Data & Policy (2024), 6: e22
doi:10.1017/dap.2024.16

TRANSLATIONAL ARTICLE

Data-driven healthcare indicators via precision gaming: With application to India

Chih-Hao Huang[1], Namita Mohandas[2], Aparna Raj[2], Susan Howard[2,3] and Feras A. Batarseh[4]

[1]School of Systems Biology, George Mason University, Fairfax 22030, VA, USA
[2]Howard Delafield International, 20007 Washington, D.C., USA
[3]School of Integrative Studies, George Mason University, Fairfax 22030, VA, USA
[4]Department of Biological Systems Engineering, Virginia Tech, Arlington 22203, VA, USA
Corresponding author: Feras A. Batarseh; Email: batarseh@vt.edu

Received: 31 May 2023; **Revised:** 28 January 2024; **Accepted:** 04 February 2024

Keywords: data-driven health; India; machine learning; precision gaming; public health policy

Abstract

Precision healthcare is an emerging field of science that utilizes an individual's health information, context, and genetics to provide more personalized diagnostics and treatments. In this manuscript, we leverage that concept and present a group of machine learning models for precision gaming. These predictive models guide adolescents through best practices related to their health. The use case deployed is for girls in India through a mobile application released in three different Indian states. To evaluate the usability of the models, experiments are designed and data (demographic, behavioral, and health-related) are collected. The experimental results are presented and discussed.

Policy Significance Statement

This research aims to analyze how data are produced to determine the outcomes that could be directly applied to increasing sexual and reproductive health knowledge among Indian adolescent girls. The presented mobile game encourages best practices related to healthcare policies in India (as compared to the United States) and the overall accepted health and hygiene norms.

1. Introduction

Adolescence is a formative time when choices and decisions can chart the course of the rest of one's life. With new awareness of bodily and emotional changes (with their parents supervision), and developing interests, including exploring relationships, most adolescents are exploring territory they have never faced before. They may experience a culture of silence and shame. India is home to the world's largest population (Jejeebhoy, 1998) of adolescents age 10–19 years at 253 million, comprising of 20% of the country's population[1].

This places the country at the center of global efforts to improve reproductive healthcare for this age group (Desai et al., 2021). Despite substantial improvement in sexual and reproductive healthcare in

[1] India has the world's largest adolescent population

India, such as the supply of a variety of contraceptive methods, adolescents continue to have low access to sexual reproductive health services, with the poorest indicators among adolescent girls (Bhan et al., 2020; Kedia et al., 2022). There is little access to accurate information and counseling on sexual reproductive health. Adolescent girls also live in a context where social and gender norms are the main driving forces (Cherewick et al., 2021).

1.1. An introduction to the game

The mobile game presented in this manuscript, namely: *Go Nisha Go* (GNG), as the first model of the The Game of Choice, Not Chance™ (GOC) (USAID, 2023) initiative, has created a novel approach to advancing adolescent sexual and reproductive health. It combines behavioral science, game-based-learning, human-centered design, predictive analytics, and interactive storytelling in a novel platform (Damaševičius et al., 2023). GOC is designed with and for adolescent girls in India, aged 15–19. The innovation uses *discovery and play* to empower players to become active decision-makers in their own lives. Evidence suggests that such interactive interventions may promise improved health outcomes, particularly in psychological and physical therapy (Primack et al., 2012). Research has shown that various antecedents significantly influence the behavioral intention of students in rural girls' schools in India when using mobile applications for the teaching–learning process (Chatterjee et al., 2020). Within the mobile game, girls learn through role-play how their choices matter and get connected to real-world resources to become better equipped to make decisions, build confidence, and ultimately, achieve positive health, safety, career, and educational outcomes in their lives (Konstantinidis et al., 2017). Together, these elements provide a virtual safe space for players to explore and navigate life choices, learn about sensitive and taboo topics, experience the simulated outcomes of their choices, and gain direct access through a direct-to-consumer approach with a strategically curated selection of information and resources designed to advance and support negotiating relationships, use of contraception, and managing menstrual health and hygiene (Johnson et al., 2016; Mayr et al., 2017; Wang and Zheng, 2021; Efe and Topsakal, 2022). Machine learning, which has found effective application in various sectors (Boukenze et al., 2016; Huang et al., 2021; Williamson, 2016), is being applied to capture in-game data, identify associations, make predictions, and create precision messaging based on in-game behaviors (Batarseh et al., 2021).

The game, developed specifically for India, is applicable to young girls in other countries as well. The game positions information about menstrual hygiene management, fertility awareness, and contraception available easily and within the context of a girl's life, along with access to and information about menstrual health management, contraceptives, and health products and services. GNG allows players to learn, engage, explore, and build confidence and agency to negotiate sex, practice consent, and delay early marriage and pregnancy. This learning and practice within a safe virtual space make GNG relevant for adolescent girls across the world.

1.2. Research questions and data collection

The interactive game, as shown in Figure 1, consists of five levels or episodes, with each addressing specific topics pertinent to Indian teenage girls. Players must make choices such as determining if the character should negotiate clothing choices with her parents or negotiate consent and contraceptive use with her boyfriend. Responses to these diverse interactions are collected for subsequent analysis. Additionally, decisions are scored based on three *vitals*: health, relationship, and confidence. The vitals aim to make a decision more "compelling" as opposed to enforcing the "right" choice. The research employed an empirical approach to examine and utilize in-game data to augment the player's experience. The study involved making observations, posing questions, formulating hypotheses, making predictions based on the hypotheses, and conducting experiments to test the outcomes.

The observation that players interacted with the game prompted the question of leveraging in-game data to enhance the gaming experience and deliver better information to players. The hypothesis

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

Figure 1. *Gameinterface example.*

suggested applying data-driven techniques to the gathered data would produce actionable insights for in-game feature development.

To test this hypothesis, a structured methodological approach was employed (Figure 2). During the data collection phase, game data were uploaded to a MongoDB database when players were connected to the internet. To address various predictive analytics methods and objectives, data preprocessing and reformatting were conducted using data wrangling techniques, ensuring compatibility with different analytical models.

Subsequently, the processed data were put into multiple predictive analytics models, each tailored to address distinct aspects of the game. The results of these models were stored in a MongoDB database for further analysis and implementations in the game.

Finally, insights derived from the predictive analytic models were applied to additional in-game features. These enhancements were integrated into the game to improve the overall gaming experience and provide precision recommendations for players.

