Original Paper

Comprehensive Clinical Decision-Support Framework for Obesity Prevention Using Large-Scale Health Checkup Data: Development and Evaluation Study

Suho Jang[1,2], MS; Yoon Ji Kim[1], MS; Ye Seul Bae[1,2,3], MD, MPH, PhD; Jae-Heon Kang[3,4,5], MD, MPH, PhD

[1]Big Data Research Center, Division of Healthcare Planning, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[2]Department of Clinical Research Design & Evaluation, Samsung Advanced Institute for Health Sciences & Technology (SAIHST), Sungkyunkwan University, Seoul, Republic of Korea

[3]Department of Family Medicine, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[4]Institute for Clinical Nutrition, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[5]Institute for Future Healthcare, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

Corresponding Author:
Ye Seul Bae, MD, MPH, PhD
Department of Family Medicine
Kangbuk Samsung Hospital
Sungkyunkwan University School of Medicine
29, Saemunan-ro, Jongno-gu
Seoul, 03181
Republic of Korea
Phone: 82 2 2001 1797
Fax: 82 2 734 2278
Email: byeye1313@gmail.com

Abstract

Background: Obesity prevention requires identifying nonobese adults at risk of developing obesity. Routine health checkups collect anthropometric, laboratory, medical history, and lifestyle information, but screening primarily describes current anthropometric status and does not directly provide individualized estimates of future risk for preventive counseling.

Objective: This study aimed to develop and evaluate a comprehensive clinical decision-support framework for obesity prevention using large-scale health checkup data.

Methods: This retrospective cohort study used repeated examination data from the Kangbuk Samsung Health Study. Adults with BMI less than 25 kg/m², at least 3 health checkup visits, a 2-year obesity-free washout period, and postindex follow-up were included. Incident obesity was defined as the first follow-up BMI of 25 kg/m² or higher. The framework integrated a single-visit prediction tier (which we termed *lite*), a survival component for time-to-transition context, a repeated-measurement updating tier (which we termed *plus*), and an explainability layer. The primary *lite*-tier and survival analyses used training-set-only imputation. A previously exported *lite*-tier Extreme Gradient Boosting (XGBoost) component was externally evaluated in 328,268 National Health Insurance Service (NHIS) participants.

Results: Among 231,683 initially nonobese participants, 38,094 developed incident obesity. In the aligned 5-year visit-1 test cohort, 4088 of 21,783 (18.8%) participants developed obesity. The *lite*-tier ensemble achieved an area under the receiver operating characteristic curve (AUROC) of 0.918, an area under the precision-recall curve (AUPRC) of 0.693, a Brier score of 0.087, and an expected calibration error of 0.005. AUPRC was approximately 3.7 times the no-information reference. AUROCs were 0.913 for BMI-only and 0.914 for BMI+age+sex. Observed 5-year incidence ranged from 0.1% (6/4357) to 64.7% (2821/4357) across accelerated failure time (AFT) risk quintiles. The *plus* tier achieved a landmark-row AUROC of 0.899 with less favorable calibration. The previously exported *lite*-tier XGBoost component achieved an external AUROC of 0.857, AUPRC of 0.575, and Brier score of 0.132.

Conclusions: Large-scale health checkup data supported an integrated framework for single-visit triage, temporal risk context, risk-domain explanation, and exploratory repeated-measurement updating, providing a basis for prevention-oriented risk

XSL•FO
RenderX

communication and follow-up prioritization. The overall gain in discrimination beyond BMI was modest, and the effects on clinical decisions and obesity prevention require prospective evaluation.

(J Med Internet Res 2026;28:e107821) doi: 10.2196/107821

KEYWORDS

obesity prevention; clinical decision support; risk prediction model; health checkup; incident obesity; machine learning; external validation; calibration; decision curve analysis; risk communication

Introduction

Effective primary prevention of obesity requires identifying currently nonobese adults who are likely to develop obesity. This preventive focus is important because sustained weight reduction after obesity develops is challenging [1], and obesity is associated with premature mortality [2] and major cardiometabolic, hepatic, and cancer outcomes [3]. In Asian populations, cardiometabolic risk can increase at relatively low BMI levels [4,5], making early risk assessment during routine health checkups clinically relevant. South Korea operates a well-established national health screening program that collects anthropometric, blood pressure, laboratory, medical history, and lifestyle information [6,7]. However, obesity screening in this setting primarily classifies current anthropometric status using BMI and waist circumference. It does not directly estimate future transition risk or provide an individualized risk context for prioritizing preventive counseling.

Previous studies have addressed obesity or overweight prediction [8,9], future weight gain [10,11], and incident obesity phenotypes among nonobese adults [12]. These studies support the feasibility of obesity-related prediction, but few have linked single-visit triage, temporal interpretation, repeated-measurement updating, and risk-domain explanation within one health checkup-based clinical decision-support framework. Such integration accommodates adults with a single visit for first-line assessment and those with repeated measurements for longitudinal risk profiling.

Risk stratification supports preventive decision-making for cardiovascular diseases and diabetes [13,14]. For obesity prevention during routine health checkups, risk information should be timely, interpretable, and usable for clinical risk communication. A practical decision-support framework should estimate risk before BMI-defined obesity occurs, summarize anthropometric, metabolic, blood pressure, renal, and lifestyle domains, and help prioritize counseling or follow-up [15,16]. Because the endpoint in this study is defined by BMI threshold crossing, the framework is intended for risk communication and follow-up prioritization rather than for providing mechanistic explanations.

We aimed to develop and evaluate a comprehensive clinical decision-support framework for obesity prevention among currently nonobese adults using large-scale health checkup data. The framework was designed for delivery through a web-based interface, and we assessed its incremental value beyond BMI-centered screening. The intended functions were triage based on a single health checkup visit, trajectory interpretation through time-to-transition modeling and repeated-measurement updating, and follow-up prioritization through explainable risk-domain outputs. We additionally evaluated the discrimination of the *lite*-tier Extreme Gradient Boosting (XGBoost) component in an independent nationwide health checkup cohort linked to the National Health Insurance Service (NHIS) claims data.

Methods

Study Design and Data Source

This study was conducted within the Kangbuk Samsung Health Study (KSHS), a large-scale cohort study of Korean adults aged 18 years or older who underwent comprehensive annual or biennial health screening examinations at the Kangbuk Samsung Hospital Total Healthcare Centers in Seoul and Suwon, South Korea [17,18]. We conducted a retrospective cohort prediction study among adults who were not obese at baseline to develop and evaluate a comprehensive clinical decision-support framework for obesity prevention using information commonly available in health screening settings. The overall clinical decision-support framework is illustrated in Figure 1.

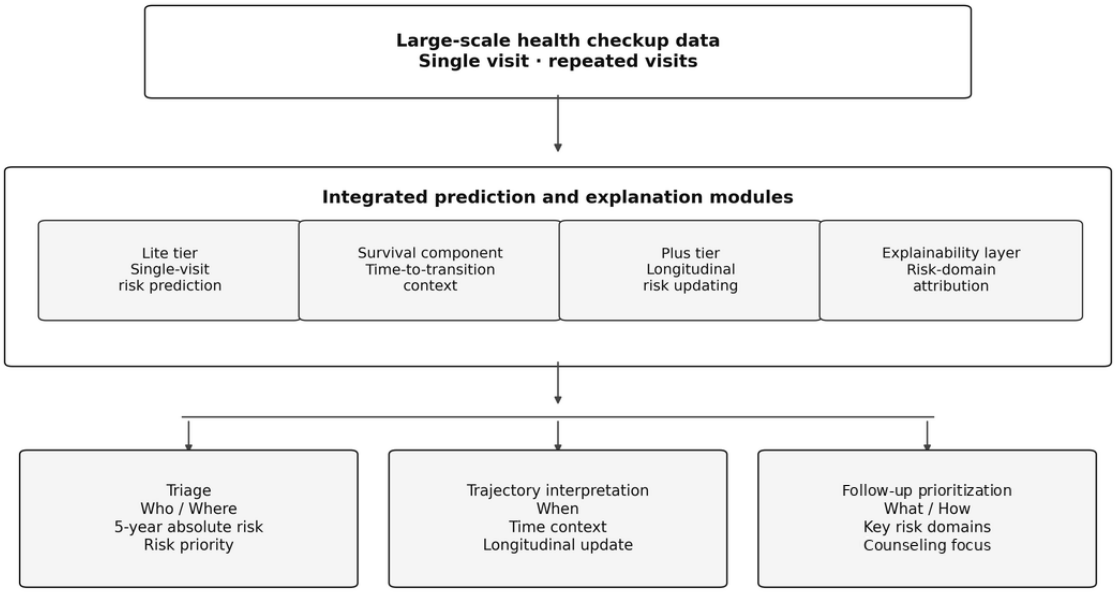
Deidentified KSHS health checkup records covering 2010 to 2024, including the 2-year washout window, were extracted and analyzed using the Samsung Medical Center Unified Big-data Engine (S-CUBE), an integrated clinical data warehouse and analytics platform at Kangbuk Samsung Hospital. Data access was authorized under institutional review board (IRB) approval and a data-use agreement, and all analyses were conducted within the institution's secure research environment. The NHIS-linked external cohort was drawn from the NHIS sample research database, accessed through the National Health Insurance Sharing Service, and covered the observation period of that database (2002-2019).

