FAMILY PLANNING AI CHATBOT BENCHMARKING

Final Presentation

Bih Moki-Suh, Muriel Konne, Madalitso Khwepeya, Cirilus Osongo, & Barclay Stewart

June 2nd, 2025

START CENTER
STRATEGIC ANALYSIS, RESEARCH & TRAINING CENTER
Department of Global Health | University of Washington

AGENDA

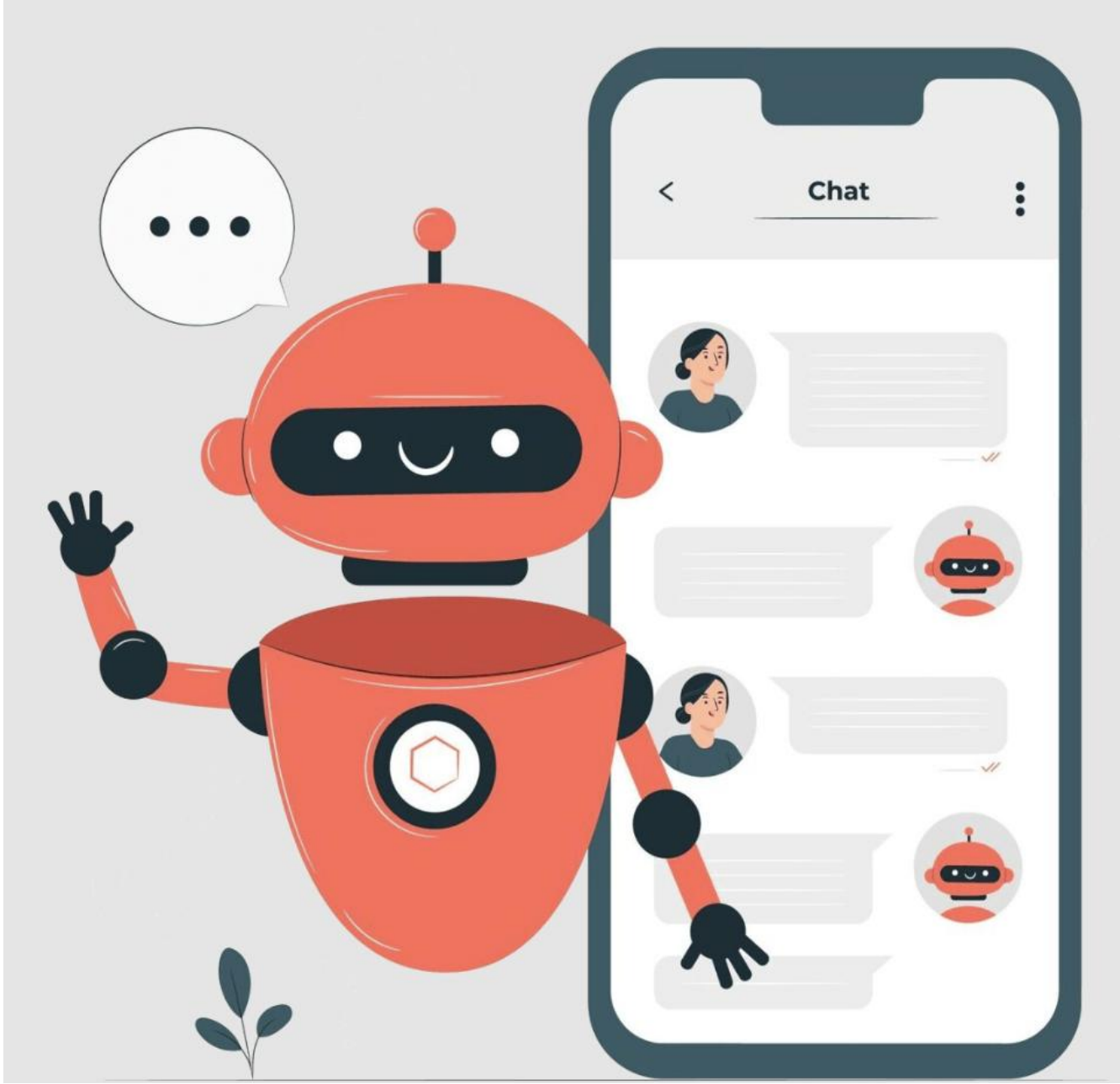
01 Introduction & Background

02 Key Project Takeaways

03 Methodology Used

04 Chatbot Evaluation & Benchmarking

05 Future Directions & Discussion



START CENTER

Photo: siamcomputing.com

PROJECT TEAM

Bih Moki-Suh, MSc

PhD Student, Implementation Science

Project Manager

Muriel Konne, MPH

PhD Student, Implementation Science

Research Assistant

Cirilus Osongo

MPH Student, Global Health

Research Assistant

Madalitso Khwepeya, RNM, MSc, PhD

MPH Student, Epidemiology

Research Assistant

Barclay Stewart, MD, PhD, MScPH

Faculty Lead

START CENTER

START OVERVIEW

Leverages leading content expertise from across the University of Washington

Provides high quality research and analytic support to the Bill & Melinda Gates Foundation and global and public health decision-makers

Provides structured mentorship and training to University of Washington graduate research assistants

BACKGROUND

MOTIVATION	Limited evidence exists on the **effectiveness, safety, and cost-efficiency** of AI-powered chatbots in improving FP outcomes for young adults
ADDITIONAL CONSIDERATIONS	In 2024, initial FP chatbot designs were tested by DIMAGI, but **further iteration and quantitative benchmarking** are needed to assess their added value and ensure safety and responsiveness for YA users.
FOCUS GEOGRAPHIES	Select young adults (18 – 24) recruited via **"Shujaaz"** and **"C'est la Vie"** multimedia youth engagement platforms in **Kenya** and **Senegal** respectively

YAs = Young Adults
FP = Family planning

PROJECT OBJECTIVES

Explore flexible chatbot evaluation methods beyond RCTs to match Gen AI advances while ensuring robust, compelling evidence for the broader community

Identify and recommend risk benchmarks for scaling FP chatbots, (*focus on data privacy, accuracy and safety*), specifying trade-offs between risks and benefits for YAs

RCTs = Randomized Control Trials
YAs = Young Adults
FP = Family planning

PROJECT DELIVERABLES

1. Final Slide-deck / Presentation summarizing feasible chatbot evaluation and benchmarking approaches

2. Excel Spreadsheets with literature findings on plausible evaluation methods and Risk benchmarking information across key metrics

3. Rapid Process Cycle Document – Capture of our decision-making process

KEY PROJECT TAKEAWAYS

START CENTER
STRATEGIC ANALYSIS,
RESEARCH & TRAINING CENTER

KEY PROJECT TAKEAWAYS 1/2

CHATBOT EFFECTIVENESS ASSESSMENT

Combine technical performance metrics with user-centered outcomes (satisfaction, trust, intent to re-use) to ensure real-world relevance.

Contextually adapt evaluation methods and benchmarks to ensure cultural relevance, equity, and practical applicability

Employ Continuous/Iterative monitoring and stakeholder feedback loops to ensure sustained chatbot impact.

Proactive identification of risk trade-offs

Ensure inclusive access for vulnerable groups

Actively track core IS outcomes (Acceptability and Feasibility)

KEY PROJECT TAKEAWAYS 2/2

CHATBOT METRICS BENCHMARKING

Data Privacy and Security

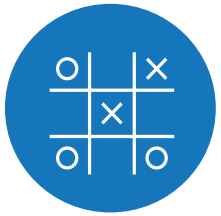
- FP chatbots must **ensure data minimization**, explicit consent, and **rigorous security measures** including encryption, anonymization, regular audits, and real-time monitoring

Safety

- SRH chatbots **must be designed to prevent harm by combining pre-deployment testing, transparent communication,** youth-friendly consent, and clear escalation to human support when risk is detected

Accuracy

- Use expert input, defined criteria, **interaction transcripts to measure and evaluate chatbot performance, response accuracy** and help combat hallucinations



Cross-cutting benchmarks for safety and accuracy

- Chatbot systems must be stress-tested for hallucinations**, transparently communicate their role and limits**, and be evaluated against expert-defined response standards & trusted sources

GENERAL PROJECT APPROACH

EVALUATION METHODS REVIEW

- Publicly available literature reviewed for chatbot effectiveness
- Data extraction from key sources

01

KEY INFORMANT INTERVIEWS

Key Informants

1. Scott Mahoney (Gates AI task force, Consultant)
2. Isabelle Amazon (Dimagi Consultant)

02

BENCHMARKING (DATA PRIVACY, CHATBOT SAFETY & ACCURACY)

Cross-validated **legally compliant criteria** with **KII insights,** and **culturally sensitive** data.

03

START CENTER

DECISION FRAMEWORK FOR EVALUATION METHODS

Methods Ruled Out for Chatbot Evaluation

Traditional RCTs

- **Too Static**
  Can't adapt to evolving LLM behaviors.
- **Ethically Problematic**
  Denies SRH access to control groups.
- **Lacks Flexibility**
  No real-time monitoring or harm mitigation.

