Reviewing Recommender Systems in the Medical Domain

Katharina Brunner[1,2], Bernhard Hametner[2*]

[1] Institute of Analysis and Scientific Computing, TU Wien, Wiedner Hauptstraße 8-10
[2] AIT Austrian Institute of Technology, Center for Health and Bioresources, Medical Signal Analysis, Vienna, Austria.
[*]*bernhard.hametner@ait.ac.at*

SNE 32(4), 2022, 203-209, DOI: 10.11128/sne.32.tn.10624
Received: 2022-04-08; Revised:2022-06-26
Accepted: 2022-07-15
SNE - Simulation Notes Europe, ARGESIM Publisher Vienna
ISSN Print 2305-9974, Online 2306-0271, www.sne-journal.org

Abstract. Medical recommender systems are increasing in popularity within the digital health sector. Two main principles for personalised support are just-in-time interventions, and adaptiveness of treatment. Intervention concepts using these principals are called JITAIs, and they aid clients in self-management for health-related issues. In this contribution, the JITAI framework is introduced, and its advantages for recommender systems are discussed. Mathematically, the JITAI concept can be interpreted as a contextual or regular multi-armed bandit problem, which is solved via a bandit algorithm. After discussing several algorithmic strategies of bandit algorithms and elaborating on their differences, the Thompson Sampling strategy is identified as a practical solution for real-life applications using the JTIAI framework. Subsequently, existing recommender systems based on the (contextual) multi-armed bandit approach are reviewed, and the disruption of the algorithm's learning process by instances of missing data is found to be a prevalent obstacle. An algorithm called Thompson Sampling with Restricted Context is put forward as a solution, where missing data is processed within the bandit setting.

Introduction

Digital health practices are expected to revolutionise the public health sector on a global scale and provide healthcare solutions to all people regardless of geographic location and social strata. In digital health, data is processed with the help of smart and connective devices using advanced computing and artificial intelligence, including machine learning as well as other data science strategies [1].

A research domain within this framework is the development of personalised recommender systems, with the aim to support the self-management of chronic illnesses, or the facilitation of building habits towards a more healthy lifestyle. Generally, two principles are combined: just-in-time assistance, and adaptiveness. Recommender systems based on this approach are called just-in-time adaptive interventions (JITAIs) [2]. Some JITAIs employ machine learning techniques to identify the best supportive intervention for the client out of a pool of possible options, and deliver it at a time when the client is most receptive, or has the highest need for it. Mathematically, this problem can be interpreted as a contextual multi-armed bandit problem, which has its origins in game theory, and aims to find the option that yields the highest reward under given circumstances [3].

This contribution introduces the most common approaches (i.e., algorithms) to solve the problem and provides an overview of the state-of-the-art mobile health applications that operate using the (contextual) multi-armed bandit approach.

1 The JITAI Framework

A pragmatic framework is provided by Nahum-Shani et al. [6], which can help developers in constructing JITAI intervention concepts, or may inform the design of a JITAI model. JITAIs are multi-component interventions, and consist of five key elements: *decision points*, *tailoring variables*, *intervention options*, *decision rules*, and *proximal outcomes*. At a decision point (i.e., a point in time at which an intervention decision is made), the decision rules determine which intervention option out of an array of possible candidates is best to facilitate a proximal outcome (i.e., the short-term goal the intervention is trying to achieve), based on the contextual data concerning the client, which is stored in tailoring variables.

Together, these components form an intervention concept, see Figure 1.

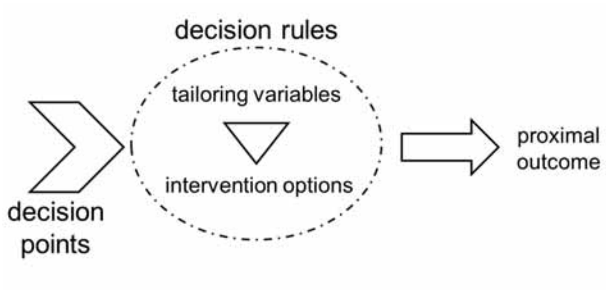

Figure 1: Intervention concept for the JITAI design, adapted from [2].

This definition of the JITAI framework allows for a variety of applicative options within each component. For example, decision points may occur at pre-specified time intervals (e.g., every three minutes), at specific times of day (e.g., daily at 9 a.m.), or following random prompts, depending on how frequent meaningful changes in the tailoring variables are expected to take place.