The data presented above and the experimental analysis performed for this work answer the following two research questions (i.e., two manuscript contributions):

- RQ#1: How can in-game data be leveraged through predictive analytics to improve players' learning through gaming?
- RQ#2: In accordance with Indian policies, how can precision gaming features improve the well-being and health awareness of teenage girls?

The next section presents the data collected and experiments designed to answer both RQs.

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

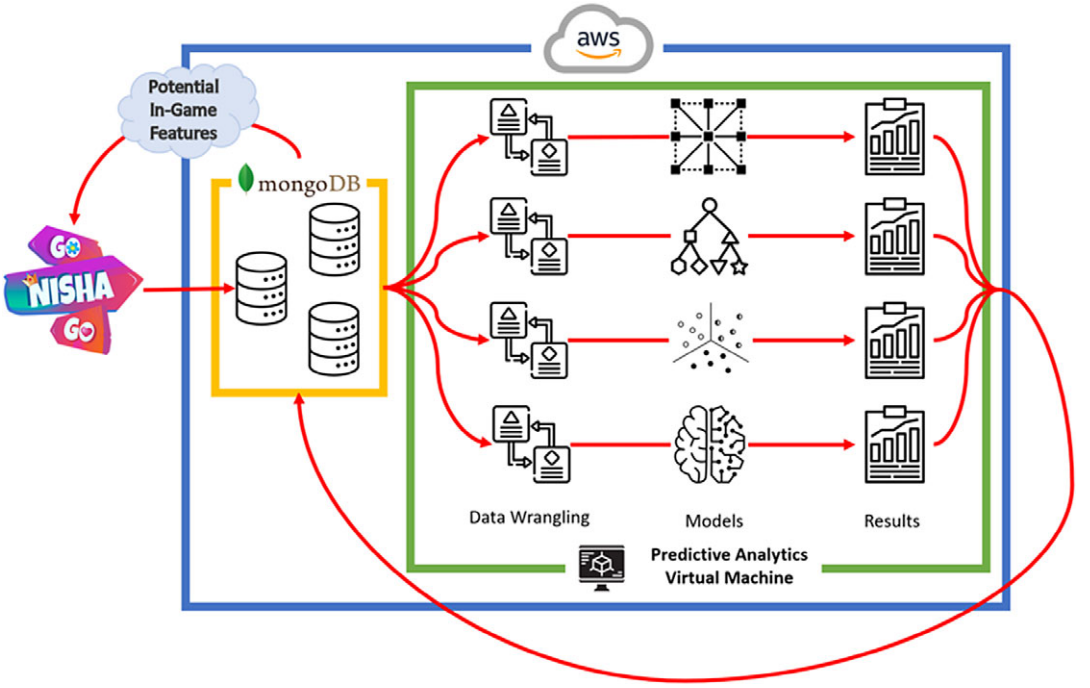

Figure 2. *Game data life cycle.*

1.3. Summary of contributions

In the presented research, we advocate for a novel clustering algorithm specifically designed for the segmentation of game users. Utilizing advanced data-driven methodologies, this model facilitates the categorization of users into discrete clusters. This, in turn, offers an enhanced granularity in the comprehension of user behavioral patterns, setting the stage for future targeted interventions and analyses within the gaming ecosystem.

Moreover, extending our scope beyond the confines of the gaming domain, we leverage the quantifiable impacts of gaming interfaces on health awareness within the Indian demographic. By using state-of-the-art artificial intelligence techniques, we ascertain measurable outcomes that underscore the pivotal role of digital platforms in health-related cognizance among diverse populations. Such findings provide a deeper understanding into the symbiotic relationship between technology and societal well-being.

Conclusively, our research inaugurates an innovative paradigm in health-related gaming, namely: "precision gaming features," which are predicated on rigorous data-driven strategies. These features aim to refine gaming experiences to an individual's contexts. The remainder of this paper is structured as follows: We begin with an overview of in-game data collection. Next, we present our experimental design with model details in the methodology section, followed by the results section. We conclude with a discussion of policy and healthcare-related impacts.

2. In-game data collection

In this study, while the potential influence of demographic factors such as socioeconomic status, age, and educational level is acknowledged, these variables were not controlled for in the research design. However, controls for geographical areas were implemented to establish a foundational understanding of the subject matter, with a strategic focus on three specific areas: Delhi, Rajasthan, and Bihar. These regions were selected to represent diverse gaming environments and user behaviors. It is crucial to note, however, that the availability of the game was not limited to users in these regions. Any individual residing

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

in India could download the game via the Google Play, the India store, ensuring widespread reach and availability across the country. Despite the delineation of these regions, it should be clarified that the analytics did not concentrate on or prioritize data from these specified areas. The primary objective of the data collection process was not only via collected demographic information but to gather comprehensive information regarding the decisions and selections made by players throughout their game play. The data aim to provide insights into player behavior, thought processes, knowledge, and decision-making patterns under various virtual situations.

Demographic data collected include age, gender, education background, current working status, number of siblings, and whether if they own their own phone. The comprehensive information collected can be further divided into three main categories and each with subcategories. The three main categories are health, confidence, and relationship. Under health, it is further divided into asking for help, pregnancy facts, menstrual hygiene management facts, and contraception negotiation. Confidence and relationship each has two subcategories; confidence is divided into openness and self-efficacy, and relationship is divided into negotiation and obligation.

Obligation measures the extent to which an individual feels compelled to reciprocate in specific ways in their relationships, with their romantic partner and their family too. This parameter encapsulates adolescent girls' drive to be health aware (Raj et al., 2023). With or without reciprocity, individuals' sense of obligation frequently seems to be the glue that holds some of their relationships together (Tedgård et al., 2018). A survey of the existing literature provides a mixed portrait of the role of obligation on both individuals' well-being and the quality of their relationships (Goldberg, 1993).

In alignment with commitment to safeguarding participant privacy and ensuring the integrity of the data collection process, all procedures strictly adhere to the Indian data privacy policies. The protocols are in full compliance with the Digital Personal Data Protection Act of 2023 (DPDP Act) and The Digital Personal Data Protection (PDP) Bill in India. These regulations mandate stringent guidelines on data storage, processing, and sharing. All participants were informed of their rights under these acts, including the right to access their data, the right to rectify inaccuracies, and the right to erasure. Since the target players are under the age of 18, to preserve player anonymity, all data collected is meticulously processed in a manner that ensures it cannot be linked back to any individual player, thereby maintaining the confidentiality and privacy of each participant's personal and game play information. Measures were taken to ensure data minimization and the principle of collecting data strictly relevant to the research objectives.