Simple Pre/Post Designs

- **Difference-in-Difference**
  Pre- intervention parallel trends may not hold between groups.
- **Regression Discontinuity**
  No eligibility cutoff to define intervention threshold.
- **They are generally susceptibility to external Influences;** concurrent FP campaigns, media exposure, or policy shifts

Simple mixed methods ONLY

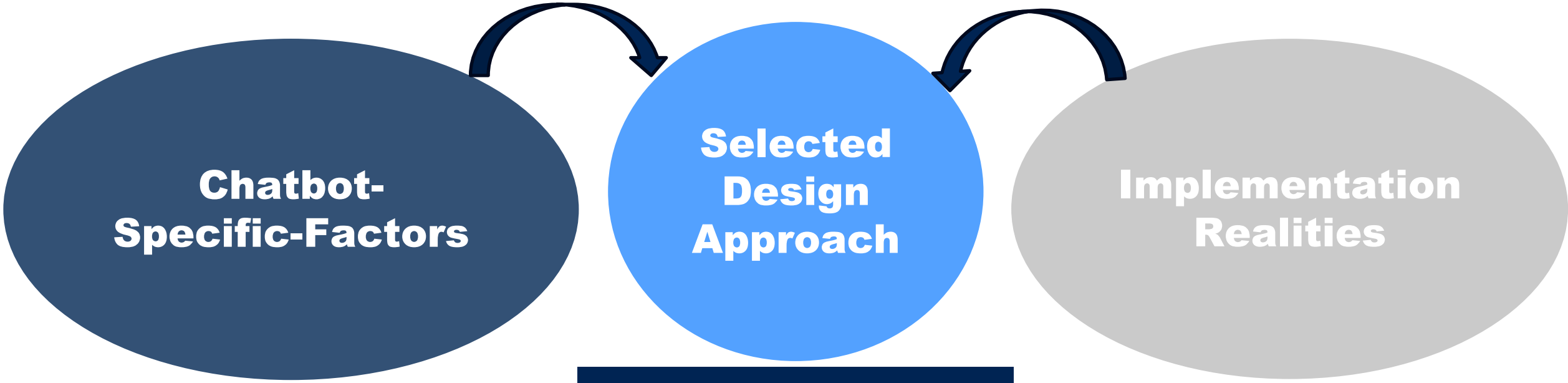
- **Convergent Design**
  Limits real-time iteration; data collected all at once.
- **Explanatory Sequential**
  Too slow for adaptive AI monitoring; qual insights come post-hoc.
- **Exploratory Sequential**
  Front-loaded; misses evolving chatbot-user dynamics during rollout.

Qualitative/ Implementation ass essment ONLY

- **No Effectiveness Evidence**
  Can't measure change in FP outcomes.
- **Descriptive Only**
  No causal inference possible.
- **Not Generalizable**
  Lacks metrics for broader scaling decisions.

DESIGNING FOR ADAPTABILITY

How We Landed on a Hybrid Evaluation for a GenAI Chatbot



- -Gen AI evolves over time
- -Requires continuous updates and finetuning
- -Not feasible/ethical to randomize
- -Needs flexible real-time

Interrupted Time Series (Pre-post trend analysis)

+

Mixed Methods (Implementation Assessment)

HYBRID TYPE I

- -Requires real world roll out timelines
- -Requires country-specific partner Input
- -Demand for actionable and timely insights

START CENTER

PHASE I

EVALUATION METHODS REVIEW (METHODS SELECTION)

START CENTER
STRATEGIC ANALYSIS,
RESEARCH & TRAINING CENTER

STUDY DESIGN

HYBRID TYPE 1 EFFECTIVENESS-IMPLEMENTATION FEASIBILITY STUDY

UNCONTROLLED INTERRUPTED TIME SERIES

+

MIXED METHODS IMPLEMENTATION ASSESSMENT

HYBRID EFFECTIVENESS IMPLEMENTATION TYPE I

DESIGN COMPONENTS

1⁰ Aim: To assess the preliminary effectiveness of a family planning chatbot in improving contraceptive self-efficacy and related behavioral outcomes among young adults in Kenya and Senegal.

Approach: Uncontrolled Interrupted Time Series (Single-group, multiple pre/post observations)

- Assess temporal changes in key outcomes such as **contraceptive self-efficacy**, family planning knowledge, and intention to use contraceptives over time (detects both level and trend shifts).

2⁰ Aim: Examine the feasibility, acceptability, and contextual factors influencing the implementation of a family planning chatbot among adolescents and young adults in Kenya and Senegal.

Approach: Mixed Methods (Convergent Parallel Design)

- **Quantitative component**: Use structured surveys and chatbot engagement metrics collected across multiple time points to measure **implementation outcomes**[1] using Likert scales and platform analytics.
- **Qualitative component:** Conduct interviews and FGDs with users and stakeholders to explore chatbot usability, trust, and contextual fit, using CFIR[2] and chatbot-specific factors (e.g GenAI trust, digital access, privacy) and chatbot-specific themes.

1. Proctor et al; 2022
2. Using CFIR; 2017

RECRUITMENT AND INTERVENTION GROUP ELIGIBILITY

Recruitment Channels

Online e-concenting via *Shujaaz* and *C'est la vie* platforms **and/or** in-person via superfan/community mobilizers

Compensation

Paid per study activity completed

Primary Intervention Group Characteristics

- **Young Adults** aged 18–24, familiar with partner platforms (*Shujazz* and *C'est La Vie*)
- Owns/uses a mobile phone with access to internet
- Consent to participate in remote surveys/interviews
- Engage with the FP chatbot during the intervention period

PROPOSED SAMPLING APPROACH

Dynamic Model: Select users based on their *stage of experience* (such as early adopters, long-term users).

Rationale: YA users' experiences evolve over time.

Example: *For a user who trusts chatbot responses today — is their trust maintained a year later?*

Cohort Progression: Successful users could move to an "Intervention 2" stage (e.g., advanced content, booster messages).

Static Model: Select a fixed cohort at one point in time, track outcomes longitudinally.

- Easier to manage, **but may miss evolving YA needs** and chatbot updates.

Proposed Approach (Dynamic)

Considerations

•Sample at multiple points to capture *trust durability*, *satisfaction shifts*, and *behavioral changes*.

• **Trust** is not a static concept (**evolves** based on both the user's changing needs and the chatbot's updates).

Key Question

Which approach best verifies real-world impact?

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

PROBABLE BASELINE, INTERMEDIATE AND DISTAL OUTCOMES

Baseline

- Engagement with chatbot
- FP knowledge and self-efficacy
- Attitudes towards FP (use, safety and accessibility)
- Intention to seek FP service or method

Intermediate

- Acceptability, usability, satisfaction, qualitative feedback
- Mental health/well-being
- Interaction quality and personalization accuracy
- Information recall support/ misinformation correction

Distal

- FP uptake
- Consistent FP use
- FP self-efficacy (Primary endpoint: Reduction in Unintended pregnancies)

Tools

- User Experience Questionnaire Short Version (UEQ-S)[1]
- Chatbot Usability Questionnaire (CUQ)[2], [2.1], [2.2]
- Post-Study Satisfaction and Usability Questionnaire (PSSUQ)[3], [3.1]
- Client Satisfaction Questionnaire (CSQ-8)[4],
- PROMIS Social Isolation Scale[5]
- PHQ-A[6]
- Social media self-efficacy[7]

Blue text = Chatbot - derived outcomes

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

SAMPLE EVALUATION SURVEYS

Outcome Indicators	Measurement type	Tools	Sample prompts
Acceptability	Perceived usefulness; intention to use; recommendation likelihood	UEQ-S (adapted for FP chatbot)	*How likely are you to recommend this chatbot to a friend seeking FP information?*
Usability	Task completion efficiency; error recovery; learnability	CUQ, PSSUQ + task-based metrics	*"I can use chatbot to find information about several contraceptive options (completion time + success rate)"*
Satisfaction	Overall experience quality; emotional response	CSQ-8 + affect measures	*"Using this chatbot made me feel..(anxious/confident/supported-semantic differential)"*
Engagement depth	Conversation quality; information seeking behavior	Chatbot analytics + conversation analysis	*"The chatbot encouraged me to ask follow-up questions"*
Digital Health Self-efficacy	Confidence using technology for health information	eHealth Literacy Scale (eHEALS) +Social media self-efficacy	*"I can tell if the health information I find online is trustworthy"*
FP Self-Efficacy	Confidence in FP decision-making; communication with providers	[1]CSESSA	*"I feel confident in my ability to choose a contraceptive method right for me"*

[1]Contraceptive Self-Efficacy among women in sub-Saharan Africa (CSESSA)

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

OPTIMIZING FP CHATBOT EVALUATION SURVEYS

Challenge	Proposed Approaches	Prospective Outcome
• Current surveys (UEQ-S, CUQ, PSSUQ, CSQ-8, PROMIS, PHQ-A) take around **20-30 minutes** • May lead to user fatigue and low response rates	• Computerized Adaptive Testing (CAT) • Ecological Momentary Assessment (EMA) • Could shorten surveys to **5-7 minutes**	• Enhanced user engagement and higher response rates • Scalable, efficient, and user-friendly

EMA, CAT

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

OUTCOME TIME MEASUREMENTS

Effectiveness Design Notation

$NR_C\ O_1\ O_2\ O_3\ O_4\ X\ O_5\ O_6\ O_7\ O_8$

NRc = Non-randomized repeated cohort
$O_1\ O_2\ O_3\ O_4$ = Pre-intervention observations (outcome measured at ≥ 4 time points before the intervention)

X = Chatbot Intervention

$O_5\ O_6\ O_7\ O_8$ = Post-intervention observations (outcome measured at ≥ 4 time points after chatbot exposure)

Best Practice

Collect ≥ 4–12 pre/post data points to detect trends and capture shifts in FP outcomes like knowledge, intent, or use

APPROPRIATE STUDY LENGTH ≥ 6 MONTHS

(Approx. 26 WEEKS)

- **Meets ITS recommended standards**
  Six months allows for **at least** 8 data points pre and 8 post intervention to detect level and slope changes.

- **Strategic Measurement**
  Designate the **"first"** 4-months as chatbot engagement assessment period with weekly measures **(14 data points) - post intervention** , **plus** a "7-month" follow-up to capture durability of chatbot's effects

Rationale

- **Captures Outcome Dynamics:** Detects both rapid (attitudes, knowledge) and gradual (self-efficacy) changes.

- **Minimizes Study Burden:** Optimizes rigor and retention without requiring a long-term follow-up.

