informatics

MDPI

Article

A Health Informatics Framework for Integrating Machine Learning and Generative AI in HIV Risk Stratification and Personalized PrEP Recommendation

Panyaphon Phiphatkunarnon [1,2], Amornphat Kitro [3,4], Benjamas Suksatit [5], Boon-Leong Neo [6], Do Tran [6] and Worawit Tepsan [1,*]

[1] International College of Digital Innovation, Chiang Mai University, Chiang Mai 50200, Thailand; panyaphon_p@cmu.ac.th
[2] Love Foundation, Chiang Mai 50300, Thailand
[3] Department of Community Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai 50200, Thailand
[4] Environmental and Occupational Medicine Excellence Center, Faculty of Medicine, Chiang Mai University, Chiang Mai 50200, Thailand
[5] Faculty of Nursing, Chiang Mai University, Chiang Mai 50200, Thailand; benjamas.s@cmu.ac.th
[6] Gilead Sciences Inc., Singapore 018983, Singapore; boon-leong.neo@gilead.com (B.-L.N.); do.tran11@gilead.com (D.T.)
* Correspondence: worawit.tepsan@cmu.ac.th; Tel.: +66-8436-44500

Abstract

Background: Although pre-exposure prophylaxis (PrEP) is highly effective for HIV prevention, identifying individuals who may benefit from PrEP and delivering personalized prevention recommendations remain challenging in routine and digital health settings. **Objective:** This study aimed to develop and preliminarily evaluate an integrated artificial intelligence framework combining machine learning (ML) for HIV risk stratification and generative artificial intelligence (GenAI) for personalized PrEP recommendation support. **Methods:** A curated dataset of 2000 de-identified client profiles from Love2Test platform was used for proof-of-concept model development. Profiles were labeled as low or high HIV acquisition risk by domain experts based on structured behavioral information. Multiple ML classifiers were trained and compared using PyCaret. The selected model was integrated with a generative AI model through structured prompting to generate personalized PrEP recommendation content. The integrated framework was evaluated through structured physician assessment by four independent medical doctors. **Results:** The selected model showed strong internal discrimination for classifying high versus low HIV acquisition risk. The integrated framework also received favorable physician evaluation for clinical accuracy, explanation validity, contextual relevance, and error minimization across fixed and randomly selected profiles. However, because expert labeling was based on structured behavioral indicators closely related to the model inputs, the high internal performance should be interpreted within the context of this proof-of-concept study. **Conclusions:** The proposed framework provides a structured approach to support HIV risk stratification and personalized PrEP recommendations in a clinician-aligned manner. However, this study was an offline proof-of-concept and did not directly evaluate patient interaction, PrEP uptake, stigma, adherence, or clinical outcomes. Prospective studies using larger and more representative real-world datasets are needed to assess implementation, generalizability, and impact on service engagement and PrEP initiation.

Check for updates

Academic Editor: Ashkan Ebadi

Received: 13 April 2026
Revised: 13 June 2026
Accepted: 23 June 2026
Published: 29 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: pre-exposure prophylaxis (PrEP); HIV prevention; machine learning; generative artificial intelligence (GenAI); health informatics; digital health platform; risk stratification; clinical decision support; human–computer interaction; automated machine learning

1. Introduction

HIV remains a global health challenge, with 39.9 million people living with the virus and 1.3 million new infections reported in 2023 [1], highlighting persistent prevention gaps. While antiretroviral therapies have substantially reduced HIV-related mortality, prevention strategies lag behind the UNAIDS 95-95-95 targets for ending the epidemic by 2030 [2]. Achieving these targets requires significant progress in HIV testing, treatment, and pre-exposure prophylaxis (PrEP) adoption. PrEP is highly effective for HIV prevention when taken as prescribed, reducing the risk of HIV acquisition from sex by about 99% [3,4]. However, real-world impact depends on effective uptake, adherence, persistence, and equitable access to services [5,6].

Current HIV risk assessment often relies on manual provider evaluations, which may fail to capture complex and evolving behavioral risk patterns, resulting in missed opportunities to identify PrEP-eligible individuals. Integrating advanced technology into healthcare systems is essential for optimizing PrEP delivery and uptake. Prior to ML adoption, traditional risk prediction models such as logistic regression–based prediction models and rule-based risk scoring systems were developed. Menza et al. developed a multivariable logistic regression model using behavioral risk factors and achieved modest predictive accuracy (AUC 0.66–0.67) for HIV acquisition among MSM [7]. Rule-based risk scoring systems, which convert regression-derived predictor weights into simplified point-based scores for clinical risk stratification, include the San Diego Early Test Score developed by Hoenigl et al. (C-statistic > 0.70) [8] and the HIV risk score proposed by Yin et al. (C-index 0.70–0.71) [9], both of which demonstrated moderate discriminatory performance.

In healthcare settings, ML approaches have been successfully validated for identifying PrEP candidates. Krakower et al. achieved an AUC of 0.86 using EHR data [10], and Marcus et al. achieved an AUC of 0.84 with 3.7 million Kaiser Permanente members [11]. Similarly, Saldana et al. reached 80% accuracy using surveillance data [12], and Ridgway et al. demonstrated AI's role in emergency departments for identifying eligible patients for counseling [13]. More recent EHR-based work has further refined this approach: Nethi et al. developed a machine learning model using EHR data from a large urban health system (n = 458,893) to predict incident HIV diagnoses and prioritize patients for proactive PrEP outreach [14], while May et al. proposed a generalizable automatic feature-engineering pipeline for HIV risk prediction across heterogeneous EHR systems, reporting consistent performance (AUC $\approx$ 0.87) and improved sensitivity for female patients [15]. In population-based settings, Balzer et al. showed that ML algorithms in rural Kenya and Uganda could capture 50% of new HIV cases by targeting only 18% of the population, a significant improvement over traditional methods [16]. ML has further expanded this scope. He et al. achieved an AUC of 0.88 for Chinese MSM using random forests [17], and Xiang et al. highlighted the potential of deep learning approaches, including recurrent neural networks and convolutional neural networks, to reduce feature engineering requirements in HIV prevention research [18]. Deep learning architectures have also been applied to relational and social-network data: Yu et al. recently demonstrated the use of explainable graph neural networks with domain adaptation to predict HIV infections among younger sexual minority men across two U.S. cohorts, illustrating both the modeling capacity and the cross-dataset transferability of such methods [19]. In addition to studies focused on PrEP

candidate identification, recent work has also examined missed opportunities for HIV diagnosis or PrEP linkage, as well as the use of deep learning and LLM-based approaches in HIV prevention support. Closely related to the objectives of the present work, several studies have specifically examined missed opportunities for earlier HIV diagnosis and PrEP linkage. Weissman et al. applied least absolute shrinkage and selection operator (LASSO) regression to linked surveillance and all-payer healthcare records and showed that a substantial proportion of individuals later diagnosed with HIV had prior healthcare encounters — most commonly in emergency department settings — that represented missed screening opportunities [20]. Such work underscores the importance of complementary, community-facing screening pathways that can intercept individuals before they present late to clinical care.

In parallel, large language models (LLMs) are emerging as decision-support and communication aids in HIV prevention. Govathson et al. developed an LLM-powered conversational app for stigma-free HIV vulnerability assessment and PrEP eligibility discussion among adults in South Africa, demonstrating feasibility and acceptability for supporting provider–patient interactions when consultation time is limited [21]. These developments motivate the integration of generative AI for personalized communication in our framework while also reinforcing the need for clinician oversight, constrained prompting, and structured evaluation. Despite these advancements, challenges remain regarding data privacy, algorithmic transparency, and integration into routine care [22,23].

Although prior studies have demonstrated the value of machine learning for identifying individuals at elevated HIV risk or potential PrEP eligibility, most have focused primarily on predictive classification performance. Many existing models were developed using electronic health record or surveillance data, which may not be readily available or transferable to community-based or digital health platforms. In addition, less attention has been given to how model outputs can be translated into personalized, context-aware recommendation content that supports communication, counseling, and real-world decision-making in digital health settings. From a practical health informatics perspective, a key challenge is not only designing end-to-end digital systems, but also ensuring that outputs from risk models can be translated into meaningful, user-facing recommendations that support real-world decision-making. While informatics research has advanced the design of clinical decision support systems across many domains, the application of such system-level integration to community-based HIV prevention—particularly through digital platforms serving key populations—remains underexplored. In particular, to the best of our knowledge, no prior study has integrated machine learning-based HIV risk stratification with generative AI to produce personalized, context-aware PrEP recommendation content within a single informatics framework. Existing ML studies in HIV prevention have focused on risk classification as an endpoint, without extending model outputs into actionable, patient-facing communication. Similarly, while generative AI has been explored for general health communication, its application to HIV-specific PrEP recommendation guided by ML-derived risk profiles has not been reported. This gap is important because prediction alone may be insufficient to support user engagement, counseling, and clinician-patient communication in routine HIV prevention practice, particularly in digital health settings where automated support may be needed. More broadly, AI in HIV prevention and care may support multiple functions, including risk prediction, decision support, workflow assistance, and personalized communication [22,23]. In this context, the present framework is positioned specifically as a health informatics-driven digital decision-support and recommendation system for HIV prevention, rather than as an autonomous clinical tool.

HIV remains a major public health challenge in Thailand. In 2023, 9100 new HIV infections were diagnosed in Thailand, and 580,000 people were living with HIV in the

country [1]. Key populations, including men who have sex with men (MSM), male sex workers (MSW), transgender women (TGW), and people who inject drugs (PWID), bear a disproportionate burden. As of June 2024, MSM accounted for 77.5% of new HIV cases, followed by MSW at 5.9% and TGW at 5.5%, highlighting the urgent need for targeted strategies [24]. Extensive efforts have focused on expanding PrEP accessibility, most notably the Princess PrEP initiative. This community-led program provides free PrEP to high-risk populations and has successfully increased uptake [25,26].

