Unleashing the Potential of Reinforcement Learning for Personalizing Behavioral Transformations with Digital Therapeutics: A Systematic Literature Review

Thure Georg Weimann[a] and Carola Gißke

Research Group Digital Health, TUD Dresden University of Technology, Dresden, Germany

Keywords: Digital Therapeutics, Reinforcement Learning, Personalized Medicine, mHealth, Health Behavior Change.

Abstract: Digital Therapeutics (DTx) are typically considered as patient-facing software applications delivering behavior change interventions to treat non-communicable diseases (e.g., cardiovascular diseases, obesity, diabetes). In recent years, they have successfully developed into a new pillar of care. A central promise of DTx is the idea of personalizing medical interventions to the needs and characteristics of the patient. The present literature review sheds light on using reinforcement learning, a subarea of machine learning, for personalizing DTx-delivered care pathways via self-learning software agents. Based on the analysis of 36 studies, the paper reviews the state of the art regarding the used algorithms, the objects of personalization, evaluation methods, and metrics. In sum, the results highlight the potential and could already demonstrate the medical efficacy. Implications for practice and future research are derived and discussed in order to bring self-learning DTx applications one step closer to everyday care.

1 INTRODUCTION

Digital Therapeutics (DTx) promise the delivery of personalized therapies independent of place or time directly to the patient (Fürstenau et al., 2023; Nahum-Shani et al., 2023). Internally, most DTx applications provide behavior change interventions via dedicated techniques (e.g., self-monitoring, feedback, goal-setting) like a virtual coach (Lee et al., 2023; Weimann et al., 2022). Therefore, DTx are, especially for behavior-modifiable diseases such as obesity, diabetes, or substance use disorders, an emerging therapy complement or alternative (Hong et al., 2021).

However, one of the grand challenges is to technically realize the idea of ”personalized medicine” within DTx. Following the definition proposed by the European Union (2015), personalized medicine generally refers to “tailoring the right therapeutic strategy for the right person at the right time,...”. Prior research on digital health interventions revealed that personalization is not only vital for the intervention outcome but also system use adherence (Kankanhalli et al., 2021; Wei et al., 2020). For example, when the patient becomes disengaged and stops using the system because it does not address personal needs, the system cannot have any further impact at all.

When implementing personalized digital health interventions, developers broadly face two options: handcrafted rules and data-driven machine learning/artificial intelligence (AI) (Nahum-Shani et al., 2018). Rule-based approaches are particularly suitable when there is enough explicit knowledge available on how to adapt the system for the user based on the values of contextual variables. However, static rules are limited by the given expert knowledge on the individual, and the system is not able to adapt the rules by itself (i.e., self-learning). For instance, these rule-based systems do not take into account that the preferences of the user might even change unforeseeably over time. In contrast, algorithms from the area of reinforcement learning (RL) do not necessarily rely on knowledge about the user beforehand and are able to capture dynamic changes over the course of the intervention.

Generally speaking, RL algorithms learn to make decisions from a continuous interaction loop with the environment, to which the patient also belongs (Sutton and Barto, 2018). Consequently, the system gets to know the user better and better over the course of interaction. This enables the system to learn which interventions led to the intended effects (i.e., were

[a] https://orcid.org/0000-0002-2762-6121

Weimann, T. and Gißke, C.
Unleashing the Potential of Reinforcement Learning for Personalizing Behavioral Transformations with Digital Therapeutics: A Systematic Literature Review.
DOI: 10.5220/0012474700003657
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2024) - Volume 2*, pages 230-245
ISBN: 978-989-758-688-0; ISSN: 2184-4305
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

beneficial for the user) and repeat them while leaving room for changes over the therapy course. Compared to the other machine learning paradigms (supervised and unsupervised learning), RL algorithms learn similarly to humans by "trial and error" and, therefore, do not necessarily rely on existing training data (Janiesch et al., 2021). Thus, RL goes beyond and provides a powerful framework for building intelligent software agents delivering personalized health behavior change interventions.

Previous literature reviews focused on the use of RL for personalization in general (across domains) (den Hengst et al., 2020), machine learning in personalized health systems (also for behavior change) (Oyebode et al., 2023) or the use of RL in the entire health domain (e.g., also medical imaging, diagnostics) (Coronato et al., 2020). However, the most recent published review by Oyebode et al. (2023) identified a mere of three papers which applied RL. This highlights the need for a more targeted literature search and analysis to capture the full spectrum of RL in the context of health behavior change applications. To the best of our knowledge, no paper provided a detailed "zoom-in" view on the state of the art here yet. The present paper aims to address this gap by answering the following research questions:

- **RQ1:** *Which RL algorithms have been applied?*
- **RQ2:** *What do the algorithms exactly personalize regarding the specific health scenario?*
- **RQ3:** *How was the efficacy, efficiency, and clinical safety of the algorithms evaluated?*

The remainder of this paper is structured as follows. Section 2 presents the theoretical background of the RL problem and available classes of learning algorithms. Afterward, the research method is described. The results of the review are presented in section 4, and trends, implications for practice, and research are discussed in section 5.

2 BACKGROUND

The reinforcement learning problem is characterized by an intelligent agent (i.e., the software application) that interacts with the environment (i.e., the patient) in a continuous feedback loop (Sutton and Barto, 2018). The intelligent agent chooses an action drawn from a set of actions (e.g., health recommendations) that might change the state of the environment observed by the agent (e.g., therapy-relevant parameters change). Afterward, the agent receives a reward produced from the environment that numerically expresses how successful the action has been regarding a particular goal (e.g., increasing the daily activity of the patient). The mathematics behind RL is grounded in Markov decision theory consisting of states, actions, a reinforcement signal (reward), and transition function (Kaelbling et al., 1996). Thus, the task of the agent is to learn an optimal policy to choose an action (based on the current state) that maximizes the reward received from the environment. However, in order to maximize the reward, the agent needs to explore different actions and their effects to be able to exploit "the best" (optimal) discovered action. To address this exploration-exploitation tradeoff, several strategies have been proposed (Lattimore and Szepesvári, 2020).

For example, the so-called epsilon-greedy strategy chooses the best action sometimes (with a probability of ε) and explores randomly otherwise (probability of 1-ε). In contrast, more advanced strategies are Upper Confidence Bound (UCB) or Thompson-Sampling (White, 2013). UCB decides on the action with the highest potential for maximizing the reward by calculating the sum of the current action value and a measure of uncertainty. The uncertainty measure decreases as the agent gains more knowledge about the user. Consequently, the balance of exploration and exploitation automatically changes towards exploitation over time instead of a constant probability (e.g., epsilon-greedy). In comparison, Thompson Sampling chooses the action with the highest expected reward by estimating the reward distribution based on Bayesian inference (Agrawal and Goyal, 2012). These strategies are fundamentally rooted in the concept of the bandit problem (Sutton and Barto, 2018).

Multi-armed bandit (MAB) algorithms assume a simplified RL setting. The basic idea of MABs can be explained in analogy to slot machines in casinos (also called "one-armed bandits"). Assuming that there are multiple bandits in a row with different reward distributions, the general goal is to explore and then exploit the bandit that maximizes payoff. Problems where these reward distributions can change over time (e.g., patient preferences change), are referred to as non-stationary (Mintz et al., 2020). MABs can be further classified into algorithms that do not consider the context at all (i.e., stateless, just learn an action-reward relationship) and contextual MABs that do. Contextual MABs (CMABs) rely on building a predictive model (e.g., logistic regression) for estimating the arm's value based on the context (Li et al., 2010). Overall, MABs and CMABs assume a one-step decision horizon and thus only maximize the immediate reward. In contrast, multi-step RL ("full RL") assumes that a decision for an action can also impact

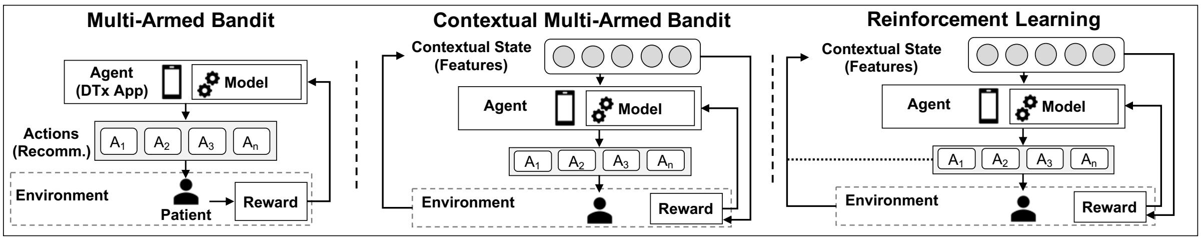


Figure 1: Multi-Armed Bandits (MABs), Contextual-Multi-Armed Bandits (CMABs) and the Full-Reinforcement Learning (RL) problem (based on McInerney et al. (2018) and Kaelbling et al. (1996)).

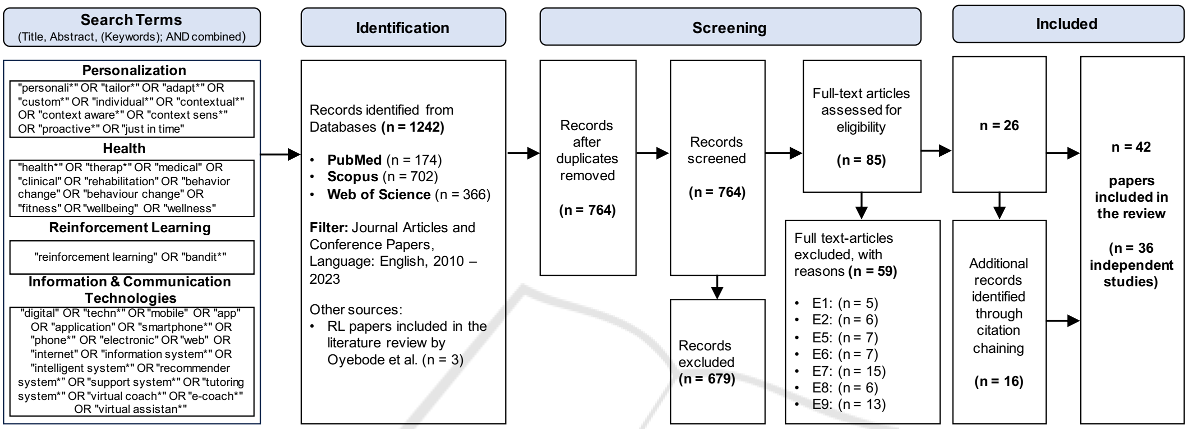

Search Terms
(Title, Abstract, (Keywords); AND combined)
Personalization
"personali*" OR "tailor*" OR "adapt*" OR "custom*" OR "individual*" OR "contextual*" OR "context aware*" OR "context sens*" OR "proactive*" OR "just in time"
Health
"health*" OR "therap*" OR "medical" OR "clinical" OR "rehabilitation" OR "behavior change" OR "behaviour change" OR "fitness" OR "wellbeing" OR "wellness"
Reinforcement Learning
"reinforcement learning" OR "bandit*"
Information & Communication Technologies
"digital" OR "techn*" OR "mobile" OR "app" OR "application" OR "smartphone*" OR "phone*" OR "electronic" OR "web" OR "internet" OR "information system*" OR "intelligent system*" OR "recommender system*" OR "support system*" OR "tutoring system*" OR "virtual coach*" OR "e-coach*" OR "virtual assistan*"
Identification
Records identified from Databases (n = 1242)
• PubMed (n = 174)
• Scopus (n = 702)
• Web of Science (n = 366)
Filter: Journal Articles and Conference Papers, Language: English, 2010 – 2023
Other sources:
• RL papers included in the literature review by Oyebode et al. (n = 3)
Screening
Records after duplicates removed (n = 764)
Records screened (n = 764)
Records excluded (n = 679)
Full-text articles assessed for eligibility (n = 85)
Full text-articles excluded, with reasons (n = 59)
• E1: (n = 5)
• E2: (n = 6)
• E5: (n = 7)
• E6: (n = 7)
• E7: (n = 15)
• E8: (n = 6)
• E9: (n = 13)
Included
n = 26
Additional records identified through citation chaining (n = 16)
n = 42 papers included in the review (n = 36 independent studies)

Figure 2: Flow diagram of the literature search and selection process (based on PRISMA (Moher, 2009)).

the choice of subsequent decisions and influence the state. For example, when the intelligent agent has prompted the patient to go out for a walk in the morning, the same recommendation for a second activity in the afternoon might not maximize the long-term cumulative reward as the patient could become bored (Philipp et al., 2019). Figure 1 illustrates the differences between MABs, CMABs, and the full RL problem.

3 METHOD

For addressing the formulated research questions, a systematic literature review based on the methodology by Webster and Watson (2002) and the PRISMA guidelines (Moher, 2009) was conducted. For the literature search, the databases Scopus, Web of Science, and PubMed were selected. The search query (see Figure 2) was composed of terms describing the concept of personalization, health, reinforcement learning, as well as digital technologies and delivery channels (concatenated with an “AND” operator). Two authors were involved in the overall literature search and analysis process. The literature search was initially conducted in February and was updated at the beginning of November 2023. The scope of the literature search was limited to the fields of title, abstract, and keywords (only Web of Science and Scopus). Only peer-reviewed journal articles and conference papers written in English and published between 2010 and 2023 were included in the review. Overall, the literature search in databases yielded 1242 results. Additionally, three studies were included that employed RL algorithms and were identified in the related work by Oyebode et al. (2023). After duplicate removal, the 764 remaining papers were screened for relevance to the research goals based on title and abstract. After screening for relevance, the full text was screened. Due to the still low maturity of this research topic, we decided also to include study protocols. However, we excluded review articles **(E1)**, articles that did not focus on long-term health behavior change **(E2)**, not report what was exactly personalized using RL **(E3)**, not describe the algorithmic approach **(E4)**, not focus (or plan to focus) on an application for the patient as end-user **(E5)**, not conducted or described an evaluation of the RL approach (e.g., mere system description) **(E6)**, and studies that only used a synthetic data set to investigate the RL algorithm (i.e., no real-world data including (states), actions and rewards) **(E7)**. When related papers described the same application and had the same research goals, we included only the paper(s) with the most comprehensive evaluation in terms of real-world evidence and excluded the others **(E8)**. This criterion aims to ensure

Table 1: Included studies (ST = Stress, PA = Physical Activity, D = Diet, HL = Health Literacy, UE = User engagement, WL = Weight loss, EM = Emotion/Mood, S = Sleep, MA = Medication adherence, QR = Questionnaire response rate, E = Effort, SAR = Substance abuse risk).

ID	Ref.	Opt.
S1	Paredes et al. (2014)	ST
S2	Rabbi et al. (2015)	PA, D
S3	Yom-Tov et al. (2017)	PA
S4	Zhu et al. (2018a)	PA
S5	Zhu et al. (2018b)	PA
S6	Zhu et al. (2018c)	PA
S7	Rabbi et al. (2018)	PA
S8	Zhou et al. (2018a,b)	PA
S9	Delmas et al. (2018)	HL
S10	Gonul et al. (2018)	UE
S11	Pelle et al. (2019, 2020)	UE
S12	Forman et al. (2019)	WL
S13	Mintz et al. (2020)	PA
S14	Liao et al. (2020)	PA
S15	Daskalova et al. (2020)	S
S16	Gasparetti et al. (2020)	WL
S17	Etminani et al. (2021)	MA
S18	Lauffenburger et al. (2021)	MA
S19	Figueroa et al. (2021)	EM
S20	Zhu et al. (2021)	PA
S21	Wang et al. (2021a,b), Sporrel et al. (2022)	PA
S22	Tomkins et al. (2021)	PA
S23	Hu et al. (2021)	QR
S24	Di et al. (2022)	HS
S25	El Hassouni et al. (2022)	UE
S26	Piette et al. (2022)	HS
S27	Ameko et al. (2020), Beltzer et al. (2022)	EM
S28	Aguilera et al. (2020), Figueroa et al. (2022)	PA
S29	Bertsimas et al. (2022)	PA
S30	Albers et al. (2022)	E
S31	Zhou et al. (2023)	UE
S32	Piette et al. (2023)	SAR
S33	Tragos et al. (2023)	UE
S34	Tong et al. (2023)	ST
S35	Gray et al. (2023)	PA
S36	Kinsey et al. (2023)	HS

that our review is based on the most robust and extensive research available. However, related papers providing complementary insights for our research questions were included. Finally, papers that did not fit within the scope of our review were excluded **(E9)**.

In total, n = 26 papers were identified from the database search and satisfied the selection criteria. To identify additional studies, a forward and backward search was conducted based on these papers, resulting in the identification of 16 additional papers. Papers were categorized as part of the same overarching study (denoted as S) when both the medical use case and the algorithm employed remained unchanged. Overall, the final analysis set comprised 42 papers referring to 36 studies (see Table 1). Figure 2 illustrates the overall literature search process.

4 RESULTS

4.1 General Characteristics

When conducting a descriptive analysis regarding the time of publication of the analyzed papers, it becomes evident that there has been a growing interest regarding RL for personalizing health behavior change interventions (see Figure 3).

When taking a look at the study participants based on the origin of the training data, n = 18 studies investigated or plan to investigate the algorithm in the general population (see Table 2). Although DTx generally go beyond primary prevention, the findings from these studies remain significant for their advancement due to the limited number of papers available and the emerging nature of the field. However, several studies also investigated RL algorithms with data from people with hypertension (n = 4), diabetes type 2 (n = 4), anxiety and/or depression (n = 3), overweight/obesity (n = 3), chronic pain (n = 2), asthma, osteoarthritis, opioid-abuse risk or tobacco-smoking behavior (n = 1).

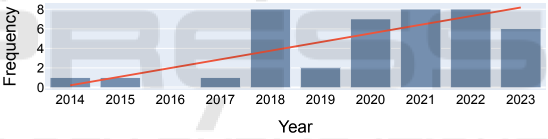


Figure 3: Number of publications by year.

Regarding the delivery platform of the RL-based intervention (see Table 3), a significant amount of studies (n = 15) delivered the intervention via a mobile app. However, several studies (n = 13) also used data collected in prior mHealth studies from electronic medical records (EMRs) or an online weight loss platform and used it to train the RL algorithm in a simulation testbed. Further delivery platforms were web applications (n = 5), SMS (n = 2), phone calls (automatic interactive voice response calls and by health professionals) (n = 2) or mobile messages and phone calls (n = 1).

4.2 Optimization Goal (Reward)

By taking the nature of the reward function variables into account, the optimization goals could be broadly classified as health- and system-usage-related. Both health- and usage-related goals were operationalized *subjectively* (e.g., perceived easiness (S7), motivation (S20), subjective stress assessment (S1)) or through more *objectively* gathered measures (e.g., step count (S4, S5, S6, S22), minutes of activity performance (S3, S7), questionnaire completion (S23)). Interestingly, one study (S12) considered economic aspects in the reward function by taking into account the time a therapist spends on an intervention as an alternative to solely computer-generated messages.

For the present literature analysis, the research was analyzed based on the nature of the immediately underlying (i.e., proximal) constructs that were operationalized in the reward function. For example, even though an app designed to promote physical activity often also leads to weight loss, it was assigned to the ”physical activity” category in this study unless body weight was also operationalized.

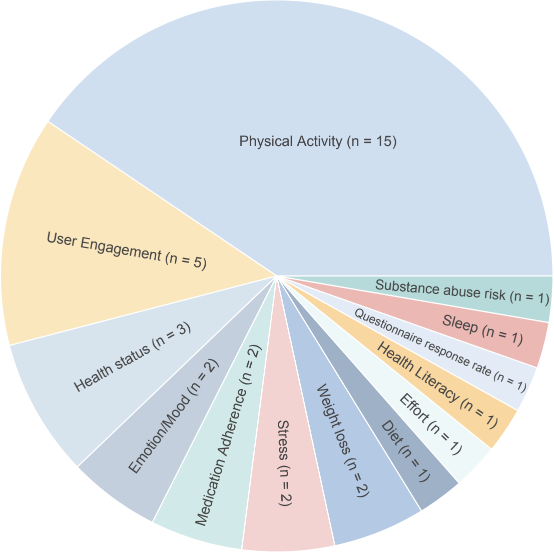


Figure 4: Studies by optimization goal.

As depicted in Figure 4, a significant proportion of the analyzed studies (n = 15) focused on optimizing the physical activity of the patient using corresponding reward measures. It should be noted that the reward of the agent can be also composed of multiple variables (e.g., in an additive or multiplicative manner). For example, in S7, a reward score was calculated by multiplying the perceived easiness (daily survey) and minutes of activity. Likewise, n = 3 studies aimed to optimize the patient’s more holistic health status and therefore combined multiple variables expressing different dimensions (e.g., physical activity, health literacy, pain, HbA1c reduction, quality of life) (S24, S26). Other studies aimed to optimize user

Table 2: Overview of the user data populations.

Population of user data	Studies
General population (n = 18)	S1, S2, S4, S5, S6, S8, S10, S15, S16, S19, S20, S21, S25, S28, S33-S36
Patients with hypertension	S14, S17, S22, S29
Patients with type 2 diabetes	S3, S18, S24, S28
Anxiety and/or depression	S23, S27, S28
Overweight/obese people	S12, S13, S31
Patients with chronic pain	S7, S26
Other	Children with asthma (S9), Patients with osteoarthritis (S11), Tobacco smoking people (S30), Patients with opoid-abuse risk (S32)

Table 3: Overview of the employed RL-based intervention delivery platforms.

Intervention delivery platform	Studies
Mobile App (n = 15)	S1, S2, S7, S8, S10, S11, S15, S17, S18, S21, S22, S25, S28, S33, S36
Simulation testbed with data from mHealth study (n = 11)	S4, S5, S6, S13, S14, S16, S21, S22, S23, S27, S29
Web application	S9, S20, S30, S34, S35
SMS	S3, S19
Phone calls (automated and by human)	S26, S32
Other	Simulation testbed with data from EMR (S24), data from online weight loss platform (S31); Mobile messages and phone calls (S12)

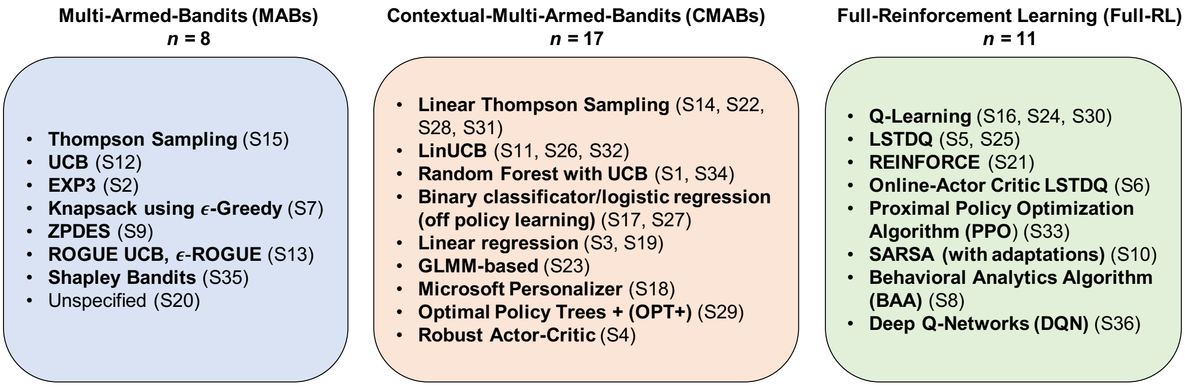

Multi-Armed-Bandits (MABs)
n = 8
Thompson Sampling (S15)
UCB (S12)
EXP3 (S2)
Knapsack using ϵ-Greedy (S7)
ZPDES (S9)
ROGUE UCB, ϵ-ROGUE (S13)
Shapley Bandits (S35)
Unspecified (S20)
Contextual-Multi-Armed-Bandits (CMABs)
n = 17
Linear Thompson Sampling (S14, S22, S28, S31)
LinUCB (S11, S26, S32)
Random Forest with UCB (S1, S34)
Binary classificator/logistic regression (off policy learning) (S17, S27)
Linear regression (S3, S19)
GLMM-based (S23)
Microsoft Personalizer (S18)
Optimal Policy Trees + (OPT+) (S29)
Robust Actor-Critic (S4)
Full-Reinforcement Learning (Full-RL)
n = 11
Q-Learning (S16, S24, S30)
LSTDQ (S5, S25)
REINFORCE (S21)
Online-Actor Critic LSTDQ (S6)
Proximal Policy Optimization Algorithm (PPO) (S33)
SARSA (with adaptations) (S10)
Behavioral Analytics Algorithm (BAA) (S8)
Deep Q-Networks (DQN) (S36)

Figure 5: Overview of the reinforcement learning algorithms used.

engagement (n = 5), emotion/mood (n = 2), weight loss (n = 2), medication adherence (n = 2), stress (n = 2), dietary change, health literacy, substance abuse risk, questionnaire response rate, the user's effort, and sleep (n = 1).

4.3 Used Reinforcement Learning Algorithms

The majority of the studies (n = 25) employed bandit algorithms (esp. CMABs or MABs), while the remaining studies (n = 11) employed algorithms designed to address the complete RL problem (see Figure 5). Overall, the studies varied in their level of detail when describing the algorithm, which made it challenging to classify and directly compare them.

MABs: The arguably most straightforward and simplified approach to address the RL problem are MABs, which have been investigated in n = 8 studies. The study by Rabbi et al. (S2) was one of the pioneering studies in this field and used the EXP3 bandit algorithm. A promising approach regarding the use of RL in chronic disease scenarios was described in a later study by the same authors (S7) and is referred to as "bandits with knapsacks". The general idea of this approach is to set an upper activity bound (e.g., 60 minutes) and select a set of recommendations that meet this constraint in total. According to the study's authors, the reason behind this approach is that the therapeutic goal is not necessarily focused on promoting maximum activity but rather on consistently adhering to a predetermined goal.

Mintz et al. (S13) proposed another sophisticated technique called ROGUE-bandits, utilizing UCB and ε-greedy strategies, specifically designed for non-stationary scenarios. In these settings, it is assumed that rewards can vary over time due to habituation phenomena, such as the patient's changing preferences during the intervention. If a particular intervention option is delivered excessively, the patient's receptiveness to this action may decrease. However, over time, the intervention option can "recover", and the patient becomes receptive again.

In the study S9, a bandit algorithm was applied that originated from the field of pedagogical agents called "Zone of Proximal Development and Empirical Success" (ZPDES). The approach explores an activity graph using a hierarchical structure comprising multiple layers for determining the health educational theme (e.g., "my body", "my tips"), the level of difficulty, and the type of delivery (e.g., game, quiz).

A promising approach from an ethical point of view was recently described in S35 and is called "Shapley Bandits". The algorithm implements fairness constraints to optimize the group performance in an exergame that presents social comparison rankings to the user. The remaining studies used Thompson Sampling or UCB. It is noteworthy that the reported exploration rates (ε) vary, ranging from 1% (due to an upfront forced exploration, see S35), 10% (S2) up to 20% (S7).

CMABs: Almost half of the analyzed studies (n = 17) used contextual bandits. Among these papers, Linear Thompson sampling was commonly employed (n = 4). In addition, several papers utilized linear regression with Boltzmann-Sampling (S3, S19), a random-forest regression algorithm with UCB (S1, S34), while other studies adapted generalized linear mixed models (S23) or trained a binary classifier/logistic regression in an offline-fashion with historic data and employed the doubly robust estimator (S17, S27). One noteworthy approach described in S18 aimed to leverage "Microsoft Personalizer", a no-code configurable CMAB, for their study. Another distinctive approach was proposed in S4, where an Actor-Critic bandit was investigated, which consists of two components (actor and critic). While the actor learns the actual policy and thus determines the actions based on the context, the critic component assesses the policy's value to drive its update.

For constructing a CMAB, one may train one model for each arm (i.e., action) as proposed in the commonly cited LinUCB algorithm (Li et al., 2010). In contrast, Yom-Tov et al. (S3) considered a concatenated feature vector (Kesler construction) encompassing the context and the actions for predicting the change in activity for each action with only one model.

Full-RL: A comparatively small amount of studies (n = 11) used algorithms designed to address full RL problems. Among these studies, several employed Q-learning, which is widely recognized as one of the most well-known and traditional RL approaches (Watkins and Dayan, 1992). The study S16 used the simplest form with Q-tables, while S24 employed Gradient Boosted Regression Trees to estimate the Q-function. In the study S25, the LSTDQ algorithm was used, which can be considered a special type of Q-learning. Building on LSTDQ, the study S6 proposed an online Actor-Critic algorithm .

The studies by Wang and colleagues (S21) used the "REINFORCE" algorithm that directly estimates the policy (i.e., policy-based). Q-learning algorithms, in contrast, estimate the value of a state-action pair and thus estimate the policy indirectly (i.e., value-based) (Watkins and Dayan, 1992). Notably, the proposed approach by Wang et al. incorporated a constraint in the algorithm on the maximum reminders per week to avoid intervention fatigue on the patient's side by sending too many reminders.

Another distinctive approach was described in S8, where the so-called "Behavioral Analytics Algorithm" was proposed combining "inverse RL" and traditional RL. Inverse RL is used to derive parameters of the patient which are the user's baseline level of steps, the user's responsiveness to the step goal, and the user's self-efficacy. Afterward, the estimated parameters are used to conduct a direct policy search by solving a mixed integer linear program to get an optimal step goal (RL step).

Remarkably, one study used deep neural networks with RL ("deep RL") and extended the DQN algorithm (S36). Other algorithms reported in the literature were proximal policy optimization (PPO) (S33) and SARSA (S10).

Personalization Approaches: The described algorithms have been used with data collected at different levels to enable personalization. Broadly speaking, three approaches are discussed in the literature that are a *"one model for all"* or *"pooled"* data approach (e.g., S1, S3, S18, S19, S24, S26), a *"group-driven"* or *"clustered"* approach (e.g., S5, S6, S25) and a *"separate"* or *"N-of 1"* approach (e.g., S2, S7, S17). However, not all papers have explicitly described on which level they used the data, leaving room for interpretation and making categorization and quantitative evaluation difficult. In general, all three approaches have benefits and drawbacks. While training one model for all patients by making use of their entire pooled data accelerates the learning process, the personalization may be too broad and not precisely enough for the individual. The underlying assumption of this approach is that the "best action" is unknown, but users tend to be homogeneous in their responses. On the other hand, building a separate model for each user addresses their heterogeneity but significantly slows down the learning process due to the sparsity and noisiness of the data. Nonetheless, the ability to build a model for the user's local device (i.e., decentralized) can have advantages in terms of privacy (Rabbi et al., 2015).

Balancing Between Pooling and Separation: Recent literature proposed sophisticated approaches to handle these two extremes. Therefore, RL is frequently combined with unsupervised learning approaches. A comparatively straightforward approach is described in the studies S5 and S25 where the k-means clustering algorithm was used to build groups of similar users and then a model was learned for each group. For dealing with the longitudinal character of the data, the study S25 used the Dynamic Time Warping approach for calculating the similarities (distances) between the users. Another approach is discussed in S6, where a network graph of users was built, and then the k-nearest neighbors algorithm was applied. In contrast to the mentioned approaches, the so-called "IntelligentPooling" approach by Tomkins et al. (S22) adaptively pools more data if homogeneity is observed in the data, while it adapts to a rather separate approach with reduced pooling when the data is heterogeneous.

Additional Algorithms: Beyond balancing between the "personalization extremes", the analyzed studies also used additional algorithms for other purposes. For example, the recent study S31 used deep learning for generating representations (embeddings) combining static patient attributes along with sequential patterns (health behavior pathways), which extract useful information. The user embedding vector then serves in conjunction with an item embedding as a recommendation context for a CMAB. In contrast to using further algorithms for pre-processing, additional algorithms may be also placed behind the RL algorithm. The study by Rabbi et al. (S2) combined human customization with RL-based suggestions. Users are allowed to remove unsuitable suggestions by themselves from a list and re-order suggestions according to their individual preferences. After the MAB has determined a set of suggestions, a Pareto-frontier algorithm is triggered. This algorithm serves to balance the activity recommendations between the user's preferences and efforts to perform the activities to maximize further the likelihood that the user follows the suggestion.

4.4 Contextual Data Inputs

Overall, those studies that used contextualized algorithms (i.e., CMAB and Full-RL) integrated a variety of variables which have been mainly measured via active and passive sensing approaches.

Actively sensed contextual data comprised self-reported data based on questionnaires such as the patient's mood (e.g., S1, S25), pain intensity (e.g., S26), the motivation to change or location (e.g., S27).

In contrast, *passively* sensed input data referred to the physical activity behavior using the pedometer or acceleration sensor (e.g., S1, S8, S14, S22, S27), the GPS sensed location (S1), system use adherence variables (esp., time since last smartphone lock event (S1), number of notifications received/read (S25), number of app screens (S14), number of times the app has been opened (S23)), health adherence variables (S18), the intervention history (S3) or variables that were directly derived from the underlying operating system (S1). Beyond actively or passively sensed data on the user's side, one study used data from the EMR of the patient (e.g., hospital admissions, emergency room visits,...), i.e., data generated and collected by health professionals (S24).

Several studies distinguished between baseline variables (e.g., age, gender, user traits and preferences) that do not change during the course of intervention and dynamic variables (e.g., S1, S3, S11, S17, S18). Dynamic variables broadly referred to the health-behavior, health-status and system usage-related metrics, the intervention history but also environment variables such as time of day, current weekday or weather (e.g., S1, S14, S25, S27). In particular, the intervention history could be expressed in terms of which intervention was delivered and when it was delivered (e.g., S3, S14, S17, S18, S19, S21, S22, S23, S31). For example, Yom-Tov et al. (S3) included the recommendation history as a contextual variable in terms of the number of days since each message category was sent. The justification for this strategy is to repetitively use the same messages if deemed advantageous by the learning algorithm or introduce more diversity. Similarly, other studies (e.g., S17, S18, S19, S31) incorporated the intervention history into the model (e.g., time when message was sent, days since message was sent, or interventions performed in the past). For incorporating the users's intervention receptivity, the studies S1 and S21 accessed the user's calendar to obtain data such as the number of free calendar records or the time until next meeting.

4.5 Objects of Personalization

For categorizing the studies regarding the objects of personalization, we followed the "Behavioral Intervention Technology Model" by Mohr et al. (2014). Based on this model, we derived the categories of personalizing the type of intervention (what?), the timing (when?), the intervention goals (why?), and the delivery channel or representation format (how?). Figure 6 depicts an overview of the objects of personalization found in the literature.

Type of Intervention: Regarding the objects of personalization, almost half of the analyzed studies (*n* = 17) used the RL algorithm to select an intervention type from a set of alternative actions. A common approach is to categorize the intervention options such as groups of activities or message types (e.g., S1, S17, S19). The RL algorithm then typically determines the intervention category from which a pre-formulated content instance (e.g., message) is randomly drawn. These categorizations of messages are often guided by theoretical and empirical findings from the field of health behavior change (e.g., S1, S3, S18, S19, S28). Beyond traditional media content (i.e., text, images, audio, or video), intervention options could be also games, questionnaires, or quizzes. However, a differentiated approach can be found in the study S20, where RL was used to optimize social comparison rankings (upward, downward, mixward) to stimulate physical activity. It should be noted that studies that used the terms "goals" or "challenges" were also assigned to this category if it was apparent that certain behavioral activities to achieve the goal were directly associated (e.g., S11, S31).

Timing: By integrating a "no intervention" or "provide nothing" action into the action space, the problems of finding the "best" intervention and time to deliver the intervention can be implicitly framed in one model. Overall, *n* = 6 studies have been identified which include a "no intervention" action of the agent among several others. If only one type of intervention exists, this "hybrid" approach is framed as a binary problem (intervention vs. no intervention) (*n* = 7). Another option to address the timing problem is to include possible time windows into the action space (e.g., 3 categories: 9 a.m. - 12 p.m.; 12 - 3 p.m.; 3 - 6 p.m.) (see S19, S23). It should be noted that although RL can be used to determine suit-

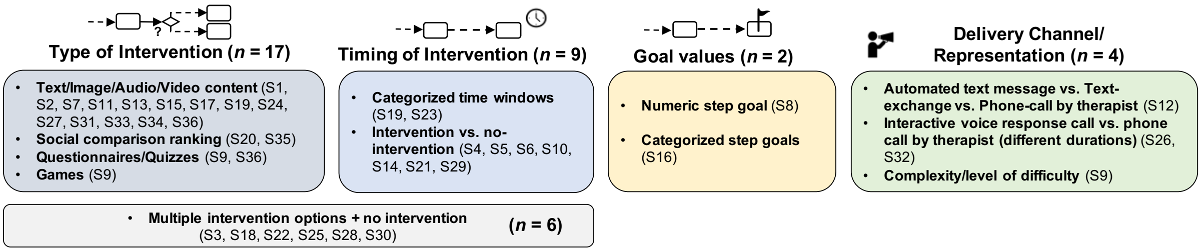

?
Type of Intervention (n = 17)
Text/Image/Audio/Video content (S1, S2, S7, S11, S13, S15, S17, S19, S24, S27, S31, S33, S34, S36)
Social comparison ranking (S20, S35)
Questionnaires/Quizzes (S9, S36)
Games (S9)
Timing of Intervention (n = 9)
Categorized time windows (S19, S23)
Intervention vs. no-intervention (S4, S5, S6, S10, S14, S21, S29)
Goal values (n = 2)
Numeric step goal (S8)
Categorized step goals (S16)
Delivery Channel/ Representation (n = 4)
Automated text message vs. Text-exchange vs. Phone-call by therapist (S12)
Interactive voice response call vs. phone call by therapist (different durations) (S26, S32)
Complexity/level of difficulty (S9)
Multiple intervention options + no intervention (S3, S18, S22, S25, S28, S30)
(n = 6)

Figure 6: Overview of the objects of personalization.

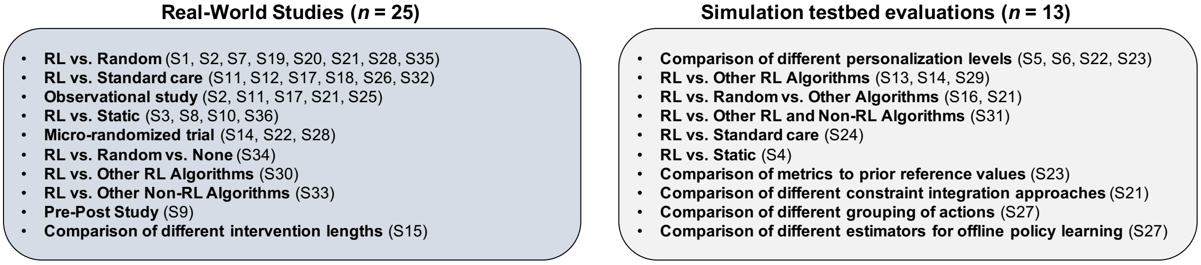

Real-World Studies (n = 25)
RL vs. Random (S1, S2, S7, S19, S20, S21, S28, S35)
RL vs. Standard care (S11, S12, S17, S18, S26, S32)
Observational study (S2, S11, S17, S21, S25)
RL vs. Static (S3, S8, S10, S36)
Micro-randomized trial (S14, S22, S28)
RL vs. Random vs. None (S34)
RL vs. Other RL Algorithms (S30)
RL vs. Other Non-RL Algorithms (S33)
Pre-Post Study (S9)
Comparison of different intervention lengths (S15)
Simulation testbed evaluations (n = 13)
Comparison of different personalization levels (S5, S6, S22, S23)
RL vs. Other RL Algorithms (S13, S14, S29)
RL vs. Random vs. Other Algorithms (S16, S21)
RL vs. Other RL and Non-RL Algorithms (S31)
RL vs. Standard care (S24)
RL vs. Static (S4)
Comparison of metrics to prior reference values (S23)
Comparison of different constraint integration approaches (S21)
Comparison of different grouping of actions (S27)
Comparison of different estimators for offline policy learning (S27)

Figure 7: Overview of the employed evaluation approaches.

able timings, the algorithm itself still needs to be invoked at certain times. These "decision points" are often pre-defined times of a day/week or repeating intervals (Nahum-Shani et al., 2018). For example, the algorithm described in S3 is triggered each morning while other studies invoke the algorithm multiple times a day (e.g., S14) or in specific intervals (e.g., S21). The time horizon between the triggers could be also larger than at the daily level (e.g., weekly (S26) or monthly (S24)). Beyond system-initiated timings, the algorithm could be also triggered by the user (e.g., S20).

Goals: A comparatively small amount of studies ($n = 2$) investigated the use of RL to optimize goal parameters in terms of specific goal values (daily steps). Therefore, the approach described in S8 uses inverse RL to predict a challenging but attainable number of daily steps that satisfy restrictions related to the user's responsiveness and self-efficacy. In contrast, the approach proposed in S16 recommends a range of steps drawn from 5 categories.

Delivery Channel/Representation: Some studies used the RL algorithm to determine the intervention delivery channel or representation of the content (how?). For example, the studies S12, S26, and S32 balanced between automated text messages or calls and human interventions by therapists. Noteworthy, the study S26 integrated the intervention duration into the action space, leading to a "hybrid" model addressing also the time aspect of the intervention (duration). In contrast, the study S9 used RL to adapt the complexity (difficulty) of the presented content.

4.6 Evaluation Approach

The evaluation approaches described in the analyzed studies were broadly categorized as real-world studies (i.e., conducted with human subjects) and testbed evaluations in a computer-simulated environment (see Figure 7).

Real-World-Studies: Overall, $n = 25$ studies were conducted or planned in real-world settings, including an experimental comparison of RL vs. Random policy (typically uniform sampling), RL vs. Static policy (fixed schedule of intervention delivery or fixed type of message), RL vs. Standard care, different advanced algorithms (RL and non-RL), the investigation of different intervention lengths (using RL), and a pre-post-study. Three studies described the use of the RL algorithm in a micro-randomized trial, which is a study design originating from the field of JITAIs (Nahum-Shani et al., 2018) for getting insights into the optimal timing, context, and effectiveness of intervention options. Furthermore, several studies reported results from observational studies.

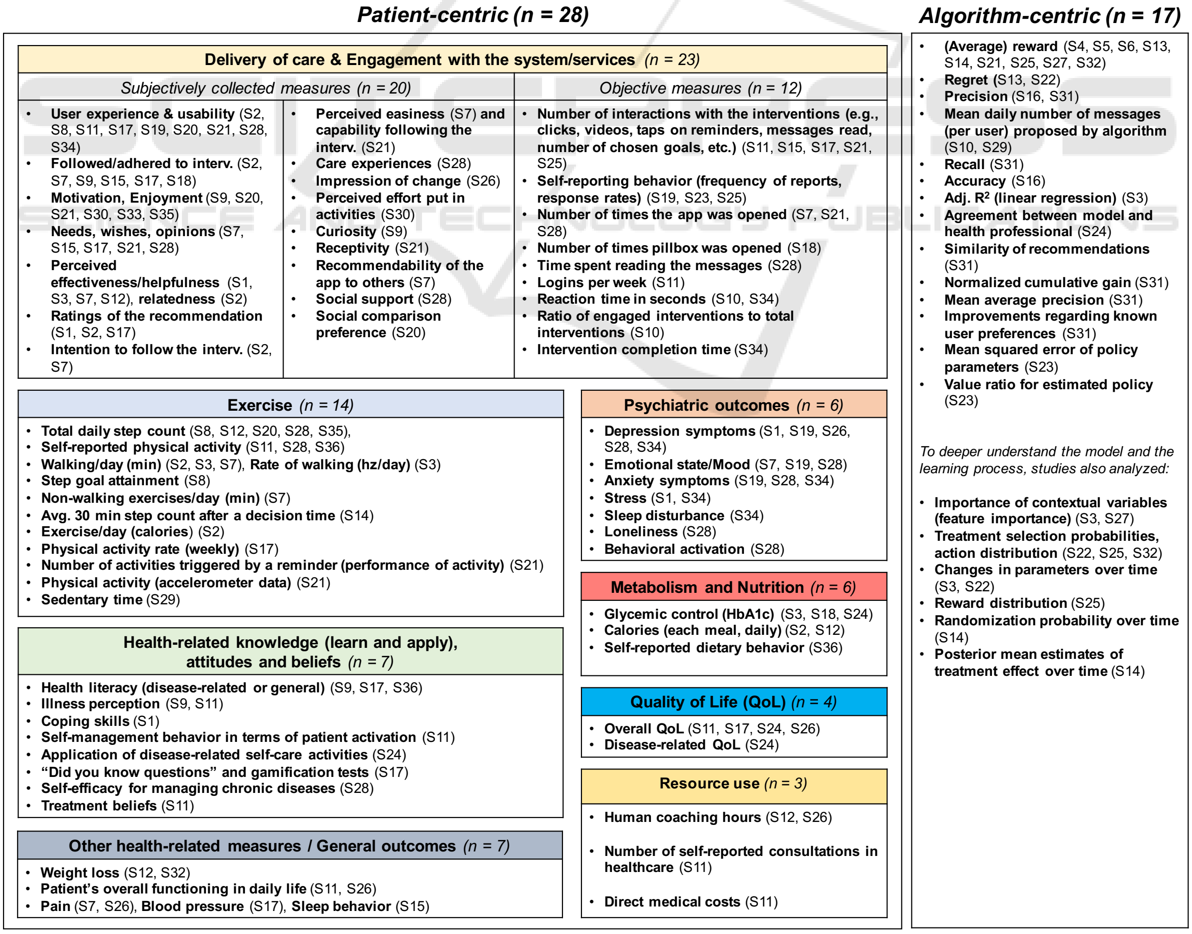
Testbed Evaluations: A significant number of papers ($n = 13$) investigated the RL algorithm in a simulated test environment using real-world datasets collected in previous studies. The comparisons include the comparison of different levels of personalization (e.g., pooled-RL vs. separated-RL), the comparison of the proposed algorithm to other RL and/or non-RL algorithm(s), comparison of different categorizations of the action space, comparison of different constraint integration approaches, comparison of estimators for offline policy learning, comparison to prior reference values, and RL vs. standard care.

A commonly cited data set used for testbed evaluations originates from the so-called “HeartSteps” (v1) mHealth study (S4, S5, S6, S14, S22, S29). The HeartSteps v1 study was a 42-day micro-randomized trial including 44 healthy adults aimed to evaluate context-aware recommendations to reduce sedentary behavior. According to Klasnja et al. (2019), the cleaned data set includes 6061 decision points linking recommendations to corresponding outcomes (cause–effect relationship). Likewise, the study S13 also used an interventional data set from a mHealth trial. Several papers also used data originating from observational studies. These observational data sets strongly differed in their sample size (from 114 to >10.000 participants) and length (5 weeks up to 4 years) (S27, S21). However, as described, e.g., in S16, inferring causal relationships from mere observational data comes with limitations, and the derivation of states-action pairs with their linkage to rewards is challenging.

4.7 Evaluation Metrics

The evaluation metrics were categorized based on the conceptualization of engagement in the context of digital behavior change interventions by Perski et al. (2017), the taxonomy of health behaviors by Nudelman and Shiloh (2015), and the taxonomy of clinical outcomes proposed by Dodd et al. (2018).

While these categorizations refer to the patient-centric perspective, an additional category was inductively introduced for metrics directly related to the RL algorithms (i.e., algorithm-centric). Algorithm-centric measures were used in n = 17 studies, including the average reward, regret, precision, and other metrics. Patient-centric metrics in terms of delivery of care and engagement with systems or services were frequently assessed (n = 23) and involved both subjective measures (such as patient-reported adherence) and objective measures (such as system-logged interaction frequency). Exercise, nutrition, quality of life, health-related knowledge, psychiatric outcomes, general health measures (like body weight and



Patient-centric (n = 28)

Delivery of care & Engagement with the system/services *(n = 23)*

Subjectively collected measures (n = 20)

- **User experience & usability** (S2, S8, S11, S17, S19, S20, S21, S28, S34)
- **Followed/adhered to interv.** (S2, S7, S9, S15, S17, S18)
- **Motivation, Enjoyment** (S9, S20, S21, S30, S33, S35)
- **Needs, wishes, opinions** (S7, S15, S17, S21, S28)
- **Perceived effectiveness/helpfulness** (S1, S3, S7, S12), **relatedness** (S2)
- **Ratings of the recommendation** (S1, S2, S17)
- **Intention to follow the interv.** (S2, S7)
- **Perceived easiness** (S7) **and capability following the interv.** (S21)
- **Care experiences** (S28)
- **Impression of change** (S26)
- **Perceived effort put in activities** (S30)
- **Curiosity** (S9)
- **Receptivity** (S21)
- **Recommendability of the app to others** (S7)
- **Social support** (S28)
- **Social comparison preference** (S20)

Objective measures (n = 12)

- **Number of interactions with the interventions (e.g., clicks, videos, taps on reminders, messages read, number of chosen goals, etc.)** (S11, S15, S17, S21, S25)
- **Self-reporting behavior (frequency of reports, response rates)** (S19, S23, S25)
- **Number of times the app was opened** (S7, S21, S28)
- **Number of times pillbox was opened** (S18)
- **Time spent reading the messages** (S28)
- **Logins per week** (S11)
- **Reaction time in seconds** (S10, S34)
- **Ratio of engaged interventions to total interventions** (S10)
- **Intervention completion time** (S34)

Exercise *(n = 14)*

- **Total daily step count** (S8, S12, S20, S28, S35),
- **Self-reported physical activity** (S11, S28, S36)
- **Walking/day (min)** (S2, S3, S7), **Rate of walking (hz/day)** (S3)
- **Step goal attainment** (S8)
- **Non-walking exercises/day (min)** (S7)
- **Avg. 30 min step count after a decision time** (S14)
- **Exercise/day (calories)** (S2)
- **Physical activity rate (weekly)** (S17)
- **Number of activities triggered by a reminder (performance of activity)** (S21)
- **Physical activity (accelerometer data)** (S21)
- **Sedentary time** (S29)

Health-related knowledge (learn and apply), attitudes and beliefs *(n = 7)*

- **Health literacy (disease-related or general)** (S9, S17, S36)
- **Illness perception** (S9, S11)
- **Coping skills** (S1)
- **Self-management behavior in terms of patient activation** (S11)
- **Application of disease-related self-care activities** (S24)
- **“Did you know questions” and gamification tests** (S17)
- **Self-efficacy for managing chronic diseases** (S28)
- **Treatment beliefs** (S11)

Other health-related measures / General outcomes *(n = 7)*

- **Weight loss** (S12, S32)
- **Patient’s overall functioning in daily life** (S11, S26)
- **Pain** (S7, S26), **Blood pressure** (S17), **Sleep behavior** (S15)

Psychiatric outcomes *(n = 6)*

- **Depression symptoms** (S1, S19, S26, S28, S34)
- **Emotional state/Mood** (S7, S19, S28)
- **Anxiety symptoms** (S19, S28, S34)
- **Stress** (S1, S34)
- **Sleep disturbance** (S34)
- **Loneliness** (S28)
- **Behavioral activation** (S28)

Metabolism and Nutrition *(n = 6)*

- **Glycemic control (HbA1c)** (S3, S18, S24)
- **Calories (each meal, daily)** (S2, S12)
- **Self-reported dietary behavior** (S36)

Quality of Life (QoL) *(n = 4)*

- **Overall QoL** (S11, S17, S24, S26)
- **Disease-related QoL** (S24)

Resource use *(n = 3)*

- **Human coaching hours** (S12, S26)
- **Number of self-reported consultations in healthcare** (S11)
- **Direct medical costs** (S11)

Algorithm-centric (n = 17)

- **(Average) reward** (S4, S5, S6, S13, S14, S21, S25, S27, S32)
- **Regret (**S13, S22)
- **Precision** (S16, S31)
- **Mean daily number of messages (per user) proposed by algorithm** (S10, S29)
- **Recall** (S31)
- **Accuracy** (S16)
- **Adj. R^2 (linear regression)** (S3)
- **Agreement between model and health professional** (S24)
- **Similarity of recommendations** (S31)
- **Normalized cumulative gain** (S31)
- **Mean average precision** (S31)
- **Improvements regarding known user preferences** (S31)
- **Mean squared error of policy parameters** (S23)
- **Value ratio for estimated policy** (S23)

To deeper understand the model and the learning process, studies also analyzed:

- **Importance of contextual variables (feature importance)** (S3, S27)
- **Treatment selection probabilities, action distribution** (S22, S25, S32)
- **Changes in parameters over time** (S3, S22)
- **Reward distribution** (S25)
- **Randomization probability over time** (S14)
- **Posterior mean estimates of treatment effect over time** (S14)

Figure 8: Overview of evaluation metrics reported in the analyzed studies.

Table 4: Overview of interventional real-world studies comparing RL to another (non-RL) condition.

ID	Comparison	Length, *n*	Key result
S1	RL vs. Random	4 weeks, $n = 95$	RL-based intervention facilitated greater stress reduction over four weeks compared to random recommendations
S2	RL vs. Random	14 weeks, $n = 16$	Significant improvements in participant's physical activity, consumed calories, perceived relatedness of suggestions and the number of suggestions followed or wanted tofollow
S7	RL vs. Random	5 weeks, $n = 10$	RL-based suggestions were easier to follow; differences regarding the physical activity and reported pain were non-significant; decreasing app use over time
S20	RL vs. Random	3 weeks, $n = 53$	No significant increase in physical activity, though motivation levels were reported higher in the RL group (small to moderate effect size)
S21	RL vs. Random	4 weeks, $n = 23$	No significant differences in physical activity behavior and perceptions, opinions, and user experience between the RL and random prompts group
S28	RL vs. Random	6 weeks, $n = 93$	RL group membership did not significantly influenced the daily steps
S34	RL, Rand., Self, None	4 weeks, $n = 69$	RL-based and self-proposed interventions had significant higher stress reduction than random interventions
S35	RL vs. Random	3 weeks, $n = 46$	Significant improvement of the user's motivation in the ShapleyBandit group compared to random intervention but not to the greedy strategy
S3	RL vs. Static	26 weeks, $n = 27$	RL-based messaging significantly improved physical activity (walking behavior); significant decrease of HbA1c levels in diabetes patients
S8	RL vs. Static	10 weeks, $n = 64$	RL-based group had a lower decrease of daily steps over time compared to the static group (10.000 steps/day)
S10	RL vs. Static	2 weeks, $n = 17$	Improvements of adapted SARSA over static policy in terms of number of engaged interventions and reaction times
S36	RL vs. Static	6 months, $n = 1049$	Improvements of the user's diabetes-related health literacy, self-reported physical activity and diet
S33	RL vs. PSO	15 weeks, $n = 69$	Significant effect of RL over particle swarm optimization (PSO) algorithm on physical activity enjoyment scale
S11	RL vs. Standard care	6 months, $n = 427$	RL-based app did not alter the frequency of consultations, no clinically relevant improvements of pain or symptoms
S12	RL vs. Standard care	12 weeks, $n = 52$	The RL-based conditions (individual- and group optimized) achieved equivalent weight loss results with reduced human coaching hours compared to the solely human delivered intervention
S26	RL vs. Standard care	10 weeks, $n = 278$	RL-based selection of delivery mode led to improvements in pain scores with less therapist time
S32	RL vs. Standard care	12 weeks, $n = 228$	Improvements of the opioid-abuse risk score, particularly for patients with the highest risk at the start of the study

blood pressure), and economic resource use were also evaluated, highlighting a comprehensive and multi-faceted approach to assessing RL-based health behavior change interventions. Figure 8 summarizes the identified evaluation metrics.

4.8 Evidence

Table 4 provides a summary of interventional real-world studies comparing RL to a non-RL condition in order to obtain a general view of the overall evidence ($n = 17$). It is noteworthy that bandit algorithms were used in $n = 12$ of the 17 real-world studies presented. Overall, the RL algorithms have been compared to random policies, static policies (e.g., always sending the same message or sending a message at the same time), other algorithms, or to the standard care involving human therapists. Despite mixed results, several studies could demonstrate the potential to improve existing traditional or digital interventions.

5 DISCUSSION

5.1 Principal Findings

The high-level findings of this literature review are summarized in a framework for integrating RL within

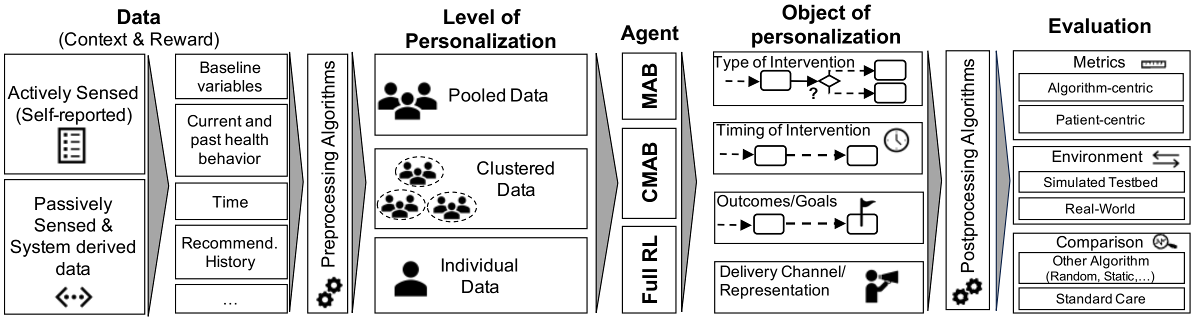


Figure 9: Framework for integrating Reinforcement Learning within DTx-supported Care Pathways.

DTx-supported care pathways (see Figure 9) and are discussed in detail below.

Data Considerations: Collectively, the findings from real-world studies underscore the potential of RL within DTx but also highlight the need for larger, longer-term studies beyond 6 months to validate these results. Despite this need, the present literature review may still clear up the misbelief that RL needs millions of data rows (data points) to obtain superior results compared to random or static policies. Obtaining such amounts of data is difficult in the context of behavior change interventions where rewards are delayed as the effects of the recommendations can be observed only several hours, days, or weeks later. However, there is still a need for publicly available data sets of health behavior change interventions. Of note is the recent study by Albers et al. (2022), which contributes in this regard by sharing their dataset consisting of 2366 persuasive messages delivered to 671 people to promote smoking cessation.

Our literature review revealed that a variety of data sources (actively and passively sensed data) may be included to capture the contextual state of the user and/or serve as a reward signal for the algorithm. In particular, self-reported data collected via questionnaires is a typical ingredient of RL-based behavior change interventions. However, when using self-reported data for training the algorithm, ensuring high data quality and mitigating biases (e.g., social desirability) is crucial as the learning process can be directly manipulated. We argue that future research should also throw an eye on advancing user interfaces to make the questionnaire completion more engaging and motivating for the user and thus improve data quality (e.g., using chatbots or gamification).

Algorithmic Considerations: In addition to considerations regarding the data, the literature analysis revealed that good results can be achieved with comparatively ”simple” algorithms (MABs and CMABs). However, given their application for one-step decision problems, the personalization of the care pathway is then rather limited to certain points than the entire path.

To address the balance between personalization at a pooled data level and on an individual level, the algorithms can be enhanced, for instance, by incorporating pre-processing methods like clustering. This approach seems promising to accelerate the learning process and mitigate data quality issues. Closely related to this trade-off between pooling and individualization are considerations regarding the privacy-preserving use of AI. Up to now, this issue has been largely unaddressed in the analyzed literature. One promising approach could be so-called ”federated” reinforcement learning where the user data remains on the local device, and only the trained models are shared and then aggregated on a centralized server (Khalid et al., 2023). Further algorithmic challenges that are not yet completely solved are related to non-stationarity (e.g., see Mintz et al. (2020), Tomkins et al. (2021)). This line of research focuses on taking into account that the learned user's preferences can vary over time, including the phenomenon that the learned model itself may change the user and consequently lose accuracy. Zhu et al. (2021) call this the ”Personalization Paradox”.

In view of the fact that DTx applications fall under the medical device legislation, patient safety must given special consideration, especially when AI techniques are used. For example, prompting a patient suffering from a chronic disease who may be on the brink of overexertion to increase exercise over and over again can cause medical damage. However, these concerns have been only addressed in a very small amount of papers by setting boundaries to the algorithm in terms of an upper time-bound of physical activity per day (e.g., 60 minutes) (S7) or by restricting the number of notifications per day (S21). We argue that future studies should take concerns regarding patient safety more deliberately into account when designing the algorithmic approach (”safety by design”). Directly associated with the need for patient safety is the need for model interpretability when AI is integrated into medical devices. Within the analyzed literature, the potential of explainable AI (XAI) has not yet been fully exploited, although most of the studies used largely transparent approximation models (esp., linear regression or regression trees). Future research could integrate explanations into the action space (e.g., see McInerney et al. (2018)) and/or employ post-processing algorithms to enhance the persuasiveness of health recommendations.

Considerations Regarding the Optimization Goal and Objects of Personalization: Overall, the results of the literature review suggest that a significant amount of papers sought to optimize physical activity. In contrast, application areas such as optimizing questionnaire response rates are comparatively understudied, although self-reporting is a typical ingredient of DTx (Lee et al., 2023). Therefore, we encourage researchers and practitioners to study the use of RL beyond physical activity promotion. For instance, future studies could focus more intensively on using RL to optimize the delivery channel, content representation, or specific goals for the patient.

Methodically Support the Development of "Intelligent" DTx Interventions: To simplify the technical process of integrating RL within DTx, a promising approach are low-/no-code development platforms. Within the literature review, one study could be identified that used such a platform to set up the RL algorithm (S18). Overall, we argue that there is a large potential in this regard as the first domain-specific platforms for DTx development are just starting to emerge (Liu et al., 2022).

5.2 Strengths and Limitations

The current literature review offers a comprehensive, multi-perspective insight into a comparatively underexplored and unstructured field. An extensive literature search was undertaken, encompassing a detailed search string and three databases, supplemented by snowballing relevant articles. However, our work also comes with several limitations. Firstly, applying RL to behavior change interventions is still in its early stage, which means that the field is rapidly evolving, and despite employing a rather inclusive search strategy, some pertinent studies might have been missed. Secondly, given the heterogeneity of this emerging field, and the exploratory nature of this review, the methodological quality of the studies was not explicitly assessed. Thirdly, the interpretation of data and results leaves room for subjective bias. However, the involvement of two authors may have helped to reduce this potential bias.

6 CONCLUSIONS

The paper aimed to offer a detailed examination of the application of RL algorithms for enhancing behavior change interventions delivered via DTx. Based on our analysis of the state of the art, we derived implications for practice and future research by elucidating the employed algorithms, the objects of personalization, and outlining their evaluation approaches and metrics. Overall, there is a need for larger studies going beyond the scope of primary prevention to gain more evidence on the suitability and efficacy of using RL within DTx. Nonetheless, several real-world studies could indicate the advantageousness compared to conventional approaches and demonstrate their impact on health outcomes. Future research perspectives may also include self-reporting approaches that are able to maintain high data quality and user engagement over time, the integration of privacy-preserving AI techniques, problems arising from the "personalization paradox", mechanisms to ensure patient safety and methods supporting the holistic development of self-learning DTx. In conclusion, this study may offer a foundational starting point for those exploring the design, implementation, and evaluation of self-learning DTx. It contributes to the growing body of research in this field.

REFERENCES

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In Mannor, S., Srebro, N., and Williamson, R. C., editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 39.1–39.26, Edinburgh, Scotland. PMLR.

Aguilera, A., Figueroa, C. A., Hernandez-Ramos, R., Sarkar, U., Cemballi, A., Gomez-Pathak, L., Miramontes, J., Yom-Tov, E., Chakraborty, B., Yan, X., Xu, J., Modiri, A., Aggarwal, J., Jay Williams, J., and Lyles, C. R. (2020). mHealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the DIAMANTE Study. *BMJ Open*, 10(8):e034723.

Albers, N., Neerincx, M. A., and Brinkman, W.-P. (2022). Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active. *PLOS ONE*, 17(12):e0277295.

Ameko, M. K., Beltzer, M. L., Cai, L., Boukhechba, M., Teachman, B. A., and Barnes, L. E. (2020). Offline Contextual Multi-armed Bandits for Mobile Health Interventions: A Case Study on Emotion Regulation. In *Fourteenth ACM Conference on Recommender Systems*, pages 249–258, Virtual Event Brazil. ACM.

Beltzer, M. L., Ameko, M. K., Daniel, K. E., Daros, A. R., Boukhechba, M., Barnes, L. E., and Teachman, B. A. (2022). Building an emotion regulation recommender algorithm for socially anxious individuals using contextual bandits. *British Journal of Clinical Psychology*, 61(S1):51–72.

Bertsimas, D., Klasnja, P., Murphy, S., and Na, L. (2022). Data-driven Interpretable Policy Construction for Personalized Mobile Health. In *2022 IEEE International Conference on Digital Health (ICDH)*, pages 13–22, Barcelona, Spain. IEEE.

Coronato, A., Naeem, M., De Pietro, G., and Paragliola, G. (2020). Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964.

Daskalova, N., Yoon, J., Wang, Y., Araujo, C., Beltran, G., Nugent, N., McGeary, J., Williams, J. J., and Huang, J. (2020). SleepBandits: Guided Flexible Self-Experiments for Sleep. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA. ACM.

Delmas, A., Clement, B., Oudeyer, P.-Y., and Sauzéon, H. (2018). Fostering Health Education With a Serious Game in Children With Asthma: Pilot Studies for Assessing Learning Efficacy and Automatized Learning Personalization. *Frontiers in Education*, 3:99.

den Hengst, F., Grua, E. M., el Hassouni, A., and Hoogendoorn, M. (2020). Reinforcement learning for personalization: A systematic literature review. *Data Science*, 3(2):107–147. Number: 2.

Di, S., Petch, J., Gerstein, H. C., Zhu, R., and Sherifali, D. (2022). Optimizing Health Coaching for Patients With Type 2 Diabetes Using Machine Learning: Model Development and Validation Study. *JMIR Formative Research*, 6(9):e37838.

Dodd, S., Clarke, M., Becker, L., Mavergames, C., Fish, R., and Williamson, P. R. (2018). A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. *Journal of Clinical Epidemiology*, 96:84–92.

El Hassouni, A., Hoogendoorn, M., Ciharova, M., Kleiboer, A., Amarti, K., Muhonen, V., Riper, H., and Eiben, A. E. (2022). pH-RL: A Personalization Architecture to Bring Reinforcement Learning to Health Practice. In Nicosia, G., Ojha, V., La Malfa, E., La Malfa, G., Jansen, G., Pardalos, P. M., Giuffrida, G., and Umeton, R., editors, *Machine Learning, Optimization, and Data Science*, volume 13163, pages 265–280. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Etminani, K., Göransson, C., Galozy, A., Norell Pejner, M., and Nowaczyk, S. (2021). Improving Medication Adherence Through Adaptive Digital Interventions (iMedA) in Patients With Hypertension: Protocol for an Interrupted Time Series Study. *JMIR Research Protocols*, 10(5):e24494.

European Union (2015). Council conclusions on personalised medicine for patients. *Official Journal of the European Union*, (C 421).

Figueroa, C. A., Deliu, N., Chakraborty, B., Modiri, A., Xu, J., Aggarwal, J., Jay Williams, J., Lyles, C., and Aguilera, A. (2022). Daily Motivational Text Messages to Promote Physical Activity in University Students: Results From a Microrandomized Trial. *Annals of Behavioral Medicine*, 56(2):212–218.

Figueroa, C. A., Hernandez-Ramos, R., Boone, C. E., Gómez-Pathak, L., Yip, V., Luo, T., Sierra, V., Xu, J., Chakraborty, B., Darrow, S., and Aguilera, A. (2021). A Text Messaging Intervention for Coping With Social Distancing During COVID-19 (StayWell at Home): Protocol for a Randomized Controlled Trial. *JMIR Research Protocols*, 10(1):e23592.

Forman, E. M., Kerrigan, S. G., Butryn, M. L., Juarascio, A. S., Manasse, S. M., Ontañón, S., Dallal, D. H., Crochiere, R. J., and Moskow, D. (2019). Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *Journal of Behavioral Medicine*, 42(2):276–290.

Fürstenau, D., Gersch, M., and Schreiter, S. (2023). Digital Therapeutics (DTx). *Business & Information Systems Engineering*.

Gasparetti, F., Aiello, L. M., and Quercia, D. (2020). Personalized weight loss strategies by mining activity tracker data. *User Modeling and User-Adapted Interaction*, 30(3):447–476. Number: 3 Publisher: Springer.

Gonul, S., Namli, T., Baskaya, M., Sinaci, A. A., Cosar, A., and Toroslu, I. H. (2018). Optimization of Just-in-Time Adaptive Interventions Using Reinforcement Learning. In Mouhoub, M., Sadaoui, S., Ait Mohamed, O., and Ali, M., editors, *Recent Trends and Future Technology in Applied Intelligence*, volume 10868, pages 334–341. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Gray, R. C., Villareale, J., Fox, T. B., Dallal, D. H., Ontanon, S., Arigo, D., Jabbari, S., and Zhu, J. (2023). Improving Fairness in Adaptive Social Exergames via Shapley Bandits. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 322–336, Sydney NSW Australia. ACM.

Hong, J. S., Wasden, C., and Han, D. H. (2021). Introduction of digital therapeutics. *Computer Methods and Programs in Biomedicine*, 209:106319.

Hu, X., Qian, M., Cheng, B., and Cheung, Y. K. (2021). Personalized Policy Learning Using Longitudinal Mobile Health Data. *Journal of the American Statistical Association*, 116(533):410–420.

Janiesch, C., Zschech, P., and Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4:237–285.

Kankanhalli, A., Xia, Q., Ai, P., and Zhao, X. (2021). Understanding Personalization for Health Behavior Change Applications: A Review and Future Directions. *AIS Transactions on Human-Computer Interaction*, pages 316–349.

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., and Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, 158:106848.

Kinsey, S., Wolf, J., Saligram, N., Ramesan, V., Walavalkar, M., Jaswal, N., Ramalingam, S., Sinha, A., and Nguyen, T. (2023). Building a Personalized Messaging System for Health Intervention in Underprivileged Regions Using Reinforcement Learning. In *32nd International Joint Conference on Artificial Intelligence, IJCAI 2023*, pages 6022–6030. International Joint Conferences on Artificial Intelligence.

Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., Hekler, E. B., and Murphy, S. A. (2019). Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-randomized Optimization Trial of HeartSteps. *Annals of Behavioral Medicine*, 53(6):573–582.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press, Cambridge ; New York, NY.

Lauffenburger, J. C., Yom-Tov, E., Keller, P. A., McDonnell, M. E., Bessette, L. G., Fontanet, C. P., Sears, E. S., Kim, E., Hanken, K., Buckley, J. J., Barlev, R. A., Haff, N., and Choudhry, N. K. (2021). REinforcement learning to improve non-adherence for diabetes treatments by Optimising Response and Customising Engagement (REINFORCE): study protocol of a pragmatic randomised trial. *BMJ Open*, 11(12):e052091.

Lee, U., Jung, G., Ma, E.-Y., Kim, J. S., Kim, H., Alikhanov, J., Noh, Y., and Kim, H. (2023). Toward Data-Driven Digital Therapeutics Analytics: Literature Review and Research Directions. *IEEE/CAA Journal of Automatica Sinica*, 10(1):42–66.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, Raleigh North Carolina USA. ACM.

Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. (2020). Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22.

Liu, S., La, H., Willms, A., and Rhodes, R. E. (2022). A "No-Code" App Design Platform for Mobile Health Research: Development and Usability Study. *JMIR Formative Research*, 6(8):e38737.

McInerney, J., Lacker, B., Hansen, S., Higley, K., Bouchard, H., Gruson, A., and Mehrotra, R. (2018). Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 31–39, Vancouver British Columbia Canada. ACM.

Mintz, Y., Aswani, A., Kaminsky, P., Flowers, E., and Fukuoka, Y. (2020). Nonstationary Bandits with Habituation and Recovery Dynamics. *Operations Research*, 68(5):1493–1516.

Moher, D. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine*, 151(4):264.

Mohr, D. C., Schueller, S. M., Montague, E., Burns, M. N., and Rashidi, P. (2014). The behavioral intervention technology model: an integrated conceptual and technological framework for eHealth and mHealth interventions. *Journal of medical Internet research*, 16(6):e146. Number: 6 Publisher: JMIR Publications Inc., Toronto, Canada.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., and Murphy, S. A. (2018). Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine*, 52(6):446–462. Number: 6.

Nahum-Shani, I., Wetter, D. W., and Murphy, S. A. (2023). Adapting just-in-time interventions to vulnerability and receptivity: Conceptual and methodological considerations. In *Digital Therapeutics for Mental Health and Addiction*, pages 77–87. Elsevier.

Nudelman, G. and Shiloh, S. (2015). Mapping health behaviors: Constructing and validating a common-sense taxonomy of health behaviors. *Social Science & Medicine*, 146:1–10.

Oyebode, O., Fowles, J., Steeves, D., and Orji, R. (2023). Machine Learning Techniques in Adaptive and Personalized Systems for Health and Wellness. *International Journal of Human–Computer Interaction*, 39(9):1938–1962.

Paredes, P., Giald-Bachrach, R., Czerwinski, M., Roseway, A., Rowan, K., and Hernandez, J. (2014). PopTherapy: Coping with Stress through Pop-Culture. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, Oldenburg, Germany. ICST.

Pelle, T., Bevers, K., Van Der Palen, J., Van Den Hoogen, F., and Van Den Ende, C. (2020). Effect of the dr. Bart application on healthcare use and clinical outcomes in people with osteoarthritis of the knee and/or hip in the Netherlands; a randomized controlled trial. *Osteoarthritis and Cartilage*, 28(4):418–427.

Pelle, T., Bevers, K., Van Der Palen, J., Van Den Hoogen, F. H., and Van Den Ende, C. H. (2019). Development and evaluation of a tailored e-self-management intervention (dr. Bart app) for knee and/or hip osteoarthritis: study protocol. *BMC musculoskeletal disorders*, 20(1):398. Number: 1.

Perski, O., Blandford, A., West, R., and Michie, S. (2017). Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Translational Behavioral Medicine*, 7(2):254–267.

Philipp, P., Merkle, N., Gand, K., and Gißke, C. (2019). Continuous support for rehabilitation using machine learning. *it - Information Technology*, 61(5-6):273–284. Number: 5-6.

Piette, J. D., Newman, S., Krein, S. L., Marinec, N., Chen, J., Williams, D. A., Edmond, S. N., Driscoll, M., LaChappelle, K. M., Kerns, R. D., Maly, M., Kim, H. M., Farris, K. B., Higgins, D. M., Buta, E., and Heapy, A. A. (2022). Patient-Centered Pain Care Using Artificial Intelligence and Mobile Health Tools: A Randomized Comparative Effectiveness Trial. *JAMA Internal Medicine*, 182(9):975.

Piette, J. D., Thomas, L., Newman, S., Marinec, N., Krauss, J., Chen, J., Wu, Z., and Bohnert, A. S. B. (2023). An Automatically Adaptive Digital Health Intervention to Decrease Opioid-Related Risk While Conserving Counselor Time: Quantitative Analysis of Treatment Decisions Based on Artificial Intelligence and Patient-Reported Risk Measures. *Journal of Medical Internet Research*, 25:e44165.

Rabbi, M., Aung, M. H., , M., and Choudhury, T. (2015). MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 707–718.

Rabbi, M., Aung, M. S., Gay, G., Reid, M. C., and Choudhury, T. (2018). Feasibility and Acceptability of Mobile Phone–Based Auto-Personalized Physical Activity Recommendations for Chronic Pain Self-Management: Pilot Study on Adults. *Journal of Medical Internet Research*, 20(10):e10147.

Sporrel, K., Wang, S., Ettema, D. D. F., Nibbeling, N., Krose, B. J. A., Deutekom, M., De Boer, R. D. D., and Simons, M. (2022). Just-in-Time Prompts for Running, Walking, and Performing Strength Exercises in the Built Environment: 4-Week Randomized Feasibility Study. *JMIR Formative Research*, 6(8):e35268.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition edition.

Tomkins, S., Liao, P., Klasnja, P., and Murphy, S. (2021). IntelligentPooling: practical Thompson sampling for mHealth. *Machine Learning*, 110(9):2685–2727.

Tong, X., Mauriello, M. L., Mora-Mendoza, M. A., Prabhu, N., Kim, J. P., and Paredes Castro, P. E. (2023). Just Do Something: Comparing Self-proposed and Machine-recommended Stress Interventions among Online Workers with Home Sweet Office. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, Hamburg Germany. ACM.

Tragos, E., O'Reilly-Morgan, D., Geraci, J., Shi, B., Smyth, B., Doherty, C., Lawlor, A., and Hurley, N. (2023). Keeping People Active and Healthy at Home Using a Reinforcement Learning-based Fitness Recommendation Framework. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6237–6245, Macau, SAR China. International Joint Conferences on Artificial Intelligence Organization.

Wang, S., Sporrel, K., Van Hoof, H., Simons, M., De Boer, R. D. D., Ettema, D., Nibbeling, N., Deutekom, M., and Kröse, B. (2021a). Reinforcement Learning to Send Reminders at Right Moments in Smartphone Exercise Application: A Feasibility Study. *International Journal of Environmental Research and Public Health*, 18(11):6059.

Wang, S., Zhang, C., Kröse, B., and Van Hoof, H. (2021b). Optimizing Adaptive Notifications in Mobile Health Interventions Systems: Reinforcement Learning from a Data-driven Behavioral Simulator. *Journal of Medical Systems*, 45(12):102.

Watkins, C. J. and Dayan, P. (1992). Technical Note: Q-Learning. *Machine Learning*, 8(3/4):279–292.

Webster, J. and Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, pages xiii–xxiii. Publisher: JSTOR.

Wei, Y., Zheng, P., Deng, H., Wang, X., Li, X., and Fu, H. (2020). Design Features for Improving Mobile Health Intervention User Engagement: Systematic Review and Thematic Analysis. *Journal of Medical Internet Research*, 22(12):e21687.

Weimann, T. G., Schlieter, H., and Brendel, A. B. (2022). Virtual Coaches: Background, Theories, and Future Research Directions. *Business & Information Systems Engineering*, 64(4):515–528.

White, J. M. (2013). *Bandit algorithms for website optimization*. Developing, deploying, and debugging. O'Reilly, Sebastopol, California.

Yom-Tov, E., Feraru, G., Kozdoba, M., Mannor, S., Tennenholtz, M., and Hochberg, I. (2017). Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical Internet research*, 19(10):e338.

Zhou, M., Fukuoka, Y., Mintz, Y., Goldberg, K., Kaminsky, P., Flowers, E., and Aswani, A. (2018a). Evaluating Machine Learning–Based Automated Personalized Daily Step Goals Delivered Through a Mobile Phone App: Randomized Controlled Trial. *JMIR mHealth and uHealth*, 6(1):e28.

Zhou, M., Mintz, Y., Fukuoka, Y., Goldberg, K., Flowers, E., Kaminsky, P., Castillejo, A., and Aswani, A. (2018b). Personalizing mobile fitness apps using reinforcement learning. In *CEUR workshop proceedings*, volume 2068. NIH Public Access.

Zhou, T., Wang, Y., Yan, L. L., and Tan, Y. (2023). Spoiled for Choice? Personalized Recommendation for Healthcare Decisions: A Multiarmed Bandit Approach. *Information Systems Research*, page isre.2022.1191.

Zhu, F., Guo, J., Li, R., and Huang, J. (2018a). Robust Actor-Critic Contextual Bandit for Mobile Health (mHealth) Interventions. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 492–501, Washington DC USA. ACM.

Zhu, F., Guo, J., Xu, Z., Liao, P., Yang, L., and Huang, J. (2018b). Group-Driven Reinforcement Learning for Personalized mHealth Intervention. In Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C., and Fichtinger, G., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, volume 11070, pages 590–598. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Zhu, F., Liao, P., Zhu, X., Yao, J., and Huang, J. (2018c). Cohesion-driven Online Actor-Critic Reinforcement Learning for mHealth Intervention. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 482–491, Washington DC USA. ACM.

Zhu, J., Dallal, D. H., Gray, R. C., Villareale, J., Ontañón, S., Forman, E. M., and Arigo, D. (2021). Personalization Paradox in Behavior Change Apps: Lessons from a Social Comparison-Based Personalized App for Physical Activity. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21.