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

TIME MEASUREMENTS – BASELINE, INTERMEDIATE & DISTAL OUTCOMES

Time Point	Timing	Measures	Assessment Focus / Rationale
T0 (Baseline)	**Month 0** Week -1 Week -2	FP Knowledge, CSESSA, PCA (ITU–FP & Chatbot)	Establish baseline for self-efficacy, assess beliefs and perceptions
T1	Week - 3 Week - 4	PCA (ITU–FP & Chatbot)	Track expectations and FP intention evolution (Early trend check)
T2	Week - 5 Week - 6	PCA (ITU–FP & Chatbot)	Validate stability or variation in intent to use chatbot/FP
T3	Week - 7 Week - 8	CSESSA, FP Knowledge, ITU–FP & Chatbot)	Final pre-trend + baseline for key outcomes
T4	**Chatbot launch (Week 1)** **Month 3 starts**	—	**Launch point for ITS** **(Chatbot rollout)**

CSESSA = Contraceptive Self-Efficacy Scale for Sub-Saharan Africa
FP = Family Planning, PCA = Perceived Chatbot Attitudes/Acceptability
ITU = Intent to use

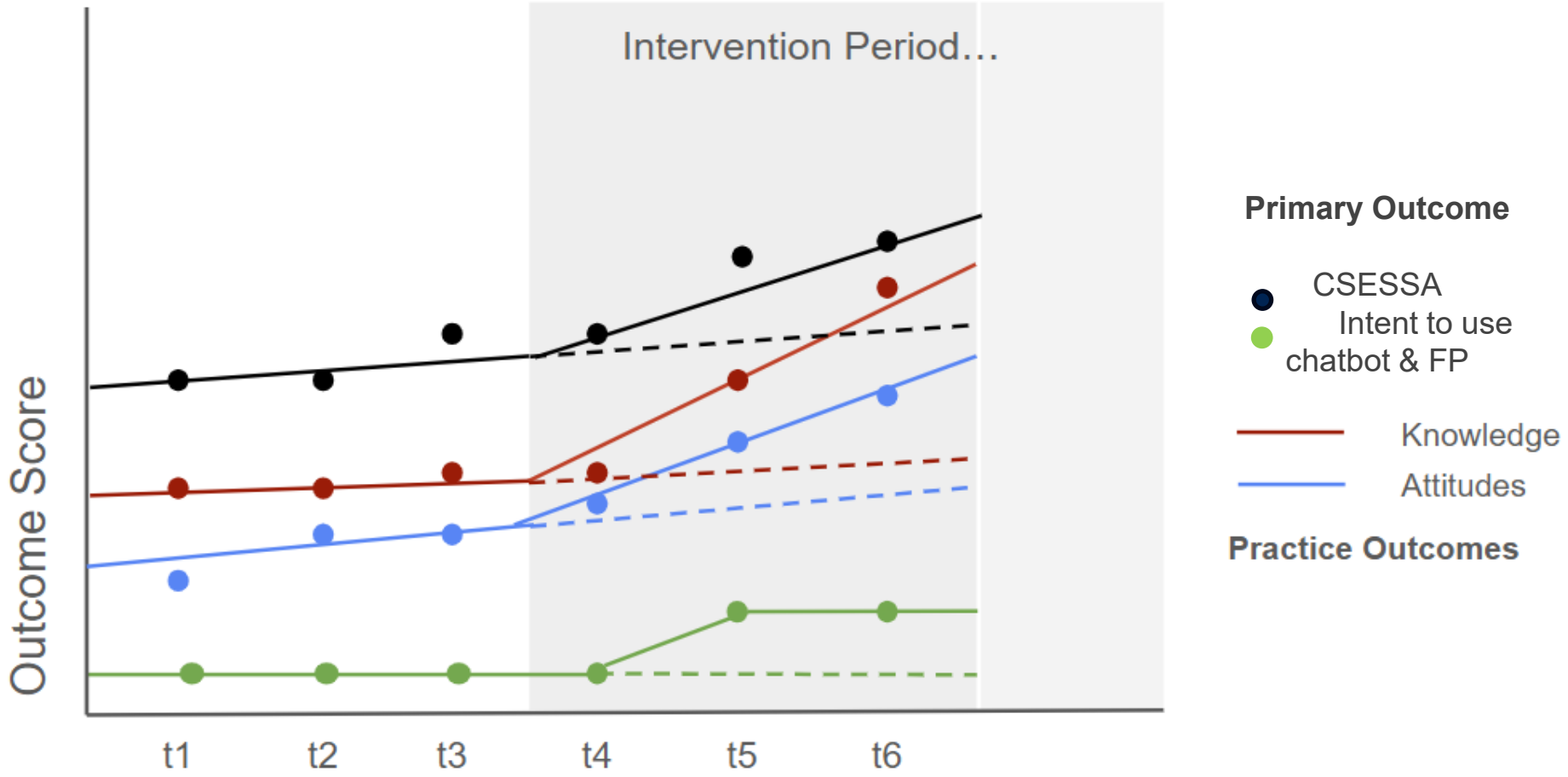
INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

TIME MEASUREMENTS – BASELINE, INTERMEDIATE & DISTAL OUTCOMES (POST CHATBOT DEPLOYMENT)

Time Point	Timing	Measures	Rationale
T5	Week 1 & 2	Chatbot Attitudes, ITU–FP & Chatbot, CSESSA	Detect rapid response to chatbot content (Immediate level change)
T6	Week 3 & 4	Chatbot Attitudes, ITU–FP & Chatbot	Assess short-term perception and intent (chatbot effect on decision-making)
T7	Week 5 & 6	FP Knowledge, Chatbot Attitudes, ITU–FP & Chatbot	Evaluate deeper change in knowledge and efficacy
T8	Week 7 & 8	Chatbot Attitudes, ITU–FP & Chatbot	Assess continued intent to use chatbot and FP
T9	Week 9 & 10 *(5 months in)*	Chatbot Attitudes, ITU–FP & Chatbot	Assess sustained engagement (Frequency + Density + Satisfaction)
T10	Week 11 & 12	Chatbot Attitudes, CSESSA	Start assessing behavior change by verifying trust/confidence in chatbot responses
T11	**Week 13 &14 (Endline Month 6)**	**FP Knowledge, CSESSA, Chatbot Attitudes, ITU–FP, FP Use**	**Distal/Final Outcome assessment**
T12	Week 16 (Follow –up, at month 7)	FP Use, CSESSA, Chatbot Attitudes (if recalled)	Capture the durability of intervention effects (Long-term outcome assessment)

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

HYPOTHETICAL OUTCOME TRENDS



APPLICABLE PROCTOR IMPLEMENTATION OUTCOMES

Implementation Outcomes[1]	Definitions and Level of Measurement	Assessment Approaches
Acceptability	Perception among users that the chatbot is satisfactory, appropriate, and engaging *(Client-level)*	Surveys, interviews, and chatbot analytics
Feasibility	The extent to which the chatbot can be successfully used within the digital/public health system *(Client & Stakeholder-level)*	User engagement rates, technical performance
Cost	Economic impact of developing, deploying, and maintaining the chatbot *(System-level))*	Development, implementation, and per-user cost estimates
Appropriateness	Perceived fit or relevance of the chatbot to user needs and system context *(Client & Stakeholder-level)*	Surveys, interviews, and stakeholder feedback asse ss cultural fit, modality preference, and alignment

Sample CFIR Adapted Questions

Acceptability: *To what extent do you believe the chatbot provides trustworthy and useful information about [family planning/HIV prevention], and how comfortable do you feel using it to seek this information*

Feasibility: How likely are you to use the chatbot as part of your routine public health activities?)

Cost: What cost will be incurred to implement the FP chatbot?)

Appropriateness: To what extent do you feel the chatbot responded to your specific FP needs

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

LESS APPLICABLE IMPLEMENTATION OUTCOMES

01 **Fidelity** assumes a prescriptive protocol; **AI – powered chatbots often offer dynamic, personalized responses**, making strict fidelity less meaningful unless specific behavioral scripts are assessed.

02 **Penetration** reflects deep integration into a system; since this is an early-phase evaluation, scaling and institutional embedding may **not yet** be observable *(Penetration is not a focus in early-stage implementation*).

03 **Sustainability** requires extended follow-up (e.g., 12+ months post-deployment) to assess continued use, funding continuity, or institutional ownership. **Not feasible within a short hybrid study**

04 **Adoption** requires initial decision and action to use the chatbot, primarily by implementing partners (at Stakeholder-level). Can be assessed using metrics such as reach, first-time use rate, and partner uptake

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

STUDY POWER CONSIDERATIONS

Effectiveness Design Notation

$NR_C\ O_1\ O_2\ O_3\ O_4\ X\ O_4\ O_5\ O_6\ O_7$

NRc = Non-randomized repeated cohort
$O_1\ O_2\ O_3\ O_4$ = Pre-intervention observations (outcome measured at ≥ 4 time points before the intervention)

X = Chatbot Intervention

$O_5\ O_6\ O_7\ O_8$ = Post-intervention observations (outcome measured at ≥ 4 time points after chatbot exposure)

Best Practice
Collect ≥ 4–12 pre/post data points to detect trends and capture shifts in FP outcomes like knowledge, intent, or use

Potential Improvement Areas

- Expand to at **least 6–12 months total** (e.g., 6 pre, 6 post) to improve effect detection and model seasonal trends
- Hybrid Type I designs **prioritize both effectiveness signals and implementation insights**, especially in real-world, evolving interventions like chatbots
- **The goal is not to produce definitive causal estimates**, but to understand:
  - *Is the intervention acceptable and feasible?*
  - *Are trends moving in the right direction?*
  - *How should it be adapted or scaled?*

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

STUDY POWER CONSIDERATIONS

Parameter	Recommendations
Total time points	≥ 12 (ideally 6 pre + 6 post)
Minimum pre/post points	≥ 4 per phase (absolute minimum)
Interval spacing	Monthly or bi-weekly
Effect size	0.5–1.0 SD (moderate)
Autocorrelation (ρ)	0.2–0.3 (typical in health behavior studies)
Sample size per time point	≥ 100–200 users
Outcome type	Continuous (e.g., FP knowledge score)
Power target	≥ 80%
Alpha (Type I error)	0.05

Key questions to think about:

- Is the outcome a single variable or a composite measure?
- Is it measured using a scale or confidence rating validated in similar settings?
- Are we powering FP knowledge, Chatbot attitude, intent, or some other outcome?
- Are we designing the study to detect a meaningful, practical change, not just statistical significance?
- How much change is actionable? Is a 10% increase enough? Is 20% more realistic?

