

Governing Artificial Intelligence in University Mental Health Services: A Population Health Intelligence Framework for Preventing Algorithmic Stratification

Effiong Ndarake Effiong

effiongneffiong@gmail.com

Nasarawa State University

Julius Barnabas Gajere

Community AIDS/HIV and Tuberculosis Foundation

Research Article

Keywords: Population Health Intelligence, algorithmic stratification, university mental health, artificial intelligence governance, student privacy, digital health equity

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11016638/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Universities worldwide are adopting artificial intelligence (AI) tools for student mental health triage, risk prediction, and counselling allocation. These tools represent a specific application of Population Health Intelligence (PHI); the use of multi-source data and analytics for surveillance, resource allocation, and policy decision-making. However, AI mental health tools may produce algorithmic stratification: the systematic sorting of students into unequal tiers of visibility, risk, and care. University populations present unique PHI governance challenges including developmental transitions, heightened privacy expectations, and the tension between institutional duty of care and student autonomy.

Objective

To develop the SAFE-STUDENT governance framework for AI mental health tools in university settings, addressing the specific PHI governance risks of student populations and preventing algorithmic stratification while respecting student privacy, autonomy, and developmental needs.

Methods

We conducted a targeted conceptual synthesis of literature on university mental health, digital health ethics, student privacy, and higher education governance. Building on algorithmic stratification theory and PHI governance principles, we used a four-stage conceptual synthesis: (1) problem identification, (2) operationalisation of four university-specific governance challenges, (3) principle development using criteria of distinctiveness, applicability, and measurability, and (4) translation into stakeholder-specific guidance with safeguards mapped across the PHI-to-decision pathway.

Results

Four PHI governance challenges salient in university contexts were identified: (1) privacy-surveillance tension, (2) developmental considerations in student-AI interaction, (3) the digital native assumption and its equity implications, and (4) institutional duty of care versus student autonomy. The SAFE-STUDENT framework comprises 11 principles (Student-Centred, Accountability, Fairness Monitoring, Escalation, Safety, Transparency, User Agency, Data Privacy, Equity, Non-Replacement) tailored with implementation indicators mapped across seven PHI stages, data generation to outcomes and stakeholder-specific guidance for seven stakeholder groups. Failure points and equity signals are mapped across the PHI pathway.

Conclusions

General AI governance frameworks are insufficient for university mental health. SAFE-STUDENT provides a tailored governance mechanism for equitable intelligence generation in university digital health settings. Empirical validation via a three-stage research agenda is required.

Background

University student mental health has become a critical global concern, with over 40% of students meeting criteria for clinically significant difficulties [1, 2]. Rates are comparable or higher in Africa, Asia and Latin America [3, 4]. Counselling services face capacity-demand gaps, leading to waitlists and triage [5, 6].

Artificial intelligence has emerged as a response, including chatbots delivering cognitive behavioural therapy [7, 8], risk prediction for self-harm [6], triage algorithms, sentiment analysis, and digital phenotyping using learning management system activity [9, 10]. Recent work shows explainable machine learning can identify distinct risk subtypes [11], demonstrating both precision potential and stratification risk.

We conceptualise university mental health AI as Population Health Intelligence (PHI): the full pathway from data generation, integration, AI analysis, intelligence generation, decision-making, resource allocation to outcomes [12, 13]. This integration of educational and health data creates distinctive governance challenges under-explored in the literature.

We define algorithmic stratification as the systematic sorting of students into unequal tiers of visibility, risk, and care through AI-mediated processes. Mechanisms include representational inequality [14], predictive triage hierarchies [15], visibility asymmetry [16], governance opacity [17], and feedback amplification [18, 19].

University contexts add four unique challenges: privacy-surveillance tension from repurposing learning data [20, 21]; developmental considerations, age 18–25 identity formation [22]; the digital native assumption obscuring digital divides [16, 23]; and duty of care versus autonomy [24]. Existing frameworks; WHO AI Ethics (2021) [25], FUTURE-AI [26], and AUGMENT for LMIC health systems [27] do not address these specifically.

Central argument: Governance of AI mental health tools in universities must address the specific PHI governance challenges and stakeholder dynamics of student populations. This paper: (1) identifies four PHI governance challenges, (2) develops the SAFE-STUDENT framework, and (3) positions it within PHI research on equitable intelligence.

Methods

Study Design

This is a conceptual governance framework development paper. The objective is to develop a framework tailored to the specific governance challenges of AI mental health tools in university student populations. The framework is intended as a conceptual contribution requiring empirical validation.