For external evaluation, the *lite*-tier XGBoost component was assessed in an independent NHIS-linked population-based health checkup cohort. Variable harmonization between the development data and the external cohort was performed by mapping shared predictors, units, and categorical definitions where available.

Reporting was guided by the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis plus Artificial Intelligence (TRIPOD+AI) guidelines for prediction-model studies and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines for observational cohort studies [19,20]. The completed checklists are provided in Multimedia Appendices 1 and 2.

XSL•FO
RenderX

Figure 1. Comprehensive clinical decision-support framework for obesity prevention. Single and repeated health checkup visits inform risk triage, trajectory interpretation, and follow-up prioritization through the *lite* tier, survival, *plus* tier, and explainability components.



Study Population and Index Visit

Eligible participants were adults who were not obese at the index health checkup examination, defined as BMI less than 25 kg/m². Participants were required to have at least 3 health checkup examinations, a 2-year obesity-free washout period before the index visit, and postindex follow-up for outcome ascertainment. Participants with missing BMI data, prevalent obesity during the washout window, early outcome events during the washout period, or insufficient follow-up information were excluded.

The 2-year washout was applied as a look-back eligibility window before the baseline index visit, and fixed prediction horizons were anchored at the index visit. Participants were split at the individual level into training, validation, and test sets in a 70:15:15 ratio. All repeated visits and landmark rows from the same participant were assigned to the same data split to prevent participant-level leakage.

Outcome Definition and Prediction Horizons

The primary outcome was BMI-defined incident obesity, defined as the first observed follow-up visit with BMI of 25 kg/m² or higher after an index visit with BMI <25 kg/m², consistent with Asia-Pacific and Korean criteria [4,5]. The primary prediction horizon was 5 years; secondary horizons were 1, 2, and 3 years. Horizon-specific classification included participants with an event within the horizon or event-free observation through the end of the horizon; participants censored earlier without an event were excluded.

Predictor Measurements and Definitions

Candidate predictors were selected from variables routinely collected during health checkups. The anthropometric measurements included height, weight, BMI, waist circumference, and waist-to-height ratio (WHtR). BMI was calculated as weight in kilograms divided by height in meters squared. WHtR was included among the final continuous predictors, whereas waist circumference itself was retained only as an auxiliary variable for imputation and for deriving WHtR and was not a model input; where abdominal obesity was referenced for clinical interpretation, sex-specific Korean waist-circumference cutoffs of ≥90 cm in men and ≥85 cm in women were used [5,21].

Blood pressure and clinical measurements included systolic blood pressure, diastolic blood pressure, heart rate, and mean arterial pressure (MAP). MAP was calculated as the diastolic blood pressure plus one-third of the pulse pressure. The laboratory predictors included fasting glucose, glycated hemoglobin A_{1c} (HbA_{1c}), total cholesterol, triglycerides, high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), aspartate aminotransferase (AST), alanine aminotransferase (ALT), gamma-glutamyl transferase (GGT), serum creatinine, and estimated glomerular filtration rate (eGFR). eGFR was calculated using the Modification of Diet in Renal Disease (MDRD) equation. Derived metabolic indices included the triglyceride-glucose (TyG) index, calculated as ln[fasting triglycerides (mg/dL) x fasting glucose (mg/dL) / 2] [22], the total cholesterol-to-HDL-C ratio, and the AST-to-ALT ratio.

Medical history variables included a history of hypertension, diabetes, and dyslipidemia. The medication variables included the use of antihypertensive, glucose-lowering, and lipid-lowering medications. Family history variables included hypertension and diabetes. The lifestyle variables included smoking status, alcohol intake, and physical activity. Smoking status was categorized as never, former, or current. Alcohol intake was analyzed as reported daily intake in grams per day. Physical activity was assessed using questionnaire-derived activity information and categorized as low, moderate, or high, consistent with the self-reported physical activity classification used in health checkup studies [23].

XSL•FO
RenderX

Preprocessing and Imputation

Missing values were imputed by predictive mean matching (PMM; 5 imputations, 5 nearest neighbors). We formed a consensus feature table by averaging continuous values and taking the modal categorical value in originally missing cells, then recalculated derived indices. The outcome was excluded from imputation. For the primary *lite*-tier and survival analyses, the imputation model was fitted on the training split and applied unchanged to validation and test data. Fitting imputation to the pooled development data was assessed as a sensitivity analysis (Table S1 in Multimedia Appendix 3). Bootstrap intervals quantify test-set sampling uncertainty but do not propagate between-imputation uncertainty. Before imputation, missingness was at most 0.23% (539/231,683) for anthropometric and laboratory variables and 0.32% (748/231,683) for history and family-history items, compared with 1.27% (2945/231,683) for physical activity, 4.73% (10,956/231,683) for smoking, and 8.22% (19,045/231,683) for reported alcohol intake.

Comprehensive Clinical Decision-Support Framework Overview

The framework combines 4 modules to convert health checkup information into interpretable risk estimates (Figure 1). The *lite* tier estimates fixed-horizon risk from a single visit for first-line triage. The survival component provides time-to-transition context across index-visit risk strata. The *plus* tier uses repeated measurements for exploratory longitudinal risk updates, and the explainability layer summarizes risk domains through feature attribution. Together, these modules support 3 intended functions: triage, trajectory interpretation, and follow-up prioritization. The outputs are intended for risk communication and decision support, not automated intervention.

The Lite-Tier Model

The *lite* tier estimates risk from predictors available at a single index health checkup. Its primary model combines XGBoost and logistic regression predictions with fixed weights of 0.6 and 0.4, respectively, which were prespecified without data-driven tuning. Component and ensemble performance are summarized in Table S2 in Multimedia Appendix 3.

The XGBoost component used 500 estimators, a maximum depth of 6, a learning rate of 0.05, column subsampling of 0.8, and row subsampling of 0.8. The logistic regression component used L2 regularization with a regularization parameter (C) of 1 and a maximum of 1000 iterations. The operating threshold was selected in the validation set using the Youden J statistic and then applied to the held-out test set.

BMI-Based Benchmarks

BMI-only and BMI+age+sex logistic regression benchmarks were fitted and evaluated on the same training and test cohorts as the *lite* tier (Table S2 in Multimedia Appendix 3). The BMI-only model used index-visit BMI as its sole predictor, and the BMI+age+sex model added age and sex. Comparisons of the *lite* tier with these benchmarks were made at 3 levels: overall discrimination in the aligned test cohort, event capture at a fixed nominal screening capacity, and discrimination within narrow baseline BMI strata. The corresponding statistical procedures are described under Data Analysis, and the results are reported in the Results subsections titled Comparison With BMI-Based Benchmarks, Fixed-Capacity Reclassification, and Incremental Discrimination Across Baseline BMI Strata.

Calibration and Decision-Curve Analysis

Raw model predictions were used for primary performance estimates. Calibration was summarized using calibration plots, the calibration slope, and calibration-in-the-large. Decision-curve analysis (DCA) evaluated 7 threshold probabilities (0.05, 0.10, 0.15, 0.168, 0.20, 0.25, and 0.30), with interpretation focused on 0.10 and 0.20. The DCA grid point of 0.168 was distinct from the validation-derived classification threshold.

Survival Component

Cox proportional hazards and log-normal accelerated failure time (AFT) models complemented fixed-horizon prediction. Event time was measured from the index visit to the first follow-up visit recording BMI ≥25 kg/m², indicating the time of detection rather than the exact time of onset. Five-year test-set performance was assessed using area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), Brier score, and expected calibration error (ECE). Test-set participants were divided into equal-sized quintiles of predicted AFT risk. Median observed event time was summarized among participants with an event.

The Plus-Tier Model

The *plus* tier was designed for participants with repeated health checkup measurements. Landmark rows were constructed from repeated visits, and each prediction window began after the corresponding landmark visit. All landmark rows from the same participant were assigned to the same data split. Predictor histories up to the landmark visit were used to estimate subsequent 5-year BMI-defined incident obesity risk. The *plus* tier was evaluated at the landmark-row level for 5-year prediction.

