

Appendix A. Cohort Definition Example

We conducted a retrospective cohort study in the All of Us Registered Tier, restricting to participants with both OMOP-standardized EHRs and connected Fitbit data. All time references are anchored to a participant-specific prediction (“index”) time defined below.

A.1. Source population and phenotyping

From 8,477 All of Us participants with valid Fitbit and EHR linkage, we identified 2,148 participants with any evidence of major depressive disorder (MDD) using the *union* of (i) OMOP concept IDs for MDD, (ii) PheCode mappings for depressive disorders, and (iii) depression-specific medications. Phenotyping was performed strictly prior to index.

A.2. Incident case definition and index time

Incident cases required the first MDD evidence to occur *at least* 180 days after Fitbit tracking began, and evidence of EHR activity both before and after Fitbit start. For positives, the index time was set to one day prior to the first MDD flag, yielding 348 incident cases. For negatives, the index time was set to one day prior to the last observed EHR record.

A.3. Control pool and analytic cohort

The control pool comprised participants with (i) no MDD evidence at any time, (ii) ≥ 180 days of valid Fitbit data, and (iii) EHR activity both before and after Fitbit start, producing 2,701 negatives. These criteria formed an initial analytic cohort of 3,049 participants (11.4% positives).

A.4. Wearable/EHR quality filters and final dataset

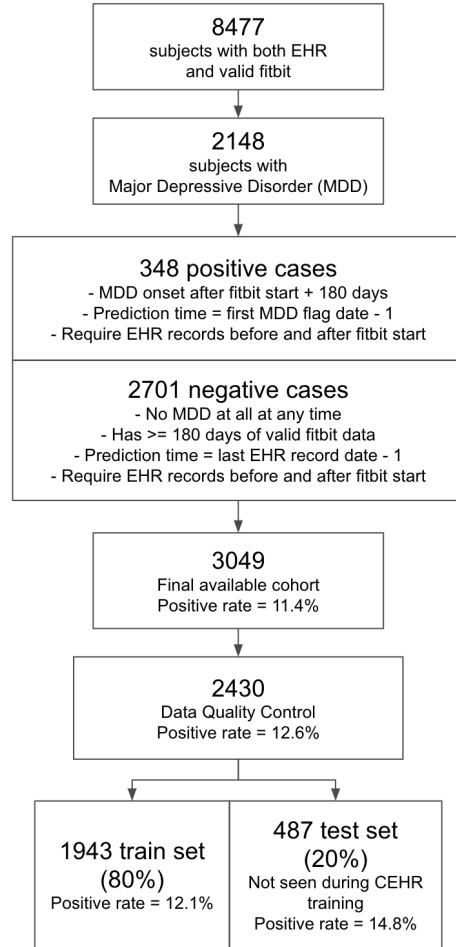
We applied pre-specified quality control (QC) on wearable data density/regularity and device sanity checks, and excluded participants failing minimal EHR coverage around index. After QC, 2,430 participants remained (12.6% positives).

A.5. Holdout protocol and leakage safeguards

To avoid leakage when comparing to an external CEHR model trained on other AoU participants, we restricted the held-out test set to participants *not*

present in CEHR training (487 participants; 14.8% positives) and used the remainder for model development (1,943; 12.1% positives). All feature extraction windows (EHR and wearable) were strictly pre-index (no look-ahead).

Summary of counts. 8,477 $\rightarrow$ 3,049 (phenotyping and design constraints) $\rightarrow$ 2,430 (data quality control) split into 1,943 development and 487 held-out test participants.



Supplementary Figure 1: Example cohort definition steps and numbers for Major Depressive Disorder

Appendix B. More results

Abbreviations used in the tables: To optimize space utilization in the performance comparison tables, we employ standardized abbreviations for health outcomes and methodological components. Health outcomes are abbreviated as follows: HLD (Hyperlipidemia), HTN (Hypertension), GERD (Gastroesophageal Reflux Disease), GAD (Generalized Anxiety Disorder), MDD (Major Depressive Disorder), OB (Obesity), SA (Sleep Apnea), T2D (Type 2 Diabetes), HF (Heart Failure), and AF (Atrial Fibrillation). Methodological abbreviations include CEHR (Clinical Element-based Health Records), OMOP (Observational Medical Outcomes Partnership), and TS Embed (Time Series Embedding). These abbreviations enable comprehensive presentation of performance metrics across all outcomes and fusion approaches while maintaining table readability and fitting within standard page constraints.

Supplementary Table 1: Performance (AUROC) of best feature fusion between EHR and wearable data across health outcomes against EHR baselines.