The design integrates three components: (1) identification of governance challenges through targeted literature engagement; (2) operationalisation of these challenges into governance principles; and (3) translation of principles into stakeholder-specific implementation guidance. This approach reflects established methodologies for developing governance frameworks in health and technology policy [28, 29].

Conceptual Foundations

The SAFE-STUDENT framework was developed through engagement with four literatures: algorithmic fairness and stratification [14, 15, 17], digital mental health governance [5, 6, 24], student privacy and higher education governance [20, 21, 24], and Population Health Intelligence [12, 13]. These literatures were selected because they address the core dimensions of the governance problem: how AI systems generate intelligence about populations, how that intelligence may be stratified, and how governance can ensure equitable outcomes.

Rather than conducting a systematic review, we engaged with these literatures to identify recurring governance challenges and to derive principles that address them. The selection of literature was purposive, drawing on foundational and contemporary sources that represent the state of knowledge in each domain. This approach is consistent with conceptual framework development in health services research [27, 28].

Framework Derivation

The SAFE-STUDENT framework was developed through a four-stage conceptual synthesis process:

Stage 1: Problem identification. Recurring governance challenges in university mental health AI were identified through thematic engagement with the literature. Four challenges were identified as particularly salient: the privacy-surveillance tension, developmental considerations, the digital native assumption and its equity implications, and the tension between institutional duty of care and student autonomy.

Stage 2: Challenge operationalisation. Each challenge was operationalised as a distinct governance problem requiring specific responses.

Stage 3: Principle development. Governance principles were derived from the literature on responsible AI, student privacy, and higher education governance. Principles were retained when they met three

criteria: (a) they addressed a distinct governance challenge, (b) they were applicable to university contexts, and (c) they could be translated into measurable indicators.

Stage 4: Implementation translation. Principles were translated into stakeholder-specific implementation guidance, including indicators and governance mechanisms mapped across the PHI pathway.

Normative Foundations

The SAFE-STUDENT framework is grounded in four normative commitments: student autonomy - students are adults with rights to privacy, informed consent, and refusal of services [24]; distributive justice - AI systems should not systematically disadvantage historically marginalised student populations [31]; procedural transparency - students and staff should be able to understand, scrutinise, and contest AI-mediated decisions [17, 32]; and dynamic accountability - fairness requires ongoing monitoring and correction [29, 18].

Note on Literature Engagement

This paper does not claim to be a systematic review. The literature engagement was purposive and conceptual, intended to inform framework development rather than to provide a comprehensive synthesis of all available evidence. The framework is offered as a conceptual contribution requiring empirical validation through the research agenda outlined in the Discussion.

Results

Algorithmic Stratification in the PHI Pathway

Population Health Intelligence is defined as an interdisciplinary scientific framework that integrates multi-source population health data, epidemiological evidence, advanced analytical methods, implementation science, and decision-support tools to generate actionable intelligence for surveillance, prevention, resource allocation, and policy decision-making at the population level under conditions of uncertainty. It encompasses the full pathway through which population health data become actionable intelligence for health-system decision-making. This pathway includes: (1) data generation, (2) data integration, (3) AI analysis, (4) intelligence generation, (5) decision-making, (6) resource allocation, and (7) outcomes [12, 13].

Algorithmic stratification occurs when underrepresentation, visibility asymmetry, differential performance, opacity, and inequitable allocation interact in a self-reinforcing cycle:

underrepresentation → lower visibility → weaker prediction → reduced support

→ altered engagement → biased retraining → amplified disparity

Research on feedback loops in algorithmic systems has demonstrated how predictions can affect the data subsequently available to the system [18, 19]. In university mental health, this means that students initially classified as low risk may receive less monitoring and support, generating less data that would allow the system to detect emerging needs, creating a self-fulfilling prophecy of reduced visibility.

Unique PHI Governance Challenges in University Mental Health

Four PHI governance challenges are particularly salient in university contexts:

Privacy-surveillance tension. Universities are data-rich environments, collecting information across learning management systems, Wi-Fi networks, library systems, and campus activities. The repurposing of academic data for mental health surveillance creates tension between institutional duty of care and student privacy rights [20, 21], with consent complexity amplified by the power asymmetry between universities and students. This is further complicated by overlapping regulatory frameworks including FERPA and HIPAA [24].

Developmental considerations. University students (ages 18–25) occupy a developmental period characterised by identity formation, increased independence from family, and often first encounters with mental health services [22]. AI tools deployed during this period may have lasting effects on help-seeking behaviour, trust in services, and self-understanding. This developmental dimension is largely absent from existing governance frameworks.