The *plus* tier used a bidirectional gated recurrent unit with decay (GRU-D) encoder with a hidden size of 64, time-gap decay, temporal self-attention, and a discrete-time hazard head. It was trained with the Adam optimizer using a learning rate of 1e-3, weight decay of 1e-4, a batch size of 4096, up to 50 epochs, and early stopping with patience 10 after self-supervised pretraining. The *plus* tier was evaluated as an exploratory extension for longitudinal risk prediction and is not ready for clinical implementation. In addition to the primary discrete-time hazard output, an alternative Breslow-based survival output formulation was evaluated; both formulations are reported with the primary *plus*-tier results (Table S3 in Multimedia Appendix 3). The *plus*-tier analyses were not repeated under training-set-only imputation and retain the original preprocessing scope.

Explainability Analysis

Tree Shapley additive explanations (SHAP) analysis was performed for the XGBoost component of the *lite* tier [24] to quantify feature contributions to model predictions. SHAP values were used as descriptive risk-domain attributions to support interpretation and risk communication.

XSL•FO
RenderX

Data Analysis

Baseline group differences were described using standardized mean differences. Approximate 2-tailed Welch *t* test statistics and degrees of freedom were reconstructed from reported means, SDs, and sample sizes; sex was compared using Pearson chi-square test. Performance measures were AUROC, AUPRC, Brier score, ECE, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) [25]. AUROC assessed discrimination; AUPRC summarized precision across recall levels and was interpreted relative to the event proportion. Brier score measured overall probability error, with lower values indicating better performance; the constant-risk reference was p(1-p), where p was the event proportion. ECE used 10 equal-frequency bins; DCA assessed net benefit at selected thresholds [26]. AUROC differences were tested using the the DeLong method. Percentile bootstrap intervals used 500 resamples for single-model metrics, 10,000 paired resamples for primary comparisons in Table S4A in Multimedia Appendix 3, and 2000 paired resamples for updated BMI-stratum and stricter-outcome comparisons in Tables S5A and S6C. Incremental value over the BMI-based benchmarks was assessed in 3 ways, as reported in corresponding Results subsections: overall differences in AUROC and AUPRC with integrated discrimination improvement; fixed-capacity reclassification, in which the *lite* tier and each benchmark flagged the same nominal top proportion of the horizon-specific test cohort (top 10% and top 20% at each of the 4 horizons, against both benchmarks; 16 comparisons) and event proportions were compared in the discordantly flagged groups; and incremental discrimination within narrow baseline BMI strata (<21, 21-22.9, and 23-24.9 kg/m²), with observed incidence described across model-specific risk tertiles (Tables S4A-S4B and S5A-S5C in Multimedia Appendix 3). Continuous net reclassification improvement was omitted because it is sensitive to predicted probabilities and calibration [27]. Discordant-group event proportions were compared using a 2-proportion z test, or the Fisher exact test if any observed cell count was below 5; binomial intervals used the Wilson method. Variance inflation factors assessed collinearity among final continuous predictors. Two-sided $P<.05$ indicated statistical significance. Analyses used Python with *numpy* 2.4.2, *pandas* 2.3.3, *scikit-learn* 1.8.0, *XGBoost* 3.2.0, *PyTorch* 2.7.1, *lifelines* 0.30.3, *SHAP* 0.51.0, *statsmodels* 0.14.6, *SciPy* 1.17.1, and *matplotlib* 3.10.8.

Sensitivity Analyses

Four groups of sensitivity analyses were conducted and are reported in the same order in the Results. First, prespecified supportive analyses assessed *lite*-tier performance across sex and baseline BMI strata, with inverse probability of censoring weighting, and after excluding underweight participants (Tables S7 and S8A in Multimedia Appendix 3). Second, the imputation fitting procedure was compared by refitting the *lite*-tier and survival models with imputation fitted to the pooled development data rather than to the training split only (Table S1 in Multimedia Appendix 3). Third, to address concerns about transient weight variation or measurement error, we repeated the *lite*-tier pipeline under 2 stricter outcome definitions while holding the source cohort and predictors fixed: BMI ≥25 kg/m² at 2 consecutive visits, and, as a complementary analysis, BMI ≥25.5 kg/m². These runs retained pooled-development imputation and regenerated participant-level splits using the same 70:15:15 ratio and random seed. For the consecutive-visit definition, the complete visit sequence was examined and the event dated at the first visit of the first confirmed pair. First-crossing events under the primary definition were classified by subsequent confirmation status. Because benchmarks were unavailable for these runs, BMI-only and BMI+age+sex models were fitted separately on each run's training split and evaluated on its test split; these were distinct fits from the primary benchmarks (Tables S6A-S6C in Multimedia Appendix 3). Fourth, a BMI-excluded *plus*-tier model was evaluated to assess whether longitudinal information beyond BMI retained a predictive signal when BMI was removed from the predictor set.

External Evaluation

External evaluation used an independent NHIS-linked health checkup cohort of 328,268 participants, with the index examination as baseline and harmonized BMI definitions for baseline nonobesity and incident obesity. Only the *lite*-tier XGBoost booster and its feature map were transferred, so the evaluation assessed this component rather than the full ensemble. The transferred booster had been trained on development data imputed using a model fitted to the pooled development data. At the external site, missing predictor values were filled using the feature map's fill_value entries; the PMM imputation pipeline was not transferred. Shared predictors used the same units and category definitions, requiring no recoding or unit conversion. HbA_{1c} was unavailable in NHIS but was not a final *lite*-tier predictor. The component was applied without refitting or recalibration, and AUROC, AUPRC, and Brier score were reported. External calibration plots, threshold-based operating characteristics, and DCA were unavailable; this evaluation therefore primarily assesses the discrimination of the previously exported XGBoost component.

Ethical Considerations

The study protocol was approved by the IRB of Kangbuk Samsung Hospital (KBSMC 2025-10-017). The requirement for informed consent was waived by the IRB because the analysis used retrospective, deidentified data from routine health checkups; for the NHIS-linked external cohort, the original approval and consent framework of the data provider applied. All data were deidentified before analysis, and access was restricted to authorized researchers within a secure research environment. No participant compensation was provided, as the study analyzed existing records and involved no direct participant contact.

Results

Study Population

The source dataset contained 960,938 participants and 3,819,655 health checkup records. Of these participants, 1407 lacked a BMI measurement, 531,112 had fewer than 3 health examinations, 122,765 did not form an eligible baseline examination sequence, and 73,971 were excluded for prevalent obesity at baseline or an early outcome event during the washout

XSL•FO
RenderX

window. The eligible development cohort therefore comprised 231,683 initially nonobese adults, including 38,094 who developed obesity (38,094/231,683, 16.4%; Figure 2). Training, validation, and test sets contained 162,177, 34,753, and 34,753 participants, respectively. Among the 145,688 participants eligible for 5-year classification, median follow-up was 7.9 (IQR 5.6-10.3 years; mean 7.6, SD 3.2 years), totaling 1,102,967 person-years. These follow-up summaries are participant-level estimates conditional on eligibility for that horizon. The primary 5-year test cohort included 21,783 participants and 4088 events (18.8%). Table 1 summarizes baseline characteristics.

Figure 2. Study flow diagram. Selection from 960,938 participants with 3,819,655 health checkup records to the development cohort of 231,683 initially nonobese adults, with exclusions at each step.



XSL•FO
RenderX

Table 1. Baseline characteristics of the development cohort (selected variables; the extended descriptive table is provided in Multimedia Appendix 4)[a].

Characteristics	Overall (N=231,683)	Nonevent (n=193,589)	Incident obesity (n=38,094)	*P* value	Test statistic (*df*)	SMD[b]
Age (years), mean (SD)	39.97 (7.66)	40.05 (7.79)	39.58 (6.95)	<.001	–11.8 (58,552)	–0.062
Male sex, n (%)	104,123 (44.9)	78,622 (40.6)	25,501 (66.9)	<.001	8918 (1)	0.548
BMI (kg/m²), mean (SD)	21.74 (2.03)	21.33 (1.93)	23.81 (0.99)	<.001	369.8 (104,837)	1.619
Waist circumference (cm), mean (SD)	77.15 (7.10)	76.01 (6.88)	82.91 (5.13)	<.001	225.6 (68,150)	1.138
WHtR[c], mean (SD)	0.46 (0.04)	0.46 (0.04)	0.49 (0.03)	<.001	168 (67,775)	1.045
SBP[d] (mmHg), mean (SD)	105.40 (11.68)	104.61 (11.58)	109.40 (11.35)	<.001	75 (54,844)	0.418
DBP[e] (mmHg), mean (SD)	68.19 (9.18)	67.68 (9.07)	70.75 (9.34)	<.001	58.9 (53,182)	0.333
Fasting glucose (mg/dL), mean (SD)	93.65 (12.22)	93.36 (12.13)	95.11 (12.58)	<.001	25 (52,958)	0.142
Triglycerides (mg/dL), mean (SD)	97.95 (62.19)	94.90 (59.81)	113.42 (71.12)	<.001	47.6 (49,246)	0.282
HDL-C[f] (mg/dL), mean (SD)	63.26 (15.74)	64.31 (15.79)	57.91 (14.36)	<.001	–78.2 (57,732)	–0.424
TyG[g] index, mean (SD)	8.29 (0.53)	8.26 (0.53)	8.45 (0.54)	<.001	63 (53,525)	0.350

[a]*P* values are descriptive given the large sample. Test statistics are approximate Welch *t* statistics (2-tailed) for continuous variables, reconstructed from the displayed group means, SDs, and sample sizes, and the Pearson chi-square statistic for male sex (*df*=1).

