Review | Received 22 February 2026; Revised 26 March 2026; Accepted 28 March 2026; Published 10 April 2026
https://doi.org/10.55092/acr20260003

Individual risk prediction and cancer prevention for the digital health era

So-Yoon Lee[1], Woo-Kyoung Shin[2], Sukhong Min[3], Hyobin Lee[1,4], Jeongheon Kim[1,3], Ji-Yeob Choi[5] and Daehee Kang[1,4,*]

[1] Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea
2 Division of Food and Pharmaceutical Technology, Mokwon University, Daejeon, Republic of Korea
3 Community Health Innovation Center, Seoul National University, Seoul, Republic of Korea
4 Integrated Major in Innovative Medical Science, Seoul National University Graduate School, Seoul, Republic of Korea
5 Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, Republic of Korea

* Correspondence author; E-mail: dhkang@snu.ac.kr.

Highlights:

- Cancer prevention is fundamentally a prediction problem.
- Calibrated absolute risk makes individual prediction actionable.
- Digital health enables monitored, updatable prevention prediction.

Abstract: Cancer prevention is, in practice, a prediction problem. Decisions about risk reduction, screening, and survivorship depend on forecasts of incident cancer, progression, recurrence, and treatment-related harms, and on whether earlier or more intensive action improves the benefit-harm balance. For prevention-oriented use, prediction tools should provide calibrated absolute-risk estimates over a defined follow-up period, be anchored to local incidence and competing mortality when relevant, and be embedded in explicit clinical or public health pathways in which prespecified thresholds trigger actionable steps. These steps may include intensified lifestyle support, eligibility assessment for chemoprevention, risk-adapted screening decisions regarding start age, interval, or modality, and referral for confirmatory evaluation. Using breast cancer as the most mature archetype, we illustrate how individualized risk estimates can be linked to multiple prevention levers while making the trade-offs of threshold-based strategies explicit, including induced follow-up procedures and the potential for overdiagnosis. We then synthesize requirements for prevention-ready prediction, spanning longitudinal follow-up with registry-linked outcome ascertainment, absolute-risk estimation anchored to local incidence and competing mortality, evaluation of calibration overall and in relevant subgroups, transparent accounting of threshold-induced follow-up procedures and harms, and planned recalibration and updating as baseline risk and practice patterns change. As cohort-enabled illustrations in the Health

Examinees (HEXA) infrastructure, we describe a breast cancer prediction benchmark, CancerFree, a questionnaire-first, registry-anchored multi-cancer framework for site-specific risk-to-action specification, alongside an AI-derived diet layer that converts food frequency questionnaire data into interpretable dietary-pattern features for scalable prevention-oriented profiling. Across emerging data layers, added model complexity should be justified not by discrimination alone, but by demonstrable gains in calibrated absolute-risk estimation and in the benefit–harm balance of the pathways they are intended to inform.

Keywords: cancer prevention; cancer risk prediction; absolute risk estimation; digital health; risk-adapted screening; prediction model calibration; risk stratification; precision prevention

1. Introduction

Cancer remains a major global health burden. The Global Burden of Disease (GBD) Study 2023 cancer analysis estimated 18.5 million incident cancer cases and 10.4 million cancer deaths worldwide in 2023 (excluding non-melanoma skin cancer), and attributed a substantial fraction of cancer deaths to modifiable risk factors. The same analysis projected further increases through 2050, with the largest growth expected in lower-resource settings [1]. These projections help explain why cancer control strategies continue to emphasize prevention and earlier detection. The WHO Guide to Cancer Early Diagnosis positions early diagnosis as a programmatic priority by strengthening timely recognition, diagnosis, and treatment once disease is suspected, thereby complementing rather than replacing screening programs. Major national initiatives also highlight prevention and early detection as policy priorities, reflecting a shared assumption that shifting effective action earlier can improve outcomes at scale [2,3].

In this context, cancer control is often a prediction problem. Many routine decisions, including whether and when to initiate screening, whether to recommend risk-reducing medications, and how intensively to follow survivors, depend on forecasts about future events such as incident cancer, progression, recurrence, second primaries, or treatment-related harms. Formal prediction models make these forecasts explicit by combining multiple predictors to estimate an individual's probability of a specified outcome within a defined time period, rather than relying on qualitative or implicit rules alone [4]. For prevention, individual risk prediction becomes actionable only when linked to a defined decision context: a target population, a defined follow-up period, for example 5 or 10 years, and a specific action triggered by a prespecified rule. Such actions may include referral to or intensification of lifestyle programs, eligibility assessment for chemoprevention, modification of screening start age or interval, or escalation to more sensitive modalities in selected groups. In practice, predicted risks may be communicated directly as probabilities and, when programs require discrete choices, summarized into risk categories for reporting, such as absolute thresholds or percentiles within age strata. These categories are best viewed as implementation constructs rather than biological thresholds; accordingly, thresholds should be defined and interpreted within the local clinical or public health pathway [4]. Because preventive actions are often triggered by such thresholds, model evaluation for prevention cannot rely on discrimination alone. Transparent reporting of prediction model development and validation, including TRIPOD and TRIPOD+AI for models using machine learning or other AI methods, and structured appraisal of risk of bias and applicability, including PROBAST and PROBAST+AI, reflect

longstanding concerns that prediction models may appear promising while remaining unreliable in specific settings or subgroups [5–8]. For prevention-oriented use, calibration, defined as agreement between predicted and observed risks over the intended follow-up period, is especially important because miscalibration can shift individuals across eligibility thresholds and change who receives screening or further diagnostic evaluation [9].

Harms from follow-up evaluation are central to prevention, particularly false positives and the subsequent follow-up investigations and procedures (hereafter, follow-up procedures), as well as overdiagnosis. Overdiagnosis, detection of cancers that would not have caused symptoms or death in the absence of detection, can increase diagnostic procedures and treatment without improving outcomes. Accordingly, prevention-oriented prediction should be discussed and evaluated with explicit attention to the consequences of applying risk categories or thresholds at scale, including false positives, additional diagnostic evaluations, and potential overtreatment [10].

Finally, the digital health era increases both the feasibility and the evidentiary expectations for prevention-oriented prediction. Linkage of cancer registries, screening databases, and electronic health record/claims data can support estimation and monitoring of predicted risks over clinically relevant periods, but it also underscores that "prediction models are not once validated, always valid" [9]. Changes in baseline incidence, population characteristics, measurement practices, and clinical protocols can lead to performance heterogeneity across settings and calibration drift over time. Recent methodological work therefore emphasizes validation as an ongoing process, with monitoring and updating when appropriate, alongside transparent reporting for studies using routinely collected health data [9,11,12]. In prevention settings, several evaluation principles recur across applications: calibrated absolute-risk estimation over a defined horizon, explicit linkage of risk thresholds to actions, attention to downstream consequences such as false positives, diagnostic work-up, and overdiagnosis, transportability and recalibration across settings and subgroups, and implementation feasibility, including equity, governance, and follow-up capacity [9,11,12].

Table 1 summarizes how individual risk prediction can support prevention actions across the primary, secondary, and tertiary prevention continuum, along with the action-relevant outputs and key cautions. Taken together, Table 1 serves as the organizing framework for this review, emphasizing that across prevention stages the key translational requirement is not discrimination alone, but action-relevant absolute risk embedded in explicit pathways and evaluated against harms, follow-up burden, and implementation constraints.

In the remainder of this review, we summarize how cancer risk prediction has evolved, from classical epidemiologic models based on readily measured risk factors to imaging, genomics, and digital data, and how these developments relate to prevention decisions. We then use breast cancer as the most mature archetypal setting in which individual risk prediction has been linked to multiple actions across primary prevention, screening, and survivorship, while recognizing that other cancers differ in pathway structure, baseline risk, screening modality, and harm–benefit profile. We next discuss why prospective cohorts and registry linkage remain foundational for prevention-relevant prediction and illustrate cohort-enabled examples using the Health Examinees (HEXA) infrastructure. Throughout, risk refers to calibrated absolute risk over a prespecified horizon, typically 5–10 years, unless stated otherwise.

Existing reviews of cancer risk prediction often emphasize model development and discrimination metrics. In contrast, we organize prediction explicitly around primary, secondary, and tertiary prevention

decisions and focus on what makes risk estimates usable in practice: absolute-risk estimates aligned to the intended follow-up period, clear decision thresholds linked to actions, and transparent reporting of the additional investigations and procedures that implementation may trigger. Using breast cancer as an archetype and cohort-enabled examples from HEXA, including performance benchmarks for a breast cancer model and illustrative modules for multi-cancer prediction and diet, we aim to provide a pathway-grounded framework for prevention-oriented translation.

Table 1. Prediction across the cancer prevention continuum: prevention actions, action-relevant outputs, and key cautions.