Digital native assumption and equity implications. The widespread assumption that university students are uniformly digitally engaged obscures significant digital divides [16, 23]. International students, first-generation students, and those from lower socioeconomic backgrounds face barriers in connectivity, device access, digital literacy, and cultural language factors [33]. This creates a visibility–need divergence: students most in need may be least visible to AI systems.

Duty of care versus student autonomy. Universities have legal and ethical duties to intervene when students are at risk. Yet students are adults with autonomy rights. AI tools may generate false positives, and overly zealous intervention can harm trust and discourage help-seeking [24]. Governance must balance appropriate escalation with respect for student agency.

The SAFE-STUDENT Governance Framework

The SAFE-STUDENT framework provides governance principles, implementation indicators mapped across the PHI pathway, and stakeholder-specific guidance for AI mental health tools in university settings. The framework is designed as an integrated system: failure in any principle threatens governance effectiveness.

The framework's name reflects its primary concerns: Safety and wellbeing, Accountability, Fairness, Escalation and human oversight, Student-centred governance, Transparency, User agency, Data privacy, Equity, Non-replacement of human care, Tailored implementation Fig. 1 maps SAFE-STUDENT safeguards across the seven PHI stages, identifying governance failure points and equity signals at each stage. This pathway-level operationalisation distinguishes SAFE-STUDENT from existing frameworks that focus primarily on AI model-level fairness or broad governance principles without operational mapping.

The SAFE-STUDENT framework is operationalised across six tables that collectively map governance safeguards onto the PHI-to-decision pathway. Table 1 identifies governance failure points at each of the seven PHI stages; data generation, data integration, AI analysis, intelligence generation, decision-making, resource allocation, and outcomes along with the mechanisms, consequences, and prevention measures associated with each. Table 2 presents the 11 SAFE-STUDENT principles (Student-Centred, Accountability, Fairness Monitoring, Escalation, Safety, Transparency, User Agency, Data Privacy, Equity, Non-Replacement, and Tailored), their meanings, the PHI stages to which they apply, and the specific governance challenges they address. Table 3 translates each principle into implementation indicators and measurement approaches across the PHI stages, providing institutions with concrete metrics for monitoring. Table 4 maps the SAFE-STUDENT safeguards against equity failure signals at each PHI stage, enabling early identification of stratification risks. Table 5 provides stakeholder-specific implementation actions for seven stakeholder groups; university administration, counselling services, IT/data governance, AI vendors, student representatives, international student services, and faculty/academic affairs, linking each action to the principles operationalised. Finally, Table 6 compares SAFE-STUDENT with existing governance frameworks; the WHO Guidance on Ethics and Governance of AI for Health (2021), FUTURE-AI (2025), and AUGMENT for LMIC health systems, demonstrating SAFE-STUDENT's distinctive focus on the privacy-surveillance tension, the developmental period of university students, student autonomy, the duty-of-care tension, the digital native assumption critique, student voice, and academic-mental health cross-domain governance. Together, these tables transform SAFE-STUDENT from a set of principles into an operational governance instrument that can be applied across the full PHI-to-decision pathway.

Table 1
PHI Governance Failure Points in University Mental Health AI

PHI Stage	Governance Failure Point	Mechanism	Consequence	Prevention
Data Generation	Data collection without consent	Passive surveillance	Privacy violation; chilling effect	Explicit informed opt-in; opt-out mechanisms
Data Integration	Function creep	Academic data repurposed for mental health	Mission blur; privacy erosion	Purpose limitation; data sharing agreements
AI Analysis	Model bias	Western/English training data	Misclassification of diverse students	Subgroup validation; diverse training data
Intelligence Generation	Opacity	Black-box vendor models	Lack of contestability	Model cards; transparency requirements
Decision-Making	Algorithmic triage bias	High-engagement prioritised	Inequitable resource allocation	Allocation audit; equity monitoring
Resource Allocation	Visibility asymmetry	Low-engagement invisible	Underservice of vulnerable	Alternative pathways; non-digital sources
Outcomes	Widening disparities	Feedback loops amplify inequities	Compounded disadvantage	Longitudinal monitoring; corrective action

Table 2
SAFE-STUDENT Principles with PHI Pathway Mapping

Principle	Meaning	PHI Stage	Challenge Addressed
S - Student-Centred	Governance prioritises wellbeing, autonomy, voice	All stages	Student autonomy
A - Accountability	Clear responsibility for AI decisions	Decision-Making; Resource Allocation	Governance opacity
F - Fairness Monitoring	Continuous subgroup performance monitoring	AI Analysis; Decision-Making	Model bias
E – Escalation	Human review for high-risk	Decision-Making	Inappropriate automation
S – Safety	No harm; appropriate oversight	Outcomes	Duty of care
T - Transparency	Students and staff understand AI role	Intelligence Generation	Governance opacity
U - User Agency	Opt-out and contestability	Data Generation; Decision-Making	Privacy-surveillance; autonomy
D - Data Privacy	Privacy protections; no passive surveillance	Data Generation; Integration	Privacy-surveillance tension
E – Equity	All students served equitably	Resource Allocation; Outcomes	Digital native assumption
N - Non-Replacement	AI augments, not replaces, counselling	Decision-Making; Resource Allocation	Diminished human connection
T – Tailored	Adapted to diverse populations	All stages	International students; digital divide