[b]SMD: standardized mean difference.

[c]WHtR: waist-to-height ratio.

[d]SBP: systolic blood pressure.

[e]DBP: diastolic blood pressure.

[f]HDL-C: high-density lipoprotein cholesterol.

[g]TyG: triglyceride-glucose.

Lite-Tier Performance

Table 2 summarizes 5-year performance for the principal framework components and BMI-based benchmarks. The *lite* tier achieved an AUROC of 0.918 and an AUPRC of 0.693, approximately 3.7 times the no-information reference (event proportion, 0.188). Its Brier score was 0.087 versus 0.152 for constant-risk prediction, and its ECE was 0.005. Extended metrics are reported in Table S9 in Multimedia Appendix 3.

Lite-tier discrimination was consistent across fixed horizons, with AUROCs of 0.939, 0.921, 0.916, and 0.918 at 1, 2, 3, and 5 years, respectively (Table 3). AUPRC increased from 0.168 to 0.693 as the event proportion increased across horizons. Table S2 in Multimedia Appendix 3 consolidates component, ensemble, and benchmark results across horizons; Table S3 in Multimedia Appendix 3 details *plus*-tier variants.

Using a threshold of 0.167, obtained by maximizing the Youden J statistic in the validation set, the *lite* tier achieved a sensitivity of 0.900 (95% CI 0.891-0.909), specificity of 0.788 (95% CI 0.782-0.794), PPV of 0.495 (95% CI 0.484-0.507), and NPV of 0.972 (95% CI 0.969-0.974) in the aligned 5-year test cohort. At this operating point 7432 of 21,783 participants were flagged, of whom 3680 developed obesity.

Collinearity among the 9 final continuous *lite*-tier predictors was limited: variance inflation factors ranged from 1.08 to 3.02 (BMI 2.91, WHtR 2.65, total cholesterol-to-HDL-C ratio 3.02), all below the conventional threshold of 5, and BMI and WHtR had the strongest pairwise correlation (r=0.78).

XSL•FO
RenderX

Table 2. Five-year performance of the principal framework components and BMI-based benchmarks[a].

Models	Cohort	Events, n	AUROC[b] (95% CI)	AUPRC[c] (95% CI)	Brier	ECE[d]
Lite tier (XGBoost[e]/LR[f] ensemble; n=21,783)	Aligned (5 years)	4088	0.918 (0.914-0.922)	0.693 (0.677-0.708)	0.087	0.005
BMI-only (n=21,783)	Aligned (5 years)	4088	0.913 (0.909-0.917)	0.679 (0.663-0.695)	0.089	0.010
BMI+age+sex (n=21,783)	Aligned (5 years)	4088	0.914 (0.910-0.919)	0.675 (0.659-0.692)	0.089	0.006
AFT[g] (log-normal; n=21,783)	Aligned (5 years)	4088	0.915 (0.911-0.919)	0.685 (0.669-0.701)	0.089	0.017
Plus tier (GRU-D[h] hazard; n=70,851)	All landmark rows	12,307	0.899	0.601	0.129	0.118
Plus tier (GRU-D, aligned visit-1; n=21,783)	Aligned visit-1	4088	0.912	0.666	0.156	0.173

[a]The *lite* tier, the BMI-based benchmarks, and AFT use the aligned 5-year participant-level cohort. The all-landmark *plus*-tier row has a different denominator and should be interpreted separately. The aligned visit-1 *plus*-tier row uses the same evaluation participants as the *lite* tier, although the models were developed using different training data structures. 95% CIs (percentile bootstrap, 500 resamples) are shown for the *lite* tier, the BMI-based benchmarks, and AFT.

[b]AUROC: area under the receiver operating characteristic curve.

[c]AUPRC: area under the precision-recall curve.

[d]ECE: expected calibration error.

[e]XGBoost: Extreme Gradient Boosting.

[f]LR: logistic regression.

[g]AFT: accelerated failure time.

[h]GRU-D: gated recurrent unit with decay.

Table 3. Lite-tier performance across fixed prediction horizons (test set)[a].

Horizon	Events, n	Event rate	AUROC[b] (95% CI)	AUPRC[c] (95% CI)	Brier	ECE[d]
1 year (n=33,376)	496	0.015	0.939 (0.933-0.945)	0.168 (0.145-0.196)	0.013	0.001
2 years (n=29,761)	1676	0.056	0.921 (0.916-0.926)	0.388 (0.366-0.414)	0.041	0.001
3 years (n=26,501)	2811	0.106	0.916 (0.912-0.921)	0.520 (0.499-0.541)	0.065	0.003
5 years (n=21,783)	4088	0.188	0.918 (0.914-0.922)	0.693 (0.677-0.708)	0.087	0.005
ever (n=34,753)	5701	0.164	0.888 (0.884-0.892)	0.588 (0.573-0.601)	0.092	0.002

[a]Values for n denote the horizon-specific test cohort with outcome ascertainment available. The ever-onset row describes events over available follow-up rather than a fixed horizon. 95% CIs are percentile bootstrap intervals (500 resamples).

[b]AUROC: area under the receiver operating characteristic curve.

[c]AUPRC: area under the precision-recall curve.

[d]ECE: expected calibration error.

Comparison With BMI-Based Benchmarks

The BMI-only benchmark achieved a 5-year AUROC of 0.913, AUPRC of 0.679, Brier score of 0.089, and ECE of 0.010. The *lite* tier's gains were modest: the AUROC difference was 0.005 (95% CI 0.003-0.006; 10,000 paired bootstrap resamples; DeLong $P<.001$), and the AUPRC difference was 0.013 (95% CI 0.006-0.020). The concordant DeLong interval was 0.004-0.006. Integrated discrimination improvement was 0.017 (95% CI 0.014-0.019), reported as supportive evidence because it depends on both models' absolute predicted probabilities. The BMI+age+sex model achieved an AUROC of 0.914 (95% CI 0.910-0.919), leaving an AUROC difference of 0.003 for the *lite* tier. WHtR-extended benchmarks were not repeated under training-set-only imputation and are reported separately (Table S10 in Multimedia Appendix 3). Table S4A provides the primary incremental comparisons. All tables are provided in Multimedia Appendix 3.

At a threshold of 0.192, the BMI-only benchmark achieved a sensitivity of 0.877 (95% CI 0.866-0.886), specificity of 0.796 (95% CI 0.790-0.802), PPV of 0.498 (95% CI 0.486-0.510), and NPV of 0.965 (95% CI 0.962-0.968). This operating point was retained from the original analysis and could not be rederived, because validation-split predictions were not retained for the benchmark models (Table S8B in Multimedia Appendix 3).

XSL•FO
RenderX

Fixed-Capacity Reclassification

At nominal top-10% capacity in the aligned 5-year cohort, the *lite* tier flagged 2179 participants and BMI-only flagged 2218 because boundary ties were retained. Obesity developed in 211/299 participants identified only by the *lite* tier (70.6%) and 209/338 identified only by BMI-only (61.8%; *P*=.02): 2 additional events with 39 fewer participants flagged. The BMI+age+sex comparison had no boundary ties, so both models flagged exactly 2179 participants. Events occurred in 232/325 identified only by the *lite* tier (71.4%) and 212/325 identified only by the benchmark (65.2%; *P*=.09), yielding 20 additional events at equal capacity. Per 1000 participants evaluated, these gains were 0.09 versus BMI-only and 0.92 versus BMI+age+sex; the BMI-only comparison also flagged 1.8 fewer participants. *Lite*-tier-only event proportions were higher in all 16 comparisons across horizons, benchmarks, and nominal capacities. Net event capture increased in 15 comparisons and decreased by 3 in 1. These overlapping comparisons provide supportive evidence rather than independent replications (Table S4B in Multimedia Appendix 3).