	HLD	HTN	GERD	GAD	MDD	OB	SA	T2D	HF	AF
OMOP	0.648 $\pm$ 0.030	0.702 $\pm$ 0.035	0.662 $\pm$ 0.034	0.733 $\pm$ 0.027	0.713 $\pm$ 0.036	0.691 $\pm$ 0.034	0.681 $\pm$ 0.042	0.751 $\pm$ 0.064	0.708 $\pm$ 0.055	0.680 $\pm$ 0.057
CEHR	0.691 $\pm$ 0.035	0.722 $\pm$ 0.030	0.669 $\pm$ 0.033	0.792 $\pm$ 0.030	0.767 $\pm$ 0.025	0.694 $\pm$ 0.036	0.746 $\pm$ 0.035	0.735 $\pm$ 0.055	0.862 $\pm$ 0.030	0.620 $\pm$ 0.049
Feature Fusion	0.808⁽¹⁾$\pm$0.027	0.829⁽¹⁾$\pm$0.029	0.785⁽¹⁾$\pm$0.034	0.848⁽³⁾$\pm$0.028	0.825⁽⁴⁾$\pm$0.023	0.805⁽⁴⁾$\pm$0.032	0.816⁽¹⁾$\pm$0.030	0.873⁽⁴⁾$\pm$0.038	0.881⁽⁴⁾$\pm$0.037	0.794⁽¹⁾$\pm$0.053
Significance	***	***	***	***	***	***	***	***	***	***

Notes: The baseline uses two types of EHR encodings: OMOP and CEHR. Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Best fusion methods for each outcome: (1) $[\mathbf{x}_{\text{EHR-OMOP}}, \mathbf{x}_{\text{wear-sum}}]$ (Concat); (3) $[\mathbf{x}_{\text{EHR-CEHR}}, \mathbf{x}_{\text{wear-sum}}]$ (Concat); (4) $[\mathbf{x}_{\text{EHR-CEHR}}, \mathbf{x}_{\text{wear-sum}}]$ (Weighted). See next table.

Supplementary Table 2: Performance comparison across all baseline models and feature fusion methods.

Encoding & Fusion Methods	HLD	HTN	GERD	GAD	MDD	OB	SA	T2D	HF	AF
$\mathbf{x}_{\text{EHR-OMOP}}$	0.648 $\pm$ 0.030	0.702 $\pm$ 0.035	0.662 $\pm$ 0.034	0.733 $\pm$ 0.027	0.713 $\pm$ 0.036	0.691 $\pm$ 0.034	0.681 $\pm$ 0.042	0.751 $\pm$ 0.064	0.708 $\pm$ 0.055	0.680 $\pm$ 0.057
$\mathbf{x}_{\text{EHR-CEHR}}$	0.691 $\pm$ 0.034	0.721 $\pm$ 0.030	0.669 $\pm$ 0.033	0.792 $\pm$ 0.030	0.767 $\pm$ 0.025	0.694 $\pm$ 0.036	0.746 $\pm$ 0.035	0.735 $\pm$ 0.055	0.862 $\pm$ 0.030	0.620 $\pm$ 0.049
$\mathbf{x}_{\text{wear-sum}}$	0.804 $\pm$ 0.029	0.807 $\pm$ 0.031	0.767 $\pm$ 0.033	0.833 $\pm$ 0.028	0.797 $\pm$ 0.026	0.762 $\pm$ 0.038	0.741 $\pm$ 0.036	0.790 $\pm$ 0.058	0.778 $\pm$ 0.059	0.791 $\pm$ 0.053
$\mathbf{x}_{\text{wear-ts}}$	0.753 $\pm$ 0.034	0.812 $\pm$ 0.026	0.792 $\pm$ 0.029	0.745 $\pm$ 0.030	0.671 $\pm$ 0.034	0.695 $\pm$ 0.041	0.687 $\pm$ 0.046	0.774 $\pm$ 0.062	0.742 $\pm$ 0.068	0.764 $\pm$ 0.058
(1) $[\mathbf{x}_{\text{EHR-OMOP}}, \mathbf{x}_{\text{wear-sum}}]$ (Concat)	0.808$\pm$0.027	0.829$\pm$0.029	0.785$\pm$0.034	0.842 $\pm$ 0.025	0.815 $\pm$ 0.020	0.793 $\pm$ 0.032	0.816$\pm$0.030	0.816 $\pm$ 0.045	0.831 $\pm$ 0.046	0.794$\pm$0.053
(2) $[\mathbf{x}_{\text{EHR-OMOP}}, \mathbf{x}_{\text{wear-ts}}]$ (Concat)	0.708 $\pm$ 0.035	0.782 $\pm$ 0.026	0.741 $\pm$ 0.031	0.805 $\pm$ 0.026	0.751 $\pm$ 0.029	0.752 $\pm$ 0.035	0.754 $\pm$ 0.032	0.578 $\pm$ 0.078	0.768 $\pm$ 0.066	0.763 $\pm$ 0.040
(3) $[\mathbf{x}_{\text{EHR-CEHR}}, \mathbf{x}_{\text{wear-sum}}]$ (Concat)	0.797 $\pm$ 0.031	0.824 $\pm$ 0.024	0.780 $\pm$ 0.030	0.848$\pm$0.028	0.823 $\pm$ 0.023	0.804 $\pm$ 0.032	0.797 $\pm$ 0.028	0.819 $\pm$ 0.045	0.880 $\pm$ 0.036	0.783 $\pm$ 0.047
(4) $[\mathbf{x}_{\text{EHR-CEHR}}, \mathbf{x}_{\text{wear-sum}}]$ (Weighted)	0.798 $\pm$ 0.031	0.818 $\pm$ 0.026	0.778 $\pm$ 0.030	0.848 $\pm$ 0.028	0.825$\pm$0.023	0.805$\pm$0.032	0.801 $\pm$ 0.029	0.873$\pm$0.038	0.881$\pm$0.037	0.785 $\pm$ 0.047
(5) $[\mathbf{x}_{\text{EHR-CEHR}}, \mathbf{x}_{\text{wear-ts}}]$ (Concat)	0.789 $\pm$ 0.036	0.815 $\pm$ 0.025	0.785 $\pm$ 0.029	0.805 $\pm$ 0.030	0.781 $\pm$ 0.024	0.724 $\pm$ 0.037	0.781 $\pm$ 0.029	0.785 $\pm$ 0.061	0.812 $\pm$ 0.050	0.749 $\pm$ 0.047
(6) $[\mathbf{x}_{\text{EHR-CEHR}}, \mathbf{x}_{\text{wear-ts}}]$ (Weighted)	0.789 $\pm$ 0.036	0.815 $\pm$ 0.025	0.785 $\pm$ 0.029	0.805 $\pm$ 0.030	0.781 $\pm$ 0.025	0.724 $\pm$ 0.037	0.781 $\pm$ 0.029	0.785 $\pm$ 0.061	0.812 $\pm$ 0.050	0.749 $\pm$ 0.047