Table 3
SAFE-STUDENT Implementation Indicators by PHI Stage

PHI Stage	Principle	Implementation Indicator	Measurement Approach
Data Generation	Data Privacy	Passive data collection audit; consent documentation	Privacy audit; consent logs
	User Agency	Opt-out mechanism documented; opt-out rates tracked	Opt-out rates; policy audit
Data Integration	Data Privacy	Purpose limitation policy; data sharing agreements	Policy audit
	Tailored	Data collection adapted for diverse populations	Accessibility audit
AI Analysis	Fairness Monitoring	Subgroup performance reports; corrective actions documented	Quarterly equity reports; incident logs
	Accountability	Named responsible person for AI decisions	Governance documentation
Intelligence Generation	Transparency	Accessible AI documentation; disclosure to students	Documentation audit; student surveys
	Safety	Incident reporting; safety monitoring	Safety reports; incident logs
Decision-Making	Escalation	Documented escalation pathway; time-to-human-review	Escalation logs; time-to-review reports
	Student-Centred	Student representation in governance	Governance documentation
Resource Allocation	Equity	Subgroup analysis; alternative pathways	Equity reports; alternative pathway usage
	Non-Replacement	Human counsellor availability; ratio tracking	Staffing reports; service utilisation
Outcomes	Safety	Mental health outcomes monitoring	Outcomes reports
	Fairness Monitoring	Disparity tracking over time	Longitudinal equity reports

Table 4
SAFE-STUDENT Governance Across the PHI-to-Decision Pathway

PHI Stage	Governance Safeguard	Equity Failure Signal
Data Generation	Population coverage and participation rates stratified by demographic, socioeconomic, linguistic, and geographic subgroups	Systematic under-representation or exclusion of specific student populations
Data Integration	Subgroup-specific completeness; missingness patterns; linkage quality audit	Differential data completeness or systematically higher missingness in underserved groups
AI Analysis	Calibration, sensitivity, specificity, and predictive values stratified by relevant subgroups	Material subgroup performance gaps or systematic miscalibration
Intelligence Generation	Transparency documentation; contestability mechanisms	Opacity; inability of students/staff to understand or challenge classifications
Decision-Making	Human-review, override, and escalation rates for high-risk classifications	High-risk cases proceeding without appropriate human review or escalation
Resource Allocation	Allocation rates, referral completion, service uptake, and time-to-service stratified by subgroup	Differential access, referral completion, or time-to-service despite comparable need
Outcomes	Longitudinal monitoring of model performance, service access, outcomes, and subgroup disparities	Persistent or widening disparities following implementation

Table 5
Stakeholder-Specific Implementation Guidance

Stakeholder	Priority Implementation Actions	Principles Operationalised
University Administration	Establish AI governance board with student representation; mandate equity reporting; ensure privacy protections	S, A, D, F
Counselling Services	Document human escalation pathways; track time-to-human-review; maintain counsellor-student ratios	E, N, S
IT/Data Governance	Implement privacy audit; restrict passive data collection; ensure data security	D, T, U
AI Vendors	Provide model cards; subgroup performance data; cultural adaptation plan; language accessibility	T, F, E, Tailored
Student Representatives	Participate in governance; provide feedback; advocate for student privacy	S, U, T
International Student Services	Review cultural adaptation; language accessibility; international student representation	Tailored, E, S
Faculty/Academic Affairs	Coordinate with mental health services; avoid academic surveillance creep	S, U, Privacy

Table 6
SAFE-STUDENT Compared to Existing PHI Governance Frameworks

Domain	WHO AI Ethics (2021) [25]	FUTURE-AI (2025) [26]	AUGMENT (LMIC Health) [27]	SAFE-STUDENT (This Paper)
Primary Context	General health	Medical imaging	LMIC health systems	University mental health
PHI Framework	Implicit	Limited	Explicit	Explicit
Privacy-Surveillance Focus	General privacy	Data governance	Accessibility without surveillance	Student privacy + surveillance tension
Developmental Considerations	Not addressed	Not addressed	Not addressed	18–25 developmental period
Student Autonomy	Not addressed	Not addressed	Not addressed	Explicit student autonomy focus
Duty of Care vs. Autonomy	General	General	General	Tension explicitly addressed
Digital Divide	General	General	Infrastructure	Digital native assumption critique
Student Voice	Not addressed	Not addressed	Community engagement	Explicit student representation
Academic-Mental Health Intersection	Not addressed	Not addressed	Not addressed	Cross-domain governance
PHI Pathway Mapping	No	No	Yes	Yes

