involve versioning of the data used by the application, retraining models on new data versions, and ensuring reproducibility of results.

7. Conclusion

Recap of Key Findings
The search for publicly available, raw, multimodal datasets encompassing EEG, ECG, PPG, and imaging data, specifically for cardiovascular web applications, reveals a landscape rich in potential yet fraught with challenges. The MIMIC-IV ecosystem (including its Waveform Database, ECG Module, and linkage to MIMIC-CXR) stands as a leading candidate, offering substantial raw ECG, PPG, and chest X-ray data within an extensive clinical cardiovascular context. However, it notably lacks systematically available raw EEG data for the same ICU cohorts and requires adherence to a strict Data Use Agreement. MC-MED on PhysioNet is another promising multimodal dataset from an emergency department setting with ECG and PPG, but clarity is needed on raw imaging access and its specific DUA. Other datasets excel in particular areas—OpenNeuro for raw EEG, CMRxRecon2024 for raw cardiac MRI, and EchoNet-Dynamic for echocardiography videos—but typically lack the full combination of required sensor and imaging modalities for the same subjects or have highly restrictive licenses unsuitable for many web application scenarios.
The primary difficulty lies in finding datasets that simultaneously satisfy all constraints: the specific quartet of raw sensor and imaging data for the same individuals, a clear cardiovascular focus, and licensing terms that are permissive enough for web application development and potential commercialization.

Dominant Trends
There is a clear and accelerating trend towards the collection and sharing of large-scale medical data, driven by the needs of AI research and the push for open science.32 The development of sophisticated AI models that leverage multimodal data for improved diagnostic accuracy and clinical decision support is a vibrant area of research, as evidenced

by numerous studies aiming to integrate imaging, physiological signals, and clinical records.1 However, persistent silos often exist between different types of data; for example, repositories strong in physiological signals may not have extensive, linked imaging data, and vice-versa. While the ambition for comprehensive multimodal datasets is high, readily available, truly raw, fully integrated datasets meeting the very specific combination requested are still maturing.

Future Outlook

The future is likely to bring more datasets that bridge these gaps. Initiatives focused on creating large, curated multimodal cohorts for specific diseases, including cardiovascular conditions, are underway. The adoption of data standards, such as BIDS (Brain Imaging Data Structure) for neuroimaging and its potential extensions or parallels in other domains 16, can significantly improve the findability, accessibility, interoperability, and reusability (FAIR principles) of medical data. As data sharing platforms like PhysioNet, OpenNeuro, and even general-purpose repositories like Hugging Face continue to grow, the likelihood of finding datasets that more closely match complex multimodal requirements will increase. Furthermore, advancements in privacy-preserving data sharing techniques and federated learning may allow for the utilization of sensitive data without compromising patient confidentiality, potentially broadening access.1

Final Word

The selection of a dataset for a web application is a critical decision with long-term implications. It is paramount that developers conduct thorough due diligence on any chosen dataset. This includes a meticulous review of its specific characteristics (modalities present, definition of "raw," data quality, sample size, cohort details), a comprehensive understanding of its licensing terms and any associated data use agreements, and a careful assessment of the ethical and privacy implications for the intended web application. Given the sensitivity of medical data, responsible data stewardship must be a core principle throughout the development lifecycle.

Works cited

1. Federated learning-based multimodal approach for early detection and personalized care in cardiac disease - PubMed Central, accessed June 9, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12055547/
2. Integrating deep learning with ECG, heart rate variability and demographic data for improved detection of atrial fibrillation, accessed June 9, 2025, https://openheart.bmj.com/content/12/1/e003185
3. GMAI-VL & GMAI-VL-5.5M: A Large Vision-Language Model and A Comprehensive Multimodal Dataset Towards General Medical AI - Hugging Face, accessed June 9, 2025, https://huggingface.co/papers/2411.14522
4. MEIT: Multi-Modal Electrocardiogram Instruction Tuning on Large Language Models for Report Generation - arXiv, accessed June 9, 2025, https://arxiv.org/html/2403.04945v3
5. MedPix 2.0: A Comprehensive Multimodal Biomedical Dataset for Advanced AI Applications, accessed June 9, 2025, https://arxiv.org/html/2407.02994v1
6. CMRxRecon2024: A Multimodality, Multiview k-Space Dataset ..., accessed June