"If your goal is feasibility and early implementation, don't over-focus on power or P-values. Instead, look at absolute changes over time and use those to inform future scale-up studies. Without baseline usage or dropout data, power calculations would be speculative." – *Brad Wagenaar, UW Faculty*

THREATS TO INTERNAL VALIDITY

01 **No Control Group:** Requiring strong assumptions that no other major changes (policies, campaigns) influenced outcomes during the study period.

02 **Time-Varying Confounding:** Participant behavior may be affected by **external influences** such as National family planning campaigns, School schedules or health facility stockouts

03 **Incomplete Seasonality Capture:** The proposed study period may miss longer-term seasonal trends (e.g., holidays, school breaks), which could confound observed effects.

04 **Concurrent Exposure to Other Platforms:** Participants may also be engaging with GPT-based tools (e.g., ChatGPT, Google Bard)

05 **Shujaaz'** and ***C'est la vie*** digital media or WhatsApp groups may influence knowledge, attitudes, or behaviors independent of the chatbot.

STRATEGIES TO MITIGATE BIAS

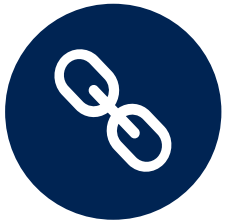
01 **Continuously track and document exposures** by including survey items on use of other digital tools and platforms during the study.

02 **Decompose the time series** into trend, seasonal, and residual components to visualize repeating patterns and isolate seasonal effects before the intervention

03 **Adjust for seasonal patterns** (e.g., school terms, holidays, stockouts) by Incorporating calendar-based indicators (e.g., capturing low engagement in December) in the ITS model.

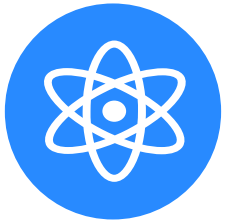
04 **Stratify or conduct sensitivity analysis**
Compare outcomes among subgroups with vs. without exposure to other FP platforms.

KEY RECOMMENDATIONS FOR STUDY DESIGN AND IMPLEMENTATION



Implement Weekly Data Collection with Biweekly Outcome Assessment

Collect engagement and interaction data weekly to capture real-time usage patterns, but aggregate and analyze key outcomes (e.g., trust, self-efficacy) on a biweekly basis.



Leverage Mixed Data Collection Modalities

Combine **in-chat surveys, behavioral prompts, and qualitative interviews** to capture both behavioral trends and nuanced user experiences, to enhance data richness while minimizing respondent fatigue.

Emphasize Feasibility Over Statistical Power

Focus on **gathering sufficient data to explore implementation challenges**, engagement patterns, and early signals of impact—without requiring formal power calculations

Monitor Dropout and Trust as Proxies for Chatbot Engagement

Addressing dropout and trust issues is critical to both the design and interpretation of this study. **High dropout rates or low user trust** may reflect broader challenges with acceptability, usability, or perceived value of the chatbot.

PHASE II

KEY INFORMANT INTERVIEWS

START CENTER
STRATEGIC ANALYSIS,
RESEARCH & TRAINING CENTER

KEY INFORMANTS

DIMAGI AND GATES AI TASK FORCE CONSULTANTS

Isabelle Amazon-Brown, MA

Inclusive, ethical service design and capacity building for chatbots & AI

- Dimagi Consultant (Norwich, United Kingdom)

Dr. Scott Mahoney, MBChB, PGDip

AI healthcare Innovations for real impact in LMICs

- Gates AI task force consultant, South Africa

EXPERT INSIGHTS

Study Design Considerations	Static RCTs are ill-suited for LLM tools—adaptive designs and expert reviews, with **ongoing monitoring** and version tracking, are essential.
Comparator Group Selection Considerations	Control selection must balance ethical risk and relevance. Direct LLM responses outperform curated links, highlighting the need for meaningful, user-centered comparators.
Chatbot Benchmarking Considerations	Benchmarking **must go beyond** model-centric metrics—integrating user-centered outcomes, using real-world, context-specific benchmarks, and adapt with LLM updates.

START CENTER

Key Finding: There was alignment between our literature review findings and KII insights across domains

PHASE III

ESTABLISHING BENCHMARKS FOR CHATBOT EVALUATION

START CENTER
STRATEGIC ANALYSIS,
RESEARCH & TRAINING CENTER

DATA PRIVACY AND SECURITY BENCHMARKING

START CENTER
STRATEGIC ANALYSIS,
RESEARCH & TRAINING CENTER

DATA PRIVACY AND SECURITY BENCHMARKING

BENCHMARKING APPROACH 1/3

EU Data Protection Regulations (GDPR): Sets high standards for data protection with principles like data minimization, consent, and impact assessments, ensuring global compliance for data controllers.

UNESCO AI Ethics Framework: Focuses on protecting human rights, privacy, and ethical use of AI, advocating for impact assessments and transparency in AI systems. *(Not included in overlap analysis – too high level)*

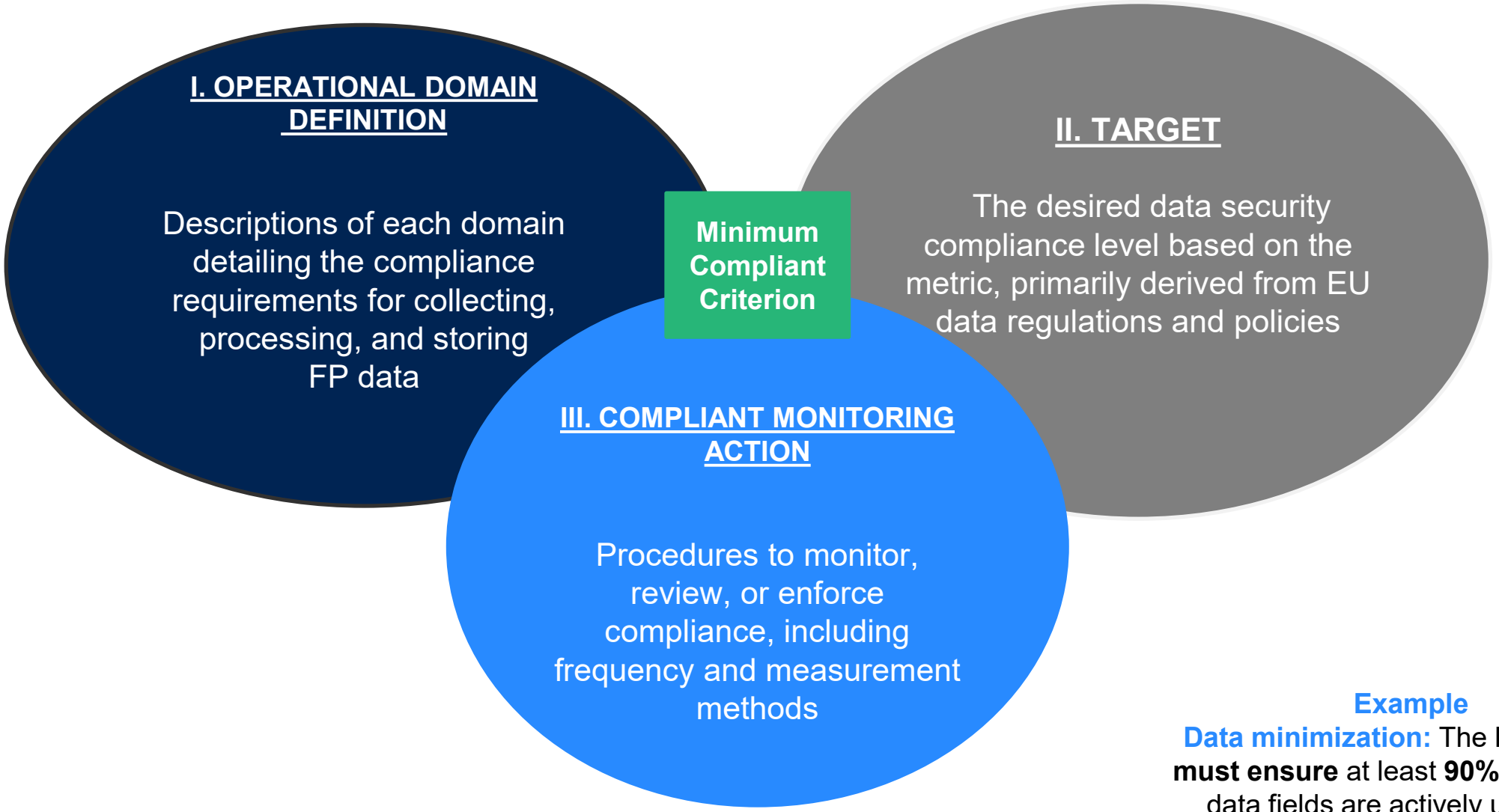
WHO AI Regulation Principles: Promotes ethical AI use in health, emphasizing autonomy, safety, transparency, and inclusivity, with clear data protection laws for health data.

Kenya Data Protection Act (2019): Outlines data rights, consent conditions, and transfer restrictions, mandating impact assessments for high-risk processing.

Senegal Data Protection Law (LDCP): Establishes an independent authority for data privacy, prohibits sensitive data collection, and mandates data anonymization for third-party use.