Discussion

Contribution to Population Health Intelligence

The SAFE-STUDENT framework contributes to the broader field of Population Health Intelligence by addressing a critical gap: the governance of AI mental health tools in university student populations. PHI frameworks have traditionally focused on data integration and analytical methods [12, 13]. However, without explicit attention to the governance of algorithmic stratification, PHI systems risk generating intelligence that systematically misrepresents the health needs of marginalised student populations.

SAFE-STUDENT provides the governance function that PHI requires in university settings: a mechanism to validate that the intelligence generated by AI mental health tools is not systematically distorted by stratification dynamics. By auditing data collection, model performance, resource allocation, human

oversight, and student agency across the PHI pathway, the framework ensures that PHI serves all students equitably.

Contribution to Digital Health Governance

This paper contributes to the digital health governance literature in three ways. First, it extends algorithmic stratification theory [14, 15, 17] from general health contexts to the specific domain of university mental health, identifying four governance challenges that are unique to this setting. Second, it provides the SAFE-STUDENT framework a governance framework tailored to the specific tensions of university digital mental health. Third, it positions digital mental health governance within the broader PHI research programme, connecting digital health innovation to population-level equity concerns.

Contribution Beyond Existing PHI Governance Frameworks

This paper builds on and extends existing PHI governance work in two ways. First, while the AUGMENT framework addresses algorithmic stratification in LMIC health systems at the population level [27], SAFE-STUDENT addresses the specific governance challenges of university student populations; a distinct population with unique characteristics including developmental transitions (age 18-25), the privacy-surveillance tension arising from the integration of educational and health data, and the dual governance of academic and mental health systems.

Second, SAFE-STUDENT provides explicit mapping across the PHI-to-decision pathway, identifying governance failure points and equity signals at each stage. Existing frameworks have focused primarily on AI model-level fairness or broad governance principles without operational pathway mapping. This pathway-level operationalisation is the specific contribution of SAFE-STUDENT to the digital health governance literature. Table 6 provides a direct comparison with WHO AI Ethics (2021), FUTURE-AI (2025), and AUGMENT, demonstrating SAFE-STUDENT's distinctive focus on privacy-surveillance tension, developmental period, student autonomy, duty-of-care tension, digital native critique, student voice, and academic-mental health cross-domain governance.

Practical Implications for Digital Health Implementation

The SAFE-STUDENT framework has several practical implications for universities implementing digital mental health tools:

Procurement must include governance requirements. Universities should require AI vendors to demonstrate transparency, auditability, equity monitoring, and cultural adaptation before deployment. Consent must be meaningful. Passive data collection for mental health purposes requires explicit, informed, opt-in consent not opt-out bundled with essential services. Student voice must be included. Students should participate in AI governance through advisory committees, feedback mechanisms, and consent processes. Alternative pathways are essential. Students with low digital engagement must have non-digital or mixed pathways to mental health services. Escalation is mandatory. High-risk classifications must trigger human review. AI should augment, not replace, clinical judgement.

Monitoring must be longitudinal. Universities should track subgroup performance, service access, and equity outcomes over time, not only at deployment.

Limitations

This paper has several limitations. Conceptual status: SAFE-STUDENT is a proposed conceptual framework requiring empirical validation. It should not be treated as a validated instrument. No stakeholder co-design: The framework was developed through literature synthesis rather than participatory design with students, counsellors, or university administrators. No empirical testing: The framework has not been tested in any university setting for feasibility, acceptability, or effectiveness. Generalizability: The framework was developed with reference to university settings in high-income and middle-income countries. Adaptation may be required for specific cultural, regulatory, or institutional contexts. Intersectionality: The framework addresses single-axis equity concerns. Future work should examine how SAFE-STUDENT applies to students with intersecting marginalised identities—for example, international female students with limited digital access who may face compounded visibility and vulnerability. Intersectional analysis of race, class, gender, disability, and international student status is underdeveloped in this version.

Research Agenda

Stage 1: Validation. Develop SAFE-STUDENT governance audit instrument; conduct expert review with university mental health stakeholders; conduct student focus groups for acceptability and feasibility; refine principles and indicators based on stakeholder feedback.