Despite these advancements, stigma and unequal access persist. Since the launch of Princess PrEP in 2016, coverage remains low; only 29,281 active PrEP users were reported in 2024, reaching just 41% of the national target [24]. Additionally, data suggest issues with persistence, as only 79,291 people have ever initiated oral PrEP [27]. To address these barriers, key population-led health services (KPLHS) have shown promise in increasing accessibility and reducing barriers to PrEP uptake in Thailand [25,28].

To support Thailand's HIV prevention efforts, this study aimed to develop and evaluate an integrated framework combining machine learning (ML) and generative artificial intelligence (GenAI) to (1) classify individuals as having high versus low HIV acquisition risk based on structured behavioral data and (2) generate personalized PrEP recommendation content. Using real-world data from Love2Test.org, we trained and compared multiple ML models through PyCaret [29] and assessed the clinical relevance and validity of the integrated framework through independent physician evaluation. More broadly, this study contributes a generalizable health informatics architecture for integrating machine learning and generative AI within user-facing digital health systems, with design principles that extend beyond HIV-specific applications to other domains requiring automated risk inference and personalized communication.

2. Materials and Methods

2.1. Data Creation and Labeling

The dataset comprised responses to a standardized behavioral questionnaire based on CDC guidance [30] for PrEP eligibility, capturing 12 key HIV-related indicators (Section 2.2.1), along with age and location (14 features in total). The data were collected from clients of Love2Test.org, a national digital sexual health platform in Thailand, through its online booking system between 5 September 2024 and 9 July 2025. A total of 16,982 records were initially extracted from the Love2Test.org database. After excluding 205 records with ages outside the predefined eligibility range (18–65 years), 16,777 records remained. Then, 1154 records with incomplete data were removed, leaving 15,623 records for further processing. Because this study focused on behavioral characteristics used for machine learning model development, duplicate behavioral profiles were identified and removed. This process excluded 11,490 duplicate behavioral records, leaving 4133 unique behavioral profiles for further processing. To preserve diversity and maintain demographic representativeness, a stratified random sampling approach was subsequently applied using age group, self-identified gender, and geographic region as stratification variables. This approach ensured that the selected subset reflected the participant characteristics of the full dataset while reducing the number of records requiring expert review. Finally, a total of 2000 profiles were selected for expert labeling and supervised machine learning model development.

Prior work has shown that datasets of 1000–5000 labeled samples can achieve stable predictive performance for structured classification tasks in healthcare [31,32]. Empirical and systematic review evidence further supports that a minimum of N = 500–1000 is recommended to mitigate overfitting and achieve stable model performance [33]. Following refinement, a stratified random sample of 2000 unique profiles, each containing 12 key

HIV-related features, was selected for supervised model development and expert review. Age group, self-identified gender, and geographic region were used as stratification variables to preserve the demographic distribution of the full dataset. The subset size was chosen to balance sufficient sample diversity for model development with the practical demands of expert annotation, as each profile required individual clinical review and risk categorization.

Key characteristics of the full dataset and the ML subset were compared, as shown in Table 1. Overall, the subset remained broadly comparable to the full dataset with respect to age and geographic distribution. The mean age was similar between the two datasets (28.8 vs. 28.6 years), and the median age was identical at 27 years. The age-group distribution was also closely aligned, particularly among users aged 18–24 years (33.1% vs. 33.3%) and 25–34 years (47.5% vs. 47.9%). Geographic coverage was likewise similar, with the central region accounting for 66.2% of both datasets. Some differences were observed in gender composition, with the subset containing a slightly lower proportion of men (51.8% vs. 56.4%) and a somewhat higher proportion of gender-diverse participants classified as other (29.5% vs. 24.8%). These findings suggest that the refined subset preserved broad demographic coverage while also reflecting some selection effects introduced by the refinement strategy. Because the 2000-record subset was curated for proof-of-concept model development rather than selected through population-based sampling, it should not be interpreted as fully population-representative of all Love2Test.org users. Accordingly, the present findings should be considered preliminary, and further validation using larger and more representative real-world datasets is required before broader implementation claims can be made.

To support clinical validity, the questionnaire structure and labeling framework were independently reviewed by two infectious disease specialists, Dr. Amornphat Kitro and Dr. Chaiwat Songsiriphan, who manually assigned HIV risk categories to the selected profiles for supervised machine learning training. Because these labels were derived from structured behavioral indicators closely related to the model input variables, the possibility of conceptual circularity should be acknowledged. Therefore, the model should be interpreted as an expert-informed proof-of-concept risk stratification framework rather than a fully independent predictive, diagnostic, or prognostic tool.

All study procedures were approved by the Research Ethics Committee (Certificate of Approval No. 179/67). Informed consent was obtained electronically from all participants, who agreed to the Love2Test Terms of Service and Privacy Policy, before completing the online questionnaire and booking services. A waiver of written consent was granted because data was collected anonymously through the digital platform. All records were fully de-identified prior to analysis to ensure participant confidentiality and data privacy. Specifically, all direct identifiers—including name, contact information, national identification number, and booking or transaction references—were removed prior to analysis. The analytic dataset retained only the 12 structured behavioral variables together with age and geographic region; no free-text fields or other potentially re-identifying attributes were included, and all storage and processing procedures were PDPA-compliant. In the proposed workflow, the personalized recommendation would be generated and displayed to the user within the active platform session immediately after questionnaire submission. By contrast, the de-identified analytic dataset used for model development was fully decoupled from any user identity.

Table 1. Comparison of participant characteristics between the initially extracted records (n = 15,623) and the ML development dataset (n = 2000).

Characteristic	Behavioral Records (n = 15,623)	ML Train/Test (n = 2000)
Age		
Mean ± SD	28.8 ± 7.4 yrs	28.6 ± 7.2 yrs
Median	27 yrs	27 yrs
Min–Max	18–65 yrs	18–59 yrs
18–24 yrs	33.1% (5176)	33.3% (666)
25–34 yrs	47.5% (7425)	47.9% (957)
35–44 yrs	15.5% (2421)	15.3% (307)
45+ yrs	3.8% (601)	3.5% (70)
Self-identified gender		
Men	56.4% (8808)	51.8% (1037)
Women	18.9% (2948)	18.6% (372)
Other (non-binary, trans, prefer not to answer)	24.8% (3867)	29.5% (591)
Geographic region		
Central (Bangkok, Nakhon Pathom, Kanchanaburi)	66.2% (10,345)	66.2% (1325)
Northern (Chiang Mai, Chiang Rai, Phayao)	12.2% (1900)	11.8% (237)
Northeastern (Khon Kaen, Ubon Ratchathani, Nakhon Ratchasima, Udon Thani)	8.0% (1244)	8.6% (171)
Eastern (Chon Buri)	9.5% (1491)	9.6% (192)
Southern (Songkhla, Surat Thani, Phuket)	4.1% (643)	3.8% (75)

2.2. Machine Learning Model

2.2.1. Feature Encoding

The ML models were trained using 12 structured behavioral variables collected through the Love2Test.org questionnaire: (1) self-identified gender, (2) sexual intercourse within the previous 6 months, (3) gender of sexual partner(s), (4) multiple sexual partners with unknown HIV status, (5) type(s) of sexual intercourse, (6) discussion of HIV status with partner(s), (7) awareness of partner's HIV status, (8) consistency of condom use, (9) bacterial STI diagnosis within the previous 6 months, (10) injection drug use, (11) participation in group sex or sex parties, and (12) engagement in transactional sex. All candidate models were multivariable classifiers trained on the same feature set.

To prepare the data for modeling, two distinct encoding methods were applied to categorical variables. For features allowing only a single selection per instance, such as partner HIV status (categorized as negative, positive, or unknown) and condom use, label encoding was used to convert categories into numerical values. In contrast, for features permitting multiple simultaneous selections, such as sex type (e.g., vaginal, oral, anal), a Multilabel Binarizer was applied to transform each category into a binary feature.

The target variable was defined as a binary HIV risk classification based on expert-labeled PrEP recommendation needs: low HIV risk (0), where PrEP may not be necessary but could still be discussed if appropriate, and high HIV risk (1), where PrEP was recommended.

2.2.2. Data Train-Test Splitting

A total of 2000 client profiles were obtained, each representing actual user responses to the PrEP screening questionnaire. Independent HIV clinical specialists subsequently reviewed and labeled these profiles based on predefined criteria, classifying 664 as low HIV risk and 1336 as high HIV risk. The dataset was then partitioned into an 80:20 ratio for model development and evaluation. The training set comprised 1600 profiles (527 low-risk and 1073 high-risk), while the test set contained 400 profiles (137 low-risk and 263 high-risk). Because the dataset was imbalanced, the Synthetic Minority Over-sampling Technique (SMOTE) was applied only to the training set to upsample the minority class and improve class balance during model development. The independent test set was not resampled and was retained in its original distribution for evaluation.

To reduce technical data leakage, model development was performed on the training set and final evaluation was conducted on an independent hold-out test set. Nevertheless, because the expert-defined labels were based on structured behavioral indicators closely aligned with the model inputs, some conceptual circularity may remain in the learning task.

2.2.3. Model Training

We used PyCaret version 3.3.2 [29], an open-source Automated Machine Learning (AutoML) framework, in classification mode to compare candidate machine learning models under a standardized workflow. During model development, candidate models were compared using 10-fold cross-validation on the training set, and model tuning was performed within the PyCaret workflow. Final model performance was then assessed on an independent hold-out test set. We acknowledge that this approach prioritizes standardized, reproducible model selection over fine-grained manual hyperparameter control, which represents an appropriate technical trade-off for this proof-of-concept study. The selected model, based on predefined evaluation metrics, was subsequently chosen for integration into the proposed framework.

2.3. System Architecture and Information Flow

To formalize the proposed framework as a reusable informatics contribution, this section describes the system architecture and information flow underlying the integrated ML–GenAI pipeline. For conceptual clarity, the architecture is organized into four functional layers, each representing a key stage of the decision-support process. In practice, each layer consists of multiple technical sub-components responsible for data transformation and communication across the system (Figure 1).