3. Methodology

3.1. Clustering

K-mode clustering is a widely used unsupervised machine learning technique for clustering categorical data based on their attributes. The algorithm is named "K-mode" because it uses modes, which are the most frequent values, to represent the clusters rather than means or medians. In Python, the KModes class from the *kmodes* library is commonly used to implement this algorithm (Cao et al., 2009; Huang, 1998). One can specify the number of clusters and the initialization method. The initialization method determines how the initial cluster centroids are chosen, which is a critical step in the clustering process. The "Huang" initialization method (Huang, 1998) is one of the available options in KModes. It utilizes a combination of random initialization and an advanced initialization strategy to select the initial centroids, which can improve the efficiency and accuracy of the clustering. K-mode clustering with the "Huang" initialization method can be a powerful tool for identifying meaningful patterns and clusters in categorical data, which can have a significant impact on scientific research in various fields (Huang, 1998).

The elbow method is a commonly used technique to determine the optimal number of clusters for K-mode clustering (Batarseh et al., 2020). The method involves calculating the within-cluster sum of squares (WCSS) for a range of cluster numbers and plotting the results against the number of clusters. The elbow point in the plot indicates the optimal number of clusters where adding more clusters does not improve the clustering performance significantly. In K-mode clustering, the elbow method can be used to

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

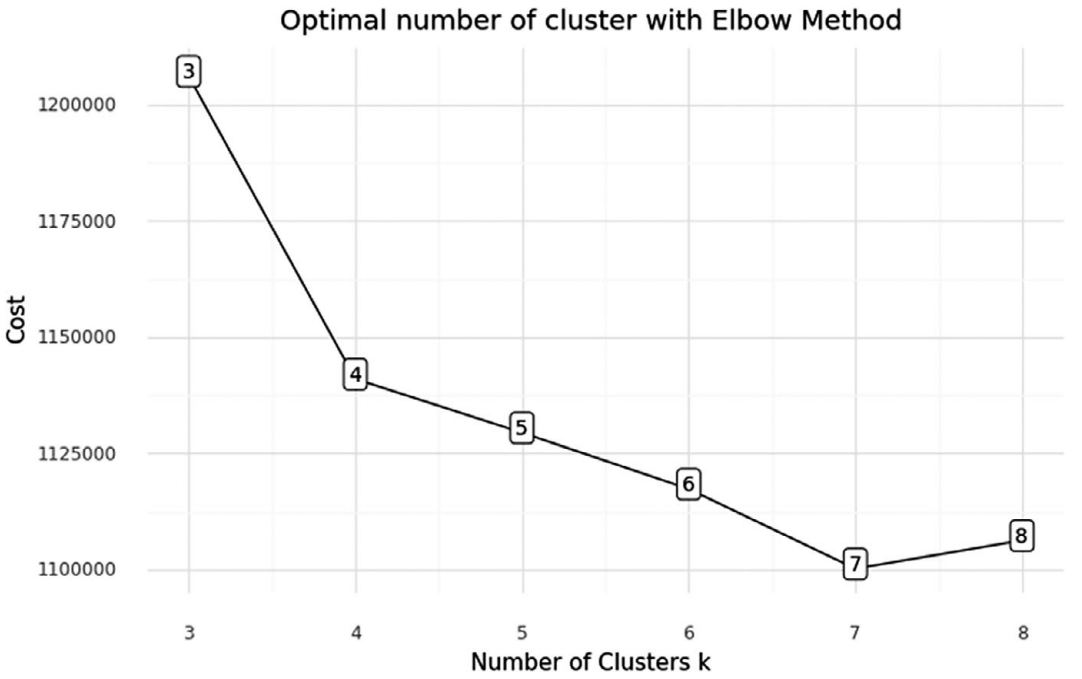

Figure 3. *The elbow diagram for K-mode clustering.*

find the best number of clusters based on the mode-based dissimilarity measure. The mode-based dissimilarity measure calculates the dissimilarity between two data points based on the number of common modes they share. The elbow method can be applied to the mode-based WCSS to determine the optimal number of clusters for K-mode clustering. By using the elbow method, the most appropriate number of clusters (in GOC, the players are clustered) for K-mode clustering can be identified, which can help in identifying meaningful patterns and insights in categorical data.

In the clustering analysis, the *kmodes* package was utilized, with the number of clusters set to seven. The determination of the optimal number of clusters was based on the elbow diagram, which examined the cost and number of clusters within the range of 3–9 (Figure 3). The Huang initiation method was employed for both the elbow diagram and k-modes clustering analysis. Huang's initialization is a frequency-based approach, selecting initial centroids based on the mode (most frequent value) for each attribute. This method is suitable to select centroids that are representative of data distribution.

The data used in this study are drawn from two types of datasets: a collection of decisions made by players depicted through 507 questions and answers combinations, and demographic data such as: age, state, sibling count, education level, employment status, gender, and phone ownership.

The hyperparameter tuning process was meticulously conducted to ensure the robustness and accuracy of the clustering model. The determination of the optimal number of clusters was made through careful consideration, resulting in the selection of seven clusters, a choice informed by the elbow diagram analysis. The clustering algorithm was initiated using the Huang method, a recognized technique for effectively initializing cluster centroids in categorical data.

To maintain the fidelity and reproducibility of our results, we set a random state of 4, which guarantees consistent outcomes across different runs. A maximum of 100 iterations was permitted to allow the algorithm to converge effectively. In the interest of precision, we refrained from enabling the "fast mode" option, prioritizing accurate clustering over speed.

It is worth noting that data scaling was deliberately omitted from the preprocessing phase before clustering. This decision was made to preserve the integrity of the categorical data attributes and avoid introducing any inadvertent bias into the analysis.

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

The "nstart.method" parameter was configured to "best," ensuring the selection of the run with the lowest cost among multiple initiations. The "cluster.only" parameter was intentionally set to FALSE, providing us with a comprehensive understanding of the clustering results beyond mere cluster assignments. The "kmodes.debug" parameter remained at its default value of FALSE, which suppressed debugging information display during the clustering process.

Notably, no weights were assigned to the model, as we aimed to maintain the integrity of the data's inherent structures. Subsequently, the resulting clusters were visually represented, and their quality was rigorously evaluated. This evaluation was based on metrics that assessed the compactness and separation of the clusters, affirming the efficacy of this method as a robust framework for extracting meaningful patterns and relationships within the categorical data.he hyperparameter tuning is conducted with seven clusters chosen based on the elbow diagram, initiation using the Huang method, five initiations, verbosity level set to 1, a random state of 4 to maintain result reproducibility, a maximum of 100 iterations, and one parallel job running to optimize the clustering process.