DATA PRIVACY AND SECURITY BENCHMARKING

BENCHMARKING APPROACH 2/3



NB: OVERLAP ANALYSIS

START CENTER

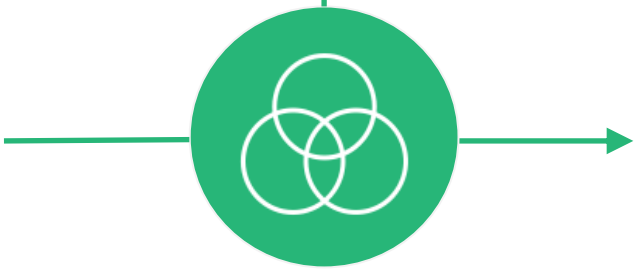
Example

Data minimization: The **FP chatbot must ensure** at least **90%** of collected data fields are actively used, with **regular reviews** to remove unnecessary data.

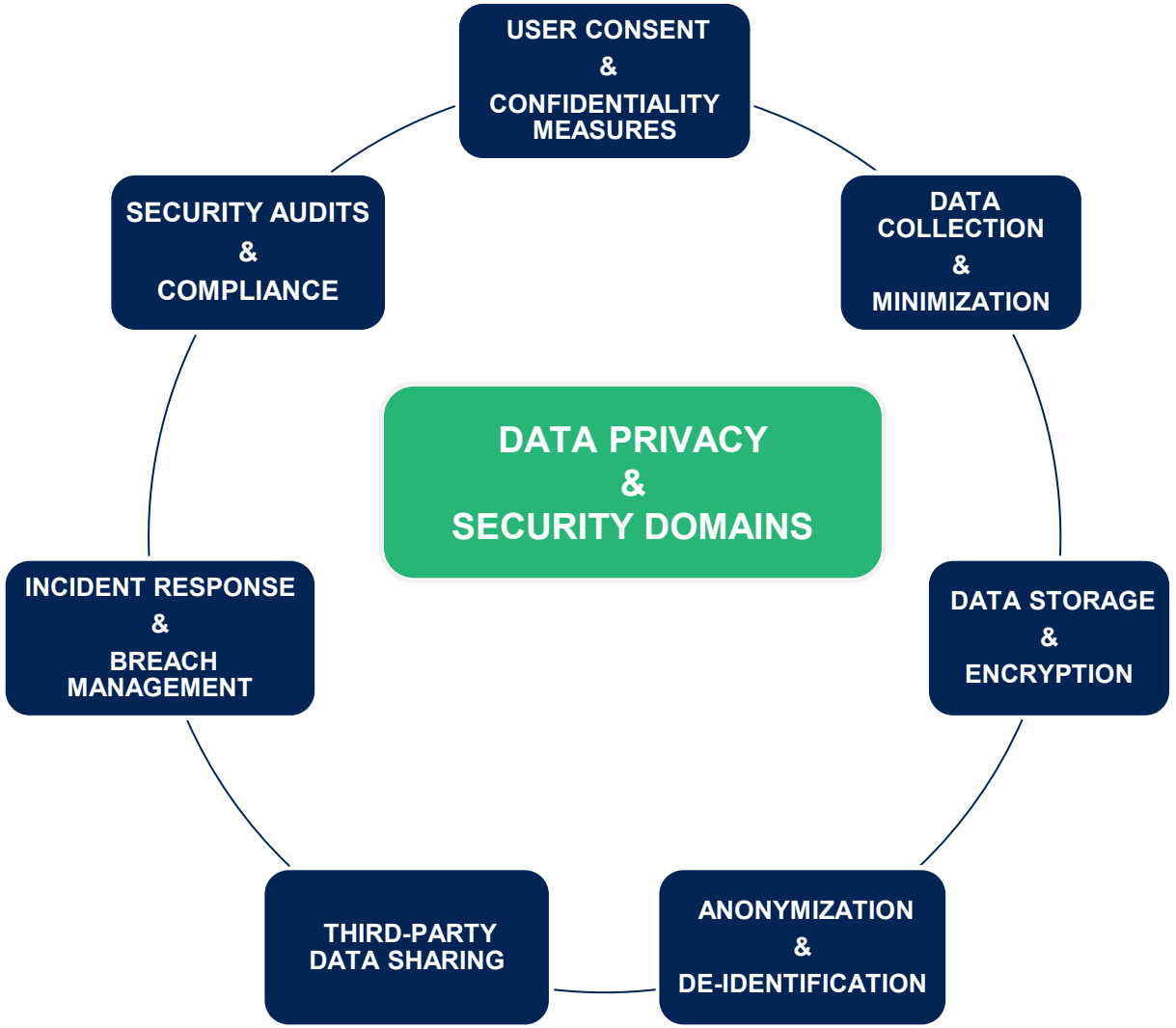
DATA PRIVACY AND SECURITY BENCHMARKING

BENCHMARKING APPROACH 3/3

DOMAINS FRAMEWORK



Adapted from WhatsApp's Data Privacy Policy



Resource spotlight: WhatsApp Privacy policy

DATA PRIVACY AND SECURITY BENCHMARKING

KEY DOMAINS 1/2

Data Minimization

- Data is only used for the explicit purpose it was collected
- Data cannot be processed in ways incompatible with those purposes

Data Storage & Encryption

- Must ensure 100% encryption and pseudonymization
- Backup success, verified by at least weekly tests
- Data retention compliance

Consent & Confidentiality

- Protecting Autonomy
- **Risks to Safety and Cybersecurity**
- Governance of Data

Anonymization & De-identification

- Regular anonymization and de-identification audits
- Non-compliance triggers immediate action

DATA PRIVACY AND SECURITY BENCHMARKING

KEY DOMAINS 2/2

Third-party Data Sharing

- Is permissible but with clear disclosure to the data subjects
- Appropriate measures should be taken while sharing data with third-party

Regular Security Audits & Compliance

- Setting up mechanism of regular security audits to ensure compliance
- This may include checking if data processes adhere to the law, investigating complaints, and sanctioning non-compliant activities

Breach Management

- Data breach may result in physical, material and non-material damage
- **Timely notification and notification of a breach**

SAFETY AND ACCURACY BENCHMARKING

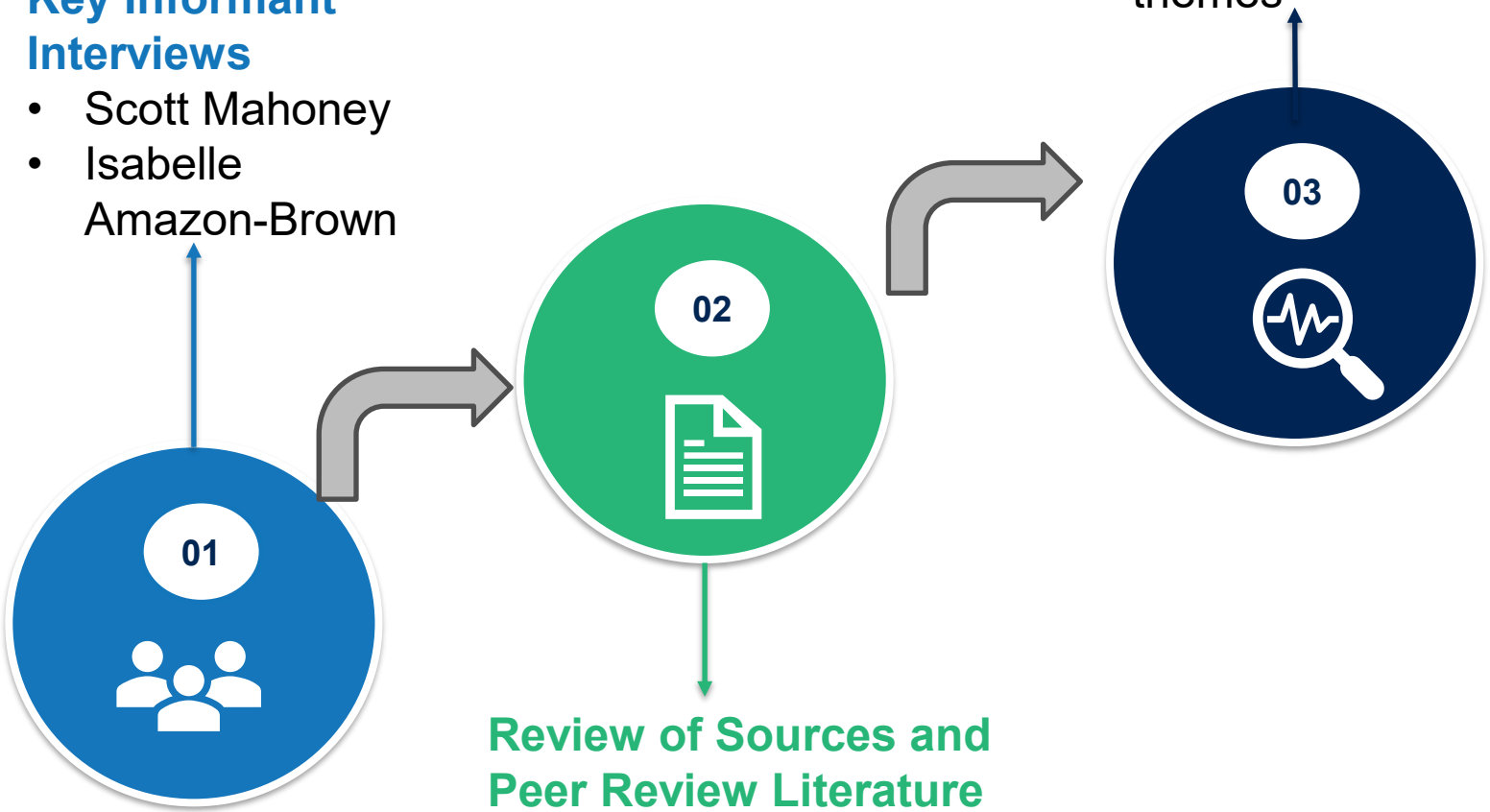
START CENTER
STRATEGIC ANALYSIS,
RESEARCH & TRAINING CENTER

SAFETY & ACCURACY BENCHMARKING APPROACH

Key Informant Interviews
- Scott Mahoney
- Isabelle Amazon-Brown

Review of Sources and Peer Review Literature
- Gen AI for Health in LMICs (Stanford University)