Layer 1 User Interaction: This layer represents the primary human–system interface. End users access the Love2Test.org platform and initiate the screening process by completing a standardized self-assessment. The interaction is designed to be simple and user-friendly, allowing individuals to provide behavioral information efficiently. The resulting user-provided input is then forwarded to the data acquisition layer for further processing.

Layer 2 Data Acquisition: In this layer, the Questionnaire UI captures 12 HIV-related behavioral variables together with age and geographic location, generating structured behavioral data. These data are subsequently processed within the Data Layer/Database sub-component, where encoding, validation, and PDPA-compliant storage are performed. This step produces an encoded feature vector suitable for machine learning inference. The interface and data pipeline are designed to ensure completeness and consistency of data while minimizing user burden.

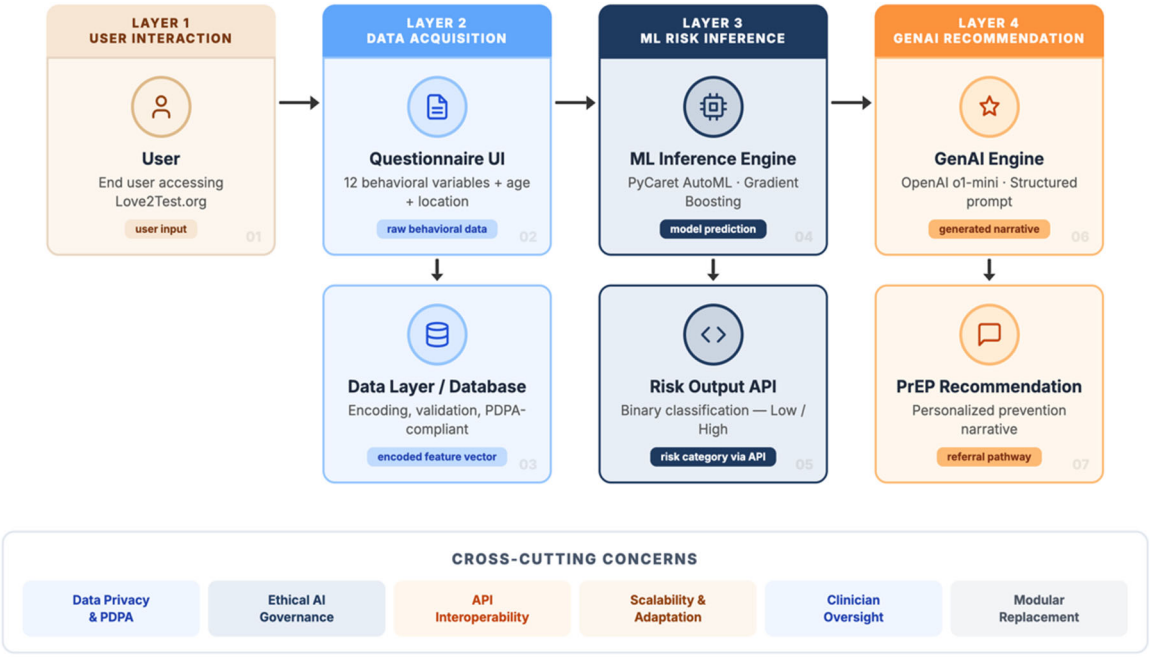


Figure 1. A modular, four-layer informatics architecture illustrating data flow, computational inference, and user interaction pathways within an integrated ML–GenAI decision-support system.

Layer 3 ML Risk Inference: The encoded feature vector is processed by a trained Gradient Boosting Classifier developed using PyCaret AutoML. The model generates a prediction that is passed to the Risk Output API, which returns a standardized binary HIV risk classification (low vs. high). This layer functions as an automated inference engine, transforming structured behavioral data into a risk category without requiring real-time clinician involvement at the point of screening.

Layer 4 GenAI Recommendation: The predicted risk category, combined with selected client attributes, is provided to the GenAI Engine (OpenAI o1-mini, 12 September 2024) through a structured prompt. The model generates a personalized PrEP recommendation in natural language. This output is delivered through the recommendation interface, translating computational results into clear, user-facing guidance. In addition, this layer supports referral pathways, enabling users to transition directly from risk awareness to preventive services. Future extensions may include interactive feedback mechanisms, adaptive explanation levels, and direct integration with PrEP service providers.

Cross-cutting concerns apply across all layers and include data privacy and compliance with the Thailand Personal Data Protection Act (PDPA), ethical AI governance, API interoperability, scalability and system adaptation, clinician oversight, and modular system design. The framework follows a modular architecture in which each layer can be independently updated, replaced, or extended. From an implementation perspective, the system can be deployed using a service-oriented architecture (SOA), where each component communicates through standardized APIs. This design supports interoperability with external clinical systems and facilitates future upgrades. For example, the machine learning model in Layer 3 can be retrained or replaced without affecting upstream data collection, while the GenAI component in Layer 4 can be updated as newer language models become available. This modular design enhances long-term maintainability and enables adaptation of the framework to other digital health applications, such as STI screening, hepatitis risk assessment, or substance use interventions.

2.4. Integration Framework

We propose an integrated health informatics framework (Figure 2) that combines machine learning (ML) and generative artificial intelligence (GenAI) to support personalized PrEP recommendations within a digital health platform. The system architecture, comprising four functional layers (data acquisition, ML risk inference, GenAI recommendation generation, and user interaction), is described in detail in Section 2.3. In the integration pipeline, the predicted risk output is combined with selected client attributes and provided as input to a generative AI model (OpenAI o1-mini, 12 September 2024), using a structured prompt-based approach without fine-tuning. Fine-tuning was not pursued at this proof-of-concept stage for several reasons. First, the available dataset consisted of structured de-identified client profiles and specialist-assigned risk categories, but it was not paired with a sufficiently large corpus of reference recommendation texts suitable for supervised fine-tuning. Second, domain-specific fine-tuning on a relatively small and sensitive clinical dataset could increase the risks of overfitting and unsafe memorization. Third, structured prompting, together with mandatory physician review, provided a more transparent and clinically controllable approach for this initial feasibility study. Fine-tuning using a larger, curated, and representative HIV prevention corpus remains an important direction for future work. Love2Test.org is a national digital sexual health platform in Thailand that provides online HIV self-test kit ordering and clinic appointment booking. The platform's questionnaire interface collects the 12 behavioral variables used in this framework, enabling seamless data flow from user interaction to risk inference without requiring additional manual data entry. The prompt was designed to present the model with the client's risk category and key behavioral indicators and to instruct it to generate a brief, prevention-focused, non-diagnostic recommendation explaining why PrEP may or may not be appropriate in that context. For transparency and reproducibility, a high-level version of the structured prompt template used for the generative component is provided in Appendix A.1. Representative de-identified examples of generated recommendation outputs for distinct risk profiles (e.g., high risk, low risk) are presented in Appendix A.2.

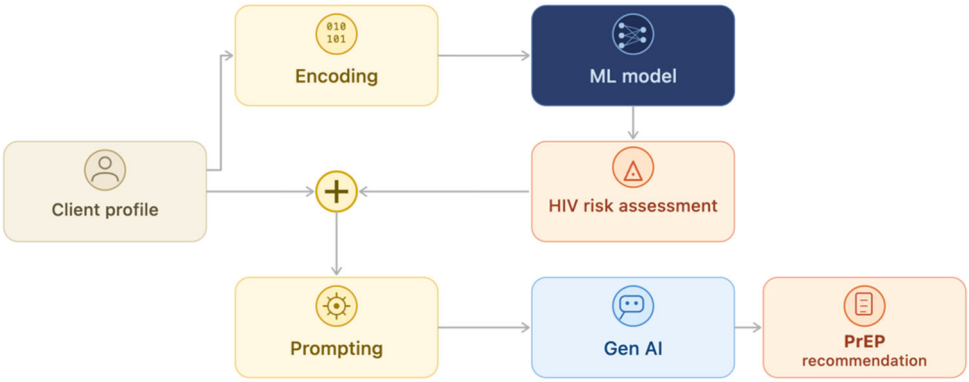


Figure 2. Framework for personalized PrEP recommendation integrating machine learning–based HIV risk assessment with generative AI–driven contextual recommendation generation. The "+" symbol indicates the combination of the client profile and the ML-derived HIV risk assessment output, which are jointly provided to the prompting component before recommendation generation.

To improve consistency and reduce inappropriate outputs, the prompting framework constrained the model to recommendation-style responses and avoided unsupported diagnostic, prescriptive, or treatment-related claims. The generated content was intended for decision-support and communication assistance only, not for autonomous clinical judgment. In this proof-of-concept study, all generated recommendations were subsequently

reviewed by independent medical doctors for clinical appropriateness, contextual relevance, explanation validity, and error minimization.

However, the present study did not formally evaluate failure modes, prompt sensitivity, or explainability across alternative prompt formulations. Detailed system prompt disclosure and formal safety benchmarking were also beyond the scope of this initial evaluation. These aspects should be addressed in future work before real-world clinical implementation.

2.5. Framework Evaluations

To evaluate the proposed framework, we assessed the performance of ML models to identify the most suitable model for integration. The final framework was evaluated for clinical face validity, content validity, and practical relevance using structured professional opinion scoring by four independent medical doctors.

2.5.1. Model Evaluation Metrics

The candidate machine learning models evaluated in this study included Extreme Gradient Boosting, Light Gradient Boosting Machine, Extra Trees Classifier, Gradient Boosting Classifier, Random Forest Classifier, Decision Tree Classifier, K Neighbors Classifier, Ada Boost Classifier, Logistic Regression, Quadratic Discriminant Analysis, Ridge Classifier, Linear Discriminant Analysis, Support Vector Machine (SVM) with linear kernel, and Naive Bayes, as summarized in Tables 2 and 3. Accuracy (ACC) measured the overall proportion of correct predictions, whereas Area Under Curve (AUC) evaluated the model's proficiency in differentiating among categories. Recall (REC) measured the ratio of actual positive cases accurately recognized; in contrast, Precision (PREC) quantified the ratio of correctly predicted positive instances relative to all predicted positive cases. The F1-score (F1) provided a balance between recall and precision. Furthermore, Cohen's Kappa Score (Kappa) assessed the agreement between predicted and actual classifications beyond random chance, while the Matthews Correlation Coefficient (MCC) offered a robust measure of classification performance, particularly for imbalanced datasets. For the selected ML model, a confusion matrix was employed to summarize predictions, outlining true positives, false positives, true negatives, and false negatives. Finally, feature importance was determined to highlight the impact of each factor within the model.