For the elbow diagram, the random state parameter is set to 0, while the number of jobs parameter (n_jobs) is set to −1, indicating the use of all available processors for parallel computation. In the k-modes clustering process, the number of initiation iterations is set to 5, and the verbosity level = 1, providing detailed outputs for the clustering analysis.

As shown in Equation (3.1), the function can represent how k-mode clustering works. Here, C_j represents the jth cluster, x is a data point, μ_j is the mode of the jth cluster, and d(x, μ_j) is the dissimilarity measure between x and μ_j, which is typically the Hamming distance for categorical data.

$$\min_{C_1,\ldots,C_k} \sum_{j=1}^{k} \sum_{x \in C_j} d\left(x, \mu_j\right) \tag{3.1}$$

Equation (3.2) represents the within clusters sum of squares of the model, whereas, k is the number of clusters, C_i is the ith cluster, x is an instance in cluster C_i, c_i is the centroid of cluster C_i, and $d(x, c_i)^2$ is the distance between instance x and centroid c_i with a dissimilarity measure appropriate for categorical data.

$$WCSS(k) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, c_i)^2 \tag{3.2}$$

The above equation is used to evaluate the quality of the unsupervised clustering algorithm.

3.2. Information gain

Information gain is a commonly used technique in machine learning to identify the most influential variables in a dataset. It measures the amount of information that a particular variable or feature contributes to the prediction of the target variable (Kent, 1983). The information gain of a variable is calculated by subtracting the entropy of the target variable before splitting the data based on the variable from the weighted sum of the entropy of the target variable after splitting the data. Variables with higher information gain are more influential in predicting the target variable. Information gain is particularly useful in decision tree algorithms, where the most influential variables are used as the root node of the tree. By identifying the most influential variables in a dataset, the dimensionality of the data can be reduced and improve the accuracy of the machine learning models. Information gain can be used in various fields, such as healthcare, social media, and finance, to identify the most important factors affecting a particular outcome or behavior. We used this technique to gain insights into the underlying patterns of the data, and to understand which variables are key indicators of different types of players.

The analysis of information gain was executed utilizing the FSselector package in R programming language. This method facilitated a comprehensive understanding of the underlying patterns in the dataset and enabled the identification of critical features for subsequent model development. By applying the

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

FSselector package, a data-driven approach was adopted, ensuring the extraction of meaningful insights that contributed to a more robust and accurate predictive model.

Equation (3.3) is the formula for information gain.

$$IG(Class, Attribute) = H(S) - \sum_i p(S_i)H(S_i) \tag{3.3}$$

In this formula, $IG(Class, Attribute)$ represents the information gain, $H(S)$ represents the entropy of the unpartitioned set S, S_i are the subsets of S induced by the attribute, and $p(S_i)$ represents the proportion of $|S_i|$ in the total size of S. The next section presents the results from the game, as it is deployed in three Indian states.

3.3. Association rules

Association rules are a powerful technique used in data mining to identify patterns and relationships between items in large datasets (Kotsiantis and Kanellopoulos, 2006). One application of association rules is in identifying the item selection preferences of individuals based on their transactional history. This technique involves analyzing the frequency of co-occurrence of items in transactions and identifying rules that describe the likelihood of an individual purchasing one item given the purchase of another. The strength of these rules can be measured using metrics such as support and confidence. Support measures the frequency of occurrence of a rule in the dataset, while confidence measures the probability of purchasing the consequent item, given that the antecedent item was already purchased. By analyzing the association rules, the common item selections made by individuals can be identified, which can help in understanding their preferences and predicting their future purchasing behavior. The use of association rules in identifying individual preferences can be applied in various fields, such as recommendation systems, marketing, and e-commerce (Kotsiantis and Kanellopoulos, 2006). This method is used in associating decisions on items that girls make in the game, and what that tells about their behavioral patterns. To add variety to the game, players are able to have a simulated shopping experiences. AR models are used to govern those choices and better inform the player of existing associations.

The association rules models were based on Andrew Brooks' R code[2]. The code has been altered to work with the data. And the data have been separated based on the completion of levels.

There are three formula associated with the association rules, support, confidence, and lift. Equation (3.4) represents the support which measures the frequency or popularity of an itemset in the dataset. Mathematically, support is the fraction of the total number of transactions in which the itemset occurs. It helps identify rules worth considering for further analysis.

$$\text{Support}(X) = \frac{\text{Number of transactions containing} X}{\text{Total number of transactions}} \tag{3.4}$$

Confidence measures the likelihood of item Y being selected when item X is selected, which is shown as Equation (3.5). It is the ratio of the support of the combined itemset (X and Y) to the support of the itemset X. A higher confidence value indicates a stronger association between the items in the rule.

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)} \tag{3.5}$$

Finally, lift is a measure of how much more likely item Y is to be selected when item X is selected, compared to when item Y is selected randomly. It is the ratio of the confidence of the rule (X ⇒ Y) to the support of the itemset Y, shown as Equation (3.6).

[2] Andrw Brooks' R code

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Confidence}(X \Rightarrow Y)}{\text{Support}(Y)} \tag{3.6}$$

A lift value greater than 1 indicates that the rule may be useful, while a lift value of 1 or below suggests that the rule is not very useful.

4. Results

4.1. Clustering

The data collected for classification can be categorized into four categories: demographics, health-related, relationship-related, and confidence-related. Each category has been further divided into two to four subcategories. Health, relationship, and confidence-related categories were collected as scores based on the score predefined for each response to different situations in the game.

For health-related scores, the scores are rated from poor, average, good, and very good to excellent, with the exception of low statistical significance (Table 1). Cluster 7 is the cluster with low statistical significance across all three scoring categories: health, relationship, and confidence-related. For health-related scores, there are four subcategories associated with it, which are asking for help, pregnancy facts, menstrual hygiene and management facts, and contraception negotiation. For the health-related scores, if a cluster scored poorly for a specific subcategory, a *needs help* flag would be raised. Overall, Clusters 1, 5, and 6 perform well in all four subcategories. All clusters (except Cluster 7) perform statistically well in menstrual hygiene and management facts. Cluster 2 requires additional assistance in the subcategory of asking for help, and Cluster 4 needs extra assistance in pregnancy facts and contraception negotiation. The game is live and will continue to create more data, we aim to improve the models further as more data are available.

The relationship score is further divided into two subcategories: obligation and negotiation (Figure 4). A score of 1 represents the highest value, and scores closer to 0 indicate lower values. Cluster 7 is the low statistical significance cluster with insufficient data. A higher score does not necessarily imply better performance, but rather that the cluster is highly obligated or highly skilled in negotiation. A cluster can exhibit high obligation and high negotiation skills simultaneously, such as Cluster 1. Clusters 1, 5, and 6 demonstrate excellent skills in negotiation, while Clusters 1, 2, and 3 exhibit extreme obligation.