Stage 2: Implementation Studies. Pilot SAFE-STUDENT in 3-5 universities with diverse characteristics; assess feasibility and acceptability; document implementation challenges and adaptations; identify contextual factors affecting implementation.

Stage 3: Effectiveness Evaluation. Examine whether SAFE-STUDENT implementation changes student awareness and understanding of AI mental health tools, reduces subgroup performance disparities, increases human escalation rates for high-risk classifications, improves student trust and satisfaction, and enhances equitable access to mental health services.

Stage 4: Cross-Cultural Adaptation. Adapt SAFE-STUDENT for university settings in different cultural contexts; address specific challenges of international students, language barriers, and cultural concepts of mental health; adapt for resource-constrained university settings.

Conclusions

Universities worldwide are adopting AI tools to address rising student mental health demands. These tools are part of a broader shift toward Population Health Intelligence; the use of multi-source data and analytics for equitable surveillance, resource allocation, and policy decision-making. However, AI mental health tools may produce algorithmic stratification: the systematic sorting of students into unequal tiers

of visibility, risk, and care. University student populations present unique PHI governance challenges that distinguish them from general health AI deployment: developmental transitions, the privacy-surveillance tension, digital divides obscured by the digital native assumption, and the tension between institutional duty of care and student autonomy.

This paper proposed the SAFE-STUDENT governance framework as a response to these challenges. The framework comprises 11 principles; Student-Centred, Accountability, Fairness Monitoring, Escalation, Safety, Transparency, User Agency, Data Privacy, Equity, Non-Replacement, and Tailored - with implementation indicators mapped across the PHI pathway and stakeholder-specific guidance. The framework is designed to prevent algorithmic stratification while respecting student privacy, autonomy, and developmental needs.

The framework should be regarded as a starting point rather than a validated instrument. The next step is empirical testing across diverse university settings, student populations, and cultural contexts. The ultimate question is not simply whether AI mental health tools are fair, but whether the governance systems surrounding them ensure that all students regardless of digital engagement, socioeconomic status, or cultural background receive the mental health support they need, when they need it, in ways that respect their autonomy, privacy, and developmental needs.

The ultimate question is not whether the AI tool is fair. It is whether the PHI governance system ensures that all students become adequately visible, appropriately supported, and meaningfully able to benefit from AI-enabled mental health services.

Declarations

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Competing Interests

The author declares that he has no competing financial or non-financial interests.

Clinical Trial Number:

Not applicable

Use of artificial intelligence tools:

In accordance with BMC policy, we declare that ChatGPT-4.0 was used for language editing only. No content was generated by AI; all intellectual contributions, conceptual development, framework design, theoretical definitions, and analytical arguments are the original work of the author. The author is fully responsible for the accuracy, originality, and integrity of all content.

Funding:

This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The author did not receive any financial support for the research, authorship, and/or publication of this article.

Author Contribution

ENE conceived the study, designed the conceptual framework, conducted the literature synthesis, developed the SAFE-STUDENT framework, and drafted the original manuscript. JBG performed the analytical synthesis, critically reviewed, revised, and proofread the manuscript. Both authors met the International Committee of Medical Journal Editors (ICMJE) criteria for authorship, contributed to the intellectual content, and approved the final version for publication.

Acknowledgement

We thank the Management and Staff of Nasarawa State University, especially the Director of Global Health and Infectious Disease Institute (GHIDI), Prof. Adamu Ishiaku Akyala, and my Project Supervisors, Prof Daniel Mairafi Gimbasan and Prof Yakubu Boyi Ngwai and all staff and students of Efficacy International School, Kaduna, Nigeria, and its dedicated research directorate, the Center for Applied Health Research and Development. The management and staff of Community AIDS/HIV and Tuberculosis Foundation are also highly appreciated.

Availability of data and materials:

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study. All data referenced are from published sources cited in the references.

Code Availability

No custom code was used in the preparation of this manuscript.

References

1. Eisenberg D, Gollust SE, Golberstein E, Hefner JL (2007) Prevalence and correlates of depression, anxiety, and suicidality among university students. *Am J Orthopsychiatry* 77(4):534–542. 10.1037/0002-9432.77.4.534
2. Lipson SK, Zhou S, Abelson S, Heinze J, Jirsa M, Morigney J, Patterson A, Singh M, Eisenberg D (2022) Trends in college student mental health and help-seeking by race/ethnicity: Findings from the national healthy minds study, 2013–2021. *J Affect Disord* 306:138–147. 10.1016/j.jad.2022.03.038
3. Wao H, Wao MA, Muriithi I, Seraii P, Syonguvi J (2022) Availability, accessibility and activation of mental health services among university students in Africa: a protocol of a mixed-methods systematic review with meta-analysis and meta-synthesis. *BMJ Open* 12(6):e059991. 10.1136/bmjopen-2021-059991
4. Kukoyi O, Orok E, Oluwafemi F, Oluwadare T, Oni O, Bamitale T, Jaiyesimi B, Ojo T, Eze H (2022) Factors affecting the utilization of mental health services among undergraduate students in a Nigerian University. *Heliyon* 8(11):e11476. 10.1016/j.heliyon.2022.e11476
5. Lattie EG, Stiles-Shields C, Graham AK (2022) An overview of and recommendations for more accessible digital mental health services. *Nat Rev Psychol* 1(2):87–100. 10.1038/s44159-021-00003-1
6. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, Carvalho AF, Keshavan M, Linardon J, Firth J (2021) The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry* 20(3):318–335. 10.1002/wps.20883
7. Fitzpatrick KK, Darcy A, Vierhile M (2017) Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment Health* 4(2):e19. 10.2196/mental.7785
8. Inkster B, Sarda S, Subramanian V, An Empathy-Driven (2018) *JMIR Mhealth Uhealth* 6(11):e12106. 10.2196/12106. Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study
9. Liu P, Tang J, Wang H, Zhang D (2026) Mapping Heterogeneity in Psychological Risk Among University Students Using Explainable Machine Learning. *Entropy (Basel)* 28(2):224. 10.3390/e28020224
10. Shaban-Nejad A, Michalowski M, Buckeridge DL (2018) Health intelligence: how artificial intelligence transforms population and personalized health. *NPJ Digit Med* 1:53. 10.1038/s41746-018-0058-9
11. Shaban-Nejad A, Michalowski M, Buckeridge DL (2018) Health intelligence: how artificial intelligence transforms population and personalized health. *npj Digital Med* 1, 53 doi.10.1038/s41746-018-0058-9
12. Benjamin R (2019) *Race after technology: abolitionist tools for the new Jim Code*. Polity, Cambridge

13. Virginia Eubanks. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: Picador, St Martin's Press. Law, Technology and Humans 1 (2018) :162 – 64.
doi.10.5204/lthj.v1i0.1386
14. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453. 10.1126/science.aax2342
15. van Deursen AJ, van Dijk JA (2019) The first-level digital divide shifts from inequalities in physical access to inequalities in material access. New Media Soc 21(2):354–375.
10.1177/1461444818797082
16. Ananny M, Crawford K Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. New Media Soc, (2018). 20, 973–989
17. Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S (2018) Runaway Feedback Loops in Predictive Policing. Proceedings of Machine Learning Research. ;81:160–171
18. Hashimoto T, Srivastava M, Namkoong H, Liang P (2018) Fairness Without Demographics in Repeated Loss Minimization. Proceedings of the 35th International Conference on Machine Learning, in Proceedings of Machine Learning Research 80:1929–1938 Available from <https://proceedings.mlr.press/v80/hashimoto18a.html>. Accessed 11th September, 2026
19. Khalil M, Slade S, Prinsloo P Learning analytics in support of inclusiveness and disabled students: a systematic review. J Comput High Educ 2023 Mar 14:1–18. 10.1007/s12528-023-09363-4
20. Williamson B (2017) *Big Data in Education: The Digital Future of Learning, Policy and Practice*. SAGE Publications Ltd. doi.10.4135/9781529714920
21. Arnett JJ (2000) Emerging adulthood. A theory of development from the late teens through the twenties. Am Psychol 55(5):469–480
22. Robinson L, Schulz J, Blank G, Ragnedda M, Ono H et al (2020) Digital in-equalities 2.0: Legacy inequalities in the information age. First Monday. Univ Ill atChicago Libr 25(7).
10.5210/fm.v25i7.10842 halshs-02889893
23. Sontag-Padilla L, Dunbar MS, Ye F, Kase C, Fein R, Abelson S, Seelam R, Stein BD (2018) Strengthening College Students' Mental Health Knowledge, Awareness, and Helping Behaviors: The Impact of Active Minds, a Peer Mental Health Organization. J Am Acad Child Adolesc Psychiatry 57(7):500–507. 10.1016/j.jaac.2018.03.019
24. World Health Organization (2021) Ethics and governance of artificial intelligence for health. WHO, Geneva
25. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, Weicken E, Asselbergs FW, Prior F, Collins GS, Kaissis G, Tsakou G, Buvat I, Kalpathy-Cramer J, Mongan J, Schnabel JA, Kushibar K, Riklund K, Marias K, Amugongo LM, Fromont LA, Maier-Hein L, Cerdá-Alberich L, Martí-Bonmatí L, Cardoso MJ, Bobowicz M, Shabani M, Tsiknakis M, Zuluaga MA, Fritzsche MC, Camacho M, Linguraru MG, Wenzel M, De Bruijne M, Tolsgaard MG, Goisauf M, Cano Abadía M, Papanikolaou N, Lazrak N, Pujol O, Osuala R, Napel S, Colantonio S, Joshi S, Klein S, Aussó S, Rogers WA, Salahuddin Z, Starmans MPA, FUTURE-AI Consortium (2025). FUTURE-AI: international consensus guideline for