Table 2. Performance of Machine Learning Models resulting from PyCaret Training.

Model	ACC	AUC	REC	PREC	F1	Kappa	MCC
Extreme Gradient Boosting	0.9560	0.9906	0.9495	0.9646	0.9549	0.9117	0.9153
Light Gradient Boosting Machine	0.9553	0.9903	0.9491	0.9638	0.9540	0.9103	0.9145
Extra Trees Classifier	0.9527	0.9716	0.9468	0.9620	0.9519	0.9050	0.9094
Gradient Boosting Classifier	0.9507	0.9876	0.9377	0.9664	0.9493	0.9009	0.9051

Table 2. *Cont.*

Model	ACC	AUC	REC	PREC	F1	Kappa	MCC
Random Forest Classifier	0.9480	0.9890	0.9346	0.9642	0.9466	0.8956	0.9005
Decision Tree Classifier	0.9447	0.9463	0.9339	0.9585	0.9432	0.8890	0.8940
K Neighbors Classifier	0.9061	0.9598	0.8720	0.9422	0.9000	0.8115	0.8214
Ada Boost Classifier	0.9020	0.9690	0.9004	0.9097	0.8996	0.8033	0.8121
Logistic Regression	0.9014	0.9672	0.8884	0.9204	0.8978	0.8019	0.8114
Quadratic Discriminant Analysis	0.8908	0.9538	0.8434	0.9385	0.8810	0.7810	0.7927
Ridge Classifier	0.8841	0.9663	0.8605	0.9108	0.8779	0.7675	0.7783
Linear Discriminant Analysis	0.8841	0.9662	0.8605	0.9108	0.8779	0.7675	0.7783
SVM—Linear Kernel	0.8768	0.9486	0.8638	0.8982	0.8725	0.7514	0.7618
Naive Bayes	0.8135	0.9151	0.6946	0.9169	0.7745	0.6244	0.6519

Table 3. Performance of Machine Learning Models on the Testing Set.

Model	ACC	AUC	REC	PREC	F1	Kappa	MCC
Extreme Gradient Boosting	0.9400	0.9895	0.9430	0.9650	0.9538	0.8682	0.8686
Light Gradient Boosting Machine	0.9450	0.9907	0.9544	0.9617	0.9580	0.8783	0.8784
Extra Trees Classifier	0.9375	0.9539	0.9544	0.9508	0.9526	0.8610	0.8610
Gradient Boosting Classifier	0.9475	0.9848	0.9506	0.9690	0.9597	0.8844	0.8848
Random Forest Classifier	0.9475	0.9866	0.9582	0.9618	0.9600	0.8836	0.8837
Decision Tree Classifier	0.9400	0.9410	0.9506	0.9579	0.9542	0.8672	0.8673
K Neighbors Classifier	0.8800	0.9449	0.8783	0.9352	0.9059	0.7408	0.7436

Table 3. *Cont.*

Model	ACC	AUC	REC	PREC	F1	Kappa	MCC
Ada Boost Classifier	0.8900	0.9563	0.9125	0.9195	0.9160	0.7566	0.7567
Logistic Regression	0.8750	0.9518	0.8935	0.9144	0.9038	0.7253	0.7257
Quadratic Discriminant Analysis	0.8825	0.9308	0.8935	0.9252	0.9091	0.7432	0.7441
Ridge Classifier	0.8675	0.8660	0.8707	0.9234	0.8963	0.7133	0.7157
Linear Discriminant Analysis	0.8675	0.9458	0.8707	0.9234	0.8963	0.7133	0.7157
SVM—Linear Kernel	0.8550	0.8425	0.8821	0.8958	0.8889	0.6803	0.6805
Naive Bayes	0.7400	0.8658	0.6996	0.8804	0.7797	0.4725	0.4913

2.5.2. Clinical Face Validity and Content Validation Using Professional Evaluation

This evaluation employed a mixed-method approach, engaging four independent medical doctors with professional expertise related to HIV to assess PrEP recommendations generated by the framework. Each medical doctor evaluated a total of 50 recommendations produced by the system. Among these, 25 recommendations were predefined based on fixed client profiles, while the remaining 25 were randomly selected to represent diverse user scenarios.

Each recommendation was rated on a 5-point Likert scale (1 = poor, 5 = excellent) using the following criteria:

1. Clinical Accuracy: How accurate is the recommendation from a clinical perspective?
2. Explanation Validity: Does the explanation correctly identify both unsafe and safe factors?
3. Error Minimization: Does the explanation minimize the likelihood of false positives (unnecessary recommendations) and false negatives (missed recommendations)?
4. Contextual Relevance: Does the recommendation appropriately align with the specific context of the client's profile?

The detailed scoring criteria used for profile-specific recommendation evaluation are provided in Appendix A.3 (Table A1).

Upon completing the evaluation, the medical doctors also responded to a set of overarching questions to assess the broader implications and feasibility of the system. These questions included:

5. Desirability: How effectively could a personalized AI-driven recommendation system address common barriers to PrEP uptake among patients?
6. Feasibility: Do healthcare systems have the capacity to effectively integrate AI/ML technologies for personalized PrEP recommendations?
7. Sustainability: Does this solution have the potential to contribute to long-term, sustainable improvements in HIV prevention and care?

The scoring framework used for broader system-level evaluation is summarized in Appendix A.4 (Table A2).

This evaluation was designed as an initial proof-of-concept assessment of clinical face validity, content validity, and practical relevance. It was not intended as a compar-

ative validation study, and no formal comparator condition, such as clinician-written recommendations or rule-based outputs, was included at this stage.

3. Results

3.1. ML Model Evaluations

Table 2 presents the performance of multiple machine learning models during training, with Extreme Gradient Boosting showing the highest training AUC (0.9906) and accuracy (0.9560). However, performance on the independent test set (Table 3) indicated that Gradient Boosting Classifier and Random Forest Classifier achieved similarly strong discrimination, each with an accuracy of 0.9475. The Gradient Boosting Classifier was retained for framework integration because it showed strong and balanced performance on the hold-out test set, including an AUC of 0.9848, Kappa of 0.8844, and MCC of 0.8848. Random Forest Classifier also performed strongly, with a slightly higher AUC (0.9866) but very similar Kappa (0.8836) and MCC (0.8837). Given the small differences among the top-performing models, the final model choice should be interpreted as a pragmatic decision rather than as evidence of a meaningful performance advantage. In addition, the high AUC values observed in internal evaluation should be interpreted cautiously, as they may partly reflect the close correspondence between the structured behavioral predictors and the expert-defined labeling criteria. In this context, the reported performance is more appropriately interpreted as strong internal discrimination within a proof-of-concept, expert-informed risk stratification task rather than as evidence of fully independent predictive inference. Confidence intervals were not computed for the reported metrics; accordingly, the point estimates presented in Table 3 should be interpreted as descriptive summaries rather than precise population-level performance measures.

For context, given the test-set class distribution (137 low-risk and 263 high-risk; n = 400), a non-informative majority-class classifier would achieve an accuracy of approximately 65.8%, and a random classifier an expected AUC of 0.50. The selected model substantially exceeded these baselines (accuracy 0.9475, AUC 0.9848), although this performance should still be interpreted in light of the conceptual circularity discussed above.

Further analysis, as shown in Figure 3, evaluates the classification performance of the Gradient Boosting Classifier using a confusion matrix and feature importance plot. The confusion matrix highlights that the model misclassified 8 low-risk cases as high risk and 13 high-risk cases as low risk out of 400 total cases, demonstrating strong overall performance but some misclassification of high-risk cases. The feature importance plot identifies "multiple partners" as the most influential predictor, followed by "anal sex" and "condom use." Other contributing factors include "vaginal sex," "STI history," and "gender of partner." These findings emphasize the role of sexual behavior in risk classification, highlighting key variables driving the model's predictions.

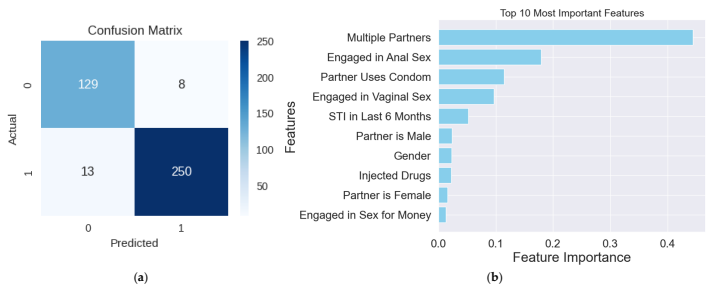


Figure 3. Performance evaluation of the selected model. (**a**) Confusion matrix showing the classification outcomes of the model on the test set; (**b**) feature importance plot showing the predictor variables included in the model. The y-axis lists the variables, and the x-axis represents their relative importance scores, where higher values indicate greater contribution to the model's predictions.

3.2. Integrated Framework of Machine Learning and Generative AI Performance Evaluation Using Professional Opinion Scoring

Table 4 presents the evaluation of AI-generated PrEP recommendations for 25 predefined client profiles, assessed by four medical doctors across multiple criteria. The framework demonstrated strong clinical accuracy (4.77), followed by contextual relevance (4.65) and explanation validity (4.62). While error minimization (4.54) scored slightly lower, the overall performance suggests that the AI effectively generates reliable recommendations. Table 5 extends this analysis to 25 randomly selected client profiles, confirming the AI's consistency across different cases.