Analogous to the relationship score, the confidence score is divided into two subcategories: openness and self-efficacy (Figure 5). Employing the same scoring system, 1 represents the highest score, with scores closer to 0 being lower. A higher score signifies that the cluster exhibits high openness or high self-efficacy, but it does not necessarily indicate superiority over other clusters. Cluster 6 achieves perfect scores of 1 in both openness and self-efficacy. All six clusters score higher than 0.5 for self-efficacy, with the exception of Cluster 7, which has low statistical significance. Cluster 2 has the lowest score of 0.154 in openness, followed by Cluster 3's score of 0.231, suggesting that these two clusters are more "traditional."

Table 1. *Healthcare scores for each cluster*

	Asking for help	Pregnancy facts	MHM facts	Contraception negotiation
Cluster 1	Average	Excellent	Excellent	Excellent
Cluster 2	Very Poor	Good	Average	Good
Cluster 3	Good	Poor	Excellent	Average
Cluster 4	Excellent	Very Poor	Average	Very Poor
Cluster 5	Average	Excellent	Good	Good
Cluster 6	Average	Excellent	Good	Excellent
Cluster 7	LSS	LSS	LSS	LSS

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

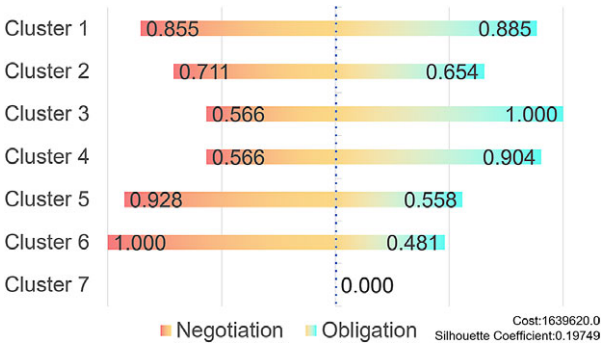

Figure 4. *Relationship scores for each clusters.*

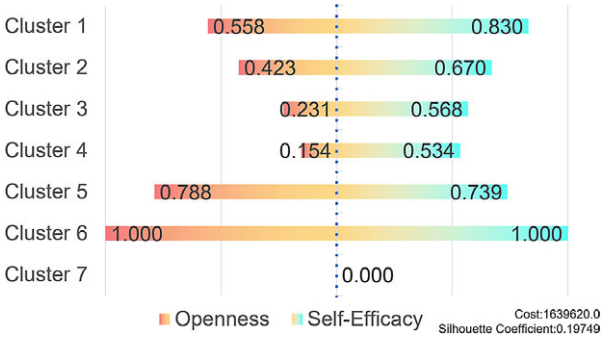

Figure 5. *Confidence scores for each clusters.*

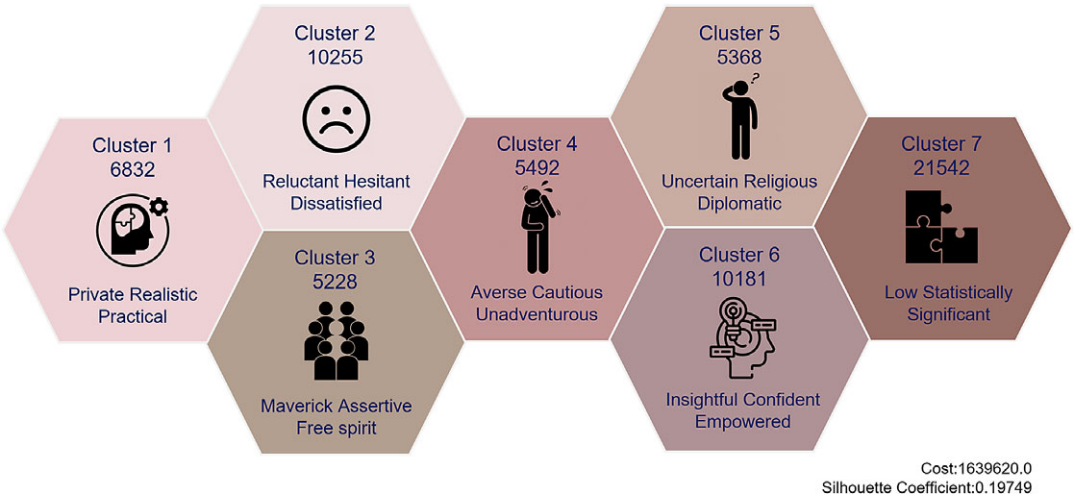

Figure 6. *Cluster summaries.*

According to the clustering results derived from demographic, health-related, relationship-related, and confidence-related data, each cluster has been assigned a description consisting of three words that best characterize and describe the cluster (Figure 6). Cluster 7 is an exception, as it is not described by three words but is instead labeled as having low statistical significance.