trustworthy and deployable artificial intelligence in healthcare. BMJ. ;388:e081554. 10.1136/bmj-2024-081554. Erratum in: BMJ. 2025;388:r340. doi: 10.1136/bmj.r340. PMID: 39909534; PMCID: PMC11795397

26. Effiong EN From model fairness to population equity: the AUGMENT framework for governing AI in LMIC health systems. [Manuscript in review: Journal of Global Health]

27. Mökander J, Floridi L (2023) Operationalising AI governance through ethics-based auditing: an industry case study. AI Ethics 3(2):451–468. 10.1007/s43681-022-00171-7

28. Raji ID, Smart A, White RN, Mitchell M, Gebru T (2020) Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 33–44. doi.10.1145/3351095.3372873

29. Morse JM (2015) Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry. Qual Health Res 25(9):1212–1222. 10.1177/1049732315588501

30. Straw I, Callison-Burch C (2020) Artificial Intelligence in mental health and the biases of language based models. PLoS ONE 15(12):e0240376. 10.1371/journal.pone.0240376

31. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, and Timnit Gebru (2019). Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. doi.10.1145/3287560.3287596

32. Kirmayer LJ, Gomez-Carrillo A, Veissière S (2017) Culture and depression in global mental health: An ecosocial approach to the phenomenology of psychiatric disorders. Soc Sci Med 183:163–168. 10.1016/j.socscimed.2017.04.034

Figures

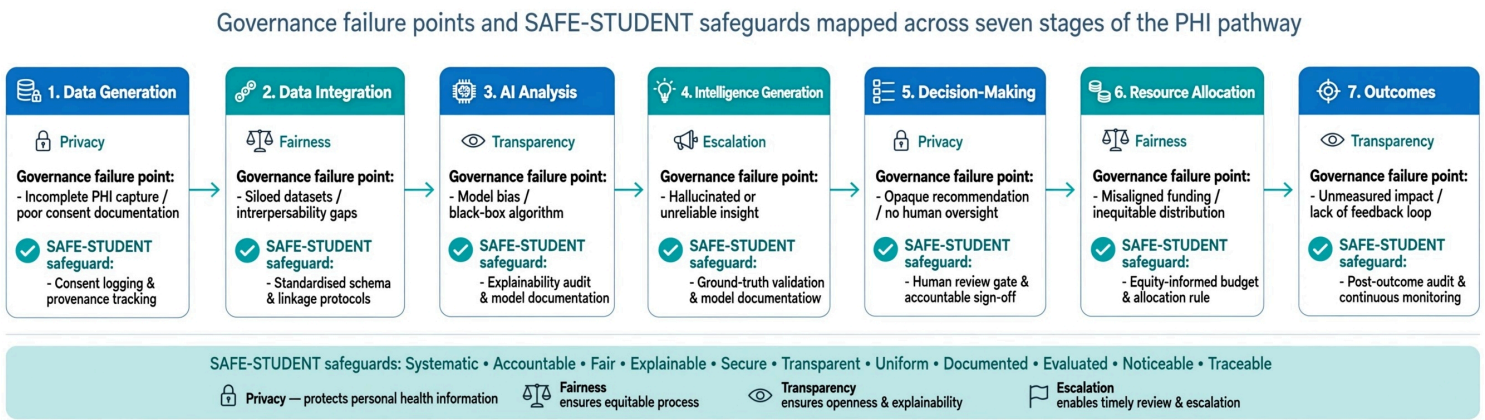


Figure 1

SAFE-STUDENT Governance Across the PHI-to-Decision Pathway

Caption: Governance failure points and SAFE-STUDENT safeguards mapped across seven stages of the PHI pathway. SAFE-STUDENT safeguards: Systematic → Accountable → Fair → Explainable → Secure → Transparent → Uniform → Documented → Evaluated → Noticeable → Traceable. PHI stages: Data Generation, Data Integration, AI Analysis, Intelligence Generation, Decision-Making, Resource Allocation, Outcomes.