Thematic Synthesis
- Capture emerging themes



Key considerations

- Benchmarks for safety are not explicitly stated in the literature, unlike for accuracy
- They vary widely depending on the geographical context, chatbot domains, end users & health area
- Accuracy domains are more straight forward compared to safety domains

START CENTER

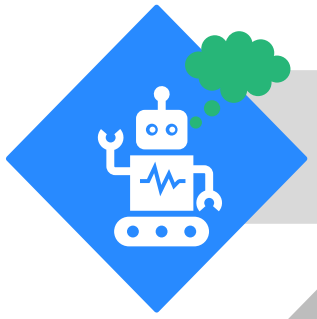
II. CHATBOT SAFETY BENCHMARKING

START CENTER
STRATEGIC ANALYSIS,
RESEARCH & TRAINING CENTER

CHATBOT SAFETY BENCHMARKING

KEY TAKEAWAYS

Chatbots must be rigorously tested to prevent confidently wrong or unsafe outputs, especially in response to sensitive SRH and FP questions



User autonomy must be protected through clear identity disclosure, youth-appropriate consent, and transparent data use

Chatbots must be designed with built-in mechanisms to detect risk and escalate users to human support when harm or distress is disclosed

All communication should be adapted to adolescents' emotional and literacy levels to reduce confusion and increase trust

CHATBOT SAFETY BENCHMARKING

RECOMMENDATIONS BY DOMAIN 1/3

01

Medical harm prevention

- Redirecting to human professionals is particularly crucial for SRH concerns where misguidance can lead to severe physical or psychological harm (e.g., contraceptive side effects or STI symptoms)[1]

02

Hallucination monitoring (Content audit)

- Hallucinations in SRH contexts, such as misinformation about fertility, pregnancy, or emergency contraception, can lead to serious consequences and require targeted stress-testing[3,4]

03

Content moderation

- FP and SRH languages are often stigmatized. Moderation tools must avoid over-censoring local slang or common expressions for contraception and sex, which may suppress valid and lifesaving content[2,4]

START CENTER

[4] Safer Chatbot Implementation guide, [1] Designing child-safe AI, [2] Girl Effect Responsible Gen AI Guidelines, [3] Chatbot for Family Planning Counseling

CHATBOT SAFETY BENCHMARKING

RECOMMENDATIONS BY DOMAIN 2/3

04

Bias and equity monitoring (algorithmic fairness)

- SRH chatbots must reflect gender and cultural diversity, avoid heteronormative bias and use inclusive language when discussing contraceptive options. This requires localizing AI tools to regional norms and marginalized users' realities[23]

05

User protection and consent

- In SRH conversations, consent must be youth-friendly, clearly explain data use, and protect confidentiality, especially when discussing sensitive issues like abortion, contraception, or sexual activity

06

Transparency of sources: trusted communication

- The FP chatbot must provide explainable, diversity-sensitive, and age-appropriate SRH information, citing trusted sources like national guidelines to build credibility among youth who may avoid formal care due to stigma

CHATBOT SAFETY BENCHMARKING

RECOMMENDATIONS BY DOMAIN 3/3

07

Escalation protocols: risk triage

- Disclosures related to sexual violence, coercion, or unsafe abortion must prompt immediate, sensitive escalation and connect users to youth-appropriate care or emergency services

08

User comprehension assurance: health literacy alignment

- FP messages must be emotionally supportive and adapted for low-literacy users. Clinical jargon or judgmental tones may discourage AYAs from engaging with the chatbot

09

Scope of practice adherence: role limitations

- SRH bots must not make clinical decisions, such as diagnosing pregnancy or advising on STI treatment, and must refer users to human professionals when needed

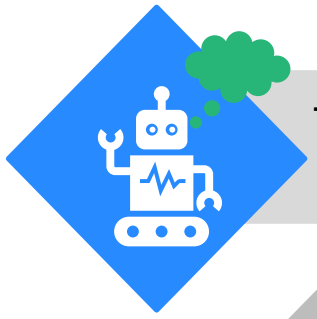
III. CHATBOT ACCURACY BENCHMARKING

START CENTER
STRATEGIC ANALYSIS,
RESEARCH & TRAINING CENTER

CHATBOT ACCURACY BENCHMARKING

KEY TAKEAWAYS

Response accuracy of family planning recommendations and guidelines can be evaluated used a set of “Golden Answers” provided by human experts



To help combat chatbot hallucination, measure its performance against a set of defined criteria and task aimed at at retrieval of accurate FP data

Clinical experts can serve as reviewers of chatbot interaction transcripts to assess its accuracy. An evaluator bot can also be employed to test accuracy

Use metrics to measure the chatbot’s alignment with trusted databases, sources, and existing protocols on FP

CHATBOT ACCURACY BENCHMARKING

RECOMMENDATIONS BY DOMAIN 1/3

01

Response Accuracy

- End-to-end (E2E) benchmark uses a set of “Golden Answers” to accurately measure chatbot performance and response by comparing and checking chatbot answers to ‘golden answers’ provided by a human experts on family planning guidelines and recommendations[1]. Ensure chatbot responses are medically accurate, up-to-date and aligned with national SRG guidelines.

- Other metrics like BLEU and ROUGE can also be used to assess quality of chatbot response.[1]

02

Hallucination rate

- AI models produce results that are not real, do not follow any data the algorithm has been trained on, or does not follow any other discernable pattern[1].

- Define a set of tasks or criteria that the chatbot must fulfil and then measure its performance against those tasks or criteria aimed at at retrieval of accurate FP data [1].

CHATBOT ACCURACY BENCHMARKING

RECOMMENDATIONS BY DOMAIN 2/3

03

Clinical Panel Validation

- Review interaction transcripts from the Chatbot and employ clinical experts to review the transcripts for accuracy by comparing the 'meaning' of each answer as opposed to comparing the exact words.[1,2]
- Tailor accuracy standards to clinical sensitivity of FP and SRH content
- Ensure chatbot responses are medically accurate, up-to-date, and aligned with national SRH guidelines
- In practice, some researchers have used an evaluator bot to compare its review with human expert review and found results to be reliable and accurate.[2]

04

Trusted Source Attribution (Rate)

- Measures how often chatbot responses cite or align with trusted databases or sources.
- Use metrics that can measure the attribution of the text generated by the chatbot

CHATBOT ACCURACY BENCHMARKING

RECOMMENDATIONS BY DOMAIN 3/3

05

Protocol Alignment

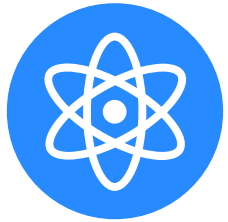
- Use of AI chatbot should not undermine the principle of protecting human autonomy

- Requires the protection of privacy and confidentiality and obtaining valid informed consent

FUTURE DIRECTIONS FOR SCALING FAMILY PLANNING CHATBOTS

Integrate Human Support for Complex Cases

Future FP chatbot models should include **escalation protocols that link users to human counselors**, such as nurses or youth champions, for cases involving contraceptive side effects, method switching, fertility concerns, or partner negotiation



Expand Accessibility through contextual & technological adaptation

Scaling requires localizing content and delivery formats to reach underserved users, & **deploying chatbots across multi-platforms** (beyond Whatsapp) to engage adolescents effectively.

Institutionalize Chatbots into National Health Systems

To ensure long-term impact, **FP chatbots should be integrated into national health systems by aligning with MOH priorities,** incorporating into referral/reporting systems. Sustainable financing (**donor-government co-financing**, telecom partnerships etc.

QUESTIONS & DISCUSSION

START CENTER
STRATEGIC ANALYSIS,
RESEARCH & TRAINING CENTER

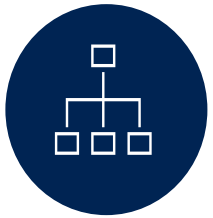
MERCI

START CENTER
STRATEGIC ANALYSIS,
RESEARCH & TRAINING CENTER

APPENDIX

START CENTER
STRATEGIC ANALYSIS,
RESEARCH & TRAINING CENTER

RECRUITMENT STRATEGY (E- CONSENTING)



Interactive, Tiered Consent Interface

Use “descriptive text” and branching logic to create layered, expandable sections.
Include *Yes/No* checks before proceeding to next sections.

Language & Accessibility Support

REDCap supports multi-language projects (*with appropriate IRB configuration*).
Can **embed audio, images, or icons** to aid low-literacy users.

Digital Self-Consent (18–24)

Include a **checkbox field + e-signature + date/time stamp**, fulfilling self-consent needs, restrict access to the survey until consent is completed.

Embed Consent Flow with Audit Trail

Consent can be the first page of a survey or part of the chatbot sign-up workflow (REDCap automatically logs consent metadata (IP address, timestamp, version).

KEY PROJECT TAKEAWAYS

CHATBOT EFFECTIVENESS METHODS REVIEW

01

Holistic Evaluation Approach

Effective assessments must combine technical performance metrics (e.g., safety, accuracy) with user-centered outcomes (satisfaction, trust, intent to reuse) to ensure real-world relevance.

02

Contextual Adaptation is Crucial

Evaluation frameworks and benchmarks should be adapted to local contexts and user needs to ensure cultural relevance, equity, and practical applicability.

03

Proactive Identification of Trade-offs

Balancing safety, accessibility, and engagement often involves trade-offs; evaluation must surface these tensions early to guide strategic refinement.

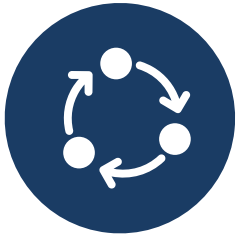
04

Iterative Learning and Feedback Loops

Continuous monitoring and stakeholder feedback are essential for adapting digital health interventions over time and ensuring sustained impact.