Prevention Stage	Prevention Action Examples	Actionable Outputs	Key Cautions†
Primary prevention (risk reduction)	• Risk-based lifestyle support (intensity, timing) • Eligibility assessment for chemoprevention‡ • Infection control (e.g., HPV vaccination, *H. pylori* eradication)	• Predicted risk over a defined follow-up period • Uncertainty around risk estimates • Calibration in the target population (and key subgroups), especially when predefined thresholds are used	• Actionability and adherence limits • Equity and access constraints
Secondary prevention (early detection)	• Risk-adapted screening (start age, interval, modality) • Referral to confirmatory testing and diagnostic evaluation (e.g., FIT→colonoscopy, LDCT, MRI)	• Predicted risk over a defined follow-up period • Calibration within subgroups relevant to screening decisions • Setting-specific validation and (when needed) recalibration	• Overdiagnosis and false positives • Follow-up procedures • Calibration drift and sensitivity to screening/diagnostic protocols
Tertiary prevention (survivorship)	• Risk-adapted survivorship care (follow-up intensity, recurrence/second primary risk, late effects/toxicity) • Monitoring and supportive care	• Prognostic risk over a defined follow-up period in treated populations • Era-specific calibration • Evaluation at defined thresholds for action (when used)	• Mixed evidence for outcome improvement • Modest incremental benefit • Changes in treatment over time • Privacy and governance

† Definitions (selected): Calibration drift—divergence between predicted and observed risks over time as baseline incidence, measurement, or practice patterns change; Transportability—performance and calibration may not generalize without local validation and recalibration; Mixed evidence—heterogeneous or limited evidence that risk-adapted survivorship strategies improve outcomes in routine practice; Modest benefit—small incremental gains relative to added complexity, cost, or follow-up tests/procedures.
‡ Chemoprevention is classified here as primary prevention when used in cancer-free high-risk individuals; pharmacologic strategies after diagnosis to reduce recurrence or second primaries fall under tertiary prevention.
Abbreviations: FIT, fecal immunochemical test; LDCT, low-dose computed tomography; MRI, magnetic resonance imaging; HPV, human papillomavirus.

2. Primary prevention: risk reduction and prevention eligibility

Primary prevention decisions concern who should receive intensified lifestyle support and who may benefit from risk-reducing medication while still cancer-free. In this setting, prediction is most useful when it provides absolute-risk estimates and supports shared decisions that weigh feasible behavior change and medication adverse effects against expected benefit.

2.1. Breast cancer as an archetype

Across Sections 2–4, we use breast cancer to illustrate how prediction informs decisions across the prevention continuum, beginning here with primary prevention levers and then turning to screening and survivorship prediction. Breast cancer has one of the most extensive evidence bases for prevention-oriented individual risk prediction because it combines high population burden, substantial heterogeneity in risk and disease course, and multiple prevention actions that can be linked to predicted risk. GLOBOCAN 2022 reported 2,296,840 incident breast cancer cases and 666,103 deaths worldwide in 2022 [13]. Survival varies markedly by stage at diagnosis. In U.S. SEER data summarized in Cancer statistics, 2026, 5-year relative

survival for female breast cancer diagnosed during 2015 to 2021 was greater than 99% for localized disease, 87% for regional disease, and 33% for distant disease [14].

Breast cancer also provides a clear setting in which the harms of screening and incidental detection have been closely examined [10,15]. False positives and diagnostic evaluations are common, and overdiagnosis is a recognized consequence of detecting indolent disease that would not have become clinically consequential during a person's lifetime [10]. Evidence for overdiagnosis is most directly assessed in randomized trials with long-term follow-up, and population signatures include rising incidence with stable mortality, or rising early-stage incidence with stable late-stage incidence [10]. These considerations support prevention strategies that use prediction to target interventions and that report the expected follow-up procedures implied by particular risk thresholds [10,15].

Over the past four decades, breast cancer prediction research has expanded from classical risk factor models intended for long-term counseling and prevention eligibility, to screening-context models that incorporate imaging and prior procedures for near-term decisions, and then to additional layers that include germline genetic information and imaging-derived features. Reviews of breast cancer prediction models document both the breadth of available tools and recurring limitations in reporting and transportability, including incomplete calibration assessment and variability in performance across settings [16–18]. Breast cancer therefore provides a concrete setting in which to review how individual risk prediction has evolved alongside prevention pathways and why model evaluation in prevention contexts often emphasizes calibration and the implications of applying risk thresholds, not discrimination alone [17,19].

2.2. Primary prevention: modifiable risk reduction

Primary prevention in breast cancer is anchored in modifiable behaviors and exposures, and risk prediction can support prevention by prioritizing counseling and structured support for those at higher predicted risk. Evidence syntheses and dose-response meta-analyses have consistently linked higher physical activity to lower breast cancer risk, alcohol intake to increased risk in a dose-response manner, and adiposity to increased risk across multiple cancer sites, including breast cancer [20–22]. These modifiable factors are routinely measured in many health examination settings, which makes them practical candidates for prevention-oriented risk prediction and for targeting behavioral programs.

More broadly, prevention-oriented prediction should not be restricted to lifestyle factors alone. Depending on the cancer site and the prevention pathway, relevant actionable domains may also include infections and vaccination [23,24], environmental and occupational carcinogen exposure [25], and circadian or sleep-related exposures when they can be measured with adequate validity and linked to feasible interventions or exposure-control strategies [26–28]. In these settings, the translational question is not simply whether a factor is associated with cancer risk, but whether incorporating it improves action-relevant, calibrated absolute-risk estimation or supports programmatic prevention. Psychosocial factors, including chronic stress and affective states, may influence cancer-relevant behaviors and healthcare engagement [29,30], and have plausible biobehavioral links to cancer progression [31,32]; however, evidence for a direct effect on incident cancer risk remains mixed [33–35]. Their inclusion in prediction models should therefore depend on repeated, validated measurement and demonstrated incremental value for calibrated risk estimation and prevention decisions.

Prospective evidence from the Health Examinees-Gem cohort aligns with several of these directions and provides setting-specific estimates using routinely measured exposures. In HEXA, weight gain after age 35 years was associated with higher breast cancer risk, with the largest relative increases observed for premenopausal breast cancer in analyses stratified by menopausal status [36]. Diet-related analyses in HEXA-Gem further support the feasibility of using routine health examination-based dietary measures for prevention-oriented modeling. In HEXA-Gem, dietary indicators were associated with breast cancer risk across staple grain type, dairy intake, and processed meat consumption [37–39]. When multiple behaviors were combined into guideline-based scores, higher adherence to the 2018 WCRF/AICR cancer prevention recommendations in HEXA-Gem was associated with lower breast cancer incidence in women, with a multivariable adjusted hazard ratio of 0.80 (95% CI 0.67 to 0.95) comparing the highest with the lowest adherence category [40]. In pooled Asia Cohort Consortium analyses, body mass index was positively associated with postmenopausal breast cancer risk, whereas associations for smoking and alcohol varied by birth cohort [41,42].

Diet is a highly modifiable factor that influences chronic disease pathways but is difficult to incorporate into prediction models because individual nutrients and foods are correlated and measured with substantial noise. Dietary pattern analysis addresses this problem by treating the overall diet as the exposure rather than isolating individual components [43]. Building on this approach, we applied machine learning methods to derive dietary patterns from food frequency questionnaire data, integrating autoencoder-based dimensionality reduction with clustering techniques. This framework was evaluated for robustness and reproducibility across datasets, including assessment against clinically relevant correlates and application in independent populations with comparable food frequency questionnaire measures [44]. The resulting patterns provide an interpretable exposure-representation layer that can function both as a standalone summary of diet and as an upstream predictor for cancer risk prediction [44]. The framework also allows individual-level assignment to derived dietary patterns, thereby supporting threshold-based prevention decisions and counseling.

A second prevention lever is pharmacologic risk reduction in cancer-free individuals at elevated risk. The USPSTF recommends offering risk-reducing medications, including tamoxifen, raloxifene, and aromatase inhibitors, to women aged 35 years or older at increased breast cancer risk who are at low risk for adverse effects, and notes that absolute benefits are higher for women with a predicted 5-year risk of about 3% or greater [45]. Real-world use remains low. In a PLCO-Medicare-linked cohort with BCRAT risk estimates, use of any risk-reducing medication during 2014 to 2019 was 3.6% overall and 5.3% among women with a 5-year risk of at least 3% [46]. In a two-site prospective study offering low-dose tamoxifen after specialist counseling, 74% initiated low-dose tamoxifen and 78.2% of initiators remained on therapy at 1 year [47].

Historically, primary prevention is closely linked to the earliest generation of breast cancer risk prediction models, which used established epidemiologic predictors to estimate long-term probabilities of developing breast cancer. The Gail model and subsequent extensions, as well as integrated models such as Tyrer-Cuzick, were explicitly designed to support counseling and prevention decisions using predicted risk over defined follow-up periods [48,49]. National adaptations such as KoBCRAT illustrate how local incidence and risk factor distributions can be incorporated to improve population alignment for counseling in specific settings [50]. Table 2 summarizes representative breast cancer prediction tools that illustrate key shifts in predictor domains and intended prevention decisions. This overview is not intended

to be exhaustive; additional inventories and comparative assessments are available in recent systematic reviews [16–18].