Table 4. Evaluation results of the integrated framework's recommendations for 25 fixed profiles, assessed by four medical doctors across clinical accuracy, explanation validity, error minimization, and contextual relevance. The average score reflects the overall performance across these metrics.

Metric	Doctor 1	Doctor 2	Doctor 3	Doctor 4	Average Score
Clinical Accuracy	4.72	4.68	4.76	4.92	4.77
Explanation Validity	4.12	4.73	4.68	4.96	4.62
Error Minimization	4.00	4.72	4.68	4.76	4.54
Contextual Relevance	4.32	4.72	4.68	4.88	4.65

Table 5. Evaluation results of the integrated framework's recommendations for 25 random profiles, reviewed by four medical doctors. The scores indicate how well the framework performs in clinical accuracy, explanation validity, error minimization, and contextual relevance when applied to randomly selected cases.

Metric	Doctor 1	Doctor 2	Doctor 3	Doctor 4	Average Score
Clinical Accuracy	4.96	4.88	4.60	4.88	4.83
Explanation Validity	4.52	4.84	4.64	4.76	4.69
Error Minimization	4.72	4.60	4.64	4.84	4.70
Contextual Relevance	4.80	4.68	4.64	4.76	4.72

The physician evaluation criteria applied in this analysis are described in Appendix A (Tables A1 and A2). Medical doctors highlighted areas for refinement, particularly in making recommendations more engaging and supportive rather than overly formal. Some explanations were flagged for redundancy or overstatement, particularly regarding bacterial STIs and chemsex. Technical improvements were also recommended, such as better integration of STI history and avoiding rigid phrasing like "ML model strongly recommends." While the AI demonstrated strong accuracy and contextual relevance, further improvements in error minimization, explanation clarity, and nuanced risk assessment would enhance its overall effectiveness.

These findings should be interpreted cautiously because the evaluation involved a small number of physicians, relied on subjective Likert-scale ratings, and did not include a

comparator condition. Therefore, the results indicate preliminary clinician acceptability rather than definitive evidence of superiority over simpler alternatives.

Key risk factors identified by medical doctors as strong indicators for PrEP use included attending sex parties, engaging in sex work, having multiple sexual partners, and recent STI diagnoses. They emphasized that an STI history within the past six months should be prioritized in AI-generated recommendations. Additionally, medical doctors stressed the need for clearer differentiation between risk levels. For instance, they considered group sex without chemsex to be lower risk, while group sex with chemsex required stronger PrEP recommendations. Despite these refinements, medical doctors recognized the AI's ability to tailor recommendations based on individual risk behaviors, such as inconsistent condom use, multiple sexual partners, and injection drug use, suggesting potential value as a clinician-aligned support tool, pending further comparative evaluation.

Table 6 provides a broader evaluation of the AI framework, assessing desirability (4.50), feasibility (4.75), and sustainability (4.50). Medical doctors acknowledged AI as a valuable tool for PrEP screening but emphasized that it should complement, not replace, clinical judgment. Given that patient-reported data may lack crucial details, clinician oversight remains essential. They also highlighted the potential for integrating referral pathways to clinics and healthcare providers to improve PrEP accessibility and streamline the care process.

Table 6. Overall performance evaluation of the integrated framework for PrEP recommendations, based on assessments by four medical doctors. The evaluation considers the desirability, feasibility, and sustainability of the recommendations, providing insights into the framework's effectiveness in practical application.

Metric	Doctor 1	Doctor 2	Doctor 3	Doctor 4	Average Score
Desirability	4.00	5.00	4.00	5.00	4.50
Feasibility	5.00	5.00	4.00	5.00	4.75
Sustainability	4.00	4.00	5.00	5.00	4.50

In addition to its technical performance, the proposed framework was designed to support HIV risk stratification and personalized PrEP recommendation within digital health settings. By providing automated risk classification and recommendation content, the system may offer practical value for clinician-aligned counseling and communication workflows. However, its effects on clinical workload, referral pathways, service engagement, and patient outcomes were not evaluated in this proof-of-concept study.

4. Discussion

This study suggests the feasibility of integrating machine learning with generative AI to support personalized, data-informed PrEP recommendations within a digital health workflow. Beyond predictive performance, the proposed framework has potential to support clinician-aligned decision-making by translating structured behavioral risk patterns into recommendation content. The identification of key predictors, including multiple sexual partners, anal sex, condom use, and STI history, is consistent with prior HIV risk prediction studies and epidemiological evidence regarding HIV acquisition risk among MSM and transgender populations [11,34].

Importantly, the generative AI component extends the value of conventional risk prediction by translating model outputs into context-aware recommendation content. This may be useful in digital health settings, where risk classification alone may be insufficient to support communication, user understanding, and patient engagement [35,36]. Emerging

evidence suggests that large language models may help bridge predictive systems and end users through personalized narrative explanations [37]. In this context, the favorable physician evaluation observed in the present study is consistent with prior literature suggesting that generative AI may complement clinical judgment within structured decision-support workflows [38].

We acknowledge that, given the limited number of structured indicators, comparable explanations could potentially be generated using predefined templates, and that generative models introduce non-determinism and a risk of hallucination. Accordingly, the present findings should not be interpreted as demonstrating the necessity or superiority of generative AI for this task. Future studies should directly compare template-based and generative approaches.

Compared with prior studies that focused primarily on predictive classification performance for HIV risk or PrEP eligibility [39], the present framework extends this area by integrating generative AI to generate personalized prevention recommendations. At the same time, the use of AI in HIV prevention raises important concerns regarding algorithmic bias and digital health equity, particularly when training data may not fully represent all populations who could benefit from PrEP [40,41]. Because the framework communicates directly with users, formal fairness evaluation is important before real-world deployment. Future work should assess subgroup performance across demographic characteristics such as gender and geographic region to identify and mitigate potential bias. These concerns are especially relevant in low- and middle-income settings, where digital innovation may expand access to prevention services but should be implemented with careful attention to fairness, inclusion, and clinician oversight [42,43].

From an implementation perspective, the framework has potential to support more standardized HIV risk stratification workflows and may, in principle, reduce some manual screening burden, particularly in resource-limited settings [42,43]. In Thailand, community-led PrEP delivery models and digital health platforms have increasingly supported linkage to HIV prevention services among key populations [44,45].

From a health informatics standpoint, this study proposes a modular architecture integrating structured data collection, machine learning inference, and generative AI within a unified digital health workflow. The proposed four-layer design is aligned with established informatics principles of modularity and interoperability and may support future adaptation to other digital health applications [46].

From a human–computer interaction perspective, the framework embeds AI-generated recommendations directly within the user's existing digital health workflow. Rather than presenting isolated risk scores, the system translates computational outputs into narrative-form recommendations intended to support user understanding and informed decision-making [47]. This design reflects a user-centered informatics approach in which AI augments, rather than disrupts, the interaction flow between users and digital health services.

Taken together, this work differs from prior ML-based HIV risk prediction studies by addressing the full information pipeline from data acquisition through automated inference to user-facing content delivery, representing a system-level contribution to health informatics that extends beyond model performance alone.

Several limitations should be considered when interpreting the findings. First, the dataset was derived from a single digital platform in Thailand, which may limit the generalizability of the findings to other populations, healthcare systems, or sociocultural contexts. Second, the behavioral variables were based on self-reported information, which may be affected by recall bias, social desirability bias, or incomplete disclosure, particularly for sensitive sexual health behaviors. Third, although the model achieved very high predictive performance, including a high AUC, this result should be interpreted cautiously, as it may

partly reflect overfitting or close alignment between the structured questionnaire items and the target labels within this dataset. Although internal validation was conducted, further external validation in independent and more diverse populations is necessary to confirm model robustness. Confidence intervals for performance metrics were not computed in this proof-of-concept study. Given the modest test set size (n = 400) and the acknowledged conceptual circularity between structured behavioral inputs and expert-defined labels, point estimates should be interpreted as descriptive summaries of internal discrimination rather than precise measures of generalizable predictive performance. Future validation studies using independent external datasets should report confidence intervals alongside point estimates.

A further limitation is that this proof-of-concept study reported only point estimates for Accuracy (ACC), Area Under the Receiver Operating Characteristic Curve (AUC), Recall (REC), Precision (PREC), F1-score (F1), Cohen's Kappa, and Matthews Correlation Coefficient (MCC), and did not include fold-level variability statistics, such as mean, standard deviation, minimum, and maximum values across cross-validation folds. These measures can provide important information regarding the stability and consistency of model performance across different training–validation partitions and help assess whether reported results may be influenced by favorable data splits.

Another limitation is that the present evaluation did not include a tailored rule-based classifier or a zero-/few-shot LLM-based comparator. Because the expert labeling criteria were themselves partly rule-based and derived from CDC-informed behavioral indicators, a simple deterministic or clinician-derived rule-based system might achieve comparable performance. Accordingly, the present findings should not be interpreted as establishing the superiority of machine learning over simpler decision-support approaches. Direct comparison against rule-based and LLM-based baselines therefore remains an important direction for future research.

The framework also relies primarily on structured behavioral data, which may not fully capture the complexity of real-world HIV risk. Important contextual, psychosocial, interpersonal, and structural factors may be difficult to represent through fixed questionnaire variables alone. This reliance on structured inputs may reduce applicability in more complex real-world settings, where risk is dynamic and shaped by nuanced social and clinical circumstances.

A clinically important limitation of the selected model is the presence of 13 false-negative classifications, in which individuals with high HIV acquisition risk were predicted as low risk. In an HIV prevention context, such errors are especially concerning because they may result in missed opportunities for PrEP counseling and referral. Future model refinement should therefore prioritize reducing false negatives through approaches such as threshold adjustment, cost-sensitive learning, class weighting, and other optimization strategies that favor recall for the high-risk group [48].

The physician evaluation also has limitations. Only four medical doctors participated, and each reviewed a limited number of recommendations. In addition, the ratings were based on Likert-scale assessments, which may involve some subjectivity. No inter-rater reliability statistics were calculated, and no comparator condition was included. Therefore, these findings should be interpreted as preliminary evidence of clinician-perceived acceptability rather than definitive evidence of superiority over simpler alternatives.