KII SUMMARY HIGHLIGHT

KEY INFORMANT: SCOTT MAHONEY (GATES AI TASK FORCE)



Study Design Best Practice

- Static evaluations (e.g., one-off RCTs) are insufficient for LLM-based tools; ongoing monitoring and flexibility for real-time adjustments are critical.

Control Groups Consideration

- Ethical comparator selection must balance risk and benefit—RCTs may be justified for high-risk chatbots but observational designs are acceptable for low-risk educational tools.

Adapting to LLM Evolution

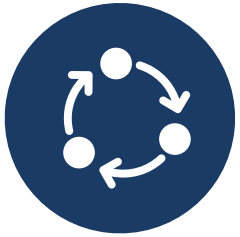
- AI models evolve rapidly; evaluations must include version tracking and periodic revalidation to stay relevant.

Benchmarking Consideration

- Current benchmarks (like MedMCQ) are inadequate; real-world, open-ended FP interaction benchmarks are needed, tailored to local context and languages.

KII SUMMARY HIGHLIGHT

KEY INFORMANT: ISABELLE AMAZON (DIMAGI CONSULTANT)



Study Design Best Practice

- Early evaluations favor A/B - split testing, human expert reviews, and rapid feedback loops rather than traditional long-term designs.

Control Groups Consideration

- Comparison between direct LLM-generated answers vs. curated article links shows **better engagement and satisfaction with direct answers**.

Adapting to LLM Evolution

- Constant reliance on tech teams and flexible evaluation frameworks are essential to keep pace with fast-changing LLM capabilities.

Benchmarking Consideration

- Key metrics **must** include both LLM auto-evaluations (safety, accuracy) and user-centered outcomes (satisfaction, likelihood to recommend).

KEY BENCHMARKING SOURCES (SAFETY AND ACCURACY)

World Health Organization — WHO guidelines on Ethics and Governance of AI for Health	unicef — Safer Chatbots Implementation Guide	NIST — NIST AI Framework	European Commission — Ethics guidelines for trustworthy AI (EC)
ICT Institute — Applying Ethical AI Frameworks in practice: Evaluating AI conversational chatbot solutions	CDH Center for Digital Health — Gen AI for Health in Low- and Middle-Income Countries	UNFPA — Adolescent and Youth-Friendly Services Toolkit	Chatbot for Family Planning Counseling
Benchmarking LLM Powered Chatbots: Methods and Metrics	The Principles for Digital Development: Widely adopted in ICT4D and global health, covers privacy, user design, scalability, and sustainability	OECD Framework for Classifying AI Systems (2022): Classifies AI systems by autonomy, interaction, and context—helpful for chatbot risk mapping	UNFPA Digital Health Platform Case Studies

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

STUDY GOALS

CHATBOT EFFECTIVENESS

Design Type: **Single-group design (ITS) without Control Group**

Objectives

- Tracks changes in FP outcomes pre/post **chatbot exposure** among users only
- Measures **within-person change over time**.
- Assesses both **level change** (immediate effect after intervention) and **trend changes post intervention**
- Controls for **time-invariant confounders** because the same individuals are observed throughout.

IMPLEMENTATION OUTCOMES ASSESSMENT

Design Type: **Mixed-methods assessment of implementation outcomes**

Implementation Outcomes: Acceptability, Usability, Feasibility, Fidelity, Appropriateness, Safety, Trustworthiness, Sustained Use, etc.

Approach: Use both quantitative surveys and Indepth interviews or Focus group discussions

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

STUDY GOALS

CHATBOT EFFECTIVENESS

PROS –SINGLE GROUP ITS

- **Equity and ethical considerations** - Ensuring all AYAs engage with the chatbot promoting fairness and reproductive autonomy
- **Better reflects real-world rollout**—*non-exposure is unrealistic due to organic sharing and access beyond study control.*
- **Aligns with LLM Improvements overtime** - *Static control groups can't account for evolving user experience over time.*
- **Less resource Intensive** - Avoids separate control group reduces costs, simplifies recruitment and tracking, and lowers attrition risk

Effectiveness Design Notation

$NR_C\ O_1\ O_2\ O_3\ X\ O_5\ O_6\ O_7$

NRc = Non-randomized Repeated Cohort
$O_1\ O_2\ O_3$ = Pre-intervention observations (outcome measured at 3 time points before the intervention)

X = Chatbot Intervention *(Month 4)*

$O_5\ O_6\ O_7$ = Post-intervention observations (outcome measured at 3 time points after chatbot exposure)

Key consideration/Best practice

Collect ≥ 4–12 pre/post data points to detect trends and capture seasonal shifts in FP outcomes like knowledge, intent, or use

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

STUDY GOALS

CHATBOT EFFECTIVENESS

CONS – SINGLE GROUP ITS

- **No control for external influences:** National FP campaigns, school re-openings, or social media trends influence FP outcomes
- **Limited causal inference**: Changes observed post-intervention could be part of a pre-existing trend, not necessarily caused by the chatbot
- **Susceptible to seasonal bias**: Without enough data points, it's hard to differentiate seasonal variation from chatbot effects.

Effectiveness Design Notation

$NR_C\ O_1\ O_2\ O_3\ X\ O_5\ O_6\ O_7$

NRc = Non-randomized Repeated Cohort
$O_1\ O_2\ O_3$ = Pre-intervention observations (outcome measured at 3 time points before the intervention)

X = Chatbot Intervention *(Month 4)*

$O_5\ O_6\ O_7$ = Post-intervention observations (outcome measured at 3 time points after chatbot exposure)

Key consideration/Best practice

Collect ≥ 4–12 pre/post data points to detect trends and capture seasonal shifts in FP outcomes like knowledge, intent, or use

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

STUDY GOALS

IMPLEMENTATION OUTCOMES ASSESSMENT

PROS – MIXED METHODS

- **Real-Time Learning**: Enables continuous feedback to improve user experience.
- **Supports Iteration**: Allows testing of new engagement strategies mid-study.
- **Explains Outcomes**: Links user experience to behavior change trends.
- **Fits Adaptive Tools**: Well-suited for evolving digital interventions like chatbots.

QUANTITIVE SURVEYS

Measure changes in key effectiveness and implementation outcomes over time (e.g., knowledge, self-efficacy, acceptability, intent to use).

QUALITATIVE INTERVIEWS

Explore ***why*** **outcomes are changing (or not),** and uncover deeper insights into user experiences, barriers, and contextual influences.

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

STUDY GOALS

IMPLEMENTATION OUTCOMES ASSESSMENT

CONS – MIXED METHODS

- **Blurs Attribution**: Iterative changes make it harder to link outcomes to a consistent version of the intervention.
- **Resource Intensive**: Requires ongoing data collection, monitoring, and coordination.
- **Risk of Over-Adaptation**: Too many mid-course changes can destabilize the intervention.
- **Timing Misalignment**: Delayed or unsynced feedback limits its usefulness for interpreting outcome trends.

QUANTITIVE SURVEYS

Measure changes in key effectiveness and implementation outcomes over time (e.g., knowledge, self-efficacy, acceptability, intent to use).

QUALITATIVE INTERVIEWS

Explore *why* outcomes are changing (or not), and uncover deeper insights into user experiences, barriers, and contextual influences.

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

STUDY GOALS

IMPLEMENTATION OUTCOMES ASSESSMENT

KEY CONSIDERATIONS / BEST PRACTICES

•**Timing Matters**: Interview at key moments (**early, midline, endline**) to capture evolving user experiences and support real-time adaptation.

•**Purposeful Sampling**: Include diverse youth across engagement levels, regions, and demographics to reflect varied experiences with the chatbot.

•**Link to Outcomes**: Align interview questions with key implementation outcomes (e.g., acceptability, trust, feasibility) to explain survey trends.

•**Context Sensitivity**: Ensure discussions are age-appropriate, culturally relevant, and account for privacy concerns around SRH topics.

START CENTER

QUANTITIVE SURVEYS

Measure changes in key effectiveness and implementation outcomes over time (e.g., knowledge, self-efficacy, acceptability, intent to use).

QUALITATIVE INTERVIEWS

Explore *why* FP outcomes are changing (or not), and uncover deeper insights into user experiences, barriers, and contextual influences.

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

STUDY GOALS

IMPLEMENTATION OUTCOMES ASSESSMENT

KEY CONSIDERATIONS / BEST PRACTICES

•**Timing Matters**: Interview at key moments (**early, midline, endline**) to capture evolving user experiences and support real-time adaptation.

•**Purposeful Sampling**: Include diverse youth across engagement levels, regions, and demographics to reflect varied experiences with the chatbot.

•**Link to Outcomes**: Align interview questions with key implementation outcomes/frameworks (e.g., acceptability (TFA), trust, feasibility) to explain survey trends.

•**Context Sensitivity**: Ensure discussions are age-appropriate, culturally relevant, and account for privacy concerns around SRH topics.

QUANTITIVE SURVEYS

Measure changes in key effectiveness and implementation outcomes over time (e.g., knowledge, self-efficacy, acceptability, intent to use).

QUALITATIVE INTERVIEWS

Explore *why* FP outcomes are changing (or not), and uncover deeper insights into user experiences, barriers, and contextual influences.

TFA = Theoretical Framework of Acceptability

INTERRUPTED TIME SERIES + IMPLEMENTATION 1/4

METHODOLOGY PAPERS

-Wagner AK, Soumerai SB, Zhang F, Ross-Regnan D. (2002). Segmented regression analysis of interrupted time seriesstudies in medication use research. Journal of Clinical Pharmacy and Therapautics; 27: 299-309

-Bernal JL, Cummins S, Gasparrini A. (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. International Journal of Epidemiology; 348-355.

-Bernal LJ, Soumerai S, Gasparrini A. (2018). A methodological framework for model selection in interrupted time series studies. Journal of Clinical Epidemiology; 103: 82-91

- Simulation-based power calculation for designing interrupted time series analyses of health policy interventions

-Interrupted time series regression for the evaluation of public health interventions: a tutorial

HYBRID EFFECTIVENESS-IMPLEMENTATION DESIGNS

Intervention Study (Hussain et al; 2019)

Objective: Explore how a mobile phone texting service can be used to improve access to information about family planning by measuring users' intention to use Chatbot to acquire information about family planning and contraceptives

Intervention Characteristics: mobile phone-based Chatbot, built using a text message service that follows a decision tree structure to provide feedback to users on specific family planning methods. Based on the Unified Theory of Acceptance and Use of Technology (UTAUT) model (7 constructs).

Control: N/A

Intended Users and Context: Age 18 – 65 years, married, living together or engaged and considering which family planning method to choose.

Applicability to FP chatbot

- Similarity in terms of use of a text-message chatbot to deliver family planning information to participants considering family planning methods.
- Identification of factors predicting behavioral intention to use family planning Chatbot
- Provides a model, UTAUT for assessing

Trade offs

- "proof –of-concept" example; hasn't been evaluated to understand whether intervention was effective and feasible
- Applicants were part of a paid-participant pool and did not include participants who are single and may not be thinking about family planning
- Purely text-based so may not be fully applicable to the proposed FP Chatbot developed by Dimagi

MIXED METHODS

Pilot RCT (Hoa et al; 2017)

Objective: Assesses the effectiveness and adherence of delivering CBT and positive psychology strategies via a chatbot interface (28 participants)

Intervention Characteristics: **Digital-only** (phone app); Daily engagement encouraged for 14 days; No long-term follow-up of all 14 participants. Personalized responses based on user input.

Control: 14 participants in a wait list who did not receive the intervention during the 14 days

Intended Users and Context: Adults (20-49 years), interested in well-being and self-development.

Applicability to FP chatbot

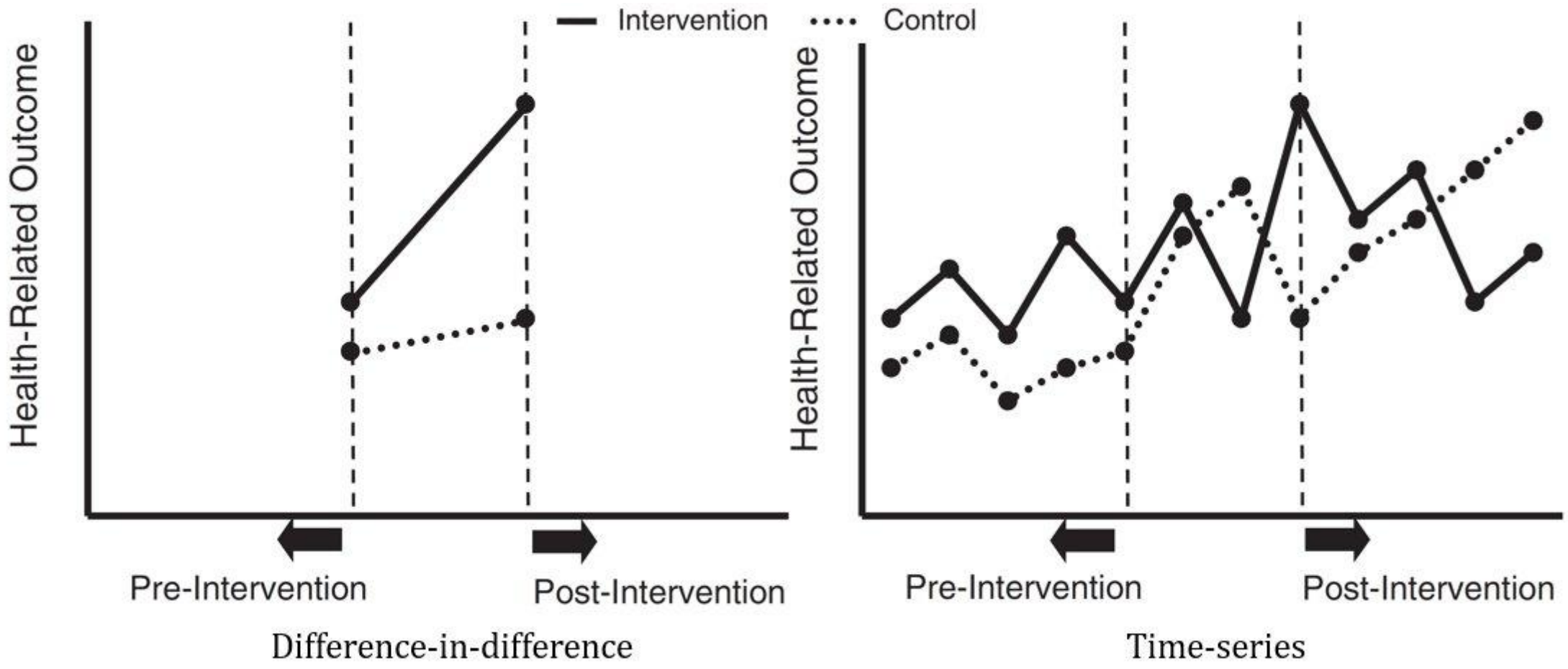
- Mixed methods allow for real-time performance tracking whereby user interviews and surveys can help uncover hidden usability barriers
- In-depth understanding of user experience
- Ongoing refinement of both **content and delivery methods**

Trade-Offs

- High technical and analytical demands – linking sentiments to chatbot usage
- Findings from small pilot studies or qualitative feedback may not generalize across all AYA groups or settings.

QUASI EXPERIMENTAL DESIGNS

Interrupted time series and Difference in Difference



INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

PROBABLE BASELINE, PROXIMAL AND DISTAL OUTCOMES

Key Indicators	How Measured	Rationale	Source Basis
Messages sent/received on the chatbot (count)	Chatbot- derived	Measures the extent of direct interaction with the intervention	Like "sending an average of 49.3 messages and average of 62.6 messages from Chatbot"
Time spent interacting with chatbot (per session or cumulatively)	Chatbot- derived	Indicates duration of exposure to content; can relate to the depth of engagement	Like "spent an average of 35.6 minutes on the bot
Conversational coherence (flow within the chatbot)	Chatbot- derived	Assesses logical flow and contextual understanding, and whether users are consuming intended content	Based on the concepts of the FP guidelines and structured interactions
Utilization of specific features (e.g., clicking links to resources, using a Q&A function)	Chatbot- derived	Measures interaction with key action- oriented components intended to facilitate behavior change or information seeking	Related to providing resources lists/videos/stories

Escobar-Viera et.al

START CENTER

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

PROBABLE PROXIMAL AND DISTAL OUTCOMES

Key Indicator	How Measured	Rationale	Source Basis
Acceptability (perceived appeal, appropriateness)	Self-reported (Survey)	Measures whether the target audience finds the chatbot suitable and relevant to their needs and context (high acceptability supports adoption and sustained use)	Measured quantitatively by the User Experience Questionnaire Short Version (UEQ-S)[1]
Usability (ease of use, learnability, efficiency)	Self-reported (Survey)	Measures how easy and intuitive the chatbot is to interact with (poor usability is a major barrier to continued engagement)	Measured quantitatively by the Chatbot Usability Questionnaire (CUQ)[2] and Post-Study Satisfaction and Usability Questionnaire (PSSUQ)[3]
Satisfaction (overall positive experience	Self-reported (Survey)	Gauges the user's overall impression and contentment with the bot experience	Measured quantitatively by the Client Satisfaction Questionnaire (CSQ-8).
Qualitative feedback on 'likes' and 'dislikes'	Self-reported (open-ended)	Provides rich context and specific areas for improvement related to usability and content	Qualitative feedback categorized into usability and content themes

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

PROBABLE PROXIMAL AND DISTAL OUTCOMES

key Indicators	How Measured	Rationale	Source Basis
FP knowledge (understanding of FP methods, where to access services, efficacy, side effects)	Sefl-reported (Survey/Quiz)	Chatbot content likely focuses on providing accurate information	FP guidelines
Perceived self-efficacy for discussing FP with partners/providers, accessing services, using FP methods correctly	Sefl-reported (Survey)	Belief in one's ability to successfully navigate FP-related situations	Social media self-efficacy
Attitudes towards FP use, safety, accessibility	Sefl-reported (Survey)	Positive attitudes are often necessary for adopting health behaviors	Optimize interactions to influence perceptions
Intention to seek FP services or use a specific FP method	Sefl-reported (Survey)	Strong predictor of future behavior, though not the behavior itself	Standard step in behavioral models

Escobar-Viera et.al

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

PROBABLE PROXIMAL AND DISTAL OUTCOMES

Key Indicators	How Measured	Rationale	Source Basis
Seeking information from other sources (e.g., healthcare provider) after chatbot interaction	Self- reported (Survey)	Measures whether the chatbot motivates users to take further action related to FP	Related to providing resources/links
Visiting a healthcare facility or resource provided by the chatbot	Self- reported (Survey)	A concrete step towards accessing FP services	Related to providing location-based resources
Perceived isolation	Self- reported (Survey)	Reduced isolation/increased connection could impact ability to seek/use FP	PROMIS Social Isolation Scale-REALbot Study
Depressive symptoms	Self- reported (Survey)	Mental health status can impact health behaviors, including FP use	PHQ-A (REALbot study0

Escobar-Viera et.al

INTERRUPTED TIME SERIES + IMPLEMENTATION ASSESSMENT

PROBABLE PROXIMAL AND DISTAL OUTCOMES

Specific Outcome	How Measured	Rationale	Source Basis
FP uptake (initiation of modern contraceptive method)	Self-reported (survey/verified through other means)	Represents the adoption of FP method	
Consistent FP use (adherence to chosen FP method)	Self- reported (Survey)	Represents the sustained use of an FP method	
FP efficacy (Reduction in unintended pregnancies)	Self-reported (survey/verified through other sources)	Represents the goal of the intervention	

Escobar-Viera et.al