9, 2025, https://pubs.rsna.org/doi/10.1148/ryai.240443
7. EchoNet Dynamic, accessed June 9, 2025, https://echonet.github.io/dynamic/
8. BrainBeats: an open-source EEGLAB plugin to jointly analyze EEG ..., accessed June 9, 2025, https://www.biorxiv.org/content/10.1101/2023.06.01.543272v2.full-text
9. PhysioNet - Laboratory for Computational Physiology, accessed June 9, 2025, https://lcp.mit.edu/physionet
10. The Cancer Imaging Archive, accessed June 9, 2025, https://www.cancerimagingarchive.net/
11. A Multimodal Dataset for Investigating Working Memory in Presence of Music - PhysioNet, accessed June 9, 2025, https://physionet.org/content/multimodal-nback-music/
12. cardiovascular-disease - Dataset - Catalog - Data.gov, accessed June 9, 2025, https://catalog.data.gov/dataset/?tags=cardiovascular-disease
13. A Multimodal Dataset for Investigating Working Memory in Presence ..., accessed June 9, 2025, https://physionet.org/content/multimodal-nback-music/1.0.0/
14. (PDF) MIMIC-IV, a freely accessible electronic health record dataset - ResearchGate, accessed June 9, 2025, https://www.researchgate.net/publication/366841127_MIMIC-IV_a_freely_accessible_electronic_health_record_dataset
15. MIMIC-IV Dataset - Papers With Code, accessed June 9, 2025, https://paperswithcode.com/dataset/mimic-iv
16. OpenNeuro, accessed June 9, 2025, https://openneuro.org/
17. MIMIC-IV Waveform Database v0.1.0 - PhysioNet, accessed June 9, 2025, https://physionet.org/content/mimic4wdb/
18. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset v1.0 - PhysioNet, accessed June 9, 2025, https://physionet.org/content/mimic-iv-ecg/
19. google/cxr-foundation - Hugging Face, accessed June 9, 2025, https://huggingface.co/google/cxr-foundation
20. Multimodal Clinical Monitoring in the Emergency Department (MC-MED) v1.0.0 - PhysioNet, accessed June 9, 2025, https://physionet.org/content/mc-med/1.0.0/
21. Datasets | Computational Medicine Laboratory, accessed June 9, 2025, https://wp.nyu.edu/cml/datasets/
22. Towards Universal Learning-based Model for Cardiac Image Reconstruction: Summary of the CMRxRecon2024 Challenge - arXiv, accessed June 9, 2025, https://arxiv.org/html/2503.03971v2
23. Towards Universal Learning-based Model for Cardiac Image Reconstruction: Summary of the CMRxRecon2024 Challenge | Request PDF - ResearchGate, accessed June 9, 2025, https://www.researchgate.net/publication/389648092_Towards_Universal_Learning-based_Model_for_Cardiac_Image_Reconstruction_Summary_of_the_CMRxRecon2024_Challenge
24. EchoNet-Dynamic Cardiac Ultrasound, accessed June 9, 2025, https://aimi.stanford.edu/datasets/echonet-dynamic-cardiac-ultrasound
25. The Cancer Imaging Archive (TCIA), accessed June 9, 2025,

https://imaging.cancer.gov/informatics/cancer_imaging_archive.htm
26. Search All Datasets - OpenNeuro, accessed June 9, 2025, https://openneuro.org/search
27. EEG Semantic Imagination and Perception Dataset - OpenNeuro, accessed June 9, 2025, https://openneuro.org/datasets/ds004306/
28. BrainBeats, an Open-Source EEGLAB Plugin to Jointly Analyze EEG and Cardiovascular Signals - bioRxiv, accessed June 9, 2025, https://www.biorxiv.org/content/10.1101/2023.06.01.543272v3.full.pdf
29. Medical Multimodal Datasets - a katielink Collection - Hugging Face, accessed June 9, 2025, https://huggingface.co/collections/katielink/medical-multimodal-datasets-6574d1db4fffc3f08b4effb8
30. openlifescienceai/medmcqa · Datasets at Hugging Face, accessed June 9, 2025, https://huggingface.co/datasets/openlifescienceai/medmcqa
31. MedPix, accessed June 9, 2025, https://medpix.nlm.nih.gov/
32. Open Data - HHS Office of the Chief Data Officer, accessed June 9, 2025, https://cdo.hhs.gov/s/open-data
33. UC Irvine Machine Learning Repository - Re3data.org, accessed June 9, 2025, https://www.re3data.org/repository/r3d100010960
34. GAF-FusionNet: Multimodal ECG Analysis via Gramian Angular Fields and Split Attention, accessed June 9, 2025, https://arxiv.org/html/2501.01960v1
35. A Comprehensive PPG-based Dataset for HR/HRV Studies - arXiv, accessed June 9, 2025, https://arxiv.org/html/2505.18165v1