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

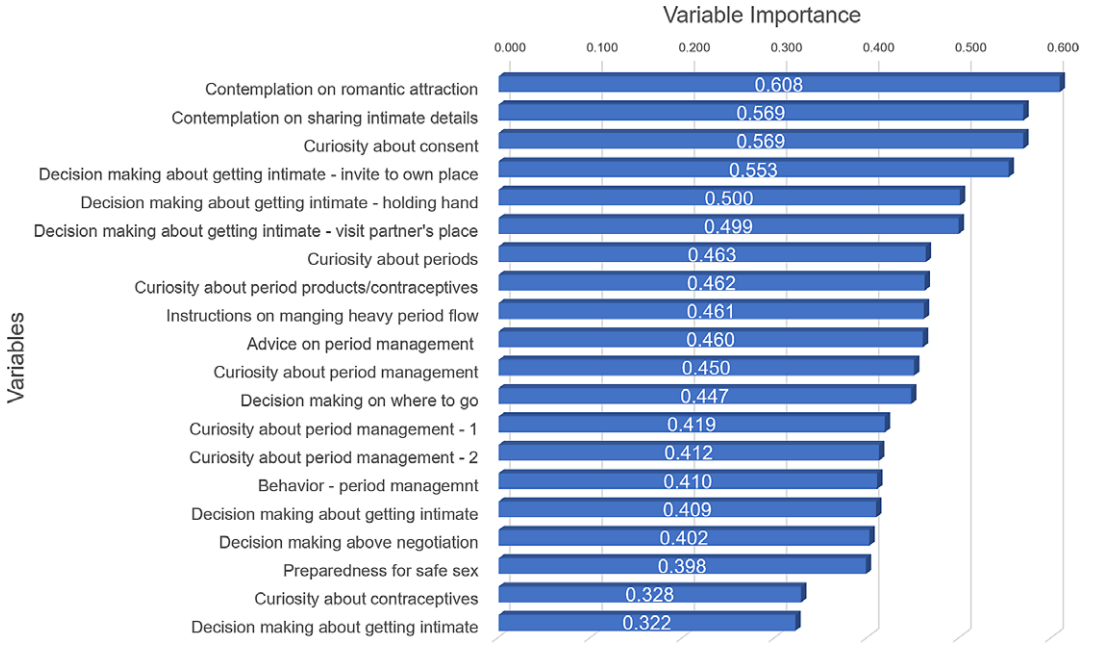

Figure 7. *Top 20 information gain.*

Cluster 7 is the largest cluster among all seven categories, with a population more than twice that of the second-largest cluster, Cluster 2, which has 10,255 players. Cluster 3 is the smallest cluster, containing 5228 players, while Clusters 4 and 5 exhibit similar population sizes, with 5492 and 5368 players, respectively.

4.2. Information gain

Variable importance reflects the degree to which a model depends on a specific variable for accurate predictions (Figure 7). It demonstrates the contribution of each variable to the model's predictive power, with higher importance signifying a greater influence on the model's performance.

The most important variable is *contemplation on romantic attraction*, with a variable importance (i.e., gain) score of 0.608. Among the top 20 variables, 8 are related to periods, followed by 6 variables associated with intimacy. Most demographic-related variables exhibit the lowest gain scores among all variables.

4.3. Association rules

The top 10 results from the association rules model reveal that laptops are the most favored item, as 7 out of the top 10 rules are related to laptops (Table 2). The three rules not associated with laptops involve items related to sexual health, such as associations between pregnancy test kits and condoms, oral contraceptive pills (OCP) and condoms, and emergency pills and OCP. Among the seven rules linked to laptops, one is associated with pregnancy test kit, another with menstrual cups, and the remaining five with non-health-related items.

4.4. Postgame evaluation

The study engaged a total of 769 players, revealing enlightening findings regarding the game's educational facets (Table 3). From the cohort, 45.3% reported gaining knowledge about contraceptive methods,

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

Table 2. *Top 10 association rules*

LHS	RHS	Support	Confidence	Coverage	Lift	Count
Makeup	Laptop	0.0370	0.4205	0.0880	1.4497	592
Diary	Laptop	0.0326	0.3508	0.0928	1.2097	521
English learning app	Laptop	0.0355	0.3653	0.0972	1.2594	568
Pregnancy test	Condom	0.0319	0.3220	0.0990	1.7522	510
Smartphone	Laptop	0.0317	0.3159	0.1003	1.0891	507
OCP	Condom	0.0338	0.3205	0.1055	1.7442	541
OCP	Laptop	0.0311	0.2950	0.1055	1.0172	498
Pregnancy test	Laptop	0.0332	0.3122	0.1063	1.6988	531
Menstrual cups	Laptop	0.0322	0.2922	0.1103	1.0074	516
Emergency pill	OCP	0.0328	0.2823	0.1162	2.3395	525

Table 3. *Game impact on player knowledge and empowerment*

Aspect	Percentage of players
Learned about contraceptive methods	45.3%
Learned about information sources and access	42.1%
Learned problem–solving abilities	38.5%
Developed self–efficacy	21.1%
Found game helpful during internships	87.1%
Believed game was beneficial for capacity building	89.5%
Valued game insights on health helplines	90.5%
Reported guidance on legal advice beneficial	89%
Negotiated with parents on career aspirations	94%
Discussed moving out for career opportunities	94.9%
Reported enhanced negotiation on contraceptive choices	90.6%

while 42.1% gleaned insights into information sources and accessibility. The game also appeared to foster problem-solving abilities, with 38.5% of the players asserting they learned this critical skill. Additionally, 21.1% felt the game bolstered their self-efficacy.

A significant proportion of participants found the game instrumental in real-life applications. Specifically, 87.1% found it advantageous during internships and a substantial 89.5% believed it was beneficial for capacity building. When it came to seeking advice, 90.5% found the game's insights on health helplines valuable, while 89% reported its guidance on legal advice beneficial.

In the realm of interpersonal negotiations, the game exhibited profound impacts. A striking 94% of players stated that the game equipped them with the skills to negotiate with parents about their career aspirations. Furthermore, 94.9% believed the game provided them with the necessary tools to discuss moving out for career opportunities.

These data robustly underscore the multifaceted advantages that players derived from the game, spanning areas of knowledge, real-world applications, and interpersonal negotiations.

5. Discussions and conclusions

This manuscript presents GNG, a game that uses data-driven methods to inform and improve on healthcare practices among adolescent girls in India using precision gaming. Such a game has both health and policy impacts, which are presented in this section.

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

5.1. Policy-related impacts

The Indian government has been making concerted efforts to ensure the health and well-being of adolescent girls. Multiple large-scale programs, such as the National Adolescent Health Program, *Rashtriya Kishor Swasthya Karyakram*, and *Kishori Shakti Yojana*[3], have been launched to provide health services, information, and counseling related to reproductive health, nutrition, mental health, and personal hygiene. These programs prove the widespread acknowledgment regarding the significance of fostering healthy behaviors, especially during the adolescent phase, and to cultivate a healthy nation overall. Nevertheless, due to the large scope of these programs, their primary emphasis lies on ensuring the provision of information and services, with limited attention given to behavior change strategies. Moreover, when such strategies are implemented, they often adopt a one-size-fits-all and didactic approach. In response to the insufficient availability of products that prioritize the needs of adolescents, games such as GNG strive to fill this gap by providing a fun and engaging experience that caters to their requirements. These games aim to adapt to the player's pace, offering a personalized approach. Integrating machine learning technology to continually learn and adjust to the evolving needs of adolescents and deliver customized content has become imperative in the current context.

5.2. Healthcare impacts

Innovative approaches like stealth learning through gaming can prove crucial for promoting preventive healthcare for adolescents. However, it is essential to note that the adolescent phase is dynamic and demands tailored content that adapts to their evolving needs. Using a clustering algorithm allows the creation of a product that serves dynamic content that matches the evolving needs of adolescents while delivering maximum impact.