As shown in Supplementary Table 1, the integration of wearable data produced strong improvements for chronic cardiovascular and metabolic conditions. For hyperlipidemia, the $[\mathbf{x}_{\text{EHR-OMOP}}, \mathbf{x}_{\text{wear-sum}}]$ (Concat) approach achieved an AUROC of 0.808, representing a +0.117 improvement over the CEHR-only baseline of 0.691 ($p = 1.6 \times 10^{-53}$). Similarly, hypertension prediction improved from a CEHR baseline of 0.721 to 0.829 using the same $[\mathbf{x}_{\text{EHR-OMOP}}, \mathbf{x}_{\text{wear-sum}}]$ (Concat), yielding a +0.107 improvement ($p = 9.3 \times 10^{-52}$). Obesity prediction demonstrated remarkable improvement through $[\mathbf{x}_{\text{EHR-CEHR}}, \mathbf{x}_{\text{wear-sum}}]$ (Weighted), rising from a CEHR baseline of 0.694 to 0.805 (+0.111, $p = 3.3 \times 10^{-61}$). Type II diabetes achieved the highest final AUROC of 0.873 through $[\mathbf{x}_{\text{EHR-CEHR}}, \mathbf{x}_{\text{wear-sum}}]$ (Weighted), improving from an OMOP baseline of 0.751 (+0.122, $p = 2.0 \times 10^{-31}$).

Gastroesophageal reflux disease showed substantial improvement as well, with $[\mathbf{x}_{\text{EHR-OMOP}}, \mathbf{x}_{\text{wear-sum}}]$ (Concat) increasing performance from a baseline CEHR AUROC of 0.669 to 0.785, representing a +0.116 gain ($p = 2.4 \times 10^{-50}$).

Mental health conditions also benefited significantly from wearable integration. Generalized anxiety disorder prediction improved from a strong CEHR baseline of 0.792 to 0.848 using $[\mathbf{x}_{\text{EHR-CEHR}}, \mathbf{x}_{\text{wear-sum}}]$ (Concat), achieving a +0.056 improvement ($p = 6.3 \times 10^{-54}$), while major depressive disorder showed similar gains from 0.767 to 0.825 with $[\mathbf{x}_{\text{EHR-CEHR}}, \mathbf{x}_{\text{wear-sum}}]$ (Weighted) (+0.058, $p = 4.3 \times 10^{-48}$).

Sleep apnea showed consistent improvement, rising from a CEHR baseline of 0.746 to 0.816 using $[\mathbf{x}_{\text{EHR-OMOP}}, \mathbf{x}_{\text{wear-sum}}]$ (Concat) (+0.070, $p = 3.6 \times 10^{-31}$).

Even outcomes with already strong baseline performance showed meaningful improvements. Heart failure, which had the highest baseline CEHR performance at 0.862, still achieved a statistically significant improvement to 0.881 using $[\mathbf{x}_{\text{EHR-CEHR}}, \mathbf{x}_{\text{wear-sum}}]$ (Weighted) (+0.019, $p = 7.5 \times 10^{-12}$). Atrial fibrillation demonstrated substantial improvement, rising from an OMOP baseline of 0.680 to 0.794 with $[\mathbf{x}_{\text{EHR-OMOP}}, \mathbf{x}_{\text{wear-sum}}]$ (Concat), representing a +0.114 gain ($p = 1.4 \times 10^{-26}$).