Incremental Discrimination Across Baseline BMI Strata

Incremental discrimination varied across baseline BMI strata. At BMI 21.0-22.9 kg/m² (6829 participants; 384 events), the *lite* tier's AUROC difference was 0.068 (95% CI 0.052-0.083) versus BMI-only and 0.035 (95% CI 0.023-0.048) versus BMI+age+sex. The increment therefore remained against the stronger benchmark; these model comparisons do not isolate individual predictor contributions. Incidence from the lowest to highest risk tertile was 1.2% (28/2276), 3.6% (81/2276), and 12.1% (275/2277) for the *lite* tier, 1.5% (34/2276), 4.3% (98/2276), and 11.1% (252/2277) for BMI+age+sex, and 1.9% (42/2256), 4.5% (101/2254), and 10.4% (241/2319) for BMI-only. The highest *lite*-tier tertile had a lower mean BMI than either benchmark's highest tertile (22.58 versus 22.67 and 22.68 kg/m²), despite higher incidence. At BMI 23.0-24.9 kg/m² (7691 participants; 3682 events), the differences were smaller: 0.009 (95% CI 0.005-0.014) versus BMI-only and 0.008 (95% CI 0.004-0.013) versus BMI+age+sex. This stratum contained 90.1% (3682/4088) of 5-year events, compared with 9.4% (384/4088) at BMI 21.0-22.9. Its tertile incidence patterns were similar for the *lite* tier and BMI-only, which does not establish agreement in individual rankings. Below BMI 21.0, only 22 events occurred and both difference intervals included zero. Restricting BMI variation reduced all 3 models' discrimination. These analyses were post hoc and remain exploratory (Tables S5A-S5C in Multimedia Appendix 3).

Calibration and DCA

The *lite* tier's internal calibration was close to ideal (slope 1.02; calibration-in-the-large 0.001; Figure 3).

Figure 3. Calibration of the *lite* tier at the 5-year horizon in the test set. Predicted probabilities are plotted against observed event rates for the Extreme Gradient Boosting/logistic regression ensemble; the dashed line denotes perfect calibration.

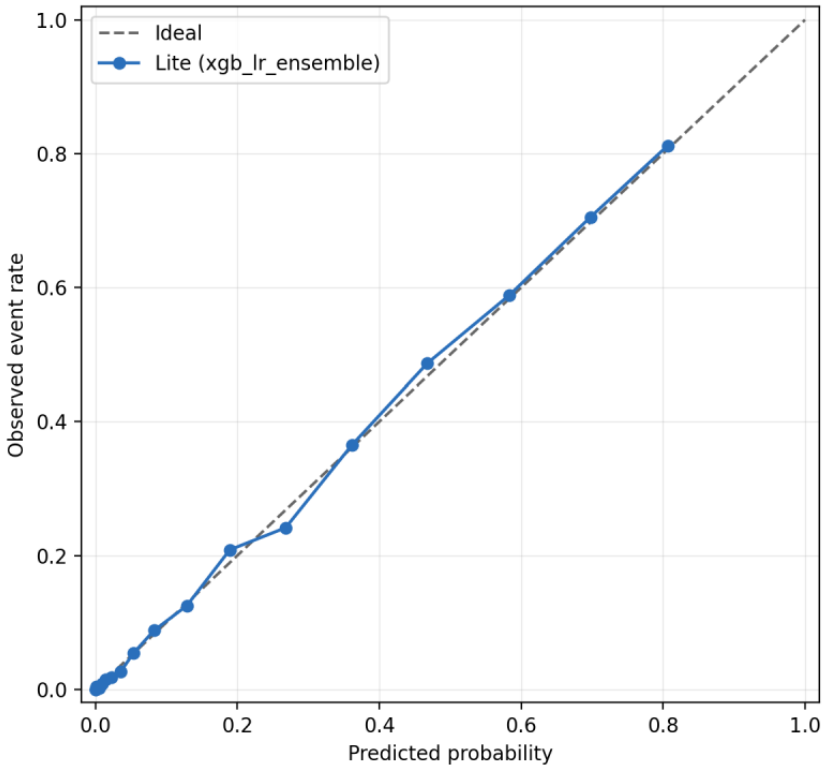
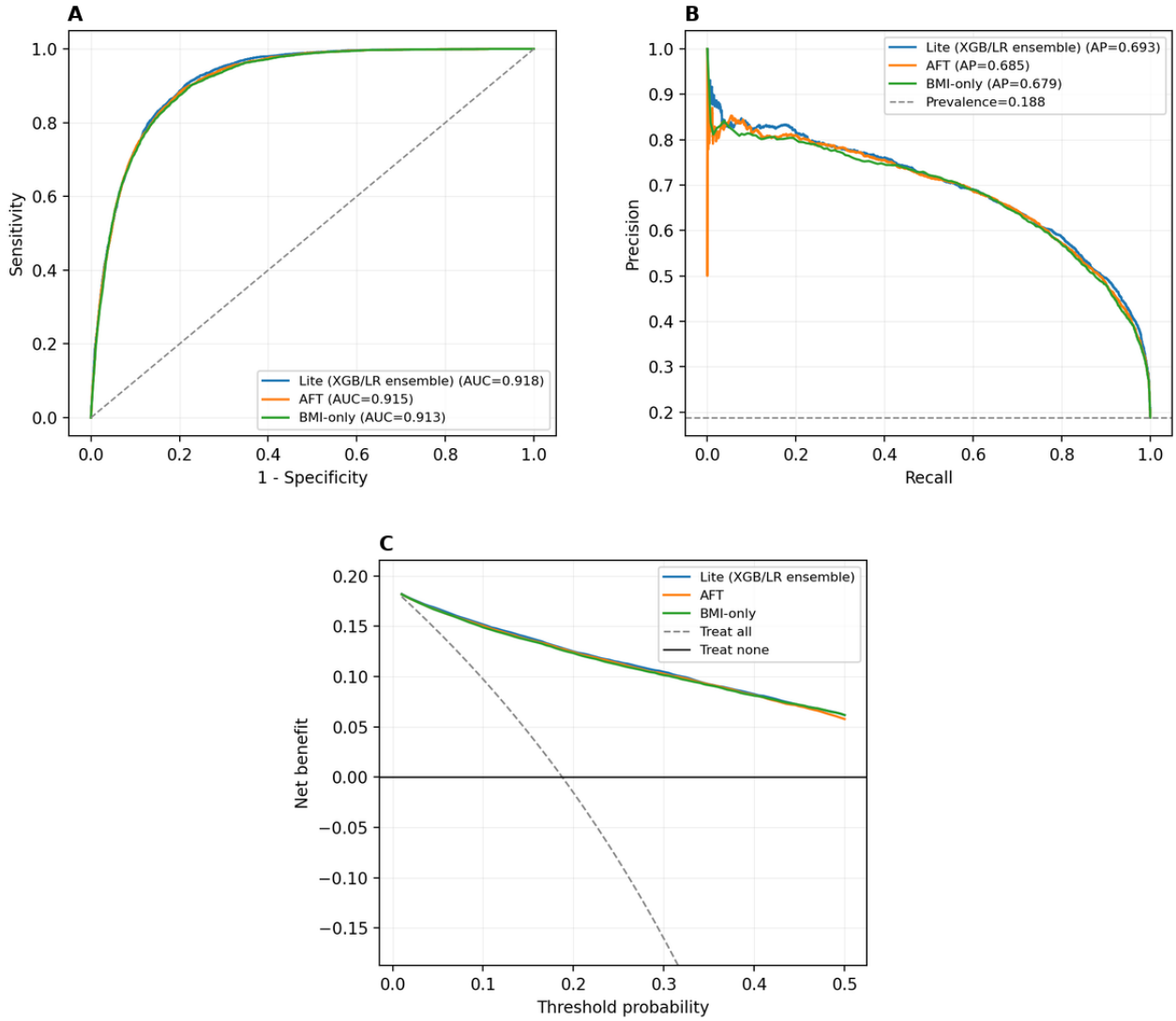


Figure 4 shows decision curves for the *lite* tier, AFT, and BMI-only. Net benefit at a threshold probability of 0.10 was 0.152 for the *lite* tier, 0.149 for BMI-only, 0.151 for BMI+age+sex, and 0.097 for treating all participants; at 0.20, values were 0.125, 0.123, 0.125, and –0.015, respectively. Across the thresholds examined, the *lite* tier's unrounded net-benefit differences were small and positive: 0.002-0.003 versus BMI-only and <0.001 to 0.002 versus BMI+age+sex. At 0.10, the *lite* tier flagged 8805 participants and captured 3857 events, compared with 9060 and 3827 for BMI-only. Fewer

XSL•FO
RenderX

participants flagged with more events captured was not a uniform pattern across thresholds. Threshold-probability and fixed-capacity analyses address different decisions and are reported separately.

Figure 4. Discrimination and decision-curve net benefit in the aligned 5-year test cohort (n=21,783). (A) Receiver operating characteristic curves, (B) precision-recall curves, and (C) decision-curve analysis for the *lite* tier ensemble, accelerated failure time model, and BMI-only benchmark. AFT: accelerated failure time, AUC: area under the curve, LR: logistic regression, XGBoost: Extreme Gradient Boosting.



