DS Data Section Draft

Tanner DeGrazia

October 2025

1 Description of the Data

To further examine personalized medicine within the scope of current technological capabilities, CoralMD contextualizes three major data flows that collectively bridge genotype, phenotype, and clinical outcome. The first pillar, **genomic data**, enables users and clinicians to explore individual genetic variation in comparison to thousands of others, identifying mutations that may influence health risk or therapeutic response. The second pillar, **wearable data**, leverages the recent "hype" of consumer and clinical-grade devices, such as the Apple Watch, Fitbit, WHOOP, and continuous glucose monitors, to capture real-time fitness and health trends, including heart rate, VO_2 max, and glucose levels. These data streams power models that will help evaluate how short-term behavioral and metabolic patterns contribute to long-term health outcomes. Finally, the third pillar, **electronic health records (EHR)**, anchors CoralMD's predictive framework in longitudinal clinical context. By integrating past diagnoses, laboratory data, and treatment history, EHR data provide the interpretive backbone for connecting molecular predispositions and physiological measurements with observed medical events. Together, these datasets form the foundation for CoralMD's multimodal goal, where biological, behavioral, and clinical information intersect to advance individualized, data-driven healthcare. All datasets used are publicly available and fully identified, ensuring ethical and reproducible research practice.

1.1 Genomic Data

The genomic component of CoralMD is contextualized by the *Homo sapiens* reference genome GRCh38.p14 (Consortium, 2022), which provides the coordinate framework for aligning and annotating variants across an individual's genome. This reference genome acts as the foundational map against which all personal genomic data are compared.

Once the mapping to the reference is established, individual-level variation is analyzed using data from the 1000 Genomes Project (Auton et al., 2015). This dataset allows CoralMD to benchmark an individual's genomic profile against thousands of others, training models to detect genetic diversity and infer metrics such as allele frequency and potential pathogenicity. In effect, this step teaches CoralMD how to "read" a personal genome in population context.

Variant interpretation is further refined through the ClinVar archive (for Biotechnology Information, 2024), which provides curated annotations linking specific variants to known clinical outcomes, categorizing them as *pathogenic*, *benign*, or of *uncertain significance*. Finally, the gnomAD database (Karczewski et al., 2020) introduces population-level allele frequencies that regularize model predictions and ensure fairness across ancestries. By weighting variant importance based on global frequency, CoralMD mitigates bias and promotes equitable genomic interpretation.

To illustrate, consider the gene *APOE*. When a user uploads their genome, CoralMD compares it against the GRCh38 reference and identifies notable variants. If the model detects the *APOE* e4 allele, the system highlights its established association with elevated Alzheimer's disease and lipid metabolism risk. While genome sequencing extends far beyond single-gene associations, this example underscores how CoralMD leverages population data to make personal risk assessment both interpretable and evidence-based.

Together, these genomic datasets provide the foundation for CoralMD's variant interpretation engine. Each

genome is represented as a vector of genetic loci mapped to GRCh38 coordinates and annotated with ClinVar and gnomAD population context. This integrated framework allows CoralMD to generate interpretable, population-aware genomic insights that link molecular variation to clinical meaning.

1.2 Wearable and Physiological Data

Wearable and metabolic datasets extend CoralMD's predictive modeling from just genomic insights to continuous physiological states. These data streams capture real-time information about an individual's activity, sleep, heart rate, and glucose regulation, providing dynamic features that can connect genomic variation to observable health patterns, making an entire person's health picture interpretable.

The Fuller (2020) Apple Watch and Fitbit dataset (Fuller, 2020) provides wearable data paired with indirect ways of measuring calorie measurements, like variables such as heart rate, energy expenditure, step count, and activity type. CoralMD uses this dataset to examine physiologic features like resting heart rate, energy balance, and circadian rhythm stability, which can serve as variables towards an individual's complete picture of health when other methods are taken into account. For example, the combination of an *APOE* e4 allele and elevated resting heart rate may suggest increased cardiometabolic stress.

Complementing this, the OhioT1DM dataset (Marling et al., 2018) supplies continuous glucose monitoring (CGM) data collected every five minutes, along with insulin bolus and meal records. This enables training of time-series models to forecast glucose fluctuations and identify periods of dysregulation, deficiencies, and absolute anomalies that shout danger. Features such as mean glucose, coefficient of variation, and time above range provide a quantitative basis for modeling metabolic risk. When integrated with genomic markers, and variables of insulin sensitivity, CoralMD can model genotype–phenotype interactions that lead to metabolic stability.

1.3 Clinical and EHR Data

The third pillar of CoralMD focuses on clinical data, which provides the essential context for understanding a person's health history and predicting future risk. The MIMIC-IV electronic health record database (Johnson et al., 2023) includes de-identified hospital encounters containing demographics, diagnoses, laboratory results, medications, and vital signs. By integrating this kind of data, CoralMD can connect a patient's past medical experiences with their current biological and physiological signals to predict potential health concerns before they become critical.

EHR data make it possible for CoralMD to model health trajectories rather than just react to events. For instance, if an individual's records indicate a history of elevated blood pressure, the system can increase attention to cardiovascular patterns seen in their wearable data or lipid markers in their genome. Similarly, prior glucose irregularities in the EHR may heighten CoralMD's sensitivity to early signs of metabolic dysregulation from continuous glucose monitoring. In this way, CoralMD learns from historical clinical information to anticipate which systems of the body may need closer observation.

Rather than serving purely as a record of past care, EHR data in CoralMD are used proactively to estimate risk, prioritize focus areas, and guide personalized recommendations. This approach enables both patients and clinicians to see how everyday physiology, genetics, and medical history converge, transforming the EHR from a static archive into a predictive and preventive tool for health management.

1.4 Integration Across Each Outlet

Thus, CoralMD's architecture integrates genomic, physiological, and clinical data through multifaceted modeling strategies. Genomic features represent biological predispositions, wearable data captures continuous behavioral and metabolic dynamics, and EHR data give us even deeper insights to a picture of health.

All datasets are publicly accessible and fully didentified, aligning CoralMD's design with ethical AI principles of transparency, fairness, and reproducibility. This integrated dataset ecosystem forms the foundation for

CoralMD's predictive analytics pipeline, enabling individualized insight into healthspan and especially chronic disease risk.

Together, these three data pillars—genomic, wearable, and clinical—form the foundation of CoralMD's approach to personalized medicine. By combining molecular predispositions, real-time physiology, and medical history, the system can model health as a living, adaptive process rather than a snapshot in time. This integration allows CoralMD to anticipate risk, reveal meaningful patterns, and support proactive healthcare decisions grounded in an individual's unique biology and lived experience.

References

Auton, A., et al. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74.

Consortium, G. R. (2022). Grch38.p14 homo sapiens reference genome.

for Biotechnology Information, N. C. (2024). Clinvar: A public archive of relationships among sequence variation and human phenotype.

Fuller, D. (2020). Apple watch and fitbit activity dataset.

Johnson, A. E., et al. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, *10*, 1–11. https://physionet.org/content/mimiciv/.

Karczewski, K. J., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*, 434–443.

Marling, C., Bunescu, R., & Schwartz, F. (2018). The ohiot1dm dataset for blood glucose level prediction. *UCI Machine Learning Repository*. https://archive.ics.uci.edu/ml/datasets/OhioT1DM.