The six clusters identified through the models discussed above have been used to create precision messaging for the girls who play GNG. The game will process the data of each player, identify the cluster they belong to, and show them the messages that best suit their needs. For example, the girls in Cluster 4 will be specifically informed about the fertile period. Additionally, they will receive nudges emphasizing the importance of taking ownership of their contraceptive choices and ensuring their use. These messages serve as additional doses of information, building upon the content discussed within the game. By aligning these messages with the gaps in knowledge and attitude identified within Cluster 4, we can enhance the likelihood of girls following healthy reproductive health behaviors.

Through the clustering model, we noticed that girls in Clusters 1, 3, and 4 required additional help refusing sexual intercourse (within the game) as they scored very high values for obligations. The game uses this information about the girls to link them with resources that would build their self-efficacy in nonconsensual sexual advances.

Identifying clusters, and providing precision messages to girls, will help girls identify gaps and receive support in areas where health awareness might be low. The in-game precision messaging provides them with the nudges and cues to focus on specific aspects of their health, or with information that might be vital in preventing early pregnancy, dysmenorrhea, seeking support from a specialist, or to encourage them to confidently negotiate for and prioritize their needs with their partner.

The game exhibited discernible positive impacts on its young female players across various facets of their lives, specifically in the areas of menstruation, self-efficacy, and decision-making. The enhancements concerning menstruation were multifaceted, spanning increased fertility awareness, an augmented understanding of menstruation, and a surge in menstrual hygiene knowledge. One of the most noteworthy metrics pertains to fertility awareness: at the outset, a mere 28% of the participants were cognizant of the days in a menstrual cycle when conception is least probable. Impressively, postgame data indicated a substantial leap to 77% in this understanding. Furthermore, there was a commendable 48% increase in players who acquired knowledge regarding the recommended frequency of pad changes, emphasizing improved menstrual hygiene practices. Additionally, the game made significant strides in rectifying

[3] Kishori Shakti Yojana

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

misconceptions: before the gameplay, only 41% of players recognized menstruation as a typical biological reproductive process, a figure that soared to 77% afterward. In contrast, the game also achieved success in dispelling prevalent myths; the belief that menstruation serves to expel impure blood, which was held by 48% of the players initially, plummeted to 18% post-gameplay. These comprehensive data underscore the game's transformative potential in not only enhancing knowledge but also reshaping deeply ingrained perceptions among young girls.

In this study, the game's profound influence on self-efficacy among its players became saliently evident. The data revealed significant upticks in the players' confidence related to career pursuits, marital timing, and personal choices—especially in the face of external pressures. Prior to engaging with the game, only 47% of the girls expressed confidence in their ability to actualize their dreams. However, this figure saw a marked increase, with postgame data revealing a surge to 90%. In tandem with this, our findings highlighted an initial 54% of girls who demonstrated confidence in their choice to delay marriage. After gameplay, this percentage rose to 75%, showcasing the game's efficacy in reshaping perceptions about marital timing. Most strikingly, in scenarios where players faced pressures from boyfriends to act contrary to their inclinations, there was a substantial shift in the participants' responses. Pregame statistics showed that only 42% of girls felt empowered to resist such pressures; postgame, this escalated dramatically to 84%. This underscores the game's transformative potential in bolstering resilience and self-assuredness among its young female players.

The impact of the game on players' proactive participation in decision-making, particularly in the context of contraceptive discussions, was notably evident from the acquired data. Prior to game exposure, 73% of the girls expressed a willingness to engage in open discussions about contraception with their boyfriends. This inclination saw a marked escalation post-gameplay, with the percentage rising to 88%. This amplification in active dialogue showcases the game's potential in fostering open communication and empowering young women in intimate discussions. Concurrently, the game also seemed to influence the girls' autonomy in contraceptive decision-making. Initially, 21% of the girls opted to defer contraceptive decisions to their boyfriends. However, after their interaction with the game, this figure witnessed a reduction, descending to 17%. This decline underscores the game's efficacy in promoting individual agency and reducing dependency on partners for crucial personal choices.

In response to the insufficient availability of products that prioritize the needs of adolescents, GNG strives to fill this gap by providing a fun and engaging experience that caters to players goals and habits. GNG aims to adapt to the player's pace, offering a personalized approach. Integrating data-driven methods into precision gaming enables adolescents to continually learn and adjust to their evolving needs and allows for customized content that has become imperative in today's world of information overload and the overwhelmingly high number of educational sources.

Data availability statement. All steps have been taken to ensure that participant information collected during this research remains fully anonymized and confidential. No identifiable data have been collected; and strict measures have been implemented to safeguard the privacy of the participants. To further guarantee the security of the collected data, all information is stored on Amazon Web Services (AWS) with multiple security features applied. Our AWS cloud server is configured to be HIPAA-compliant and meets all the necessary security requirements in India for storing in-game data. Data samples are available upon request from the team.

Acknowledgments. The authors are grateful for the support of HDI team members: Kavita Ayyagari, Lalita Shankar, Joy Pollock, and Jeannette Cachan.

Author contribution. C.H. performed the overall writing, data curation, and formal analysis; N.M. contributed to writing, overall validation and project administration; A.R. contributed to resources, writing, and methodology aspects; S.H. contributed to funding acquisition and design; and F.B. performed writing, conceptualization, supervision, and methodology.

Funding statement. This research was supported by grants from the United States Agency for International Development (USAID) and FHI360 (Grant number: 7200AA18CA00046), with management support from HDI.

Competing interest. The authors declare no competing interests exist.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country. Prior to engaging in data collection, all members of the research team underwent approved research training with a strong emphasis

on the protection of privacy and confidentiality. This training equips our team with the knowledge and skills necessary to handle participant data responsibly and ethically. In compliance with Indian law, ethical standards, and best practices, we have obtained approval from our Institutional Review Board (IRB) for this research. The IRB protocol outlines the ethical guidelines and procedures we have followed to protect the rights and well-being of our human subjects. It underscores our commitment to conducting this research with the highest ethical standards, ensuring the safety and privacy of our participants. By collecting all participant information anonymously and adhering to rigorous privacy protection measures, we have minimized any potential ethical or data-related risks associated with this research. The IRB (10030/IRB/22-23) was performed by Sigma-IRB; India (https://www.sigma-india.in/).

References

Batarseh FA, **Freeman L and Huang C-H** (2021) A survey on artificial intelligence assurance. *Journal of Big Data 8*(1), 60. https://doi.org/10.1186/s40537-021-00445-7

Batarseh FA, **Ghassib I**, **Chong DS and Su P-H** (2020) Preventive healthcare policies in the US: Solutions for disease management using big data analytics. *Journal of Big Data 7*(1), 38. https://doi.org/10.1186/s40537-020-00315-8

Bhan N, **McDougal L**, **Singh A**, **Atmavilas Y and Raj A** (2020) Access to women physicians and uptake of reproductive, maternal and child health services in India. *EClinicalMedicine 20*, 100309. https://doi.org/10.1016/j.eclinm.2020.100309

Boukenze B, **Mousannif H and Haqiq A** (2016) Predictive analytics in healthcare system using data mining techniques. In *Computer Science & Information Technology (CS & IT)*, pp. 1–9. https://doi.org/10.5121/csit.2016.60501. https://airccj.org/CSCP/vol6/csit65201.pdf

Cao F, **Liang J and Bai L** (2009) A new initialization method for categorical data clustering. *Expert Systems with Applications 36* (7), 10223–10228. https://doi.org/10.1016/j.eswa.2009.01.060

Chatterjee S, **Majumdar D**, **Misra S and Damaševičius R** (2020) Adoption of mobile applications for teaching-learning process in rural girls' schools in India: An empirical study. *Education and Information Technologies 25*(5), 4057–4076. https://doi.org/10.1007/s10639-020-10168-6

Cherewick M, **Lebu S**, **Su C**, **Richards L**, **Njau PF and Dahl RE** (2021) Promoting gender equity in very young adolescents: Targeting a window of opportunity for social emotional learning and identity development. *BMC Public Health 21*(1), 2299. https://doi.org/10.1186/s12889-021-12278-3

Damaševičius R, **Maskeliūnas R and Blažauskas T** (2023) Serious games and gamification in healthcare: A meta-review. *Information 14*(2), 105. https://doi.org/10.3390/info14020105

Desai S, **Pandey N**, **Singh RJ and Bhasin S** (2021) Gender inequities in treatment-seeking for sexual and reproductive health amongst adolescents: Findings from a cross-sectional survey in India. *SSM - Population Health 14*, 100777. https://doi.org/10.1016/j.ssmph.2021.100777

Efe H and Topsakal ÜU (2022) Digital Games and Health Education: Meta-Synthesis Study. https://doi.org/10.2139/ssrn.4045896

Goldberg LR (1993) The structure of phenotypic personality traits. *American Psychologist 48*(1), 26–34. https://doi.org/10.1037/0003-066X.48.1.26

Huang C-H, **Batarseh FA**, **Boueiz A**, **Kulkarni A**, **Su P-H and Aman J** (2021) Measuring outcomes in healthcare economics using artificial intelligence: With application to resource management. *Data & Policy 3*, e30. https://doi.org/10.1017/dap.2021.29

Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery 2*(3), 283–304. https://doi.org/10.1023/A:1009769707641

Jejeebhoy SJ (1998) Adolescent sexual and reproductive behavior: A review of the evidence from India. *Social Science & Medicine 46*(10), 1275–1290. https://doi.org/10.1016/S0277-9536(97)10056-9

Johnson D, **Deterding S**, **Kuhn K-A**, **Staneva A**, **Stoyanov S and Hides L** (2016) Gamification for health and wellbeing: A systematic review of the literature. *Internet Interventions 6*, 89–106. https://doi.org/10.1016/j.invent.2016.10.002

Kedia S, **Verma R and Mane P** (2022) Sexual and reproductive health of adolescents and young people in India: The missing links during and beyond a pandemic. In Pachauri S and Pachauri A (eds.), *Health Dimensions of COVID-19 in India and beyond*. Singapore: Springer Nature, pp. 203–217. https://doi.org/10.1007/978-981-16-7385-6_10.

Kent JT (1983) Information gain and a general measure of correlation. *Biometrika 70*(1), 163–173. https://doi.org/10.1093/biomet/70.1.163

Konstantinidis EI, **Billis AS**, **Paraskevopoulos IT and Bamidis PD** (2017) The interplay between IoT and serious games towards personalised healthcare. In *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pp. 249–252. https://doi.org/10.1109/VS-GAMES.2017.8056609. https://citeseerx.ist.psu.edu/document?repid=rep1&type&pdf&doi=73a19026fb8a6ef5bf238ff472f31100c33753d0

Kotsiantis S and Kanellopoulos D (2006) Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering 32*, 71–82.

Mayr S, **Schneider S**, **Ledit L**, **Bock S**, **Zahradnicek D and Prochaka S** (2017) Game-based cultural competence training in healthcare. In *2017 IEEE 5th International Conference on Serious Games and Applications for Health (SeGAH)*. Perth, Australia: IEEE, pp. 1–5. https://doi.org/10.1109/SeGAH.2017.7939113

https://doi.org/10.1017/dap.2024.16 Published online by Cambridge University Press

Primack BA, **Carroll MV**, **McNamara M**, **Klem ML**, **King B**, **Rich MO**, **Chan CW and Nayak S** (2012) Role of video games in improving health-related outcomes. *American Journal of Preventive Medicine 42*(6), 630–638. https://doi.org/10.1016/j.amepre.2012.02.023

Raj A, **Quilter I**, **Ashby E**, **Dixit A and Howard S** (2023) Psychographic profiling—A method for developing relatable avatars for a direct-to-consumer mobile game for adolescent girls on mobile in India. *Oxford Open Digital Health 1*, oqad001. https://doi.org/10.1093/oodh/oqad001

Tedgård E, **Råstam M and Wirtberg I** (2018) Struggling with one's own parenting after an upbringing with substance abusing parents. *International Journal of Qualitative Studies on Health and Well-Being 13*(1), 1435100. https://doi.org/10.1080/17482631.2018.1435100

USAID (2023) How a Mobile Game is Normalizing Periods and Creating Access to Products Through Gameplay. Retrieved May 26, 2023, from https://medium.com/usaid-2030/how-a-mobile-game-is-normalizing-periods-and-creatingaccess-to-products-through-gameplay-532afee035e2.

Wang M and Zheng X (2021) Using game-based learning to support learning science: A study with middle school students. *Asia-Pacific Education Researcher 30*(2), 167–176. https://doi.org/10.1007/s40299-020-00523-z

Williamson B (2016) Digital education governance: Data visualization, predictive analytics, and 'real-time' policy instruments [Publisher: Routledge _eprint. *Journal of Education Policy 31*(2), 123–1412. https://doi.org/10.1080/02680939.2015.1035758

Cite this article: Huang C-H, Mohandas N, Raj A, Howard S and Batarseh FA (2024). Data-driven healthcare indicators via precision gaming: With application to India. *Data & Policy*, 6: e22. doi:10.1017/dap.2024.16