Survival Component

For the 5-year fixed-horizon prediction in the test set, the AFT model achieved an AUROC of 0.915 and an AUPRC of 0.685, with a Brier score of 0.089 and an ECE of 0.017. The Cox model achieved an AUROC of 0.885, an AUPRC of 0.594, a Brier score of 0.115, and an ECE of 0.083. The survival model calibration-bin estimates are provided in Table S11 in Multimedia Appendix 3.

Across equal-sized quintiles of predicted AFT risk in the test set, 5-year obesity incidence increased monotonically. In the lowest-risk quintile, 6 of 4357 participants developed obesity within 5 years, corresponding to 0.1% (6/4357). In the highest-risk quintile, 2821 of 4357 participants developed obesity within 5 years, corresponding to 64.7% (2821/4357). These findings showed clear separation of 5-year obesity risk across AFT-derived strata (Table 4). Detailed AFT-derived risk-quintile characteristics are provided in Table S12 in Multimedia Appendix 3.

XSL•FO
RenderX

Table 4. Five-year obesity incidence across AFT-derived risk quintiles (test set)[a]. AFT: accelerated failure time.

Risk quintile	Events, n	Event rate	Median 5-year risk	Median BMI (kg/m²)	Median waist circumference (cm)	Median observed event time (years)
1 (n=4357)	6	0.001	0.001	18.97	69	3.3
2 (n=4356)	35	0.008	0.012	20.75	74	2.9
3 (n=4357)	230	0.053	0.057	22.11	78.5	3.2
4 (n=4356)	996	0.229	0.211	23.30	82	2.9
5 (n=4357)	2821	0.647	0.580	24.40	84.8	2

[a]Quintiles are equal-sized groups defined by predicted AFT risk within the test set; cut points were not derived in the validation set. Median 5-year risk, BMI, and waist circumference are within-quintile medians. Median observed event time summarizes time from the index visit to the first qualifying follow-up visit among participants with an event in each quintile; it is conditional on an event occurring during observation and is not a survival-analysis estimate of median time to onset. This estimate is unstable when few events occur (6 events in quintile 1). All values are from the primary analysis with training-set-only imputation.

Plus-Tier Performance

The *plus*-tier hazard formulation achieved an AUROC of 0.899, an AUPRC of 0.601, a Brier score of 0.129, and an ECE of 0.118 across 70,851 landmark rows with 12,307 events. On the same rows, the Breslow-based survival formulation achieved an AUROC of 0.902 and an AUPRC of 0.612, with lower probability error and calibration error (Brier 0.095; ECE 0.035). Table S3 (Multimedia Appendix 3) reports both formulations across cohort views.

At its original operating threshold, the *plus*-tier hazard formulation had sensitivity 0.889, specificity 0.774, PPV 0.453, and NPV 0.971 (Table S8B in Multimedia Appendix 3). These landmark-row estimates use a different denominator from the *lite* tier; the *plus* tier retains the original preprocessing scope. In the aligned visit-1 cohort (n=21,783), the hazard formulation achieved an AUROC of 0.912, Brier score of 0.156, and ECE of 0.173. Its Brier score exceeded the constant-risk reference of 0.152, supporting recalibration before use. Breslow-based estimates in the same cohort were more favorable (Brier 0.090; ECE 0.032; Table S3), although ECE remained higher than the *lite* tier's 0.005.

Explainability Findings

Tree SHAP analysis from the initial pooled-development analysis, not repeated under training-set-only imputation, attributed 66.7% of the total mean absolute SHAP value to BMI in the *lite*-XGBoost component. The next-ranked predictors were the total cholesterol-to-HDL-C ratio, WHtR, eGFR, TyG index, LDL-C, MAP, smoking status, sex, and age category. These attributions describe the model's reliance on BMI and additional risk domains; they do not isolate independent predictor effects. Feature-attribution results are in Tables S13A and S13B in Multimedia Appendix 3, with visual summaries in Multimedia Appendices 5-7.

Sensitivity Analyses

Lite-Tier Robustness Analyses

In the original pooled-development analysis, which was not repeated under training-set-only imputation, *lite*-tier discrimination varied across sex and baseline BMI strata (Table S7 in Multimedia Appendix 3). Discrimination was higher in women than in men (AUROC 0.934 versus 0.889) and decreased markedly among participants already close to the obesity threshold (baseline BMI 23.0-24.9 kg/m², AUROC 0.757, versus 0.860 for BMI below 23.0 kg/m²), consistent with the BMI-defined outcomes. The calibration-in-the-large remained near zero across most strata. In the inverse probability of censoring weighting analysis, the 5-year AUROC was 0.913, compared with 0.918 among participants with 5-year outcome ascertainment. Excluding underweight participants (BMI <18.5 kg/m²; 6.6%, 1428/21,783) yielded an AUROC of 0.910, indicating that discrimination was not driven by easily classified underweight individuals. Robustness and threshold-based sensitivity analyses are provided in Tables S8A and S8B in Multimedia Appendix 3.

Imputation Fitting Procedure

The 5-year *lite*-tier AUROC was 0.918 under both imputation procedures; the maximum difference across horizons was <0.001. The AFT and Cox models gave identical 5-year AUROCs under both procedures (0.915 and 0.885, respectively), with Brier scores unchanged to the reported precision (Table S1 in Multimedia Appendix 3).

Stricter Outcome Definitions

The stricter-definition runs retained pooled-development imputation. Under the 2-consecutive-visit definition, eligible 5-year landmarks decreased from 477,975 to 454,350, and the test cohort contained 20,886 participants and 2249 events (10.8%; AUROC 0.911). Under BMI ≥25.5 kg/m², 458,570 landmarks were eligible, and the test cohort contained 21,105 participants and 2261 events (10.7%; AUROC 0.906). These estimates describe performance in different eligible cohorts, not paired differences in a common sample. Of 38,094 first-crossing events, 48.5% (18,465/38,094) were confirmed at the next visit, 8.3% (3177/38,094) later, 24.2% (9213/38,094) were transient unconfirmed elevations, and 19% (7239/38,094) occurred at the final observed visit. Incremental discrimination also persisted against benchmarks fitted separately for each definition. Under the consecutive-visit definition, AUROC differences were 0.005 (95% CI 0.003-0.006) versus BMI-only and 0.003 (95% CI 0.002-0.005) versus BMI+age+sex. Under BMI ≥25.5 kg/m²,

XSL•FO
RenderX

the corresponding differences were 0.009 (95% CI 0.007-0.011) and 0.005 (95% CI 0.003-0.006). These benchmarks were refitted on each run's training split and differ from the primary fitted models. Tables S6A-S6C in Multimedia Appendix 3 report eligibility, events, confirmation status, and incremental comparisons.

Plus-Tier Sensitivity Analysis

In the BMI-excluded sensitivity model, the *plus* tier achieved an AUROC of 0.881 and an AUPRC of 0.561, with a Brier score of 0.125 and an ECE of 0.102. This suggested that repeated measurements other than BMI retained predictive information, although calibration remained less favorable than that of the *lite* tier.

External Evaluation

In the independent NHIS-linked external cohort, 65,491 of 328,268 participants developed incident obesity, corresponding to an event proportion of 20% (65,491/328,268). The externally evaluated *lite*-tier XGBoost component achieved an AUROC of 0.857 (95% CI 0.856-0.859). The AUPRC was 0.575, and the Brier score was 0.132.

These results provide component-level evidence of discrimination in an independent population-based cohort. They do not externally validate the revised full ensemble or the survival and *plus*-tier components.

Discussion

Principal Results

The framework integrates single-visit triage, temporal risk interpretation, repeated-measurement updating, and risk-domain explanation. The *lite* tier of our framework showed strong internal discrimination and calibration at 5 years. Its AUROC exceeded the conventional 0.9 benchmark for outstanding discrimination [28], although this alone does not establish clinical usefulness. AUPRC exceeded the event-proportion reference, while the Brier score was lower than the constant-risk reference; these comparisons assess absolute performance rather than the increment over BMI. Survival modeling described observed incidence across risk strata, while the *plus* tier explored repeated-measurement prediction. These complementary outputs provide a basis for prospective evaluation of risk communication in routine health checkups.

Comparison With Prior Work

Prior studies have examined obesity or overweight prediction [8,9], future weight gain [10,11], and incident obesity phenotypes [12], with longitudinal prediction also studied in other clinical settings [29-34]. Our main contribution is the integration of these related tasks within one health checkup-based framework. Selected obesity- and weight-gain studies reported best-model AUROCs of approximately 0.68 to 0.87 across different prediction tasks [9-11], including 0.75 for a 10% or greater BMI increase over 1 year in United States electronic medical records [10]. The *lite* tier achieved 0.918 internally; the XGBoost component achieved 0.917 internally and 0.857 externally. Differences in study populations, outcome definitions, and prediction horizons preclude direct comparison of these values. We therefore assessed calibration, decision-curve net benefit, and reclassification alongside discrimination.