3. Secondary prevention: early detection and screening-context prediction

Secondary prevention decisions include when to start screening, how often to screen, which modality to use, and when to refer for confirmatory testing after an abnormal screen. In this setting, prediction is useful only when risk estimates are linked to explicit screening actions and when evaluation reports not only cancer detection, but also the false positives and follow-up procedures triggered by threshold-based strategies, such as supplemental imaging and biopsies [10,15,51].

Large pragmatic trials have been designed to test risk-adapted breast screening at scale [52,53]. In the WISDOM randomized clinical trial, women aged 40 to 74 years were assigned either to annual mammography or to a risk-based strategy incorporating sequencing of 9 susceptibility genes, a polygenic risk score, and the Breast Cancer Surveillance Consortium version 2 model [53]. The risk-based arm used four recommendations: highest-risk women, defined as those with a 5-year risk of at least 6% or a high-penetrance pathogenic variant, received alternating mammography and MRI every 6 months plus counseling; elevated-risk women, defined as those in the top 2.5% of the age-specific risk distribution, received annual mammography plus counseling; average-risk women received biennial mammography; and low-risk women aged 40 to 49 years with a 5-year risk below 1.3% received no screening until their risk reached 1.3% or they turned 50 years old [53]. Among 28,372 randomized participants, the trial reported noninferiority for stage IIB or higher cancers in the risk-based group compared with annual screening, fewer mammograms in the risk-based group, and no statistically significant reduction in biopsy rates [53]. These findings support the feasibility of pathway-based risk stratification, while also underscoring that the practical value of such strategies depends on how risk thresholds change downstream testing and follow-up burden.

Breast cancer also illustrates why prediction methods changed when screening programs expanded. Risk tools can be grouped into population-based models intended for general-population counseling and eligibility decisions, and screening-context models developed in mammography cohorts that incorporate imaging and prior procedures to estimate near-term risk among screened women [16,17]. Pedigree-based genetic counseling models are not included here because their target setting and input requirements differ from those of population-based and screening-context prevention tools. The Breast Cancer Surveillance Consortium models, for example, incorporate age, family history, BI-RADS mammographic density, and history of breast biopsy or procedures to estimate predicted risk over time periods that are directly relevant to screening pathways [54,55]. These models were developed within screening cohorts and are intended to inform questions such as whether supplemental screening is warranted or whether a given screening interval is appropriate for a person with high density and prior biopsy history.

Table 2 summarizes representative breast cancer prediction tools that illustrate these shifts in predictor domains and intended prevention decisions.

More recent work has extended screening-context prediction using imaging-derived features and learning-based approaches. A digital breast tomosynthesis risk model reported validation discrimination around 0.82 (AUC) when combining vendors and projected risk using U.S. incidence and competing mortality rates [61]. Deep learning models that use screening mammograms to estimate future breast cancer risk have also been evaluated across multiple populations. For example, the Mirai model was

evaluated in independent test sets from the United States, Sweden, and Taiwan and reported C-indices ranging from 0.76 to 0.81, with higher 5-year AUC than Tyrer-Cuzick in the reported evaluations [62]. For prevention-oriented use, however, the central question is not simply whether these approaches improve rank-order discrimination, but whether they provide well-calibrated time-specific risk estimates in the intended setting and improve screening decisions without disproportionate increases in follow-up procedures [17].

Table 2. Selected breast cancer risk prediction tools for prevention, by chronology and predictor domain (population- and screening-based tools).

Year	Model / tool	Time period	Data source	Key predictors	Primary use	Reference
1989	Gail/BCRAT	5y, lifetime	Case–control + population incidence rates	Age; reproductive history; biopsy; first-degree FH	Counseling; chemoprevention eligibility	[48]
1996	Rosner–Colditz (NHS log-incidence)	5y/10y, lifetime	Prospective cohort	Life-course reproductive and hormonal history	Absolute-risk estimation	[56]
2004	IBIS/Tyrer–Cuzick	10y, lifetime	Model synthesis	FH; personal risk factors	Risk-adapted screening planning	[49]
2006	BCSC screening model	1y	Screening-based cohort	Age; screening-context factors (BI-RADS density, FH, prior biopsy/procedure)	Screening-context risk prediction	[54]
2008	BCSC density model	5y	Screening-based cohort	Age; BI-RADS density; FH; prior biopsy (± race/ethnicity)	Supplemental screening discussion	[55]
2013	KoBCRAT	5y, lifetime	Case–control + population incidence rates	Classical risk factors	Population-calibrated counseling	[50]
2016	BCSC + PRS integration	5y	Nested case–control	BCSC factors; PRS	Screening reclassification	[57]
2019	Subtype-aware PRS (BCAC)	Relative risk; projection to absolute risk	Consortium case–control	Genome-wide PRS (subtype refined)	Risk prediction	[58]
2020	PRS portability (Asian women)	Relative risk; calibration needed for absolute risk	Multi-study evaluation (case–control + cohort)	European PRS evaluated in Asian ancestry	Portability assessment	[59]
2021	PRS + classical risk factors (BCAC)	Relative risk; projection to absolute risk	Consortium case–control	PRS; classical risk factors	Joint stratification	[60]
2022	DBT short-term imaging risk model	Short-term	Screening-based cohort	DBT imaging features	Short-term triage	[61]

Abbreviations: BCAC, Breast Cancer Association Consortium; BCRAT, Breast Cancer Risk Assessment Tool (Gail model); KoBCRAT, Korean Breast Cancer Risk Assessment Tool (Gail/BCRAT-based model recalibrated using Korean risk-factor estimates and Korean incidence/mortality data); BCSC, Breast Cancer Surveillance Consortium; DBT, digital breast tomosynthesis; FH, family history; PRS, polygenic risk score.
Note: Pedigree-based genetic counseling models (e.g., BRCAPRO, BOADICEA) are not included because they are designed for high-risk family settings and require detailed pedigree information, which differs from the data and decision context of population- and screening-based prevention tools. KoBCRAT is a Korean adaptation of the Gail/BCRAT framework, recalibrated with Korean data for population-calibrated absolute-risk counseling.

Breast cancer also provides a clear example of how germline genetics can shape prevention pathways, while also illustrating that population impact depends on both effect size and prevalence. High-penetrance pathogenic variants such as BRCA1 and BRCA2 identify individuals at very high risk and support clearly defined prevention actions, including intensified surveillance and consideration of risk-reducing surgery in selected settings. Large prospective analyses of BRCA1 and BRCA2 carriers estimate cumulative breast cancer risk to age 80 years of approximately 72% for BRCA1 and 69% for

BRCA2 [63]. For PALB2, estimates of absolute breast cancer risk by age 70 years vary by family history, ranging from roughly one-third without family history to more than one-half with strong family history [64]. Risk-reducing mastectomy has been associated with substantial reductions in breast cancer incidence among high-risk women, supporting the clinical relevance of these pathways [65].

At the population level, genome-wide association studies and large consortia expanded discovery beyond rare variants and established that common susceptibility is polygenic. BCAC meta-analyses have identified numerous susceptibility loci and enabled subtype-aware discovery [66]. Polygenic risk scores translate these loci into an individual-level measure that can be combined with classical predictors, and methodological requirements for development and evaluation in stratified prevention contexts have been summarized [67]. For prevention-oriented use, two issues recur. First, calibration and setting-specific evaluation remain essential because absolute risk depends on baseline incidence and competing mortality. Second, cross-ancestry performance can vary, so portability and subgroup calibration must be examined to avoid widening disparities [59,68]. In Korean women, joint models integrating polygenic risk scores with non-genetic predictors have been developed and show graded risk across polygenic risk score percentiles, supporting feasibility while also reinforcing the need for local validation before any threshold-based use in practice (Figure 1) [69]. Extending risk-adapted screening beyond a single cancer site requires explicit risk-to-action specification because confirmatory testing and downstream procedures differ by site; we return to this issue in Section V using the CancerFree example.

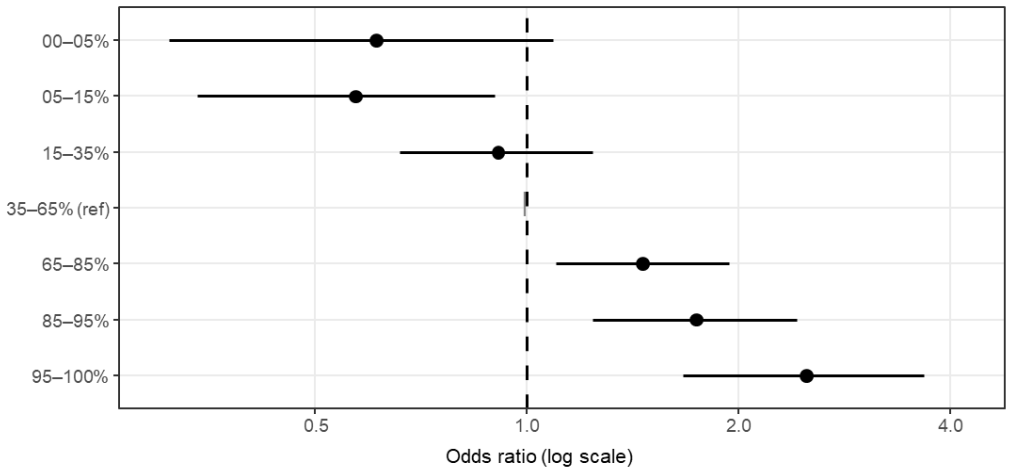


Figure 1. Associations between polygenic risk score (PRS) percentile tiers and breast cancer risk in Korean women. Points show odds ratios (ORs) and horizontal bars show 95% confidence intervals from unconditional logistic regression comparing each PRS percentile tier with the reference tier (35%–65%). The vertical dashed line indicates OR = 1. Adapted from Table 2 of Choi *et al.* [69].

Although breast cancer provides one of the clearest mature examples of risk-adapted screening, the same framework extends to other cancers with distinct prevention pathways. In colorectal cancer, risk may inform fecal immunochemical test-based or colonoscopy-based screening strategies, including colonoscopy prioritization and risk-adapted starting ages [70–73]. In lung cancer, it may refine low-dose computed tomography eligibility within smoking-exposed populations [74]. In liver cancer, risk may help guide surveillance eligibility or intensity in people with chronic hepatitis or chronic liver disease [75,76].

These settings differ in modality, confirmatory work-up, and benefit-harm balance, but the common translational requirements remain calibrated absolute risk, explicit risk-to-action specification, and transparent accounting of downstream burden. As in breast cancer, the usefulness of risk stratification in these settings ultimately depends on pathway design and capacity, not just statistical separation.

4. Tertiary prevention: prognostic prediction and survivorship follow-up

Tertiary prevention in breast cancer, including prevention of recurrence, metastasis, second primaries, and treatment-related harms, has one of the most mature evidence bases for prognostic prediction in oncology. Tumor gene-expression assays are now routinely used to guide adjuvant therapy decisions, helping to avoid overtreatment while identifying patients more likely to benefit from treatment escalation. The 21-gene recurrence score established this approach, and large prospective trials support de-escalation of chemotherapy for many patients with specific clinical risk profiles and low to intermediate scores, while also identifying groups with clearer benefit from chemotherapy [77–79]. The 70-gene signature has likewise been supported by prospective evidence for treatment decision support [80]. More recently, a next-generation sequencing-based multigene assay using a 179-gene expression panel was validated in an independent cohort and stratified distant metastasis-free survival over 5 and 10 years [81]. A subsequent multicenter evaluation compared its decision index with Oncotype DX and reported correlation with context-specific risk separation in selected groups [82]. Regulatory translation is also occurring. A Ministry of Food and Drug Safety announcement in January 2026 described approval of a domestically developed gene-expression *in vitro* diagnostic intended to classify 10-year distant recurrence risk as low *versus* high using surgical tissue from early invasive breast cancer, reporting a negative predictive value of 97.36% in those classified as low risk [83].

Survivorship prediction extends beyond recurrence to contralateral breast cancer risk and selected late effects or toxicities, where germline genetics has been used to identify higher-risk subgroups and, in principle, to inform follow-up intensity. However, evidence for routine-practice outcome improvement remains heterogeneous, and changes in treatment over time can shift baseline risks [84,85]. Era-specific calibration therefore remains essential for prognostic tools, because evolving systemic therapies and changing practice patterns can materially alter the clinical meaning of historical predictors.

More broadly, breast cancer illustrates how individual cancer prediction has evolved for prevention-oriented use across the continuum. Classical epidemiologic models enabled long-term risk estimation for counseling and prevention eligibility. Screening programs then motivated models that incorporate imaging and prior procedures for near-term decisions, and recent work has evaluated germline genetics and imaging-derived features for added value in risk-adapted screening. Prognostic tools and multigene assays show a parallel trajectory in tertiary prevention, where risk estimates inform adjuvant therapy decisions and follow-up planning. Across stages, prevention-oriented translation depends on the use of predicted risk within explicit pathways, demonstration of calibration in the target setting, and transparent accounting of the follow-up procedures induced by implementation. More generally, tertiary-prevention prediction becomes relevant whenever risk estimates are used to tailor recurrence surveillance, second-primary monitoring, toxicity follow-up, or supportive care intensity. The same cautions apply: calibration is era-specific, thresholds must be explicit, and incremental benefit must justify added complexity.

Work on breast cancer provides a mature template for prevention-oriented individual risk prediction, while also clarifying what is required for translation beyond a single cancer site. When predicted risk is used to guide screening intensity, eligibility assessment, or referral for confirmatory evaluation, the core requirements are standardized measurement of predictors, consistent outcome ascertainment, and calibration in the intended setting. These requirements become more salient when extending prediction beyond breast cancer, because the relevant pathways, baseline risks, and diagnostic evaluation differ across cancer sites. This motivates an explicit focus on epidemiologic infrastructure and cohort linkage as the foundation for deployable prevention-oriented prediction tools. Section V therefore turns to cohort and registry-linkage infrastructure required to estimate and maintain calibrated absolute risks over defined follow-up periods, especially when extending prediction from a single cancer to multi-cancer prevention.

5. Cohort linkage for prevention-ready risk prediction

Prospective cohorts remain the most direct framework because they define a baseline population at risk, measure predictors in a standardized manner, and accrue outcomes over time, enabling evaluation in the same temporal direction as real-world decisions. Their limitations motivated mega-cohorts and biobank-scale resources designed for linkage and scale, including UK Biobank, the Million Veteran Program, and China Kadoorie Biobank [86–88]. For population-scale use, follow-up increasingly extends beyond cohorts to routinely collected data. Electronic health records, screening databases, and claims can provide timeliness and coverage, but they also reflect care-seeking patterns and coding processes, so transparent reporting and design-aware evaluation are essential [12,89]. Deep-measurement resources, including genomics and other assay-based layers, are often most useful as modular components that can be linked to longitudinal follow-up frameworks for outcome accrual and calibration in the target setting [90]. Together, longitudinal follow-up, absolute-risk estimation over defined follow-up periods, and linkage-based outcome ascertainment form the epidemiologic infrastructure for the cohort-enabled examples discussed below.

Prevention-ready data integration requires more than simple record linkage. It depends on prespecified linkage methods, harmonized predictor and outcome definitions, reconciliation rules for discrepant dates or codes across cohort, registry, and routine clinical sources, and explicit assessment of linkage quality, completeness, and temporal plausibility. Because preventive decisions are threshold-based, errors in linkage, coding, or outcome ascertainment can directly affect calibration and downstream work-up [12,89,91,92]. Standardization should therefore include predictor definitions, outcome definitions, calendar-period alignment, and monitoring for coding or ascertainment shifts over time. In digitally enabled or routinely collected data settings, additional safeguards are also needed for informative missingness, device or platform drift, and unequal participation or follow-up across socioeconomic and demographic subgroups [93].

For multiyear absolute-risk estimation, competing mortality can materially affect calibrated risk, particularly as follow-up lengthens. Prevention-oriented tools should therefore specify whether absolute-risk projection is based on cause-specific hazards with competing mortality, yielding cumulative incidence, or on alternative competing-risk formulations such as subdistribution hazard approaches, and should report whether this choice meaningfully affects threshold-based decisions.

The HEXA study provides infrastructure that supports prevention-oriented prediction. HEXA recruited adults aged 40 to 69 years through health examination centers, collected standardized

questionnaire and clinical measurements, and enables long-term follow-up through linkage-based outcome ascertainment, absolute-risk estimation, and calibration assessment within a well-defined baseline population [94].

As an example of what registry-linked follow-up enables, an internal HEXA-based breast cancer prediction analysis (manuscript under review) evaluated both discrimination and calibration over follow-up in 73,245 women, with 862 incident breast cancer cases identified over 9.2 years of follow-up [95]. Harrell's C-statistic was 0.835, whereas the time-dependent AUC over follow-up was 0.602 [95]. Calibration was closest to unity in the middle quintiles but showed overprediction in the lowest-risk group (O/E 0.55) and underprediction in the highest-risk group (O/E 1.49), indicating that threshold-based prevention actions would require setting-specific recalibration and monitoring to limit inappropriate downstream diagnostic work-up [95]. This example illustrates why prevention translation cannot be judged by discrimination alone: even a model with reasonable rank-order performance may require recalibration before implementation if miscalibration at the tails alters screening, referral, or diagnostic decisions. The next challenge is to extend this logic beyond a single cancer site, where a reported risk estimate no longer maps to one uniform action.

5.1. CancerFree: multi-cancer risk-to-action specification

In multi-cancer prevention, a given risk threshold does not imply a single uniform action because confirmatory testing, follow-up procedures, and the balance of benefits and harms differ substantially by cancer site. Extending prediction from a single cancer site to multiple cancer sites therefore requires explicit specification of what a reported risk estimate means for each site and what action it is intended to trigger. For example, an elevated colorectal cancer risk may point to FIT followed by colonoscopy, whereas an elevated lung cancer risk may support low-dose computed tomography, an elevated gastric cancer risk may support endoscopic screening, and an elevated liver cancer risk may support surveillance in people with chronic hepatitis or chronic liver disease.

To illustrate this problem using routinely collected inputs, we describe CancerFree, a questionnaire-first, registry-anchored multi-cancer prediction tool developed within the HEXA study. CancerFree links health examination questionnaire and measurement data to cancer registry outcomes and generates site-specific risk groups together with a brief, interpretable summary of major contributors, such as family history and modifiable behaviors, and lifestyle guidance intended to support prevention discussions. We present CancerFree as an illustrative cohort-enabled approach rather than as an implementation-ready risk program; site-specific discrimination and calibration metrics will be reported in dedicated validation studies. CancerFree is included here not as a single omnibus predictor, but as a framework for site-specific risk-to-action specification in multi-cancer prevention, where the meaning of elevated risk depends on the confirmatory test, the follow-up procedures it may trigger, and the acceptable balance of benefits and harms for each cancer type.

The seven CancerFree cancer sites were selected using explicit criteria: high national burden, together accounting for approximately 68% of incident cancers in Korea; availability of scalable early-detection or surveillance pathways, relevance of modifiable risk factors for personalized counseling, and the need for harm-aware strategies in overdiagnosis-prone settings such as thyroid cancer [96,97]. Table 3 summarizes the epidemiologic context for these sites and the rationale for site selection in Korea, and Figure 2 illustrates the corresponding site-specific risk-to-action framework. Representative mobile implementation examples

of these modular tools, including a CancerFree multi-cancer risk report and an AI-derived dietary questionnaire, are shown in Figure 3.

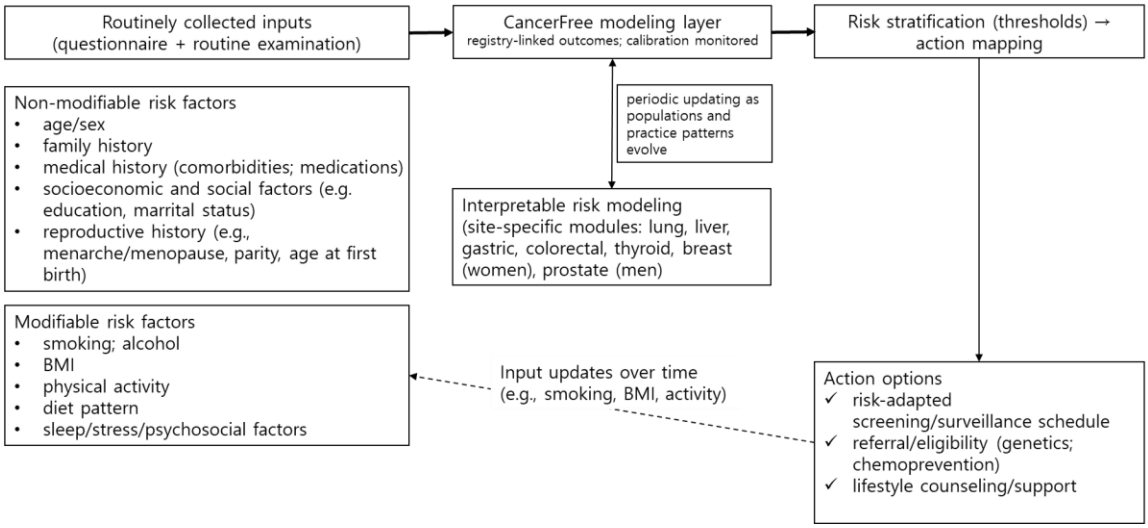


Figure 2. CancerFree example: a precision prevention framework for risk grouping and risk-to-action specification. Routinely collected questionnaire and examination inputs feed a modular, interpretable risk-modeling layer linked to registry-ascertained outcomes, producing reportable site-specific risk groups with key contributors and a lifestyle guidance module to support risk-to-action specification (e.g., risk-adapted screening or surveillance, referral or eligibility for targeted services, and lifestyle counseling). Performance is periodically checked for changes over time and threshold-induced follow-up procedures, with planned periodic recalibration and updating under appropriate validation, governance, and privacy safeguards.

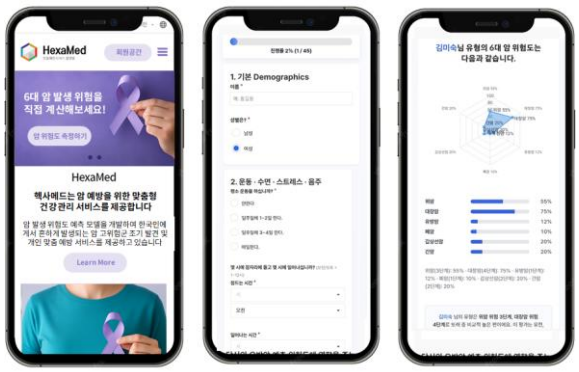
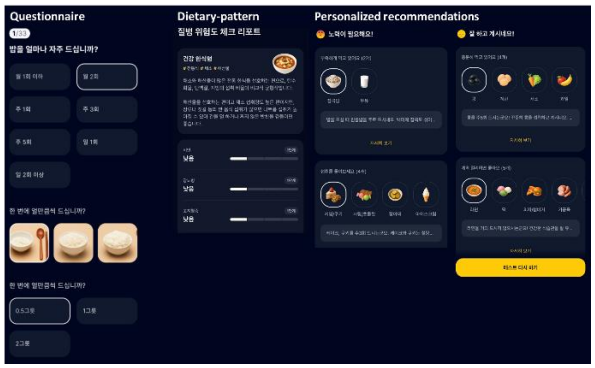
CancerFree multi-cancer risk report (mobile example)	AI-diet questionnaire (mobile example)
	

Figure 3. Mobile implementation examples for deployable, modular risk grouping in the digital health era. Representative mobile screenshots illustrate (left) a CancerFree multi-cancer risk report that summarizes site-specific risk groups and key contributors based on routinely collected questionnaire and health-examination inputs with registry-linked outcomes, and (right) an AI-derived dietary questionnaire module that generates interpretable dietary-pattern features for scalable prevention-oriented profiling. These examples highlight a mobile-first delivery format that can support periodic updating and integration of modular predictor layers under appropriate validation, governance, and safeguards for induced follow-up procedures.

Table 3. Cancer sites included in CancerFree and epidemiologic context in Korea.

Cancer site (CancerFree)	Incidence rank (overall)	Incident cases, n (%)	ASR per 100,000	5-year relative survival, %	Recent incidence trend	Site selection rationale	Major modifiable risk factors/preventable causes (selected)
Thyroid	1	35,440 (12.3)	68.9	100.1	↑	Overdiagnosis-prone; prioritize harm-minimizing pathways and avoid unnecessary follow-up procedures	Obesity
Lung	2	32,953 (11.4)	57.5	40.6	↓ (↑ women)	High mortality; inform LDCT eligibility/interval	Smoking
Colorectal	3	32,610 (11.3)	58.7	74.6	↑ (since 2018)	Optimize FIT/colonoscopy strategy	Alcohol; obesity; physical inactivity; processed/red meat
Breast (women)	4	29,871 (10.3)	56.8	94.3	↑	Risk-adapted screening/prevention decisions	Alcohol; adiposity; physical inactivity
Gastric	5	28,943 (10.0)	51.4	78.4	↓	Risk-based screening and H. pylori management	*Helicobacter pylori* infection; smoking; high-salt/salt-preserved foods
Prostate (men)	6	22,640 (7.8)	39.2	96.4	↑	Rising incidence; triage work-up amid overdiagnosis	Obesity (advanced/high-grade in some evaluations); diet/physical activity (evidence mixed)
Liver	7	14,707 (5.1)	26.1	39.4	↓	Surveillance in chronic hepatitis/liver disease	Chronic HBV/HCV infection; alcohol; obesity/metabolic dysfunction; smoking

Notes: Incidence rank, cases, proportions, and ASR (2023) were obtained from KOSIS (Statistics Korea), based on national cancer incidence statistics [96]. Five-year relative survival (2018–2022) was obtained from the national cancer statistics report [97]. Trend arrows summarize incidence patterns over 2018–2023 based on annual incidence series in KOSIS [96]. The seven CancerFree sites accounted for approximately 68% of incident cancers in 2023 [96]. Preventable causes include infections (e.g., H. pylori, HBV/HCV); for thyroid cancer, incidence is strongly influenced by detection practices, including imaging and diagnostic evaluation that can increase detection of indolent disease. This motivates an overdiagnosis-aware pathway emphasis and harm-minimizing follow-up procedures. Predictor candidates were prespecified based on authoritative evidence syntheses and IARC evaluations, supplemented by evidence from Asian populations and large Korean prospective cohorts, including HEXA, and restricted to variables routinely collected at scale to support deployable implementation and periodic updating.

6. Future directions

Together, CancerFree and the AI-diet module illustrate a cohort-enabled approach to prevention-oriented prediction that prioritizes scalable measurement and longitudinal outcome linkage. Registry-based follow-up supports consistent endpoint definition and absolute-risk estimation over defined follow-up periods. Questionnaire-based predictors facilitate population-scale use, while modular exposure layers such as dietary patterns allow incremental additions to be evaluated for transportability and action relevance. This framework emphasizes prevention models that can be monitored and periodically recalibrated as baseline incidence, measurement practices, and screening patterns evolve, while also allowing stepwise integration of additional predictors, including polygenic risk scores, as evidence and feasibility permit. Against this baseline, this section outlines future directions for prevention-oriented prediction in the digital health era.

The priority is no longer additional stand-alone discrimination studies, but pragmatic evaluation of prevention programs. Future studies should test whether threshold-guided pathways improve net benefit, reduce follow-up burden per cancer detected or prevented, and remain feasible across different workforce and diagnostic-capacity settings. They should also address transportability and recalibration across

populations and over time, because prevention-oriented models remain useful only if they continue to support valid decisions in the settings where they are deployed [9,11].

This modular approach also aligns with an emerging shift toward more frequent measurement in prevention. Digital devices and home-based sensors increasingly enable repeated or continuous capture of behaviors and physiologic correlates in daily life, including activity, sleep, glycemic excursions, and markers related to autonomic balance and recovery. These data streams make time-updated exposure assessment more feasible and may support revision of predicted risk estimates and prevention recommendations as behaviors and intermediate markers change. Realizing this potential still requires careful attention to calibration over time, protocol sensitivity, and governance, but the overall direction is consistent with a move from static baseline risk scores toward monitored and updatable prevention-oriented prediction in the digital health era.

Equity and access are central to digital prevention. Mobile-first delivery can expand reach but may also exclude groups with limited digital access or health literacy; prevention pathways should therefore specify alternative access routes and assess uptake and follow-up procedures across socioeconomic strata. For genomic and imaging layers, subgroup evaluation should include ancestry-related portability and the possibility that identical thresholds may trigger different follow-up procedures across populations, alongside clear communication and consent for linkage-based follow-up.

Routine electronic data capture and linkage increasingly make it feasible to estimate absolute risk over clinically relevant follow-up periods, assess calibration longitudinally, and plan model updating as baseline incidence and clinical practice evolve [9,12,89]. In this setting, added model complexity is justified for prevention only when it improves action-relevant performance, including calibration, threshold performance, and benefit-harm trade-offs, rather than discrimination alone [9].

In resource-limited settings, the greatest near-term public-health value may come from low-burden, questionnaire-first or registry-anchored tools that prioritize limited high-intensity services for those most likely to benefit, rather than from workflow-intensive or omics-heavy systems. Under such conditions, threshold choice depends not only on statistical performance but also on available diagnostic capacity, follow-up infrastructure, and affordability. Economic feasibility should therefore be assessed as a property of the full prevention pathway rather than of test performance alone. Decision-analytic approaches such as decision curve analysis can help determine whether gains in prediction translate into net benefit across clinically plausible threshold ranges [98,99].

A recurring reason cancer risk prediction has underdelivered for prevention is not the absence of algorithms, but the mismatch between what models can measure and what prevention decisions require. Key exposures are often measured sparsely and with substantial error, even though cancer risk accumulates over long latency and through time-varying trajectories. Intermediate biological or clinical states that mediate exposure-cancer pathways are also rarely captured in scalable, repeatable ways, limiting the ability to distinguish stable susceptibility from evolving near-term risk. In addition, many models are evaluated without being embedded in a defined pathway, so improvements in AUC do not translate into actionable threshold decisions, and calibration errors can directly increase unnecessary follow-up procedures [9,100].

Accordingly, a pragmatic roadmap is to strengthen prediction by strengthening both measurement and pathway specification. Prevention-relevant models should prioritize scalable and repeatable exposure assessment, including time-updated measures where feasible, and evaluation of intermediate

states that lie between exposures and clinically manifest cancer, using clearly defined lead times to reduce near-diagnosis effects and reverse causation [90,100]. Digital tools, imaging, and molecular assays may contribute to these aims, but their value for prevention depends on demonstrating temporal validity, external calibration, and net benefit within the pathway they would trigger, rather than on signal discovery alone [9,100].

Genetic susceptibility identified through GWAS and summarized by polygenic risk scores remains an important but bounded layer. Polygenic risk scores are most defensible when used as modular additions to a stable baseline model anchored to the target population, with explicit attention to ancestry portability, subgroup calibration, and whether they meaningfully change prevention decisions at predefined thresholds [68,69]. The central question is not whether a new layer improves discrimination, but whether it improves calibrated absolute-risk estimates and prevention outcomes without disproportionate increases in follow-up procedures and related harms [9,100].

Because prevention-oriented models may alter testing and follow-up in otherwise asymptomatic people, governance is part of translational validity rather than an external add-on [101]. Trustworthy deployment requires privacy-preserving data use, transparent consent and communication processes, and clear policies for the return of results and accountability for downstream actions [101–104). It also requires life-cycle performance monitoring, subgroup evaluation, and prespecified updating or recalibration plans as incidence, practice, and measurement change over time [7,105]. Public trust depends not only on data security, but also on transparent communication of uncertainty, possible harms, and the follow-up consequences of threshold-based decisions [106,107].

Finally, deployable prevention-oriented prediction requires predefined decision thresholds linked to explicit actions, quantification of follow-up procedures, longitudinal monitoring for calibration drift and protocol sensitivity, and governance that protects privacy while auditing subgroup performance and access [9,12]. With these elements in place, cancer prevention can move from static baseline risk scores toward a monitored, updatable, and pathway-grounded prediction approach that is both clinically defensible and scalable.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used generative AI tools solely to improve language and readability. The authors reviewed and edited the output as appropriate and take full responsibility for the final content of the manuscript.

Authors' contribution

Conceptualization, J.Y.C. and D.K.; investigation, S.Y.L. and J.K.; resources, W.K.S. (breast cancer prediction model), S.M. (CancerFree), and H.L. (AI-based diet component); writing—original draft preparation, S.Y.L.; writing—review and editing, W.K.S., S.M., H.L., J.K., J.Y.C., and D.K.; supervision, J.Y.C. and D.K. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

Daehee Kang holds the position of Editorial Board Members for *Advanced Cancer Research* and has not peer reviewed or made any editorial decisions for this paper. Daehee Kang is the founder and owner

of, and an equity holder in, HexaMed, which develops CancerFree and the AI-diet module described in this review; the AI-diet module has been commercially implemented in the Kakao Pasta application. Woo-Kyoung Shin, Sukhong Min, and Hyobin Lee are named inventors on intellectual property related to CancerFree and/or the AI-diet module described in this review. All other authors declare no competing interests.

References

[1] GBD Cancer Collaborators. The global, regional, and national burden of cancer, 1990–2023, with forecasts to 2050: a systematic analysis for the Global Burden of Disease Study 2023. *Lancet* 2025, 406(10512):1565–1586.

[2] The White House. Fact sheet: President Biden reignites Cancer Moonshot to end cancer as we know it. 2022. Available: https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2022/02/02/fact-sheet-president-biden-reignites-cancer-moonshot-to-end-cancer-as-we-know-it/ (accessed on 24 January 2026).

[3] World Health Organization. Guide to cancer early diagnosis. 2017. Report No.: 978-92-4-151194-0. Available: https://www.who.int/publications/i/item/guide-to-cancer-early-diagnosis (accessed on 12 January 2026).

[4] Vickers AJ. Prediction models in cancer care. *CA Cancer J. Clin.* 2011, 61(5):315–326.

[5] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024, 385:e078378.

[6] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* 2015, 162(1):55–63.

[7] Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, *et al.* PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* 2025, 388:e082505.

[8] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, *et al.* PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* 2019, 170(1):51–58.

[9] Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Med.* 2023, 21(1):70.

[10] Welch HG, Bergmark R. Cancer screening, incidental detection, and overdiagnosis. *Clin. Chem.* 2024, 70(1):179–189.

[11] Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TPA, *et al.* Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn. Progn. Res.* 2021, 5(1):1.

[12] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, *et al.* The reporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* 2015, 12(10):e1001885.

[13] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 2024, 74(3):229–263.

[14] Siegel RL, Kratzer TB, Wagle NS, Sung H, Jemal A. Cancer statistics, 2026. *CA Cancer J. Clin.* 2026, 76(1):e70043.

[15] Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, *et al.* The benefits and harms of breast cancer screening: an independent review. *Br. J. Cancer* 2013, 108(11):2205–2240.

[16] Louro J, Posso M, Hilton Boon M, Roman M, Domingo L, *et al.* A systematic review and quality assessment of individualised breast cancer risk prediction models. *Br. J. Cancer* 2019, 121(1):76–85.

[17] Clift AK, Dodwell D, Lord S, Petrou S, Brady SM, *et al.* The current status of risk-stratified breast screening. *Br. J. Cancer* 2022, 126(4):533–550.

[18] Re F, Manaboriboon N, Raza IGA, Shipley A, Thompson LE, *et al.* A systematic review of prediction models for risk of breast cancer. *BMC Cancer* 2025, 25(1):1650.

[19] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019, 17(1):230.

[20] Kyu HH, Bachman VF, Alexander LT, Mumford JE, Afshin A, *et al.* Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013. *BMJ* 2016, 354:i3857.

[21] Cao Y, Willett WC, Rimm EB, Stampfer MJ, Giovannucci EL. Light to moderate intake of alcohol, drinking patterns, and risk of cancer: results from two prospective US cohort studies. *BMJ* 2015, 351:h4238.

[22] Kyrgiou M, Kalliala I, Markozannes G, Gunter MJ, Paraskevaidis E, *et al.* Adiposity and cancer at major anatomical sites: umbrella review of the literature. *BMJ* 2017, 356:j477.

[23] Lei J, Ploner A, Elfstrom KM, Wang J, Roth A, *et al.* HPV vaccination and the risk of invasive cervical cancer. *N. Engl. J. Med.* 2020, 383(14):1340–1348.

[24] Chang MH, Chen CJ, Lai MS, Hsu HM, Wu TC, *et al.* Universal hepatitis B vaccination in Taiwan and the incidence of hepatocellular carcinoma in children. *N. Engl. J. Med.* 1997, 336(26):1855–1859.

[25] International Agency for Research on Cancer. Preamble to the IARC Monographs (amended January 2019). 2019. Available: https://monographs.iarc.who.int/iarc-monographs-preamble-preamble-to-the-iarc-monographs/ (accessed on 3 January 2026).

[26] Wong ATY, Heath AK, Tong TYN, Reeves GK, Floud S, *et al.* Sleep duration and breast cancer incidence: results from the Million Women Study and meta-analysis of published prospective studies. *Sleep* 2021, 44(2):zsaa166.

[27] Von Behren J, Goldberg D, Hurley S, Clague DeHart J, Wang SS *et al.* Prospective analysis of sleep characteristics, chronotype, and risk of breast cancer in the California Teachers Study. *Cancer Causes Control* 2024, 35(4):597–604.

[28] Richmond RC, Anderson EL, Dashti HS, Jones SE, Lane JM, *et al.* Investigating causal relations between sleep traits and risk of breast cancer in women: mendelian randomisation study. *BMJ* 2019, 365:l2327.

[29] Kummer S, Walter FM, Chilcot J, Scott S. Measures of psychosocial factors that may influence help-seeking behaviour in cancer: a systematic review of psychometric properties. *J. Health Psychol.* 2019, 24(1):79–99.

[30] McCutchan G, Weiss B, Quinn-Scoggins H, Dao A, Downs T, *et al.* Psychosocial influences on help-seeking behaviour for cancer in low-income and lower middle-income countries: a mixed-methods systematic review. *BMJ Global Health* 2021, 6(2):e004213.

[31] Eckerling A, Ricon-Becker I, Sorski L, Sandbank E, Ben-Eliyahu S. Stress and cancer: mechanisms, significance and future directions. *Nat. Rev. Cancer* 2021, 21(12):767–785.

[32] Chang A, Sloan EK, Antoni MH, Knight JM, Telles R, *et al.* Biobehavioral pathways and cancer progression: insights for improving well-being and cancer outcomes. *Integr. Cancer Ther.* 2022, 21:15347354221096081.

[33] Mohan A, Huybrechts I, Michels N. Psychosocial stress and cancer risk: a narrative review. *Eur. J. Cancer Prev.* 2022, 31(6):585–599.

[34] van Tuijl LA, Basten M, Pan KY, Vermeulen R, Portengen L, *et al.* Depression, anxiety, and the risk of cancer: An individual participant data meta-analysis. *Cancer* 2023, 129(20):3287–3299.

[35] Basten M, Pan KY, van Tuijl LA, de Graeff A, Dekker J, *et al.* Psychosocial factors, health behaviors and risk of cancer incidence: testing interaction and effect modification in an individual participant data meta-analysis. *Int. J. Cancer* 2024, 154(10):1745–1759.

[36] De la Torre K, Shin WK, Lee HW, Huang D, Min S, *et al.* Weight gain after 35 years of age is associated with increased breast cancer risk: findings from a large prospective cohort study. *Cancer Biol. Med.* 2024, 21(8):651–655.

[37] Shin WK, Lee HW, Shin A, Lee JK, Lee SA, *et al.* Multi-grain rice diet decreases risk of breast cancer in korean women: results from the health examinees study. *Nutrients* 2020, 12(8):2273.

[38] Shin WK, Lee HW, Shin A, Lee JK, Kang D. Milk consumption decreases risk for breast cancer in korean women under 50 years of age: results from the health examinees study. *Nutrients* 2019, 12(1):32.

[39] Lee H, Shin WK, De la Torre K, Min S, You SH, *et al.* Meat intake and breast cancer risk from the Health Examinees study. *Clin. Nutr.* 2025, 53:109–118.

[40] Lee J, Shin A, Shin WK, Choi JY, Kang D, *et al.* Adherence to the World Cancer Research Fund/American Institute for Cancer Research and Korean Cancer Prevention Guidelines and cancer risk: a prospective cohort study from the Health Examinees-Gem study. *Epidemiol. Health* 2023, 45:e2023070.

[41] Wada K, Kuboyama K, Abe SK, Rahman MS, Islam MR, *et al.* Body mass index and breast cancer risk in premenopausal and postmenopausal East Asian women: a pooled analysis of 13 cohort studies. *Breast Cancer Res.* 2024, 26(1):158.

[42] Nabila S, Choi JY, Abe SK, Islam MR, Rahman MS, *et al.* Differential patterns of reproductive and lifestyle risk factors for breast cancer according to birth cohorts among women in China, Japan and Korea. *Breast Cancer Res.* 2024, 26(1):15.

[43] Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr. Opin. Lipidol.* 2002, 13(1):3–9.

[44] Lee H. Development and validation of dietary patterns using machine learning methods and their association with cancer risk in a prospective cohort study. Doctoral Thesis, Seoul National University, 2026.

[45] US Preventive Services Task Force. Medication use to reduce risk of breast cancer: US preventive services task force recommendation statement. *JAMA* 2019, 322(9):857–867.

[46] Pinsky PF, Sauter E, Samimi G. Use of breast cancer risk reducing medications by breast cancer risk level in an older U.S. cohort. *Cancer Epidemiol.* 2025, 97:102856.

[47] Cornell LF, Klassen CL, Ghosh K, Ball C, Advani P, *et al.* Improved uptake and adherence to risk-reducing medication with the use of low-dose tamoxifen in patients at high risk for breast cancer. *Cancer Prev. Res.* 2024, 17(12):565–570.

[48] Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, *et al.* Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* 1989, 81(24):1879–1886.

[49] Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat. Med.* 2004, 23(7):1111–1130.

[50] Park B, Ma SH, Shin A, Chang MC, Choi JY, *et al.* Korean risk assessment model for breast cancer risk prediction. *PLoS One* 2013, 8(10):e76736.

[51] Pashayan N, Morris S, Gilbert FJ, Pharoah PDP. Cost-effectiveness and benefit-to-harm ratio of risk-stratified screening for breast cancer: a life-table model. *JAMA Oncol.* 2018, 4(11):1504–1510.

[52] Roux A, Cholerton R, Sicsic J, Moumjid N, French DP, *et al.* Study protocol comparing the ethical, psychological and socio-economic impact of personalised breast cancer screening to that of standard screening in the “My Personal Breast Screening” (MyPeBS) randomised clinical trial. *BMC Cancer* 2022, 22(1):507.

[53] Esserman LJ, Fiscalini AS, Naeim A, van’t Veer LJ, Kaster A, *et al.* Risk-based vs annual breast cancer screening: the WISDOM randomized clinical trial. *JAMA* 2025, 335(9):763–774.

[54] Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, *et al.* Prospective breast cancer risk prediction model for women undergoing screening mammography. *J. Natl. Cancer Inst.* 2006, 98(17):1204–1214.

[55] Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, *et al.* Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann. Intern. Med.* 2008, 148(5):337–347.

[56] Rosner B, Colditz GA. Nurses’ health study: log-incidence mathematical model of breast cancer incidence. *J. Natl. Cancer Inst.* 1996, 88(6):359–364.

[57] Shieh Y, Hu D, Ma L, Huntsman S, Gard CC, *et al.* Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast Cancer Res. Treat.* 2016, 159(3):513–525.

[58] Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, *et al.* Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* 2019, 104(1):21–34.

[59] Ho WK, Tan MM, Mavaddat N, Tai MC, Mariapun S, *et al.* European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. *Nat. Commun.* 2020, 11(1):3833.

[60] Kapoor PM, Mavaddat N, Choudhury PP, Wilcox AN, Lindstrom S, *et al.* Combined associations of a polygenic risk score and classical risk factors with breast cancer risk. *J. Natl. Cancer Inst.* 2021, 113(3):329–337.