An additional limitation concerns the transparency of the generative AI component. Although a structured prompt-based approach was used to combine the predicted HIV risk category with selected client profile variables, this study did not formally evaluate prompt sensitivity, alternative prompt formulations, or failure modes of the generated recommendations. Moreover, explainability of the generative component was addressed only at a

practical design level rather than through formal explainable AI methods. Recent work has highlighted that explainability, usability, and user trust remain critical challenges for AI-based clinical decision support systems, particularly when such systems are intended for patient-facing or clinician-facing deployment [49]. These considerations should be examined more systematically in future work, including the application of formal explainability frameworks, before any real-world clinical implementation.

Finally, this study was conducted entirely as an offline proof-of-concept evaluation. No patients directly interacted with the system, and no measures related to stigma, usability, uptake, adherence, or real-world service engagement were collected. Accordingly, claims regarding stigma reduction, scalable deployment, or real-world effectiveness should be interpreted cautiously and remain to be tested in prospective implementation studies.

While the proposed framework demonstrated strong predictive performance and favorable clinical evaluation, this study did not directly measure real-world outcomes such as PrEP uptake, adherence, or sustained engagement in care. The current evaluation focused on classification performance and clinician assessment of recommendation quality. Therefore, any potential effect on PrEP uptake or adherence should be interpreted as hypothetical rather than empirically demonstrated.

Finally, this study should be interpreted within its commercial and ethical context. PrEP is a commercial product, the study was funded by Gilead Sciences, and two co-authors are employees of Gilead Sciences. Accordingly, systems that may influence PrEP recommendation require careful attention to conflict-of-interest transparency, clinician oversight, and scientific independence. In this study, the generated recommendations were non-branded, non-prescriptive, and intended only for decision-support communication rather than promotion of any specific product. In addition, all analytic data were de-identified, the analytic workflow was led by the academic investigators, and the source code is publicly available for independent scrutiny.

Future research should address these limitations by incorporating more diverse datasets, integrating broader clinical and social determinants, and evaluating system performance in real-world implementation settings. Prospective studies should assess measurable outcomes such as PrEP initiation, persistence, adherence, and follow-up engagement with healthcare services. In addition, ensuring ethical AI governance, data privacy, fairness, and human oversight will be essential for safe and scalable deployment.

5. Conclusions

This study presents a preliminary health informatics framework integrating machine learning and generative AI for HIV risk stratification and personalized PrEP recommendation within digital health platforms. The four-layer architecture—spanning data acquisition, risk inference, recommendation generation, and user interaction—demonstrates a modular approach to embedding AI-driven clinical decision support into community-based digital prevention services. In this proof-of-concept evaluation, the framework showed strong internal discrimination and received favorable physician ratings for recommendation quality. Future research should evaluate the system using larger and more representative datasets and assess its performance, usability, interoperability, and real-world impact in prospective implementation settings.

Future work should focus on real-world implementation and prospective evaluation, including integration with clinical workflows, referral systems, and digital health platforms. Measurable outcomes such as PrEP initiation, persistence over time, adherence to prescribed regimens, and follow-up engagement with healthcare services should be assessed to determine whether AI-assisted recommendations translate into meaningful improvements in HIV prevention practice. Future model refinement should also explore

threshold adjustment, cost-sensitive learning, and related optimization strategies to reduce false-negative classifications and prioritize recall for the high-risk group. In addition, expanding the framework across more diverse populations and settings will be important to strengthen its generalizability and implementation potential.

Author Contributions: Conceptualization, P.P. and W.T.; methodology, P.P. and W.T.; software, P.P. and W.T.; validation, A.K., B.S., and W.T.; formal analysis, P.P.; investigation, P.P.; resources, W.T., B.-L.N., and D.T.; data curation, P.P.; writing—original draft preparation, P.P.; writing—review and editing, A.K., B.S., B.-L.N., D.T., P.P., and W.T.; visualization, P.P.; supervision, W.T.; project administration, P.P. and W.T.; funding acquisition, B.-L.N. and D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Gilead Sciences.

Institutional Review Board Statement: This study was approved by the Research Ethics Committee of Chiang Mai University under Certificate of Approval (COA) No. 179/67 as part of research project code CMREC 67/258 (approval date: 13 September 2024). All data analyzed were de-identified prior to use to maintain participant confidentiality and privacy.

Informed Consent Statement: Informed consent waived due to secondary data.

Data Availability Statement: De-identified datasets generated and analyzed during this study are available from the corresponding author upon reasonable request due to ethical and privacy restrictions. The source code and implementation materials supporting the findings of this study are publicly available at: https://github.com/wtepsan/PrEP-AI-Recommendations (accessed on 11 February 2026).

Acknowledgments: We sincerely thank Wisawa Uboncharoen, Wachiraphan Thaimit, and Chaiwat Songsiriphan for their invaluable contributions in labeling client profiles and evaluating the framework. We also extend our gratitude to the Erawan HPC at the Information Technology Service Center, Chiang Mai University, Thailand, for providing the computational resources that made this work possible.

Conflicts of Interest: P.P. is affiliated with Love Foundation, which operates the Love2Test digital health platform from which the de-identified data used in this study were derived. P.P., A.K., B.S., and W.T. declare no competing interests. B.-L.N. and D.T. are employees of Gilead Sciences, which funded this study. The funder, as an institution, had no role in the independent data curation, formal analysis, model development, or publication decision. Any contributions from B.L.N. and D.T. were made in their capacity as named co-authors and are reported transparently in the Author Contributions statement.

Abbreviations

The following abbreviations are used in this manuscript:

PrEP	Pre-exposure prophylaxis
HIV	Human Immunodeficiency Virus
AI	Artificial intelligence
ML	Machine learning
GenAI	Generative artificial intelligence
CDC	Centers for Disease Control and Prevention
STI	Sexually transmitted infection
MSM	Men who have sex with men
MSW	Male sex workers
TGW	Transgender women
PWID	People who inject drugs
KPLHS	Key population-led health services
EHR	Electronic health record

Appendix A

Appendix A.1

High-Level Prompt Template Used for the Generative AI Component

The generative artificial intelligence (GenAI) component was implemented using a structured prompt-based approach. The prompt combined the machine learning (ML)–derived HIV risk classification with selected client behavioral risk factors and instructed the model to generate a brief, supportive, and non-diagnostic recommendation regarding the potential appropriateness of pre-exposure prophylaxis (PrEP). A high-level version of the prompt template used in this study is provided below.

Prompt Template

You are an AI decision-support assistant specialized in HIV prevention. Please respond in {setLang} and include the titles "Explanation" and "Final Thought".
Your task is to provide brief, supportive, non-diagnostic guidance on whether HIV pre-exposure prophylaxis (PrEP) may be appropriate for a client based on the following:

ML Model Recommendation: {ml_recommendation}
Client Behavioral Risk Factors: {client_profile}

Risk factors that can support PrEP consideration include:

- Condomless vaginal or anal sex with partners of unknown HIV status
- HIV-positive partner, especially if the partner's viral load is detectable or unknown
- Recent diagnosis of a bacterial sexually transmitted infection
- Injection drug use involving shared needles or equipment
- Engagement in transactional or survival sex
- Desire to conceive with a partner who is HIV-positive or whose HIV status is unknown

Instructions:

- Summarize the client's relevant risk factors and explain whether PrEP may be appropriate based on both the ML model recommendation and the behavioral risk factors.
- Highlight both factors that may increase HIV risk and factors that may lower risk.
- Keep the response supportive, clear, and concise.
- Do not diagnose HIV, prescribe treatment, or claim to replace a clinician.
- If the information is incomplete, inconsistent, or borderline, state that further clinical assessment is recommended.
- If the apparent risk level is low, provide a shorter explanation.

{optional_concerns_block}
Format your response as follows:
Explanation: Briefly explain the relevant factors contributing to the recommendation.
{optional_concerns_output}
Final Thought: Provide a brief, warm, and supportive conclusion. Emphasize that the decision should be made with appropriate clinical guidance and that PrEP is a highly effective HIV prevention option when used as prescribed.

Prompt Design Notes. The prompt was designed to constrain the model to recommendation-style, non-diagnostic outputs. It explicitly avoided unsupported diagnostic, prescriptive, or clinician-replacing claims and instructed the model to recommend further clinical assessment when the available information was incomplete, inconsistent, or borderline. The output structure was standardized into an "Explanation" section and a "Final Thought" section to improve consistency, readability, and clinical interpretability. This prompt was

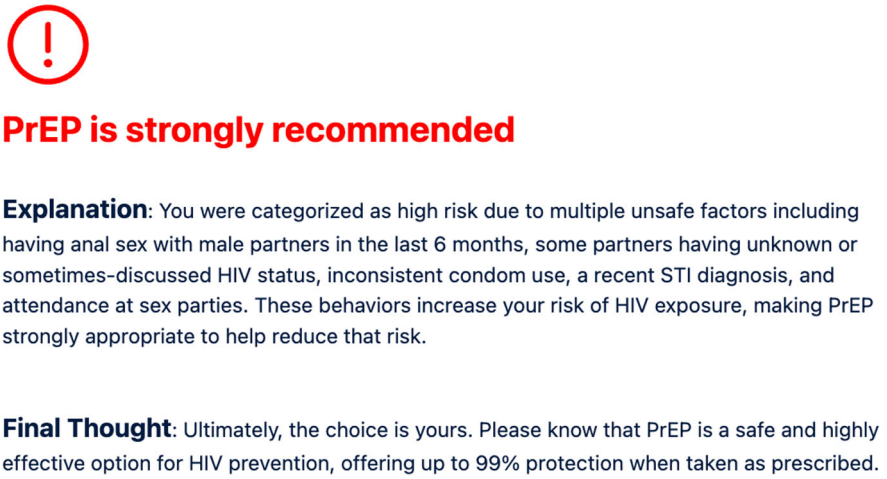
used for decision-support and communication assistance only and was not intended for autonomous clinical judgment.

Appendix A.2

Example 1: High-Risk Profile—MSM with Multiple Risk Factors

Client Profile (de-identified): Male; men who have sex with men (MSM); reported anal sex with male partners within the past 6 months; multiple sexual partners with unknown or undisclosed HIV status; inconsistent condom use; recent bacterial STI diagnosis; participation in sex parties; no injection drug use.

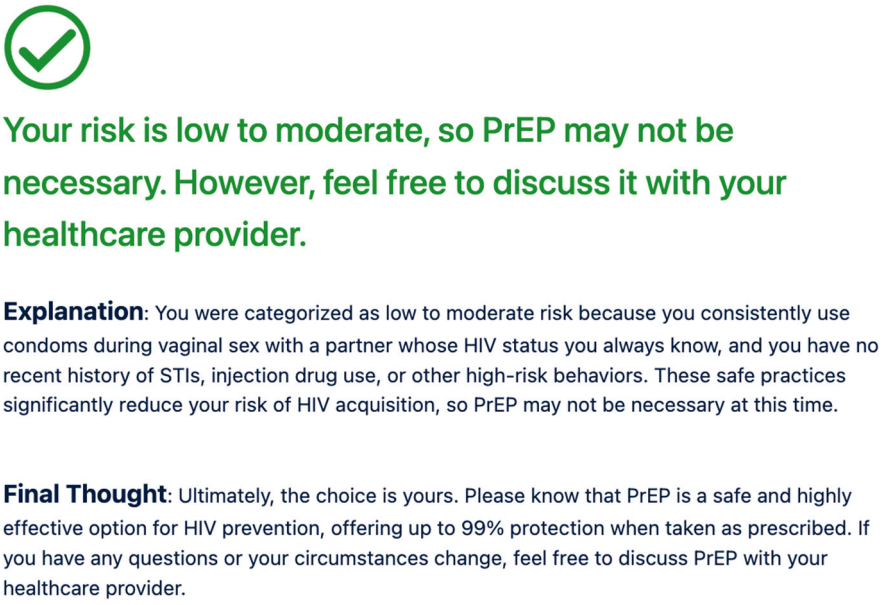
ML Model Recommendation: High HIV Risk—PrEP Recommended



Example 2: Low-Risk Profile—Consistent Protective Behaviors

Client Profile (de-identified): Female; reported vaginal sex within the past 6 months; one sexual partner with known HIV-negative status; consistent condom use; no recent history of bacterial STI; no injection drug use; no transactional sex; no group sex participation.

ML Model Recommendation: Low HIV Risk—PrEP May Not Be Necessary



Appendix A.3

Survey Questions for Healthcare Professionals to Evaluate Each Profile Recommendation

- Clinical Accuracy: How accurate is the recommendation from a clinical perspective?
- Explanation Validity: Does the explanation correctly identify both unsafe and safe factors?
- Error Minimization: Does the explanation effectively minimize the likelihood of false positives (unnecessary recommendations) and false negatives (missed recommendations)?
- Contextual Relevance: Does the recommendation appropriately align with the specific context of the client's profile?

Table A1. Scoring Metrics for Profile-Specific Recommendations.

Metric	Description	Score Range
Clinical Accuracy	How accurate is the recommendation from a clinical perspective?	1 (Low)–5 (High)
Explanation Validity	Does the explanation correctly identify both unsafe and safe factors?	1 (Low)–5 (High)
Error Minimization	Does the explanation minimize the likelihood of false positives (unnecessary recommendations) and false negatives (missed recommendations)?	1 (Low)–5 (High)
Contextual Relevance	Does the recommendation appropriately align with the specific context of the client's profile?	1 (Low)–5 (High)

Appendix A.4

Survey Questions for Healthcare Professionals to Evaluate the Overall System Integration

- Desirability: How effectively could a personalized AI-driven recommendation system address common barriers to PrEP uptake among patients?
- Feasibility: Do healthcare systems have the necessary capacity, infrastructure, and expertise to integrate AI/ML technologies for personalized PrEP recommendations effectively?
- Sustainability: Does this solution have the potential to contribute to long-term, sustainable improvements in HIV prevention and care?

Table A2. Scoring Metrics for Broader System Impact.

Metric	Description	Score Range
Desirability	How effectively could a personalized AI-driven recommendation system address common barriers to PrEP uptake among patients?	1 (Low)–5 (High)
Feasibility	Do healthcare systems have the capacity to integrate AI/ML technologies for personalized PrEP recommendations effectively?	1 (Low)–5 (High)
Sustainability	Does this solution have the potential to contribute to long-term, sustainable improvements in HIV prevention and care?	1 (Low)–5 (High)

References

1. UNAIDS. *The Urgency of Now: AIDS at Crossroads*; UNAIDS: Geneva, Switzerland, 2024. Available online: https://www.unaids.org/sites/default/files/media_asset/2024-unaids-global-aids-update_en.pdf (accessed on 1 February 2025).
2. World Health Organization. *Progress Towards 95-95-95 Targets*; WHO: Geneva, Switzerland, 2023. Available online: https://cdn.who.int/media/docs/default-source/hq-hiv-hepatitis-and-stis-library/j0294-who-hiv-epi-factsheet-v7.pdf (accessed on 7 June 2026).
3. Centers for Disease Control and Prevention. *Pre-Exposure Prophylaxis (PrEP): Clinical Guidance*; HIV Nexus; CDC: Atlanta, GA, USA, 2024. Available online: https://www.cdc.gov/hivnexus/hcp/prep/index.html (accessed on 7 June 2026).
4. Fonner, V.A.; Dalglish, S.L.; Kennedy, C.E.; Baggaley, R.; O'Reilly, K.R.; Koechlin, F.M.; Rodolph, M.; Hodges-Mameletzis, I.; Grant, R.M. Effectiveness and safety of oral HIV preexposure prophylaxis for all populations. *AIDS* **2016**, *30*, 1973–1983. [CrossRef] [PubMed]
5. Haberer, J.E.; Bangsberg, D.R.; Baeten, J.M.; Curran, K.; Koechlin, F.; Amico, K.R.; Anderson, P.; Mugo, N.; Venter, F.; Goicochea, P.; et al. Defining success with HIV pre-exposure prophylaxis: A prevention-effective adherence paradigm. *AIDS* **2015**, *29*, 1277–1285. [CrossRef] [PubMed]
6. Garrison, L.E.; Haberer, J.E. Pre-exposure prophylaxis uptake, adherence, and persistence: A narrative review of interventions in the U.S. *Am. J. Prev. Med.* **2021**, *61*, S73–S86. [CrossRef] [PubMed]
7. Menza, T.W.; Hughes, J.P.; Celum, C.L.; Golden, M.R. Prediction of HIV acquisition among men who have sex with men. *Sex. Transm. Dis.* **2009**, *36*, 547–555. [CrossRef] [PubMed]
8. Hoenigl, M.; Weibel, N.; Mehta, S.R.; Anderson, C.M.; Jenks, J.; Green, N.; Gianella, S.; Smith, D.M.; Little, S.J. Development and validation of the San Diego Early Test Score to predict acute and early HIV infection risk in men who have sex with men. *Clin. Infect. Dis.* **2015**, *61*, 468–475. [CrossRef] [PubMed]
9. Yin, L.; Zhao, Y.; Peratikos, M.B.; Song, L.; Zhang, X.; Xin, R.; Sun, Z.; Xu, Y.; Zhang, L.; Hu, Y.; et al. Risk prediction score for HIV infection: Development and internal validation with cross-sectional data from men who have sex with men in China. *AIDS Behav.* **2018**, *22*, 2267–2276. [CrossRef] [PubMed]
10. Krakower, D.S.; Gruber, S.; Hsu, K.; Menchaca, J.T.; Maro, J.C.; Kruskal, B.A.; Wilson, I.B.; Mayer, K.H.; Klompas, M. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: A modelling study. *Lancet HIV* **2019**, *6*, e696–e704. [CrossRef] [PubMed]
11. Marcus, J.L.; Hurley, L.B.; Krakower, D.S.; Alexeeff, S.; Silverberg, M.J.; Volk, J.E. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: A modelling study. *Lancet HIV* **2019**, *6*, e688–e695. [CrossRef] [PubMed]
12. Saldana, C.S.; Burkhardt, E.; Pennisi, A.; Oliver, K.; Olmstead, J.; Holland, D.P.; Gettings, J.; Mauck, D.; Austin, D.; Wortley, P.; et al. Development of a machine learning modeling tool for predicting HIV incidence using public health data from a county in the Southern United States. *Clin. Infect. Dis.* **2024**, *79*, 717–726. [CrossRef] [PubMed]
13. Ridgway, J.P.; Almirol, E.A.; Bender, A.; Richardson, A.; Schmitt, J.; Friedman, E.; Lancki, N.; Leroux, I.; Pieroni, N.; Dehlin, J.; et al. Which patients in the emergency department should receive pre-exposure prophylaxis? Implementation of a predictive analytics approach. *AIDS Patient Care STDS* **2018**, *32*, 202–207. [CrossRef] [PubMed]
14. Nethi, A.K.; Karam, A.G.; Alvarez, K.S.; Luque, A.E.; Nijhawan, A.E.; Adhikari, E.; King, H.L. Using Machine Learning to Identify Patients at Risk of Acquiring HIV in an Urban Health System. *J. Acquir. Immune Defic. Syndr.* **2024**, *97*, 40–47. [CrossRef] [PubMed]
15. May, S.B.; Giordano, T.P.; Gottlieb, A. Generalizable pipeline for constructing HIV risk prediction models across electronic health record systems. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 666–673. [CrossRef] [PubMed]
16. Balzer, L.B.; Havlir, D.V.; Kamya, M.R.; Chamie, G.; Charlebois, E.D.; Clark, T.D.; Koss, C.A.; Kwarisiima, D.; Ayieko, J.; Sang, N.; et al. Machine learning to identify persons at high risk of human immunodeficiency virus acquisition in rural Kenya and Uganda. *Clin. Infect. Dis.* **2020**, *71*, 2326–2333. [CrossRef] [PubMed]
17. He, J.; Li, J.; Jiang, S.; Cheng, W.; Jiang, J.; Xu, Y.; Yang, J.; Zhou, X.; Chai, C.; Wu, C. Application of machine learning algorithms in predicting HIV infection among men who have sex with men: Model development and validation. *Front. Public Health* **2022**, *10*, 967681. [CrossRef] [PubMed]
18. Xiang, Y.; Du, J.; Fujimoto, K.; Li, F.; Schneider, J.; Tao, C. Application of artificial intelligence and machine learning for HIV prevention interventions. *Lancet HIV* **2022**, *9*, e54–e62. [CrossRef] [PubMed]
19. Yu, E.; Du, J.; Xiang, Y.; Hu, X.; Feng, J.; Luo, X.; Schneider, J.A.; Zhi, D.; Fujimoto, K.; Tao, C. Explainable artificial intelligence and domain adaptation for predicting HIV infection with graph neural networks. *Ann. Med.* **2024**, *56*, 2407063. [CrossRef] [PubMed]
20. Weissman, S.; Yang, X.; Zhang, J.; Chen, S.; Olatosi, B.; Li, X. Using a machine learning approach to explore predictors of healthcare visits as missed opportunities for HIV diagnosis. *AIDS* **2021**, *35*, S7–S18. [CrossRef] [PubMed]

21. Govathson, C.; Chetty-Makkan, C.; Greener, R.; Frade, S.; Rech, D.; Morris, S.; Richard, Y.; Mendonca, R.; Maricich, N.; Long, L.; et al. Breaking barriers: Harnessing artificial intelligence for a stigma-free, efficient HIV prevention assessment among adults in South Africa. *Front. Digit. Health* **2026**, *7*, 1731002. [CrossRef] [PubMed]
22. Ngcobo, S.; Mntla, E.M.; Shock, J.; Louw, M.; Mbonambi, L.; Serite, T.; Rossouw, T. Artificial intelligence for HIV care: A global systematic review of current studies and emerging trends. *J. Int. AIDS Soc.* **2025**, *28*, e70045. [CrossRef] [PubMed]
23. Jin, R.; Zhang, L. AI applications in HIV research: Advances and future directions. *Front. Microbiol.* **2025**, *16*, 1541942. [CrossRef] [PubMed]
24. Ministry of Public Health Thailand. *2024 Report on PrEP Usage*; Ministry of Public Health: Nonthaburi, Thailand, 2024.
25. Phanuphak, N.; Sungsing, T.; Jantarapakde, J.; Pengnonyang, S.; Trachunthong, D.; Mingkwanrungruang, P.; Sirisakyot, W.; Phiayura, P.; Seekaew, P.; Panpet, P.; et al. Princess PrEP program: The first key population-led model to deliver pre-exposure prophylaxis to key populations by key populations in Thailand. *Sex. Health* **2018**, *15*, 542–555. [CrossRef] [PubMed]
26. Cheewanan, L.; Chomnad, M.; Nittaya, P.; Deondara, T.; Thana, K.; Tharee, P.; Supabhorn, P.; Patcharaporn, P.; Prin, V.; Surang, J.; et al. Providing HIV pre-exposure prophylaxis to men who have sex with men and transgender women in hospitals and community-led clinics in Thailand: Acceptance, patterns of use, trends in risk behaviors, and HIV incidence. *AIDS Care* **2023**, *35*, 524–537. [CrossRef] [PubMed]
27. PrEPWatch. Thailand Country Profile. 2025. Available online: https://www.prepwatch.org/country/thailand (accessed on 1 October 2025).
28. Chautrakarn, S.; Rayanakorn, A.; Intawong, K.; Chariyalertsak, C.; Khemngern, P.; Stonington, S.; Chariyalertsak, S. PrEP stigma among current and non-current PrEP users in Thailand: A comparison between hospital and key population-led health service settings. *Front. Public Health* **2022**, *10*, 1019553. [CrossRef] [PubMed]
29. Ali, M. PyCaret: An Open Source, Low-Code Machine Learning Library in Python. 2020. Available online: https://pycaret.org (accessed on 1 October 2025).
30. Centers for Disease Control and Prevention; US Public Health Service. *Preexposure Prophylaxis for the Prevention of HIV Infection in the United States—2021 Update: A Clinical Practice Guideline*; CDC: Atlanta, GA, USA, 2021. Available online: https://stacks.cdc.gov/view/cdc/112360 (accessed on 7 June 2026).
31. Silvey, S.; Liu, J. Sample size requirements for popular classification algorithms in tabular clinical data: Empirical study. *J. Med. Internet Res.* **2024**, *26*, e60231. [CrossRef] [PubMed]
32. Mitsakakis, N.; Liu, D.; Walters, T.; El Emam, K. Sample size calculation for training ensemble machine learning models on health data. *Patterns* **2026**, *7*, 101498. [CrossRef] [PubMed]
33. Dhiman, P.; Ma, J.; Qi, C.; Bullock, G.; Sergeant, J.C.; Riley, R.D.; Collins, G.S. Sample size requirements are not being considered in studies developing prediction models for binary outcomes: A systematic review. *BMC Med. Res. Methodol.* **2023**, *23*, 188. [CrossRef] [PubMed]
34. Albernas, A.; Patel, M.D.; Cook, R.L.; Vaddiparti, K.; Prosperi, M.; Liu, Y. HIV Risk Score and Prediction Model in the United States: A Scoping Review. *AIDS Behav.* **2025**, *29*, 2388–2407. [CrossRef] [PubMed]
35. Pool, J.; Indulska, M.; Sadiq, S. Large Language Models and Generative AI in Telehealth: A Responsible Use Lens. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 2125–2136. [CrossRef] [PubMed]
36. Rabbani, S.A.; El-Tanani, M.; Sharma, S.; Rabbani, S.S.; El-Tanani, Y.; Kumar, R.; Saini, M. Generative Artificial Intelligence in Healthcare: Applications, Implementation Challenges, and Future Directions. *BioMedInformatics* **2025**, *5*, 37. [CrossRef]
37. Roshani, M.; Zhou, X.; Qiang, Y.; Suresh, S.; Hicks, S.; Sethuraman, U.; Zhu, D. Generative Large Language Model—Powered Conversational AI App for Personalized Risk Assessment: Case Study in COVID-19. *JMIR AI* **2025**, *4*, e67363. [CrossRef] [PubMed]
38. Dullabh, P.; Zott, C.; Gauthreaux, N.; Peterson, C.; Aronoff, A.; Monkhouse, K.; Sittig, D. Integrating Generative AI into Patient-Centered Clinical Decision Support: Viewpoint on Research and Practice Considerations. *J. Med. Internet Res.* **2026**, *28*, e81628. [CrossRef] [PubMed]
39. Fieggen, J.; Smith, E.; Arora, L.; Segal, B. The Role of Machine Learning in HIV Risk Prediction. *Front. Reprod. Health* **2022**, *4*, 1062387. [CrossRef] [PubMed]
40. Joseph, J. Algorithmic Bias in Public Health AI: A Silent Threat to Equity in Low-Resource Settings. *Front. Public Health* **2025**, *13*, 1643180. [CrossRef] [PubMed]
41. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* **2019**, *366*, 447–453. [CrossRef] [PubMed]
42. Olayiwola, Q.B.; Sanusi, O.M.; Amoo, G.S.; Agboola, O.J.; Adeyemi, J.A.; Suleiman, H.A.; Ibrahim, M.A.; Hassan, T.A. Barriers to Digital Health Implementation in Low- and Middle-Income Countries: A Narrative Review. *Discov. Public Health* **2026**, *23*, 499. [CrossRef]
43. Wahl, B.; Cossy-Gantner, A.; Germann, S.; Schwalbe, N.R. Artificial Intelligence (AI) and Global Health: How Can AI Contribute to Health in Resource-Poor Settings? *BMJ Glob. Health* **2018**, *3*, e000798. [CrossRef] [PubMed]

44. Versteegh, L.; Amatavete, S.; Chinbunchorn, T.; Thammasiha, N.; Mukherjee, S.; Popping, S.; Triamvichanon, R.; Pusamang, A.; Colby, D.J.; Avery, M.; et al. The Epidemiological Impact and Cost-Effectiveness of Key Population-Led PrEP Delivery to Prevent HIV among Men Who Have Sex with Men in Thailand: A Modelling Study. *Lancet Reg. Health Southeast Asia* **2022**, *7*, 100097. [CrossRef] [PubMed]
45. Phanuphak, N.; Anand, T.; Jantarapakde, J.; Nitpolprasert, C.; Himmad, K.; Sungsing, T.; Trachunthong, D.; Phomthong, S.; Phoseeta, P.; Tongmuang, S.; et al. What Would You Choose: Online or Offline or Mixed Services? Feasibility of Online HIV Counselling and Testing among Thai Men Who Have Sex with Men and Transgender Women and Factors Associated with Service Uptake. *J. Int. AIDS Soc.* **2018**, *21*, e25118. [CrossRef] [PubMed]
46. World Health Organization; International Telecommunication Union. *Digital Health Platform Handbook: Building a Digital Information Infrastructure (Infostructure) for Health*; World Health Organization: Geneva, Switzerland, 2020. Available online: https://www.who.int/publications/i/item/9789240013728 (accessed on 2 June 2026).
47. Campos, H.; Wolfe, D.; Luan, H.; Sim, I. Generative AI as Third Agent: Large Language Models and the Transformation of the Clinician–Patient Relationship. *J. Particip. Med.* **2025**, *17*, e68146. [CrossRef] [PubMed]
48. Araf, I.; Idri, A.; Chairi, I. Cost-Sensitive Learning for Imbalanced Medical Data: A Review. *Artif. Intell. Rev.* **2024**, *57*, 80. [CrossRef]
49. Abbas, Q.; Jeong, W.; Lee, S.W. Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges. *Healthcare* **2025**, *13*, 2154. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.