Clinical Implications

The framework was designed to support complementary clinical functions. The *lite* tier provides an absolute 5-year risk estimate from a single visit to inform triage for earlier counseling or closer monitoring. Survival modeling adds temporal context, while the *plus* tier explores changes in risk as further visits become available. Feature attribution summarizes the risk domains contributing to the model output, providing contextual information for risk communication and follow-up discussions. A risk communication interface can present these outputs together. Their intended role is to inform clinical judgment; effects on decisions and outcomes require prospective evaluation.

Illustrative Web-Based Prototype

The framework itself consists of prediction and explanation modules; its intended use is within clinical decision support during routine health checkups, where risk estimates and their explanations would be delivered to examinees and clinicians through a web-based interface. To illustrate what such an implementation could look like, we built a web-based prototype as a demonstration rather than a deployed service. It accepts health checkup inputs and displays 5-year BMI-defined obesity risk, observed incidence in similar risk-score groups, and plain-language risk-domain explanations. Researcher screens show *lite*- and *plus*-tier performance metrics. Representative English-language screens are provided in Multimedia Appendix 8. The prototype illustrates presentation of the outputs; implementation effectiveness was not evaluated.

Incremental Value Over BMI-Based Benchmarks

The *lite* tier's overall advantage over BMI-based benchmarks was modest. Sensitivity analyses using pooled-development imputation showed small within-run AUROC gains over separately refitted BMI-based benchmarks under stricter definitions. At exactly equal top-10% capacity, the *lite* tier captured 20 additional events versus BMI+age+sex (0.92 per 1000 participants evaluated). The BMI-only comparison captured 2 additional events while flagging 39 fewer participants because boundary ties made capacities unequal. These absolute gains (Table S4B in Multimedia Appendix 3) and small net-benefit differences define the contribution's scale. SHAP describes predictor contributions to model outputs, not independent clinical benefit, although the represented domains are relevant to risk communication [4,5,15,16,22].

The strongest observed within-stratum increment occurred at BMI 21 to 22.9 kg/m² and remained against BMI+age+sex (AUROC difference 0.035; Tables S5A-S5C in Multimedia Appendix 3). This stratum included 31.4% (6829/21,783) of participants but 9.4% (384/4088) of events, whereas 90.1% (3682/4088) occurred at BMI 23.0 to 24.9, where gains were smaller. Risk differentiation before individuals approach the diagnostic threshold is a plausible application for preventive assessment. These exploratory findings support prospective testing of earlier risk communication; they do not establish a

XSL•FO
RenderX

prevention benefit. Predictors were recorded in comprehensive health checkups, but availability varies across settings and cost-effectiveness was not evaluated.

Interpretation of the Survival and Plus-Tier Components

Observed 5-year incidence ranged from 0.1% (6/4357) to 64.7% (2821/4357) across AFT strata. These estimates describe risk groups, not individual onset dates: BMI was measured at visits, so onset between visits was unobserved and visit intervals limit temporal precision. The *plus* tier provides exploratory updates from repeated measurements; its landmark evaluation depends on visit frequency and prediction-window definition. In the aligned visit-1 cohort, hazard-based probabilities required recalibration; the Breslow formulation improved calibration (Table S3 in Multimedia Appendix 3). Independent validation and calibration assessment remain necessary before clinical use.

Strengths

The study integrates complementary risk outputs within one framework using routinely recorded health checkup information. Its large cohort supports evaluation across multiple horizons, with participant-level separation of training, validation, and test data. Assessment extends beyond overall AUROC to parsimonious benchmarks, BMI-stratum comparisons, calibration, decision curves, reclassification, and stricter outcome definitions. Training-set-only imputation strengthens the primary internal *lite*-tier and survival analyses. The nationwide cohort of 328,268 participants provides independent evidence for the previously exported XGBoost component. Together, these analyses establish strong internal *lite*-tier performance and specify the additional information, evaluation scope, and remaining requirements for clinical application.

Limitations

The development cohort comprised regular health screening attendees and may differ from the general Korean adult population in risk distribution and obesity incidence. Although the *lite* tier uses 1 visit as input, eligibility required repeated examinations for washout and follow-up; performance in first-time or irregular attendees remains uncertain. BMI-defined obesity cannot distinguish fat mass, muscle mass, or fat distribution, and a single threshold crossing may reflect transient variation or measurement error. Confirmation-status analyses and 2 stricter definitions addressed this concern, with supportive discrimination and incremental-value findings in their respective eligible cohorts under pooled-development imputation (Tables S6A-S6C in Multimedia Appendix 3). The retrospective design and illustrative prototype do not establish improvements in counseling, follow-up, behavioral outcomes, or obesity prevention; these require prospective evaluation.

The consensus PMM table does not propagate between-imputation uncertainty into performance intervals, and eGFR was calculated using MDRD rather than the more recent Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) 2021 equation. External evidence concerns a previously exported *lite*-tier XGBoost component trained with pooled-development imputation and applied using fixed fill values. It does not establish transportability of the revised full ensemble. External calibration curves, decision-curve net benefit, threshold-based and stratum-specific performance, and independent evaluation of the survival component and the *plus* tier remain necessary. The NHIS data-use period has expired, preventing repeat evaluation under training-set-only imputation and construction of an external BMI-only benchmark. Whether the increment over BMI persists externally is therefore unknown. Internal stability across imputation procedures does not replace these external assessments.

Conclusions

Routine health checkup data supported an integrated framework for single-visit triage, temporal risk interpretation, exploratory repeated-measurement updating, and risk-domain explanation. The *lite* tier showed strong internal performance and modest additional discrimination beyond BMI-based benchmarks, with supportive findings under stricter outcome definitions. These findings support prospective evaluation of the integrated outputs for risk communication and follow-up prioritization, with clinical and preventive benefits yet to be established.

Funding

This research was supported by the Institute for Information and Communications Technology Planning and Evaluation (IITP; grant number: RS-2025-25443209).

Data Availability

The individual-level data used in this study are not publicly available because they contain sensitive health screening information. Aggregate supplementary tables and figures supporting this manuscript are included in this article and its additional files. Analysis code that does not contain individual-level health data may be made available from the corresponding author upon reasonable request and subject to institutional approval.

Generative AI Use

The authors declare that generative AI was not used in the preparation of this manuscript.

Authors' Contributions

Conceptualization: SJ, YSB, JHK
Data curation: SJ

XSL•FO
RenderX

Formal analysis: SJ
Funding acquisition: YSB
Investigation: JHK
Methodology: SJ, YJK
Project administration: YSB
Software: SJ
Supervision: YSB
Validation: YJK
Visualization: SJ
Writing – original draft: SJ
Writing – review & editing: YJK, YSB, JHK
YSB and JHK contributed equally as co-corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Completed TRIPOD+AI checklist for prediction model reporting.
[PDF File (Adobe PDF File), 155 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Completed STROBE checklist for observational cohort reporting.
[PDF File (Adobe PDF File), 112 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Tables S1-S13B: model performance, benchmarks, sensitivity analyses, feature attribution, and discrimination and incidence across body mass index strata.
[PDF File (Adobe PDF File), 848 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Extended baseline characteristics and standardized mean differences for the development cohort.
[PDF File (Adobe PDF File), 139 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Global Shapley additive explanations feature attribution for the lite-tier Extreme Gradient Boosting component at the 5-year horizon, retained from the original analysis and not repeated with imputation fitted on the training split only.
[PDF File (Adobe PDF File), 246 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Permutation importance of the plus-tier gated recurrent unit with decay model, retained from the original analysis and not repeated with imputation fitted on the training split only.
[PDF File (Adobe PDF File), 261 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Input saliency of the plus-tier gated recurrent unit with decay model, retained from the original analysis and not repeated with imputation fitted on the training split only.
[PDF File (Adobe PDF File), 282 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Representative English-language screens of the illustrative web-based prototype for risk communication within the clinical decision-support framework.
[PDF File (Adobe PDF File), 362 KB-Multimedia Appendix 8]