[61] Eriksson M, Destounis S, Czene K, Zeiberg A, Day R, *et al.* A risk model for digital breast tomosynthesis to predict breast cancer and guide clinical care. *Sci. Transl. Med.* 2022, 14(644):eabn3971.

[62] Yala A, Mikhael PG, Strand F, Lin G, Smith K, *et al.* Toward robust mammography-based models for breast cancer risk. *Sci. Transl. Med.* 2021, 13(578):eaba4373.

[63] Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, *et al.* Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA* 2017, 317(23):2402–2416.

[64] Antoniou AC, Casadei S, Heikkinen T, Barrowdale D, Pylkas K, *et al.* Breast-cancer risk in families with mutations in PALB2. *N. Engl. J. Med.* 2014, 371(6):497–506.

[65] Hartmann LC, Schaid DJ, Woods JE, Crotty TP, Myers JL, *et al.* Efficacy of bilateral prophylactic mastectomy in women with a family history of breast cancer. *N. Engl. J. Med.* 1999, 340(2):77–84.

[66] Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* 2017, 551(7678):92–94.

[67] Chatterjee N, Shi J, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* 2016, 17(7):392–406.

[68] Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 2019, 51(4):584–591.

[69] Choi J, Ha TW, Choi HM, Lee HB, Shin HC, *et al.* Development of a breast cancer risk prediction model incorporating polygenic risk scores and nongenetic risk factors for Korean women. *Cancer Epidemiol. Biomarkers Prev.* 2023, 32(9):1182–1189.

[70] Hull MA, Rees CJ, Sharp L, Koo S. A risk-stratified approach to colorectal cancer prevention and diagnosis. *Nat. Rev. Gastroenterol. Hepatol.* 2020, 17(12):773–780.

[71] Thomsen MK, Pedersen L, Erichsen R, Lash TL, Sorensen HT, *et al.* Risk-stratified selection to colonoscopy in FIT colorectal cancer screening: development and temporal validation of a prediction model. *Br. J. Cancer* 2022, 126(8):1229–1235.

[72] Chen H, Shi J, Lu M, Li Y, Du L, *et al.* Comparison of colonoscopy, fecal immunochemical test, and risk-adapted approach in a colorectal cancer screening trial (TARGET-C). *Clin. Gastroenterol. Hepatol.* 2023, 21(3):808–818.

[73] Dong X, Luo Z, Wu Z, Hang D, Xia C, *et al.* Risk-adapted starting age for personalized colorectal cancer screening: validated evidence from national population-based studies. *Clin. Gastroenterol. Hepatol.* 2023, 21(3):819–826.e13.

[74] Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, *et al.* Selection criteria for lung-cancer screening. *N. Engl. J. Med.* 2013, 368(8):728–736.

[75] Lee YT, Fujiwara N, Yang JD, Hoshida Y. Risk stratification and early detection biomarkers for precision HCC screening. *Hepatology* 2023, 78(1):319–362.

[76] Gu W, de Ledinghen V, Aube C, Krag A, Strassburg C, *et al.* Hepatocellular cancer surveillance in patients with advanced chronic liver disease. *NEJM Evidence* 2024, 3(11):EVIDoa2400062.

[77] Paik S, Shak S, Tang G, Kim C, Baker J, *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* 2004, 351(27):2817–2826.

[78] Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, *et al.* Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N. Engl. J. Med.* 2018, 379(2):111–121.

[79] Kalinsky K, Barlow WE, Gralow JR, Meric-Bernstam F, Albain KS, *et al.* 21-gene assay to inform chemotherapy benefit in node-positive breast cancer. *N. Engl. J. Med.* 2021, 385(25):2336–2347.

[80] Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, *et al.* 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *N. Engl. J. Med.* 2016, 375(8):717–729.

[81] Lee HB, Lee SB, Kim M, Kwon S, Jo J, *et al.* Development and validation of a next-generation sequencing-based multigene assay to predict the prognosis of estrogen receptor-positive, her2-negative breast cancer. *Clin Cancer Res.* 2020, 26(24):6513–6522.

[82] Kang E, Cheun JH, Lee J, Koh J, Lee H, *et al.* Prognostic performance of the next-generation sequencing-based multigene assay in early breast cancer patients treated according to the 21-gene assay results. *Cancer Res. Treat.* 2025, 57(3):749–759.

[83] Ministry of Food and Drug Safety. Press release: MFDS approves a domestically developed novel medical device intended to classify distant recurrence risk in patients with early invasive breast cancer. 2026. Available: https://www.mfds.go.kr/brd/m_99/view.do?company_cd=&company_nm=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&page=1&seq=49664&srchFr=&srchTo=&srchTp=&srchWord= (accessed on 19 January 2026).

[84] Kramer I, Hooning MJ, Mavaddat N, Hauptmann M, Keeman R, *et al.* Breast cancer polygenic risk score and contralateral breast cancer risk. *Am. J. Hum. Genet.* 2020, 107(5):837–848.

[85] Morton LM, Kerns SL, Dolan ME. Role of germline genetics in identifying survivors at risk for adverse effects of cancer treatment. *Am. Soc. Clin. Oncol. Educ. Book* 2018, 38(38):775–786.

[86] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018, 562(7726):203–209.

[87] Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, *et al.* Million veteran program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* 2016, 70:214–223.

[88] Chen Z, Chen J, Collins R, Guo Y, Peto R, *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* 2011, 40(6):1652–1666.

[89] Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* 2016, 183(8):758–764.

[90] Jiang P, Sinha S, Aldape K, Hannenhalli S, Sahinalp C, *et al.* Big data in basic and translational cancer research. *Nat. Rev. Cancer* 2022, 22(11):625–639.

[91] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.* 2013, 20(1):144–151.

[92] Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. *Ann. Hum. Biol.* 2020, 47(2):218–226.

[93] Darke P, Cassidy S, Catt M, Taylor R, Missier P, *et al.* Curating a longitudinal research resource using linked primary care EHR data-a UK Biobank case study. *J. Am. Med. Inform. Assoc.* 2022, 29(3):546–552.

[94] Health Examinees Study Group. The Health Examinees (HEXA) study: rationale, study design and baseline characteristics. *Asian Pac. J. Cancer Prev.* 2015, 16(4):1591–1597.

[95] Shin W-K, De la Torre K, Huang D, Min S, Choi JY, *et al.* Development of the HexaMed breast cancer prediction model using the health examinees study. Manuscript submitted for publication.

[96] Statistics Korea. Incidence by cancer site (61 sites) and sex: cases, proportion, crude rate, and age-standardized rate. Korean Statistical Information Service (KOSIS). 2026. Available: https://kosis.kr/statHtml/statHtml.do?sso=ok&returnurl=https%3A%2F%2Fkosis.kr%3A443%2FstatHtml%2FstatHtml.do%3Fconn_path%3DMT_ZTITLE%26list_id%3DF_35%26obj_var_id%3D%26seqNo%3D%26tblId%3DDT_117N_A0024%26vw_cd%3DMT_ZTITLE%26itm_id%3D%26language%3Dkor%26lang_mode%3Dko%26orgId%3D117%26 (accessed on 14 January 2026).

[97] Park EH, Jung KW, Park NJ, Kang MJ, Yun EH, *et al.* Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2022. *Cancer Res. Treat.* 2025, 57(2):312–330.

[98] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* 2006, 26(6):565–574.

[99] Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016, 352:i6.

[100] Fitzgerald RC, Antoniou AC, Fruk L, Rosenfeld N. The future of early cancer detection. *Nat. Med.* 2022, 28(4):666–677.

[101] World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. 2021. Report No.: 9789240029200. Available: https://www.who.int/publications/i/item/9789240029200 (accessed on 4 January 2026).

[102] Muller SH, van Thiel GJ, Mostert M, van Delden JJ. Dynamic consent, communication and return of results in large-scale health data reuse: Survey of public preferences. *Digital Health* 2023, 9:20552076231190997.

[103] Health Canada, U.S. Food and Drug Administration, Medicines and Healthcare products Regulatory Agency. Good machine learning practice for medical device development: guiding principles. 2021. Available: https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles/good-machine-learning-practice-for-medical-device-development-guiding-principles (accessed on 6 January 2026).

[104] Health Canada, U.S. Food and Drug Administration, Medicines and Healthcare products Regulatory Agency. Predetermined change control plans for machine learning-enabled med ical devices: guiding principles. 2023. Available: https://www.gov.uk/government/publications/predetermined-change-control-plans-for-machine-learning-enabled-medical-devices-guiding-principles/predetermined-change-control-plans-for-machine-learning-enabled-medical-devices-guiding-principles (accessed on 6 January 2026).

[105] Davis SE, Greevy Jr RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. *J. Biomed. Inform.* 2020, 112:103611.

[106] Fehr J, Citro B, Malpani R, Lippert C, Madai VI. A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. *Front. Digital Health* 2024, 6:1267290.

[107] Riley RD, Collins GS, Kirton L, Snell KI, Ensor J, *et al.* Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. *BMJ* 2025, 388:e080749.