In mathematics and many mathematics-adjacent fields of application, the JITAI framework can be constructed as a contextual or regular multi-armed bandit problem (see below), due to both concepts being translatable in a natural way.

2 The Multi-Armed Bandit Problem

The multi-armed bandit (MAB) problem is perhaps the simplest model for sequential decision making where the aim is to maximise the cumulative sum of rewards over a certain time horizon [4].

The original setup describing the MAB problem shows a player who is faced with k slot machines (colloquially known as one-armed bandits) [5]. Identically, one can imagine a single slot machine with several arms, thus obtaining the term *multi-armed bandit*. The player aims to maximise the cumulative reward from playing the machines over a certain number of plays, which holds the intrinsic dilemma of the MAB problem: the exploration-exploitation trade-off. The player needs to balance between trying out all arms sufficiently often to discover which ones are most lucrative (exploration), while concurrently playing the arms that they have found to yield the highest rewards (exploitation).

General approaches of the MAB problem call the playing entity the *agent*, and the process of systematically playing towards a specific goal is done by a MAB algorithm. The contextual MAB (CMAB) setting is an extension of the MAB problem where the player views additional information about the current situation before deciding which arm to play, thereby avoiding unnecessary exploration, and is guided towards the arms that need to be explored. Therefore, CMABs lend themselves to applications in the medical setting, where any recommender system should base their intervention decisions on the observed health data. In general, CMAB algorithms are derived from MAB algorithms, so the algorithmic strategies of the MAB problem are a good entry point into applied bandit algorithms.

2.1 Bandit Algorithms

A possible option for MABs is to consider the stochastic bandit setting, where the agent chooses an arm at an iteration point t and subsequently receives a reward drawn from an arm-specific distribution unknown to the agent. The agent then improves its arm selection strategy based on the observation, with the goal to estimate parameters that describe the distributions linked to the arms, see Algorithm 1 [7]. Then, the agent will exploit the arm that is estimated to yield the highest reward. Usually, fixed distributions are assumed for all arms, and the performance of the algorithm can be quantified by observing how quickly the optimal arm is identified through the obtained reward, or, alternatively, the concept of regret [4].

Algorithm 1: Multi-Armed Bandit Algorithm

```
Input: 𝒜 = {A_1, ..., A_k},  ℛ_1, ..., ℛ_k
for t = 1,2,... do
    choose arm a_t ∈ 𝒜
    receive reward r_t ~ ℛ_i(t)
    improve arm-selection strategy with new
      observation (a_t, r_t)
end
```

Since the true reward distributions are unknown to the agent, MAB algorithms estimate the expected reward [8], and their overall performance can be evaluated by regarding the cumulative sum of rewards over all iterations. Instead of observing how well the algorithm has done, one can also investigate how often the algorithm has missed out on the optimal arm.

At iteration t, the optimal arm is defined as the arm with the (currently) highest expected reward estimate, and the regret at t is the difference between the optimal choice and the arm that the MAB algorithm has actually chosen.

As an example for a bandit setting, consider the Bernoulli bandit problem (a special case of the MAB problem), where all k reward distributions are Bernoulli distributions. The reward $r_{i(t)}$ the agent receives when choosing arm $a_t \in \mathscr{A}$ at time t is:

$$r_{i(t)} = \begin{cases} 1 & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i \end{cases}, \quad p_i \in (0,1)$$

The parameters p_i, $i \in \{1,\dots,k\}$ are unknown to the agent [9]. In case of a Bernoulli bandit, the expected reward for chosen action a_t at t is the success probability for the arm:

$$\mathbb{E}\left[r_{i(t)}\right] = p_{a_t}$$

After choosing an arm, the agent either succeeds ($r_{i(t)} = 1$) with probability p_i, or fails ($r_{i(t)} = 0$) with probability $(1 - p_i)$. Since the expected reward of the optimal arm is $r_{a_{opt}} = \max_i(p_i)$, the regret $R_{a(t)}$ of choosing a suboptimal arm at time t is denoted by:

$$R_{a(t)} = \max_i(p_i) - p_{a(t)}$$

Different simulation runs of the Bernoulli bandit can be rated in their efficiency: the lower the cumulative regret, the more efficient the algorithm is in finding and exploiting optimal options, thus exemplifying how regret can be practical when comparing different algorithmic strategies. Even though the Bernoulli bandit gives the most basic bandit setting, it can be found in many practical applications, and more complex MAB problems may be simplified by assuming a Bernoulli distribution instead of a more advanced one [8].

2.2 Algorithmic Strategies

The main obstacles for solving the Bernoulli bandit problem are the unknown success probabilities, because the expected regret cannot be computed. However, when determining algorithmic performance in a simulation environment, the success probabilities are known, even though they are assumed not to be, in order to assess the quality of different algorithmic strategies. For a good strategy, the regret rapidly decreases to zero, thus the best option is identified quickly, and subsequently exploited. In contrast, if the value of regret decreases slowly or never reaches zero, the strategy is considered poor. Figure 2 shows a sketch of possible regret curves.

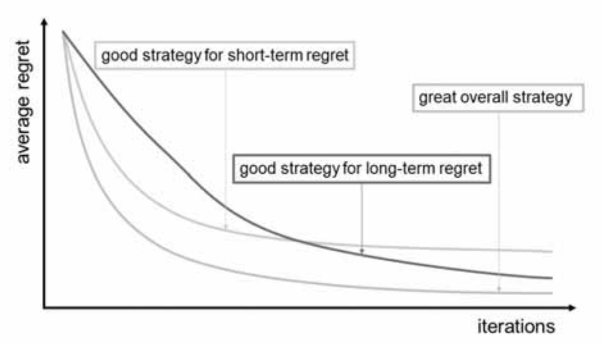

Figure 2: Possible regret curves for different MAB algorithm strategies.

Different strategies are distinguished by how the exploration-exploitation trade-off is handled. Three potential strategies are: the *e-Greedy strategy*, the *Upper Confidence Bound* (UCB) *strategy*, and the *Thompson Sampling* (TS) *strategy*.

The e-Greedy Strategy. The e-Greedy strategy explicitly trades off between exploration and exploitation, by using the exploration parameter e. A greedy strategy refers to exploitation without exploring, i.e., choosing the arm with the highest current reward estimate. This bears the risk of missing the optimal arm forever, see the following example:

Let $\mathscr{A} = \{A_1, A_2, A_3\}$ be a Bernoulli bandit with three arms, and let the (true) success probabilities be:

$$p_1 = 0.3, \quad p_2 = 0.7, \quad p_3 = 0.8$$

Furthermore, let the initial success estimates be equal for all three arms:

$$e_1^0 = 0.5, \quad e_2^0 = 0.5, \quad e_3^0 = 0.5$$

Here, arm A_1 is overestimated, whereas arms A_2 and A_3 are underestimated. Since all arms have equally high success estimates, the agent picks one arm at random. Let A_2 be the agent's choice, and let $r_1 = 1$. The estimates are updated to:

$$e_1^1 = 0.5, \quad e_2^1 = 0.75, \quad e_3^1 = 0.5$$

Following a greedy strategy, the agent picks A_2 again, and A_3 (the optimal arm) will only be explored if the estimate for A_2 drops down to 0.5, which is not likely to happen, due to the true success probability being 0.7.

This problem is solved by introducing an exploration parameter $e \in (0,1)$, which sets the probability of performing an exploration step at t, wherein one arm is chosen randomly. However, the risk remains that the exploration continues after having identified the optimal arm, since the algorithm forces the agent to select a (known) suboptimal arm during each exploration step, thus the regret will never converge to zero.

An alternative is presented by the decaying e-Greedy strategy, where e is not fixed, but decays over time. However, an accurate value for the decay is difficult to determine.

The Upper Confidence Bound Strategy. The UCB strategy deals with the exploration-exploitation trade-off in an implicit way. In the previous strategy, the agent's knowledge at time t is modelled as a point estimate, which does not reflect the uncertainty regarding this value. In contrast, the UCB strategy explicitly models the knowledge uncertainty as confidence intervals, where both the current knowledge (i.e., the mean) and the related uncertainty (i.e., the width of the confidence interval) are used to guide the arm selection process.

In case of the Bernoulli bandit example, the probability estimates e_1^t, e_2^t, e_3^t, are substituted by UCBs for each arm. This principle is called "optimism in front of uncertainty": the uncertainty about the expected reward is expressed as a confidence interval, and the expected reward is estimated optimistically as the upper bound of that confidence interval.

Thus, there are two reasons why the UCB is high: the arm has not yet been explored, resulting in much uncertainty about the success probability, or the arm has been found to be a good choice, thus there is little uncertainty about the (high) success probability. This way, the agent keeps exploring arms that have not yet been proven to yield low rewards instead of arms that produce low rewards with high certainty [10].

There are different ways to derive UCBs, for example via the Hoeffding equation, or the Bayes theorem. However, computing UCBs can be difficult, depending on the assumed distributions.

The Thompson Sampling Strategy. The TS strategy works similarly to the previous strategy, but an agent following this strategy picks an arm randomly, according to its probability to be the best. The Bayesian update rule, which is derived from the Bayes theorem, lies at the heart of the TS strategy. A general formulation of the Bayesian update rule is

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

In practice, it means that, once an arm is chosen, the estimates representing the distribution of the arm are updated with the help of the reward observation. The conjugate property of prior-likelihood combinations plays an important part when updating distributions. The use of the Bayesian update rule is only recommended if the updated distribution (i.e., the posterior) is easily calculated, which is the case for conjugate prior-likelihood combinations. In a Bernoulli bandit, the conjugate combination is given by the Beta distribution.

TS is best explained when investigating the Beta-Bernoulli prior-likelihood combination. The Bernoulli distribution depends on the success parameter p, which needs to be estimated. The Beta distribution as the prior (and posterior) represents the uncertainty about p. Its parameters α and β correspond to successful and failed draws: if $r_t = 1$, α is upped by 1, and if $r_t = 0$, β is upped by 1, according the Bayesian update rule. The density function of the Beta distribution changes with each reward observation whenever the arm is chosen, and congregates around the estimated success parameter $e^t \sim p$. Instead of calculating statistical quantities concerning p, it is sufficient to draw a random variable from the Beta distributions at t, and the arm with the highest sample value is chosen by the agent, see Algorithm 2.

Empirical evaluation has shown that TS algorithms are more robust against delayed, or batched, feedback in applications for advertising and news article recommendations modelled as a CMAB problem [11], and that it has lower regret in the long run compared to UCB algorithms [12]. Even though the theoretical understanding of TS is still limited, optimal regret bounds on the expected regret exist for the MAB problem with Bernoulli distributions [8], and theoretical guarantees are provided for a TS algorithm equipped to solve the CMAB problem [13].

Algorithm 2: Thompson Sampling strategy

Input: $\mathscr{A} = \{1,\ldots,k\}$, initial parameters α_i, β_i, auxiliaries $S_i = F_i = 0$
for $t = 1,2,\ldots$ **do**
  for $i = 1,\ldots,k$ **do**
    Draw θ_i according to Beta$(\alpha_i + S_i, \beta_i + F_i)$
  end
  choose arm $a_t = j = \text{argmax}_i\,\theta_i$
  receive reward $r_t \sim \text{Bernoulli}(e_j)$
  if $r_t = 1$ **then**
    $S_j = S_j + 1$
  else
    $F_j = F_j + 1$
  end
end

Furthermore, the applicability of the TS strategy to any conjugate prior-likelihood combination facilitates the increase of model complexity beyond Bernoulli distributions without increasing computational complexity [9].

3 JITAIs and CMABs

The CMAB problem setup provides a natural model for developing digital health interventions of the JITAI design. In the previous section, MAB algorithms are discussed as a way of solving the MAB problem. From this, three main elements in a MAB algorithm can be derived: points in time (also *trials*), a set of arms, and respective reward distributions. The contextual bandit setting adds one more element: a context vector $x_t \in \mathbb{R}^d$, which holds the additional information that the agent views before selecting an arm during each trial. Depending on the chosen method of implementation, exploration and exploitation are balanced in order to minimise the cumulative expected regret.

Section 1 introduces the five components of the JITAI design, and explains how they interact. There are similarities in the formulation of both concepts, making the translation easy: decision points denote the trials, tailoring variables represent any contextual information in the form of a vector, and the possible intervention options serve as the arms of a bandit. Reaching the proximal outcome is equal to minimising the regret at t, or, in case of a Bernoulli bandit, succeeding at a trial [14]. The decision rules can be viewed as a mapping between the current values of the tailoring variables and the intervention options. This mapping is done by the bandit algorithm. Table 1 summarises the analogies.

JITAI framework	CMAB concept
Decision points	Trials
Tailoring variables	Context vector
Intervention options	Arms
Proximal utcome	Rewards
Decision rules	Bandit algorithm

Table 1: Representation of the analogy between the elements of the CMAB approach and the components of the JITAI design.

Due to its setup and adaptability, the JITAI design is favoured for personalised recommender systems. Since CMABs are most convenient for algorithmic implementation of JITAIs, by transitivity, recommender systems are best realised via implementing and solving a CMAB problem.

4 Medical Applications using (C)MABs

It is already common practice to use (C)MAB algorithms for researching personalised adaptive interventions in digital health. In the past decade significant progress has been made in creating functional applications that work in a (C)MAB setting, adapting to a client's intervention preferences in real time, as part of the trend towards reinforcement learning methods.

Medical recommender systems are multi-faceted, and the algorithm for intervention decisions is only one cog in a delicate machine. Mechanisms must be in place so that the decision rules can still function in case of missing, or erroneous, data, without compromising the algorithm's learning process. Clients must be kept engaged beyond an initial novelty period, so behavioural psychology plays an important part in the delivery of intervention suggestions to ensure that intervention engagement prevails over intervention fatigue [2]. Recommender systems intended for client use address these issues in different ways.

HeartSteps. HeartSteps is a mobile phone application currently available for download in the United States. Originally tested during at trial for improving physical activities of individuals with blood pressure in the stage 1 hypertension range, it delivers activity suggestions to encourage walking while monitoring the client's daily step count with a Fitbit tracker.

The intervention decisions are made via a CMAB algorithm that uses TS and the application is designed to include the delayed effect of treatment. However, the chosen algorithm cannot deal with missing data within the decision rules. Instead, a lack of data is compensated outside of the bandit algorithm [15].

MyBehavior. MyBehavior is a mobile phone application that delivers personalised interventions for promoting physical activity and dietary health as a JITAI, via a MAB algorithm. Phone sensory data is used to design unique recommendations for a client, with the goal to find activity suggestions that maximise the chance of daily calorie burns. The application records data every minute, and issues an activity suggestion once each morning. It then analyses the location tagged activities to find patterns that are representative of the client's behaviour. Additionally, MyBehavior allows clients to self-report exercise and food intake, which is backed by a crowd-sourcing database. Like HeartSteps, the decision rules cannot compensate for instances of missing data [16].

PopTherapy. PopTherapy is a mobile phone application that helps clients cope with stress and depression-related symptoms based on cognitive behavioural theory technology. Intervention suggestions are issued after a request is prompted by the client, and the goal is to maximise stress reduction. The application uses a CMAB algorithm combined with the UCB approach to select an intervention from a series of stress management strategies. However, the bandit algorithm requires knowledge of the correct model for the reward function [17] and, as is intrinsic to the UCB strategy, bases its arm selection process deterministically on historical data [15].

5 Conclusion and Outlook

Digital health is currently at the forefront of biomedical research, with recommender systems promising easier access to treatment for a variety of chronic illnesses, mental health challenges, and general life improvements. Intervention concepts that combine just-in-time support and adaptiveness of treatment (JITAIs) are aiming to provide personalised support to clients at points in time when they need it most, or are most receptive to it, based on the processing of health data.

As can be seen from existing recommender systems, the mathematical concept of (C)MABs is a convenient way of implementing JITAIs as real-life applications.

Different strategies for (C)MAB algorithms are available, and a thorough comparison of the most common approaches shows that the TS strategy stands out in terms of arm selection performance, adaptivity, and computational effort. Theoretical knowledge about TS is still sparse. However, in setups where regret is bounded, the algorithm is guaranteed to identify arms that are close to optimal options eventually. Thus, when investigating algorithmic strategies for JITAI-backed recommender systems, consideration should be given to TS above others, when applicable to the research question.

Furthermore, consideration should also be given to an extension of TS called *Thompson Sampling with Restricted Context* by Bouneffouf et al. [18], where the agent only observes limited context (i.e., a restricted context vector) at the cost of a slight decrease in performance. This setting offers a natural way of dealing with instances of missing data by simply disregarding it, and the lack of data can be addressed within the bandit setting while the learning process of the bandit algorithm remains mostly uncompromised.

Acknowledgement

Katharina Brunner is supported by the program 'Talente' by the Federal Ministry of the Republic of Austria for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK).

Bernhard Hametner is supported by a grant of the government of Lower Austria and the European Regional Development Fund (ERDF), contract No. WST3-F-5030665/010-2021

References

[1] WHO, *Global strategy on digital health 2020-2025*. ISBN: 78-92-4-002092-4, License: CC BY-NC-SA 3.0 IGO

[2] Nahum-Shani I, Smith S, Spring B, Collins L, Witkiewitz K, Tewari A, Murphy S. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. In: *Ann. Beh. Med.*, 52(6):446-462, 2018. doi: 10.1007/s12160-016-9830-8.

[3] van Otterlo M, Wiering M. Reinforcement learning and markov decision processes. In: *Reinforcement Learning: State-of-the-Art*. 2012. p 3-42. doi: 10.1007/978-3-642-27645-3_1.

[4] Tewari A, Murphy S. From Ads to Interventions: Contextual Bandits in Mobile Health. In: *Mobile Health: Sensors, Analytic Methods, and Applications*. 2017. p 495-517. doi: 10.1007/978-3-319-51394-2_25

[5] Auer P, Cesa-Bianchi N, Fischer P. Finite-time Analysis of the Multiarmed Bandit Problem. In: *Mach. Learn.*, 47:235-256, 2002. doi: 10.1023/A:1013689704352

[6] Nahum-Shani I, Heckler EB, Sprujitz-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. In: *Health Psychol.*, 34(0):1209-1219, 2015. doi: 10.1037/hea0000306

[7] Li L, Chu W, Langford J, Schapire RE. A contextual-bandit approach to personalized news article recommendation. In: *Proceedings of the 19th International Conference on World Wide Web*; 2010 Apr; Raleigh. 661–670. doi: 10.1145/1772690.1772758

[8] Agrawal S, Golya N. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In: *Proceedings of the 25th Annual Conference on Learning Theory*; 2012 Jun; Edinburgh. JMLR. 39.1-39.26.

[9] Rocca J, Rocca B. *The exploration-exploitation trade-off: intuitions and strategies*. `https://towardsdatascience.com/the-exploration-exploitation-dilemma-f5622fbe1e82`, accessed on 2021-07-16

[10] Auer P, Using confidence bounds for exploitation-exploration rade-offs. In: *JMLR*, 3:397-422, 2002.

[11] Chapelle O, Li L. An empirical evaluation of Thompson Sampling. In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*; 2011 Dec; Granada. 2249-2257.

[12] Kaufmann E, Cappe O, Garivier A. On bayesian upper confidence bounds for bandit problems. In Lawrence ND, Girolami M, editors. *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*; 2012 La Palma. 22:592-600.

[13] Agrawal S, Goyal N. Thompson sampling for contextual bandits with linear pay-offs. In Dasgupta S, McAllester D, editors. *Proceedings of the 30th International Conference on Machine Learning*; 2013 Jun; Atlanta. 28:127-135.

[14] Rabbi M, Klasnja P, Choudhury T, Tewari A, Murphy S. Optimizing mHealth interventions with a bandit. In Baumeister H, Montag C, editors. *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*. Cham: Springer; 2019. 277-291. doi: 10.1007/978-3-030-31620-4.

[15] Liao P, Greenewald K, Klasnja P, Murphy S. Personalized HeartSteps: A reinforcement learning algorithm for optimizing physical activity. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*; 2020 Mar. 4(1):1-22. doi: 10.1145/3381007.

[16] Rabbi M, Aung MH, Zhang M, Choudhury T. MyBehavior: automatic personalized health feedback form user behaviors and preferences using smartphones. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*; 2015 Sep. 707-718. doi: 10.1145/2750858.2805840.

[17] Paredes P, Gilad-Bachran R, Czerwinski M, Roseway A, Rowan K, Hernandez J. PopTherapy: Coping with stress in pop culture. In: *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*; 2014 May; Oldenburg. 109-117. doi: 10.4108/icst.pervasivehealth.2014.255070.

[18] Bouneffouf D, Rish I, Cecchi G, Féraud. Context attentive bandits: Contextual bandit with restricted context. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*; 2017 Aug; Melbourne. 1468–1475. doi: 10.24963/ijcai.2017/203