References

XSL•FO
RenderX

1. Hall KD, Kahan S. Maintenance of lost weight and long-term management of obesity. Med Clin North Am. 2018;102(1):183-197. [FREE Full text] [doi: 10.1016/j.mcna.2017.08.012] [Medline: 29156185]
2. Global BMI Mortality Collaboration, Di Angelantonio E, Bhupathiraju S, Wormser D, Gao P, Kaptoge S, et al. Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. Lancet. 2016;388(10046):776-786. [FREE Full text] [doi: 10.1016/S0140-6736(16)30175-1] [Medline: 27423262]
3. GBD 2015 Obesity Collaborators, Afshin A, Forouzanfar MH, Reitsma MB, Sur P, Estep K, et al. Health effects of overweight and obesity in 195 countries over 25 years. N Engl J Med. 2017;377(1):13-27. [FREE Full text] [doi: 10.1056/NEJMoa1614362] [Medline: 28604169]
4. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. Lancet. 2004;363(9403):157-163. [doi: 10.1016/S0140-6736(03)15268-3] [Medline: 14726171]
5. Haam J, Kim BT, Kim EM, Kwon H, Kang J, Park JH, et al. Diagnosis of obesity: 2022 update of clinical practice guidelines for obesity by the Korean Society for the Study of Obesity. J Obes Metab Syndr. 2023;32(2):121-129. [FREE Full text] [doi: 10.7570/jomes23031] [Medline: 37386771]
6. Kang H. Current status of the national health screening programs in South Korea. Korean J Fam Med. 2022;43(3):168-173. [FREE Full text] [doi: 10.4082/kjfm.22.0052] [Medline: 35610963]
7. Choi EK. Cardiovascular research using the Korean National Health Information Database. Korean Circ J. 2020;50(9):754-772. [FREE Full text] [doi: 10.4070/kcj.2020.0171] [Medline: 32725984]
8. Ferreras A, Sumalla-Cano S, Martínez-Licort R, Elío I, Tutusaus K, Prola T, et al. Systematic review of machine learning applied to the prediction of obesity and overweight. J Med Syst. 2023;47(1):8. [doi: 10.1007/s10916-022-01904-1] [Medline: 36637549]
9. Lin W, Shi S, Huang H, Wen J, Chen G. Predicting risk of obesity in overweight adults using interpretable machine learning algorithms. Front Endocrinol (Lausanne). 2023;14:1292167. [FREE Full text] [doi: 10.3389/fendo.2023.1292167] [Medline: 38047114]
10. Choong C, Xavier N, Falcon B, Kan H, Lipkovich I, Nowak C, et al. Identifying individuals at risk for weight gain using machine learning in electronic medical records from the United States. Diabetes Obes Metab. 2025;27(6):3061-3071. [doi: 10.1111/dom.16311] [Medline: 40069847]
11. Murtha JA, Birstler J, Stalter L, Jawara D, Hanlon BM, Hanrahan LP, et al. Identifying young adults at high risk for weight gain using machine learning. J Surg Res. 2023;291:7-16. [doi: 10.1016/j.jss.2023.05.015] [Medline: 37329635]
12. Eftekharzadeh A, Asghari G, Serahati S, Hosseinpanah F, Azizi A, Barzin M, et al. Predictors of incident obesity phenotype in nonobese healthy adults. Eur J Clin Invest. 2017;47(5):357-365. [doi: 10.1111/eci.12743] [Medline: 28294315]
13. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357:j2099. [FREE Full text] [doi: 10.1136/bmj.j2099] [Medline: 28536104]
14. Kengne AP, Beulens JW, Peelen LM, Moons KG, van der Schouw YT, Schulze MB, et al. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. Lancet Diabetes Endocrinol. 2014;2(1):19-29. [FREE Full text] [doi: 10.1016/S2213-8587(13)70103-7] [Medline: 24622666]
15. US Preventive Services Task Force, Curry SJ, Krist AH, Owens DK, Barry MJ, Caughey AB, et al. Behavioral weight loss interventions to prevent obesity-related morbidity and mortality in adults: US Preventive Services Task Force recommendation statement. JAMA. 2018;320(11):1163-1171. [FREE Full text] [doi: 10.1001/jama.2018.13022] [Medline: 30326502]
16. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002;346(6):393-403. [FREE Full text] [doi: 10.1056/NEJMoa012512] [Medline: 11832527]
17. Chang Y, Noh J, Cheon JY, Kim Y, Kwon YD, Ryu S. Self-rated health and risk of incident non-alcoholic fatty liver disease: a cohort study. Sci Rep. 2020;10(1):3826. [FREE Full text] [doi: 10.1038/s41598-020-60823-8] [Medline: 32123241]
18. Cho Y, Chang Y, Ryu S, Wild SH, Byrne CD. Baseline and change in serum uric acid level over time and resolution of nonalcoholic fatty liver disease in young adults: the Kangbuk Samsung Health Study. Diabetes Obes Metab. 2024;26(5):1644-1657. [doi: 10.1111/dom.15466] [Medline: 38303100]
19. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. [FREE Full text] [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
20. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453-1457. [FREE Full text] [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
21. Lee SY, Park HS, Kim DJ, Han JH, Kim SM, Cho GJ, et al. Appropriate waist circumference cutoff points for central obesity in Korean adults. Diabetes Res Clin Pract. 2007;75(1):72-80. [doi: 10.1016/j.diabres.2006.04.013] [Medline: 16735075]
22. Simental-Mendía LE, Rodríguez-Morán M, Guerrero-Romero F. The product of fasting glucose and triglycerides as surrogate for identifying insulin resistance in apparently healthy subjects. Metab Syndr Relat Disord. 2008;6(4):299-304. [doi: 10.1089/met.2008.0034] [Medline: 19067533]

XSL•FO
RenderX

23. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International Physical Activity Questionnaire: 12-country reliability and validity. Med Sci Sports Exerc. 2003;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
24. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017. Presented at: Proceedings of the Thirty-First Annual Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017:4765-4774; Long Beach Convention Center in Long Beach, California, United States. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
25. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-138. [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
26. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. [FREE Full text] [doi: 10.1177/0272989X06295361] [Medline: 17099194]
27. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. Epidemiology. 2014;25(1):114-121. [FREE Full text] [doi: 10.1097/EDE.0000000000000018] [Medline: 24240655]
28. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression 3rd ed. Hoboken, NJ. John Wiley & Sons; 2013.
29. Weng SF, Vaz L, Qureshi N, Kai J. Prediction of premature all-cause mortality: a prospective general population cohort study comparing machine-learning and standard epidemiological approaches. PLoS One. 2019;14(3):e0214365. [FREE Full text] [doi: 10.1371/journal.pone.0214365] [Medline: 30917171]
30. Artzi NS, Shilo S, Hadar E, Rossman H, Barbash-Hazan S, Ben-Haroush A, et al. Prediction of gestational diabetes based on nationwide electronic health records. Nat Med. 2020;26(1):71-76. [doi: 10.1038/s41591-019-0724-8] [Medline: 31932807]
31. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572(7767):116-119. [FREE Full text] [doi: 10.1038/s41586-019-1390-1] [Medline: 31367026]
32. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Health. 2020;2(4):e179-e191. [FREE Full text] [doi: 10.1016/S2589-7500(20)30018-2] [Medline: 33328078]
33. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18. [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
34. Saux P, Bauvin P, Raverdy V, Teigny J, Verkindt H, Soumphonphakdy T, et al. Development and validation of an interpretable machine learning-based calculator for predicting 5-year weight trajectories after bariatric surgery: a multinational retrospective cohort SOPHIA study. Lancet Digit Health. 2023;5(10):e692-e702. [FREE Full text] [doi: 10.1016/S2589-7500(23)00135-8] [Medline: 37652841]

Abbreviations

AFT: accelerated failure time
ALT: alanine aminotransferase
AST: aspartate aminotransferase
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
CKD-EPI: Chronic Kidney Disease Epidemiology Collaboration
DCA: decision-curve analysis
ECE: expected calibration error
eGFR: estimated glomerular filtration rate
GGT: gamma-glutamyl transferase
GRU-D: gated recurrent unit with decay
HbA1c: glycated hemoglobin
HDL-C: high-density lipoprotein cholesterol
IRB: institutional review board
KSHS: Kangbuk Samsung Health Study
LDL-C: low-density lipoprotein cholesterol
MAP: mean arterial pressure
MDRD: Modification of Diet in Renal Disease
NHIS: National Health Insurance Service
NPV: negative predictive value
PMM: predictive mean matching
PPV: positive predictive value

XSL•FO
RenderX

S-CUBE: Samsung Medical Center Unified Big-data Engine
SHAP: Shapley additive explanations
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus Artificial Intelligence
TyG: triglyceride-glucose
WHtR: waist-to-height ratio
XGBoost: Extreme Gradient Boosting

Edited by L MacNeill; submitted 24.Jul.2026; peer-reviewed by J-W Chun, G Jerome, O Reges; comments to author 01.Sep.2026; revised version received 13.Oct.2026; accepted 21.Oct.2026; published 19.Nov.2026

Please cite as:
Jang S, Kim YJ, Bae YS, Kang J-H
Comprehensive Clinical Decision-Support Framework for Obesity Prevention Using Large-Scale Health Checkup Data: Development and Evaluation Study
J Med Internet Res 2026;28:e107821
URL: https://www.jmir.org/2026/1/e107821
doi: 10.2196/107821
PMID:

©Suho Jang, Yoon Ji Kim, Ye Seul Bae, Jae-Heon Kang. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 19.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX