Mobile Health Interventions through Reinforcement Learning

Alexander Galozy

Mobile Health Interventions through Reinforcement Learning

Halmstad University Dissertation No. 102
ISBN 978-91-89587-17-5 (printed)
ISBN 978-91-89587-16-8 (pdf)
Publisher: Halmstad University Press, 2023 | www.hh.se/hup

Abstract

This thesis presents work conducted in the domain of sequential decision-making in general and Bandit problems in particular, tackling challenges from a practical and theoretical perspective, framed in the contexts of mobile Health. The early stages of this work have been conducted in the context of the project "improving Medication Adherence through Person-Centred Care and Adaptive Interventions" (iMedA) which aims to provide personalized adaptive interventions to hypertensive patients, supporting them in managing their medication regimen. The focus lies on inadequate medication adherence (MA), a pervasive issue where patients do not take their medication as instructed by their physician. The selection of individuals for intervention through secondary database analysis on Electronic Health Records (EHRs) was a key challenge and is addressed through in-depth analysis of common adherence measures, development of prediction models for MA, and discussions on limitations of such approaches for analyzing MA. Providing personalized adaptive interventions is framed in several bandit settings and addresses the challenge of delivering relevant interventions in environments where contextual information is unreliable and full of noise. Furthermore, the need for good initial policies is explored and improved in the latent-bandits setting, utilizing prior collected data to optimal selection the best intervention at every decision point. As the final concluding work, this thesis elaborates on the need for privacy and explores different privatization techniques in the form of noise-additive strategies using a realistic recommendation scenario.

The contributions of the thesis can be summarised as follows: (1) Highlighting the issues encountered in measuring MA through secondary database analysis and providing recommendations to address these issues, (2) Investigating machine learning models developed using EHRs for MA prediction and extraction of common refilling patterns through EHRs, (3) formal problem definition for a novel contextual bandit setting with context uncertainty commonly encountered in Mobile Health and development of an algorithm designed for such environments. (4) Algorithmic improvements, equipping the agent with information-gathering capabilities for active action selection in the latent bandit setting, and (5) exploring important privacy aspects using a realistic recommender scenario.

To my wife and family

Acknowledgments

I want to thank my principal supervisor, Sławomir Nowaczyk, for his continuing guidance during my PhD studies. He helped me to think not only critically about my work but also see value where I sometimes did not, which helped me recognize my continuing growth as a researcher. I'm grateful for the many engaging discussions and ideas, sometimes of a personal nature, that keep me interested in science and research.

Furthermore, I would like to thank Anita Sant'Anna for being a motivating presence in the initial phases of my PhD. Her interest and involvement in my work have led to fruitful ideas. I want to thank my co-supervisor, Mattias Ohlsson, for his perspective and health data-related expertise, helping me keep practical considerations of my research in mind. I also want to express my sincere gratitude to Sadi Alawadi, my latest co-supervisor, who has shown a great deal of involvement, even though it was only for a short time. I really appreciated our long discussions about work and life. I would like to believe that your energy has given me the final boost at the end of my studies.

I want to thank both Björn Avgall and Markus Lingman for always being ready to help me with my questions from the clinical side of my research. Through my project work, they provided me with an indispensable source of knowledge and expertise, helping me understand the issues and challenges healthcare professionals face today.

I also would like to thank all my lab colleagues who provided a friendly and relaxed research environment. I became a significantly better researcher just by our (often too long) discussions.

I thank my sister, nephew, and mother for being there for me, even though a significant distance may separate us. I also would like to thank my great parents-in-law, who always had my back and provided sincere guidance and cheers. Last but certainly not least, I would like to thank my loving wife for always being there for me providing encouragement to achieve my goals and taking care of myself. I take great solace in knowing you, and all those years with you have made me, in all, a better human being.

List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I: **Pitfalls of medication adherence approximation through EHR and pharmacy records: Definitions, data and computation**
Alexander Galozy, Sławomir Nowaczyk, Anita Sant'Anna, Mattias Ohlsson, Markus Lingman. **International Journal of Medical Informatics**, *published* 31 January 2020.

PAPER II: **Prediction and pattern analysis of medication refill adherence through electronic health records and dispensation data**
Alexander Galozy, Sławomir Nowaczyk. **Journal of Biomedical Informatics**, *published* 13 June 2020.

PAPER III: **A New Bandit Setting Balancing Information from State Evolution and Corrupted Context**
Alexander Galozy, Sławomir Nowaczyk, Mattias Ohlsson. *submitted* 21 November 2022.

PAPER IV: **Information-gathering in Latent Bandits**
Alexander Galozy, Sławomir Nowaczyk, **Knowledge-Based Systems**, *published* 25 January 2023

PAPER V: **Beyond Random Noise: Insights on Anonymization Strategies from a Latent Bandit Study**
Alexander Galozy, Sadi Alawadi, Victor R. Kebande, Sławomir Nowaczyk, *submitted* 30 September 2023

Papers that are not included in the thesis.

PAPER VI: **Improving Medication Adherence Through Adaptive Digital Interventions (iMedA) in Patients With Hypertension: Protocol for an Interrupted Time Series Study**
Kobra Etminani, Carina Göransson, Alexander Galozy, Margaretha Norell Pejner, Sławomir Nowaczyk. **JMIR Research Protocols**, *published* 5 December 2021.

PAPER VII: **Patterns and Predictors Associated With Long-Term Glycemic Control in Pediatric and Young Adult Patients with Type 1 Diabetes**
Johan Jendle and Björn Agvall and Alexander Galozy and Peter Adolfsson. **Journal of Diabetes Science and Technology**, *published* 5 December 2022.

PAPER VIII: **Better Glycemic Control and Higher Use of Advanced Diabetes Technology in the Age Group 0-17 Yrs Compared to 18-25 Yrs With Type 1 Diabetes**
Johan Jendle and Björn Agvall and Alexander Galozy and Peter Adolfsson. **Diabetes Technology & Therapeutics**, *published* 2022.

Contents

List of Figures

1. INTRODUCTION

Imagine a scenario of a patient diagnosed with hypertension. To manage her condition, the patient's doctor prescribes her daily medication to control her blood pressure. However, due to her busy schedule and forgetfulness, she occasionally misses taking her medication, leading to fluctuations in her blood pressure levels and negatively impacting her health. During her routine check-up, the patient's doctor reviews her medical records, indicating an adverse health progression, and suspects inadequate adherence to treatment. Recognizing the importance of addressing this issue, the doctor decides to assess the patient's medication adherence and determine if she needs intervention or further education about her condition.

To gain insights into the patient's medication adherence patterns without burdening her, the doctor decides to utilize her electronic health records (EHR). The EHRs contain information about the patient's prescription refills and past appointments, which can indirectly measure her adherence. Leveraging the data from the EHR, the doctor goes a step further and employs predictive analysis to estimate patients' future medication refills. This analysis helps identify early warning signs of potential non-adherence and enables the doctor to intervene before the situation worsens. The doctor decided that the patient would benefit from additional support to improve her medication adherence. Subsequently, her physician recommends a mHealth app specifically designed to assist patients with chronic health conditions in managing their regimens. Understanding the importance of personalized care, the mHealth app starts by gathering relevant information about the patient's lifestyle, preferences, and daily routines. This data helps the app provide interventions tailored to her needs, ensuring higher patient engagement.

As the patient starts using the mHealth app, it monitors her medication-taking behavior and response to interventions. The app employs machine learning algorithms to adapt its interventions dynamically based on patients' adherence patterns and feedback. As the patient keeps using the mHealth app, her medication adherence significantly improves. With better management and monitoring of her chronic condition, the patient experiences fewer complications and improves overall quality of life.

The presented scenario serves as an illustrative example of an application motivating this doctoral thesis. It demonstrates an idealistic scenario where all

aspects of measurement, prediction and offering personalized interventions to support the patient's regimen are adequately implemented. As one can imagine, these aspects are not easily solved in real-world implementation, and we will encounter challenges and limitations.

The core focus of the thesis is to investigate these challenges. In particular, this thesis investigates **(1)** the measurement of medication adherence through secondary database analysis and provide solutions and recommendation if such an analysis is carried out. Measurement via such databases is prone to pitfalls that would affect the pool of candidates for intervention significantly [1]. **(2)** The prediction of adherence using EHRs. Previous studies utilizing sociodemographic factors, clinical factors, or purchasing information had only marginal success in predicting long-term adherence to medication [2–5]. While these efforts did not result in accurate prediction models, they point out the importance of analyzing patients' refilling behaviors and patterns of healthcare utilization. Given the comprehensive EHRs and pharmacy records available, we identified a research gap concerning the predictability of adherence using EHRs, focusing specifically on healthcare utilization factors and analyzing temporal patterns of refill medication adherence. **(3)** formulating the task of tailored intervention in a Reinforcement Learning (RL) framework, specifically in the domain of Bandits - a simplified RL setting that lends itself to mobile health due to not only good performance in many real-world applications, particularly in mHealth but also allows a more straightforward analysis of developed new algorithms. **(4)** Learning bandit algorithms from scratch often has the undesired effect of providing irrelevant intervention during the initial learning phase. We aim to address this issue by investigating settings that allow the use of prior collected data to inform initial policy generation, specifically in the domain of latent bandits. And finally, **(5)** investigating privacy aspects in the context of latent bandits in a realistic deployment scenario.

1.1 Research Questions

We investigated in total four research questions that would frame the doctoral thesis. Research questions II - IV constitute significant parts of the thesis, mainly concerned with the RL perspective on mHealth. I was of significant interest for the project iMedA project in the first part of the thesis and whose answer would guide part of the research path in the later stages.

- **RI: How can mediation adherence be measured and predicted through secondary database analysis?**

Timely identification of patients requiring support is crucial for effective pharmacological treatment in cardiovascular diseases, especially for secondary or tertiary prevention [6]. Current approaches aim to measure Medication Adherence (MA) throughout treatment and intervene when nonadherence is detected. However, accurately and cost-effectively measuring adherence poses challenges, such as high patient burden and missing or incorrect data in secondary databases to name a few. These challenges require tailored solutions we aim to investigate.

Predicting medication adherence is clinically valuable as it helps estimate future nonadherence probability, enabling early interventions by physicians. Exploring measurement and prediction methods within the idiosyncrasies of administrative EHR is of interest. It poses open challenges, such as linkage of different data sources, missingness not at random, input errors, and non-iid data.

- **RII: How can the commonly occurring requirements in mHealth, such as fast learning in noisy contexts, accommodation of model misspecification, and non-stationarity, be formulated in a Bandit framework?**

  Bandits are a mature and robust framework for sequential decision-making. It provides an ideal fit for mHealth applications due to its ability to learn intervention strategies individually tailored to patients. Furthermore, the capacity of bandit algorithms for lifelong learning equips autonomous agents with the ability to adapt to the changing circumstances of the patient quickly, promising higher patient engagement and better health outcomes through more relevant interventions. To our knowledge, there is a lack of formal settings that incorporate the properties of mHealth that would allow practical and theoretical evaluation of algorithms designed for mHealth. Some of the challenges include the requirement of a good initial policy avoiding too frequent or irrelevant interventions, assessing the usefulness of features for decision-making, robustness to failure of algorithmic assumptions, and dealing with noisy or missing data [7].

- **RIII: How can prior collected data be exploited for intervention selection in Bandit settings where online learning remains a critical component to adjust for a changing environment?**

  Learning rewards and environmental dynamic models from scratch can lead to unfavorable interventions during the initial stages of bandit learning. Therefore, it is desirable to leverage prior collected data to guide

initial intervention selection while continuously personalizing the interventions for each patient. Previous research has primarily focused on selecting interventions that maximize immediate reward without explicitly considering the intervention's effectiveness in uncovering the current state of the patient, which would allow an agent to select the best intervention significantly more often. Thus, we believe it important to investigate alternative strategies that better exploit existing data to effectively uncover states while retaining strong online learning capabilities to adjust to changing patient preferences.

- **RIV: How effective are additive noise schemes in keeping data private?** In order to ensure the widespread acceptance of AI solutions, it is imperative to provide patients with a sufficient degree of privacy, safeguarding their data from third-party entities, which may also include the service provider hosting the AI solution. The importance of privacy in our increasingly interconnected world cannot be overstated, as the interlinking of personal and public information can lead to the deanonymization of privatized data, resulting in catastrophic repercussions for individuals. We investigate the effectiveness of popular privatization schemes in their ability to achieve privacy goals, especially when additional public data might be available that renders these privatization techniques potentially ineffective. We are not only interested in ensuring privacy but also in the potential impact on the recommendation performance of privatization schemes. We investigate common schemes and suggest alternatives that promise a better privacy-performance trade-off.

1.2 Contributions

The overall theme of the doctoral thesis is framed from the perspective of tackling the issue of providing personalized adaptive interventions in the domain of mHealth through AI. In the following, we list the individual contributions of this thesis toward better data-driven interventions:

- We have shown, through extensive comparative experimentation, how data quality influences refill adherence estimates and provided recommendations on how to remedy data quality issues (Paper I).
- We show that while predictive performance is high in selecting patients for intervention, using EHRs and dispensation records for refill adherence prediction introduces a data bias towards patients with high healthcare utilization. Choosing patients for intervention based on prediction models that use these data sources is potentially unreliable (Paper II).

- We extracted common longitudinal medication pickup patterns through cluster analysis. Furthermore, we cross-correlated these patterns using simple simulation models of mediation consumption. We show that several different consumption patterns can result in similar pickup patterns, potentially misleading when determining the necessity and type of interventions (Paper II).
- We formulate a new problem setting for mHealth in a Bandit framework, a sub-field of RL, exhibiting action order that can be exploited in case of significant context corruption. This setting simulates commonly occurring issues like context uncertainty. (Paper III).
- We develop a meta-algorithm for this setting that shows superior empirical performance compared to state-of-the-art algorithms (Paper III).
- We provide a regret lower bound for the problem setting and analyze our algorithm's regret upper bound (Paper III).
- We frame learning with initial policies in the latent bandit problem, where a hidden state governs rewards. We develop an algorithm based on posterior sampling that chooses arms via an information-directed criterion and trajectory roll-outs (Paper IV).
- We conduct a theoretical analysis of the regret of our algorithm for the stationary and non-stationary settings, showing that sublinear regret is achieved (Paper IV).
- We test our algorithm on a large-scale real-world recommendation data set, showing the benefit of using information-gathering in improving cumulative regret. Additionally, we conduct experiments on synthetic and real-world data to determine when we benefit from using our approach, particularly when the latent state cannot be effectively uncovered without information-gathering arms (Paper IV).
- We investigate important privacy aspects in the context of latent bandits, using different strategies for providing a transition matrix to a user in a realistic deployment scenario (Paper V).
- We demonstrate the susceptibility to attacks of simple additive noise strategies and investigate alternatives with better privacy-performance trade-offs (Paper V).
- We conduct our experiments using real-world datasets, highlighting the importance of implementing tailored privatization strategies for each dataset (Paper V).

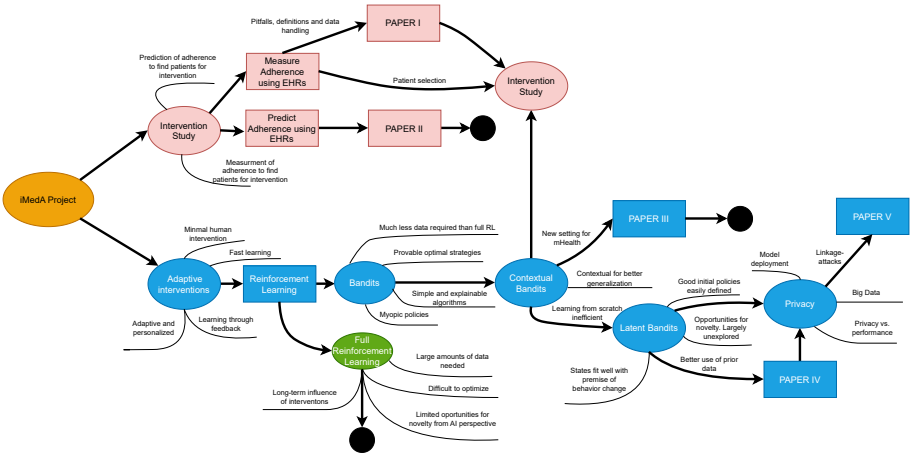


Figure 1.1: Research direction.

- We are the first to explore privacy concerns within a learning setup in the latent bandit context, specifically in a realistic and frequently encountered linkage attack scenario (Paper V).

Figure 1.1 depicts the research direction followed during the doctoral journey, illustrating the evolution of research ideas and the paths taken. The initial research phase was inspired by the iMedA project, which necessitated addressing specific requirements. The project involved an intervention study with actual hypertensive patients, requiring the selection of an appropriate patient pool by identifying those with inadequate levels of refill adherence. During our analysis, we identified data quality and measure definition issues, which we promptly addressed and published in an article, providing recommendations for resolving these challenges.

As the project progressed, we gained access to a substantial amount of electronic health record (EHR) data on hypertensive patients. However, a notable challenge was the infrequent updating of the research database, occurring only once every three months. This led us to explore the development of prediction models for selecting patients at risk of non-adherence in the future. The results of this study were published in a journal article. Nonetheless, we discovered significant data bias that rendered the prediction models unreliable, ultimately abandoning the idea since the task of patient selection based solely on this type of data was not possible before the intervention study.

Subsequently, our focus shifted towards the need for providing adaptive tailored interventions for the intervention study. During this phase, we contemplated various research avenues for accomplishing this goal. Ultimately, we decided to employ Reinforcement Learning (RL) for this approach, as it

aligned well with the concept of a healthcare app and patient feedback for intervention selection. To enhance efficiency, we chose the Bandit framework over the full RL settings due to its better sample complexity, better interpretability of algorithms, and provable performance guarantees.

Further investigation into behavior change issues revealed that myopic policies are still effective in driving patient engagement, a critical factor for intervention success. Consequently, we committed to using the Bandit framework for further research, as it presented ample opportunities for novelty in exploring mHealth-specific properties, which had not been extensively studied before. Our initial studies in integrating mHealth properties into the Bandit setting resulted in the third paper, where we developed an algorithm for the setting and analyzed the regret bounds associated with both the setting and the algorithm.

During this phase, we recognized the inefficiency of learning from scratch and directed our attention toward bandit approaches that utilize prior data to more quickly provide effective interventions while retaining online learning capabilities. This led us to investigate latent bandits, wherein we developed a more efficient method of exploiting data structure for selecting the best intervention and analyzed our algorithm in terms of expected regret performance.

Toward the final stages of the thesis, we shifted the focus toward privacy concerns, which became the central theme of the last paper. We developed a realistic deployment scenario and specifically focused on the linkage-attack patterns, wherein an attacker uses third-party data to re-identify records from a dataset that has been transformed with several privacy-ensuring methods. The outcome of this research emphasized the necessity for tailored privatization strategies and improved schemes to address linkage-attack scenarios.

1.3 Ethical Approval

All studies using patient data from electronic health records had approval from the Ethics Committee in Lund (Dnr. 2018/294). Prediction studies and adherence analysis were carried out on data between 2012 and 2019 obtained from the Regional Healthcare Information Platform [8]. Consent from individual patients was obtained through opt-out. That is, an opportunity was given for patients to request the removal of their data from the analysis.

1.4 Disposition

The remainder of this thesis is organized as follows. The work is divided into two chapters, chapter 2 discusses the background for each part of the work to

familiarize the reader with relevant knowledge and prior work to understand the results better. The results are discussed in specificity in section 3. The summary of the papers is presented in chapter 4. Finally, chapter 5 concludes the thesis.

2. BACKGROUND

Significant improvements in healthcare over the last decades increased lifespan, and with it, the number of people that live with chronic conditions [9]. This development has led to allocating a significant portion of healthcare resources for the treatment and prevention of chronic illnesses, such as hypertension, stroke, coronary artery disease, and heart failure[10]. While knowledge in effective treatment of chronic conditions is ever increasing, a promising and arguably necessary approach to improve patient well-being and reduce cost is through self-management, empowering patients to manage their illness by educating and teaching individuals how to identify and solve problems related to their condition [9]. The increasing pervasiveness of digital technologies in modern societies provides fertile ground for the successful growth of new approaches that involve the individual patient in a more holistic healthcare framework towards Person-Centered-Care (PCC). One such approach focuses on using the advances in mobile phone technology to deliver tailored interventions that aim to educate, remind or help patients change their habits and attitudes towards their illness to improve outcomes.

Through the 20th century, PCC has become the de facto mantra of modern healthcare internationally, promising improved patient outcomes and increased care satisfaction [11]. PCC has been described as "understanding the patient as a unique human being" [12] and doing away with the notion that the patient is a passive receiver of care with the sole purpose of medicine being in the diagnosis, treatment, and prevention of individual diseases [13]. Through a growing body of evidence, it became more evident for healthcare providers, researchers, and policymakers that the premise of PCC constitutes a cultural shift towards the full integration of patients into medical treatment, focusing on their unique needs, goals, and experiences [14].

This thesis focuses on TeleHealth applications and Mobile Health (mHealth) in particular. The WHO defines TeleHealth as "[The] delivery of healthcare services, where patients and providers are separated by distance. TeleHealth uses information and communication technologies to exchange information for the diagnosis and treatment of diseases and injuries, research and evaluation, and for the continuing education of health professionals". MHealth focuses on delivering healthcare services and reminders through mobile phone applications and has seen a significant increase in interest in the past decade, allowing

the easy monitoring and exchange of individual health information at any time and anywhere. Because of the popularity of health applications, researchers have focused on using this new technology to aid the vision of PCC further, integrating information about the patient's life outside the clinical setting. This holistic view of healthcare opens many new opportunities to provide patients with needed support throughout their daily lives and reduce costly human involvement using Artificial Intelligence (AI). Some examples of mHealth applications in healthcare are: Monitoring patients' health status [15], allowing early detection of patient deterioration [16]; Wellness applications supporting patients in leading a healthier lifestyle [17]; Applications that deliver behavior modifications to help patients change habits to improve health outcomes [18] or applications that support patients in staying adherent to their treatment plan[19]. For a comprehensive survey on mHealth applications, we refer to [20].

Through targeted digital interventions and reminders, mHealth applications provide an effective pathway to support patients with their daily medication regimens in a more personalized manner. While the idea is promising, implementing effective mobile interventions is challenging. From a patient's perspective, intervention fatigue and engagement are important processes for intervention adherence and retention [21]. Intervention fatigue is the emotional or cognitive overload or burden associated with treatment engagement. Patients feel overwhelmed with the constant effort of managing their disease and subsequently become nonadherent to their treatment. Intervention engagement is a "multifaceted state of motivational commitment or investment in the client role over the treatment process" [22]. Both mechanisms play an important role in adherence to interventions and point towards the need to provide individualized support to which the patient is receptive given internal (e.g., current emotional state) and context-specific (e.g., location, time-of-day) factors [21; 23].

From the technological perspective, systems developed for automatic intervention support need to be robust to issues idiosyncratic to both user and mobile technology. Interest in Reinforcement Learning (RL) as a framework has been growing in recent years, showing great promise for applications in mHealth. Wide-scale adoption of RL is stunted due to the sensitivity of contemporary algorithms to assumption-mismatch between environments the algorithms are designed for and the environment of mHealth. For example, when considering survey or self-report data for decision-making, systems have to deal with the completeness of the records driven by the engagement level of the users[24]. Using contextual information through mobile phone sensors might be incomplete due to an unstable wireless connection. Sensory information might also be noisy, too costly to acquire or otherwise unavailable due to

privacy concerns [7]. These issues can significantly delay learning such that a good intervention strategy might never be learned at a time when it matters. This highlights the need for algorithms and methods to act under the information uncertainty commonly encountered in mHealth. We aim to define and investigate a formal problem setting for RL that captures this information uncertainty. We investigate the effect of corrupted contextual information that prohibits the effective selection of relevant interventions and how to deal with this problem for more effective learning.

2.1 Medication Adherence

Medication nonadherence, commonly defined as the "failure of patients to take their medication as prescribed", is a pervasive issue and significant public health concern, contributing to an increase in levels of morbidity and mortality [25]. It is estimated that the rate of medication nonadherence among people with chronic diseases lies between 30-50%, increasing the overall cost of care due to avoidable hospitalisations [26]. Medication adherence is a complex, multi-dimensional phenomenon that is influenced by a multitude of societal, health system, and personal factors, often unique to the individual circumstances of the patient. The WHO defines adherence as the interplay between five different factors or dimensions of adherence: Health system/healthcare team factors, social/economic factors, therapy-related factors, patient-related factors, and condition-related factors. The complex problem of MA is not amendable to "one-size-fits-all" solutions and requires a targeted, individual approach taking the profile of the patients along these five dimensions into account when creating the optimal intervention or treatment plan [27].

Significant efforts have been undertaken to develop interventions targeted at nonadherent patients to support and improve outcomes [28]. The challenge is identifying patients at risk of nonadherence as early as possible to maximize interventions' effectiveness. For many cardiovascular diseases, pharmacological treatment's timeliness and effectiveness are tightly linked, especially for secondary or tertiary prevention [6]. The cost-effective and less intrusive measurement of MA remains a key challenge. This thesis investigates indirect measures that approximate medication adherence through medication refill adherence using information such as pharmacy dispensation data. This proxy information provides a low-cost, low-burden solution, with the caveat that actual medication consumption remains unknown [29; 30]. Many medication refill adherence measures, henceforth called refill adherence, have been developed, with the Medication Possession Ratio (MPR) and Proportion of Days Covered (PDC) being the most popular ones.

MPR is one of the most commonly used adherence measures. MPR is

computed as the ratio between the number of dispensed pills over the total number of pills to be dispensed in the measurement window:

$$MPR = \frac{\#\,dispensed\ pills}{\#\,total\ pills} \tag{2.1}$$

The MPR belongs to the Continuous Measures of medication Acquisition (CMA) since the timeliness of dispensations is not considered, only the total dispensed supply.

The PDC measure is computed as the ratio of the number of days theoretically covered by medication over the total number of days in the measurement window. A common definition is "percentage of days covered by medication" [31; 32]:

$$PDC = 1 - \frac{\#gap\ days}{measure\ window} \tag{2.2}$$

Where *#gaps days* refers to days without medication and *measure window* is the number of days over which MA is measured.

For digital interventions to be most effective, it is important to identify patients needing support technologies early during treatment. A key challenge is to estimate the level of adherence accurately and cost-effectively. We aim to investigate approximate measures using secondary databases, prescription and pharmacy records that promise a valid and cost-effective analysis of medication adherence.

2.2 Predictive Modeling in EHRs

EHRs are inherently retrospective data sources. That is, they contain data collected in the context of routine clinical operations and do not contain data from controlled clinical trials. As such, they are primarily used in retrospective studies to analyze and predict patient outcomes.

The outcome variable is directly extracted or purposefully constructed from EHRs. Direct extraction is done through conditional statements. For example, if primary medication adherence[1] is the outcome of interest, we might define the target variable in the following way *if prescription date - dispensation date < 30 days; target = 1 (adherent) else target=0*. Some outcomes are not contained natively in EHRs and thus are not directly accessible through conditional statements but may require post-processing after extraction. For instance, refill medication adherence might be defined on individual prescriptions with varying prescription lengths. In this thesis, we define the outcome variable through external adherence measures.

[1] medication is picked up at the pharmacy within a set time window

EHRs are usually comprised of information about all care visits of a patient through time. Each visit contains clinical and demographic information. For example, a patient might receive a prescription at a primary care unit. The visit would contain patient and care provider demographic information such as age, gender, primary care unit, prescriber's age, and more. The clinical information for a visit is represented in codes representing different clinical concepts such as diagnosis, performed procedures, prescribed medications, lab tests, and vital signs. Furthermore, each visit entry may contain examination information as free text input.

While the visit data can be used directly, the resulting dimensionality of the data representation is usually very high and thus results in inefficient learning of statistical models due to the curse of dimensionality [33]. One method for dealing with high dimensional data from EHRs relies on computing features using expert knowledge to summarize or condense the high dimensional representation in human-understandable concepts. Another approach uses the power of deep neural networks to learn features that abstract higher-level concepts. While appealing from the perspective of avoiding time-consuming feature engineering, these representations are often difficult to interpret by a human. In this thesis, we mainly consider human-derived features. We leave the investigation of learned representations to future work.

2.2.1 Defining Refill Adherence

Part of the purpose of the study conducted in Paper II was to investigate the predictability of adherence as measured by the PDC via machine learning algorithms. Patients with a high probability of nonadherence could be considered for early intervention to mitigate the potential impact of long-term effects of uncontrolled hypertension [34; 35].

There are a variety of clinically relevant outcomes about adherence we can investigate for the prediction study. The adherence literature identified two types of adherence:

- *Primary Adherence*

  Patients who fill their first prescription are called *primary adherent*. Treatment initiation is the first important step and one of the instances where refill adherence directly maps to real medication adherence. If the patient did not pick up their medication, we can be somewhat sure that they will not take them, cases aside where the patient might procure the medication through other means. Prediction of whether patients will fill their first prescription or not might provide care providers with the ability to intervene early and address the reasons for primary non-adherence.

- *Secondary Adherence*

  Patients who take their medication as prescribed by the physician are *secondary adherent*. Here, we already see a difference compared to primary adherence. It is not enough to fill the medication, but it has to be taken as prescribed. In the context of refill adherence, patients would fill and refill their prescriptions regularly and on time. Naturally, the information on whether patients take their medication or not is not available in secondary database analysis. Still, irregular or a refill stop might indicate patterns of real nonadherence. Like primary nonadherence, predicting whether or not patients will inadequately fill their prescriptions can help to facilitate timely investigations into the reasons for secondary nonadherence.

Most adherence studies using secondary database analysis use these two types of adherence for prediction. An 80% threshold is commonly used in adherence and prediction studies for arbitrary or historical reasons, above which it is assumed that pharmacological treatment is effective [36; 37].

2.3 Sequential Decision Making

Sequential Decision Making is a formalism that allows the modeling of processes that require decisions or "actions" to be made within an evolving environment to achieve a specific goal. These decisions may or may not influence the state of the environment. The agent needs to be aware of what actions lead to desirable states while avoiding actions that lead to undesirable states. Many, if not all, decisions humans make daily can be viewed as sequential decision-making processes. For example, buying and selling stocks at the stock market to maximize short- or long-term profit.

Particularly interesting for the theme of this thesis, we could imagine a hospital setting where the physician needs to decide on the patient's treatment plan. Choosing a particular step in the treatment plan will affect the health state of the patient and thus determine the health outcome positively or negatively. The physician must be aware of each step's risks and benefits, personalized to the treated individual, and decide appropriately. In a mHealth setting, providing automated and personalized interventions is vital to keeping the patient's interest and engagement. The decision-maker or agent needs to decide what series of digital interventions is interesting to the individual patient while considering the long-term effects on a performance metric, such as medication adherence level or blood pressure.

Given the ubiquitous nature of decision processes, significant focus has been placed on sequential decision-making to devise methods and algorithms

to find optimal or near-optimal decisions automatically. Great success has been achieved in complex and challenging environments, particularly in the domain of competitive games, such as Go [38], Starcraft [39], DotA 2 [40]. These environments are often unpredictable, requiring the player to perform a mixture of short-term and long-term strategies under incomplete information to be successful.

In this thesis, we focus on the bandit formulation, specifically *multi-armed bandits*, *contextual bandits*, and *latent bandits*, of RL. In this simplified setting, the agent is only concerned with maximizing cumulative *immediate* reward. The agent chooses the intervention that would result in the highest immediate reward, contrary to the setting where interventions might provide no (or negative) reward in the immediate but result in future states where more reward can be achieved following the agent's policy. In the following sections, we explain common nomenclature in RL in more detail.

2.3.1 The Agent

The entity or system that interacts with the world and makes decisions is the so-called "agent". The agent takes the information about the environment and decides what actions to take based on its future belief or estimate of the outcome. This information is also known as the "state" that evolves due to the agent's action or outside influences that the agent has no control over. In the context of mobile health, the agent would be the background system that observes the state of each patient. This state could be, for example, demographic information or health-related information. Based on this information, the agent decides which intervention to deliver at a particular moment.

2.3.2 The Environment

The environment the agent finds itself in is, roughly speaking, everything that is external to the agent. The agent may or may not influence the environment through its actions, but in many real-world scenarios, this is often the case, such as in the hospital example described above. If the agent can affect the state of the environment, there are several scenarios in which the environment dynamics may be affected. In a stationary setting, the environment transitions into states according to a fixed set of rules to respond to the agent's actions. These rules can be probabilistic, but the probabilities stay fixed. An agent can discover these rules over time and exploit them to optimize decision-making. In the nonstationary setting, these rules can change over time, making the problem significantly harder since experience becomes obsolete such that the agent chooses potentially sub-optimal actions.

Nonstationary is not only encountered as a feature of the environment but may manifest itself indirectly through incomplete information available to the agent for decision-making. States might look similar but need different actions. In the mHealth setting, the agent has to potentially contest with a combination of nonstationarity, incomplete information, and varying degrees of influence of actions, requiring complex approaches to action selection to perform optimally.

2.3.3 The Reward

For the agent to decide if it needs to change its behavior, it requires feedback on the "goodness" or "utility" of actions to achieve its goal. This feedback is often encoded in the so-called reward, indicating if actions lead to desirable states to solve a particular problem. Specific to our setting, improvement in blood pressure regulation is the primary goal. Given that blood pressure measurements might be infrequent and prone to errors, we might look at other metrics that we can use to define the reward. For example, the percentage of medication taken. Furthermore, patients might provide feedback for certain types of interventions that are interesting to them, allowing the agent to customize intervention selection further to improve patient engagement.

2.3.4 Behaviour or Policy

The policy is the formal description of the behavior of the agent. In essence, it operationalizes as a set of (possibly complicated) rules the agent follows at every step of interaction with the environment. There are several ways of generating policies depending on the knowledge of the environment.

One straightforward policy is not necessarily learned but provided to the agent through external means. The agent does not need external information and achieves its goal by executing a fixed *plan*. Fixed plans without environmental feedback are often not robust in real-world scenarios due to the environment's occasional uncertainties. In a mHealth application, the patient's lifestyle might change over time such that reminders at fixed times may become a nuisance. The agent must occasionally execute a plan to ask the user for an updated reminder schedule before continuing the reminder plan.

If the environmental state can change at every time step, we have the so-called stationary policy, or "universal plan" [41]. The agent uses a stationary policy, either deterministic or stochastic, to evaluate the current state of the environment and performs the action that would maximize the immediate or future reward. Complex real-world environments exhibit randomness, prohibiting exact reward prediction. In such scenarios, it is more beneficial to

consider *stochastic policies*, where the agent chooses an action from a probability distribution to maximize the expected reward. As mentioned earlier, in a mHealth setting, we face changing patient behaviors and must adopt a strategy that performs well under nonstationarity and noise.

2.3.5 Exploration and Exploitation

One fundamental issue in most sequential decision-making problems is the balance between the exploration and exploitation of actions. This need for balancing both aspects arises due to incomplete reward feedback. The agent only receives the reward of the chosen action; the reward of other actions is not revealed. This type of reward mechanism is often described as *bandit feedback*. This significantly delays learning since the agent needs to explore all actions for all states sufficiently often to ensure the optimal action has been chosen for the given state. When the agent explores, it deliberately chooses actions that might seem noncompetitive to confirm or revise its belief about the explored actions' utility. When the agent exploits, it selects the action it believes would maximize its reward.

Naturally, exploring suboptimal actions will result in less reward obtained and carries the risk of affecting the environment in ways that would come with significant penalties. For instance, in mHealth applications, user engagement is paramount for interventions to be successful. Interventions that are irrelevant or timely inconvenient might cause early abandonment of the application. Furthermore, insufficient exploration can lead to habituation that significantly diminishes the effectiveness of the interventions. While the latter problem can be addressed by ensuring diversity among interventions, the former problem requires efficient exploration schemes.

2.3.6 mHealth Specific Challenges for Reinforcement Learning

Focus is placed on the problem of providing these interventions under domain-specific constraints that need special consideration. While promising, several challenges in mHealth make the straightforward application of contemporary RL algorithms difficult. Some challenges include the requirement of a good initial policy avoiding too frequent or irrelevant interventions, robustness to failure of algorithmic assumptions, and dealing with noisy or missing data [7]. Recently, four major types of challenges have been identified [42].

Fast learning in noisy contexts. Recently, authors in [42] identified four major types of challenges in RL. These challenges can be divided into two categories: Technical challenges stem from domain-specific constraints in mHealth applications. For example, observational variables may be missing intermittently due to factors such as a lack of GPS signal or unavailability of the

phone [7]. Ensuring privacy is paramount for building trust in the system that patients are expected to use. Thus, privacy concerns may also restrict the collection of certain data or variables relevant for optimal intervention selection, negatively affecting the performance of RL algorithms susceptible to contemporary data privatization techniques.

The other category is patient-related challenges that arise from the nature of patient observations. Observations may be noisy due to erroneous feedback or adversarial behavior. That is, a patient may deliberately choose to sabotage the system. Or, more likely, the absence of feedback altogether. These noise modalities pose a significant challenge for algorithms relying on such feedback to choose interventions.

Accommodation of model misspecification and non-stationary. Incorrect assumptions about models that underlie the system's behavior may lead to suboptimal performance of RL algorithms. These assumptions include but are not limited to, linear or non-linear relationships between observations and rewards. The thesis explores methods to handle such model misspecification and improve the accuracy of choosing the correct interventions. Further, several kinds of environmental non-stationarity need to be taken into account. Patients' behaviors can change over time, leading to a drift in the data collected. This drift can be attributed to variations in patient characteristics or external factors. As a result, the input distribution of the model may change, requiring the RL algorithm to generalize and adapt.

Further, non-stationarity may manifest as concept drift, referring to the phenomenon where the mapping between input and output changes over time. This change can occur due to various factors, such as shifts in environmental conditions or evolving treatment strategies. Dealing with concept drift in mHealth applications involves developing mechanisms to detect and adapt to changes in the relationship between variables. To address this challenge, the thesis investigates adaptation techniques and continual learning methods that enable the RL algorithm to adapt to changing mappings between input and output.

2.3.7 Multi-Armed Bandits and Contextual Bandits

The multi-armed bandit (MAB) problem in the area of sequential decision-making has attracted significant attention due to its applicability in many real-world areas such as clinical trails [43; 44], finance [45; 46], routing networks [47; 48], online-advertising [49; 50] and movie [51] or app recommendation [52]. The agent aims to select from a set of available actions (also known as arms) to maximize the cumulative immediate reward. Furthermore, the bandit formulation assumes no influence of actions on future rewards or states. While

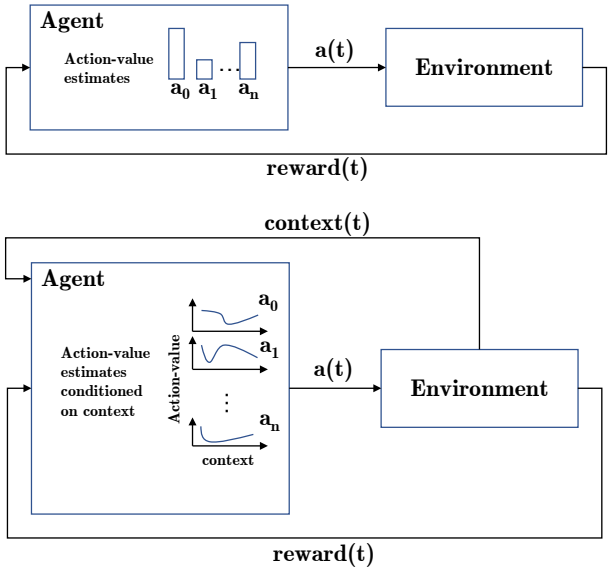


Figure 2.1: Top: Multi-Armed Bandit. Bottom: Contextual Bandit

limiting for problems where actions may significantly influence future states and rewards, this simplified setting works well in various practical settings and enjoys provable regret guarantees and good sample complexity. The fact that behavior change is a long process such that any single intervention has a temporally limited effect gives us some leeway in estimating the long-term rewards for particular interventions [53]. This allows us to consider simpler methods with good convergence, optimality guarantees, and better interpretability than more general methods from the full RL setting that lack these properties.

Significant work has been done to design algorithms that provide an optimum or near optimum exploitation/exploration trade-off for various problem settings. Previous works have explored the context-free MAB-setting such as the stochastic variant using upper confidence bounds (UCB) operating under the principle of "optimism in the face of uncertainty" [54; 55] or Bayesian treatments using Thompson sampling[56; 57] a so-called probability matching technique.

One particular formulation of the MAB has seen significant attention in the past. This extension to the MAB problem is the Contextual Bandit (CB) problem, also known as the Multi-Armed-Bandit problem with side information or associative reinforcement learning [58]. The agent receives a context, some description of the environmental state in the form of a feature vector, before

choosing an action, effectively solving a separate[1] MAB-problem conditioned on each context. Figure 2.1 illustrates the differences between MAB and CB schematically.

2.4 Privacy Challenges: Keeping Data Private in the Age of Big Data

In today's increasingly interconnected and data-driven world, protecting personal information has become a critical concern. As individuals and organizations generate and store vast amounts of data, maintaining data privacy has become a challenging task. The rapid proliferation of digital platforms, coupled with advancements in data collection, storage, and analysis techniques, has given rise to numerous data privacy challenges. Maintaining data privacy requires a delicate balance between the need for data-driven insights and protecting individual privacy rights. Achieving this balance necessitates implementing robust security measures, stringent regulations, and ethical considerations.

Data privacy challenges encompass a broad spectrum of issues related to the collection, use, and storage of personal data. They include but are not limited to unauthorized access to sensitive information [59], data breaches [60], re-identification attacks [61], and the loss of individual control over personal data [62].

One specific data privacy challenge that has gained prominence in recent years is the linkage attack. This attack exploits the presence of multiple data sources or databases to identify and connect seemingly unrelated pieces of information, leading to the re-identification of individuals and the compromise of their privacy. The repercussions of data privacy breaches and linkage attacks are not merely theoretical; numerous real-world incidents have highlighted their profound impact on individuals and organizations. One notable example is the Facebook-Cambridge Analytica Scandal: In 2018, the social media giant Facebook faced a massive data breach when the personal information of over 87 million users was harvested without consent by the political consulting firm Cambridge Analytica. The firm exploited a third-party app that collected users' data and used it to create psychographic profiles for targeted political advertising, showcasing the alarming consequences of unauthorized data linkage and usage. In the healthcare sector, where patient privacy is paramount, linkage attacks have been a major concern. A notable example occurred in 1990 when researchers demonstrated the ability to identify individuals in sup-

[1] Strictly speaking, the reward predictor generalizes over a space of MAB-problems.

posedly de-identified medical datasets by combining the data with publicly available information. A claims database containing information on 135,000 patients was subject to a re-identification attack. The attack focused on the discharge record of the Governor of Massachusetts at that time. By leveraging basic demographic details available in the Cambridge voter registration list, which was obtained for a mere $20, the governor's record was successfully re-identified [63].

To mitigate the problem, Differential Privacy (DP) has gained widespread acceptance as a fundamental metric for quantifying and ensuring provable privacy [64]. However, DP has failed, for example, in federated learning setting [65]. Therefore, there is a need to investigate alternative or complementary privacy-preserving techniques and anonymization strategies beyond the conventional use of noise, striking a balance between privacy protection and system performance.

While there are many interesting avenues for further research, in this thesis, we focus on linkage-attack scenarios, one of the most accessible and insidious of privacy attacks—especially considering the abundance of useful auxiliary information collected in the wild that may render contemporary privatization techniques less effective.

3. RESULTS

3.1 Medication Adherence: Common Pitfalls and Prediction

From the data analysis perspective, evaluating refill adherence using administrative databases is prone to methodological pitfalls, affecting the resulting adherence values significantly enough to warrant special consideration [1]. The study conducted in Paper I highlights several measure-related and data-related pitfalls that can affect the interpretation of medication adherence (MA) and the selection of patients for intervention.

Regarding measure-related pitfalls, inconsistent operational definitions of MA measures can lead to confusion when comparing adherence values between studies. For example, the commonly used Medication Possession Ratio (MPR) measure has been defined in different ways [66], while the Prescription Dispensing Coverage (PDC) measure has also seen alternative definitions [30]. Overall, we have discovered issues with consistently naming particular measures and emphasizing in our research their important differences regarding MA measurement in EHRs.

The inclusion or exclusion of oversupply in measuring refill adherence can significantly impact the accuracy of results in both prospective and retrospective adherence studies. Including oversupply can mask nonadherence patterns, while excluding it may result in lower adherence estimates, mainly when patients delay refills due to adequate supply. However, excluding oversupply in long-term adherence monitoring can skew adherence values at individual and population levels. This is especially relevant for patients with chronic illnesses who often have access to oversupply, and targeting interventions based on inaccurate adherence measures can be inefficient and harm the provider-patient relationship.

Measurement windows and gaps also play a role in MA assessment. The fixed observation window in prospective studies and the choice of the measurement window in secondary database analysis can influence adherence estimates significantly. Ignoring that patients often delay first pickups, thus removing the crucial initiation phase from consideration, can lead to overestimating refill adherence [67]. Additionally, the operationalization of adherence

measures, particularly those with measurement windows based on dispensations, can exclude patients and skew adherence estimates. The choice of measure should be carefully considered to avoid excluding a significant number of patients and compromising accuracy.

Accurate computation of refill adherence requires addressing common data-related pitfalls. Incomplete, missing, or incorrect data entries must be corrected or imputed. Simply removing missing entries is incorrect, as patients with incomplete records should be removed from the analysis to avoid underestimating adherence [68]. Imputation of missing or incorrect values is not as straightforward as using mean or most common values. The assumption of data being missing at random or missing at completely random is often not satisfied in modern EHR databases. Imputation strategies should consider the specific characteristics and patterns of missing data.

Data correction and imputation should focus on relevant factors for adherence computation, such as the prescribed daily dose and prescription periods. Missing or incorrect prescribed daily doses can significantly affect adherence values, and simple heuristics can be used to impute missing or identify incorrect values based on the number of dispensed pills. Prescription periods, including missing end dates, require careful imputation strategies, considering the specific context and policies. Duplications, often caused by manual data entry, must be identified and addressed to prevent bias in adherence estimates, especially in multi-drug therapy. Understanding and addressing measure-related and data-related pitfalls is crucial for accurate and meaningful assessment of medication adherence in EHRs.

Resolving these issues is the first important step when predicting MA with data-driven methods. Unfortunately for the Data Scientist, the reasons for nonadherence to medication can be manifold; many are not contained in EHRs. This limitation makes it significantly harder to predict refill adherence with any specificity. While demographic factors are available, critical patient-related factors or behavioral factors such as “stress”, “being busy”, “healthcare satisfaction” or “treatment burden” are not directly available. Some of these individual factors can be approximated to some extent through healthcare utilization patterns in the form of visits to primary care centers, hospitals, and emergency rooms. The number of concurrent prescriptions and drug variety might approximate the treatment burden.

Our study found a surprisingly high overall model discriminability, with the area under the curve ranging from 0.78 to 0.91 on the test sets. The presence of an oversupply of medication had the most significant impact on model decisions, as higher oversupply values correlated with increased adherence. Notably, patients who regularly refilled their prescriptions tended to accumulate a substantial supply, further contributing to their adherence. Reserving the

latest prescription of a patient in the test set and taking the rest for the training set, which excludes new patients, achieved the best performance on the test set. Still, it showed a bias towards patients already several years into treatment and likely to be adherent. Predictive performance was modest for patients with a single prescription, and clinical predictors showed a weak correlation with future medication adherence. Models struggled to predict adherence for patients with missing previous information or a sudden drop in adherence, showing a bias towards patients with regular follow-ups. However, adding historical information improved prediction performance, emphasizing past adherence as one of the strongest predictors of future adherence.

These results also show a clear bias towards patients who remain in treatment with regular follow-ups, i.e., patients who remain refill adherent continue to utilize healthcare resources, generating a data bias towards adherence. The fact that patients had regular checkups in the past provides increased confidence that they remain refill adherent, as evidenced by past adherence being among the strongest predictors. While predicting adherence might provide a way of selecting patients for intervention, we have discovered that models developed on comprehensive EHRs may be unreliable for patients without treatment history or sudden changes in adherence.

3.2 A Problem Setting for mHealth

The contextual setting we primarily consider is mHealth, where users' needs and wants are partially determined by a hidden and evolving state, for example. Depending on the state, different interventions might be required, e.g., some users are significantly affected by stress and may need a particular type of stress-coping techniques. In contrast, others may experience stress less severely, where general advice might be enough [69]. The underlying state may induce a natural ordering of actions as the users transition through different "levels" or "stages" that require specific interventions. Additionally, we can see the hidden underlying state as the different stages people may go through when forming habits, such as initiation, learning, or maintenance phase, requiring different interventions or intervention strategies [70].

Furthermore, we have mentioned earlier that the inherent nonstationary in mHealth poses a problem for simple MAB approaches. We presented one solution to this problem using contextual information leading to the CB formulation. In our setting of mHealth, we expect users not to know their state perfectly and provide the agent with a noisy estimate of the state or potentially completely irrelevant context that can significantly affect the agent's decision-making. This introduced nonstationary through incomplete information, masking changes in the context to reward mapping or misleading the agent to pro-

vide an intervention that is not appropriate anymore.

To tackle the issue of decision-making in environments that exhibit the properties mentioned above of context uncertainty and action order, we extend a previously described problem setting. In the previous setting, see [71], the agent needs to deal with corrupted contexts where the information content is entirely and irreversibly lost. With probability p, the agent receives a corrupted context. The arbitrary corruption function $v : X \to X$ governs how the context is corrupted and is unknown and non-retrievable. The context the agent receives at every time step is defined as

$$\hat{\mathbf{x}}_t = \begin{cases} v(\mathbf{x}_t) & \text{with probability} \quad p \\ \mathbf{x}_t & \text{with probability} \quad 1-p. \end{cases}$$

We extend this problem setting to include several users that can provide context at varying degrees of corruption. Additionally, to incorporate the intuition of stages patients might go through, the hidden state evolves in a Markovian manner; that is, the previous state at $t-1$ fully determines the current state at t. Each state is associated with a specific action. Coupled with the Markovian state evolution, this defines a sequence of actions as patients transition through different stages or levels. In protocol 1, we present a high-level description of our problem setting.

Protocol 1 Problem Protocol

procedure PROTOCOL
  for $t = 1, 2, \ldots, T$ **do**
    for user $i \in I$ **do**
      the environment generates context $\mathbf{x}_{i,t}$ from state $s_{i,t}$
      the context is corrupted $\hat{\mathbf{x}}_{i,t} = v(\mathbf{x}_{i,t})$ with probability $p_{i,t}$
      the agent chooses an action $a_{i,t} = \pi(\hat{\mathbf{x}}_{i,t})$
      the environment reveals the reward $r_{a_{i,t}}$
      the state s_i is updated: $s_{i,t+1} = \phi(r^i_{a_{i,t}}, \mathbf{s}_{i,t})$ ▷ "Markovian sampling" of next state
      policy π of the agent is updated

At each iteration and for every user, the environment generates the context $\mathbf{x}_{i,t}$ of user i from the underlying state $s_{i,t}$. It corrupts it with probability $p_{i,t}$. The corrupted context is observed by the agent, which chooses an action $a_{i,t}$ to play according to its policy. The environment reveals the action-reward $r^i_{a_{i,t}}$ and updates the state for user i. Finally, the agent updates its policy π [72].

3.2.1 Regret Bounds of the Problem Setting

From a scientific view, particularly in the literature on bandit algorithms, we are quite interested in what we can expect regarding algorithmic performance. More specifically, we would like to ask: "Given the definition of regret, what is

the minimum amount of regret I can expect in this setting provided that we can find the best algorithm?". Answering this question and comparing the regret bounds with our developed algorithms gives us a hint as to whether we are on the right track to solve the problem perfectly. While these are purely theoretical questions, they give a unique and mathematically rigorous insight into why certain decision problems are hard to solve. Thus, it is of scientific interest to analyze our developed problem setting regarding minimum achievable regret bounds. Detailed proofs are given in the respective paper. Here, we discuss the results and their implications.

Regret Lower Bound

We generally expect the number of state changes to be of the order less than T, justified by the domain-specific phenomenon that people recurrently traverse different stages of behavior change, spending longer and longer time in the "Maintenance" or "Terminal" stage [73]. This assumption is essential for context-free algorithms to achieve sublinear regret. It is particularly relevant in our setting with corrupted contexts, where we may need to rely on context-free algorithms to make decisions.

In essence, our problem setting exhibits both properties of the contextual bandit problem and the restless Markov bandit problem. We expect either the contextual bandit lower bound or switching Markov bandit lower bound to hold. We first analyze the contextual and context-free settings and combine them later in our final result.

Theorem 1. *For any context-free algorithm, the regret after T time steps in the CBCCAC setting, for any $K > 1$, minimum sub-optimally gap Δ_{min} for all actions, number of state changes L, number of states S and minimum detectable change $\varepsilon_{min} > \Delta_{min}$ in mean reward, is lower bounded by*

$$\mathbb{E}[R(T)] \geq \mathcal{O}\left(\sqrt{SKT} + \frac{\Delta_{min}\sqrt{KT\tilde{L}}}{\varepsilon_{min}}\right),$$

with

$$\mathbb{E}[\tilde{L}] \leq L.$$

Proof Sketch. We have S distinct learning problems in the latent bandit setting. Each problem incurs a regret of order $\Omega\left(\sqrt{KT/S}\right)$ if the learner spends T/S time steps in each state. Summing over all states gives a regret bound of $\Omega(\sqrt{SKT})$. However, our setting lacks direct state observation, requiring the detection of state switches.

To detect a change, approximately $n_\varepsilon = O(1/\varepsilon^2)$ samples are needed. Before detecting a change, the regret incurred is $\Delta\sqrt{KT/L}/\varepsilon$, where Δ represents the sub-optimally gap, or the difference in rewards between two actions that are compared. ε is the minimum detectable change. Summing over all changes gives a regret contribution of $\Delta\sqrt{KTL}/\varepsilon$.

By storing information about action reward distributions for each visited state, the problem reduces to determining the current hidden state. Utilizing the collected history of action rewards when revisiting states avoids relearning from scratch. The number of state switches L becomes the number of switches into unique states $\tilde{L}$, which can be much smaller than L. Combining both contributions with a minimum sub-optimally gap Δ_{min} and minimum detectable change ε_{min} yields the final result.

We now consider the case where policies can exploit the context, therefore "detecting" changes in the reward distributions of actions. For the case where $p = 0$, the context is a perfectly reliable way of detecting whether the state has changed. We use the result from [71] for this scenario.

Lemma 1. *[71] For any algorithm solving the CBCC problem with context size d, with $(1-p_\nu)$ and $0 \leq p_\nu \leq 1$ there exists a constant $\gamma > 0$, such that the lower bound of the expected regret accumulated by the algorithm over T iterations is lower-bounded as follows: $E[R(T)] > \gamma\sqrt{Td}$. where p_ν is the probability that the context is corrupted by an unknown function ν.*

We arrive at our final result by combining the results from the analysis of context-free and contextual settings. For the CBCCAC setting, the lower bound depends on the context size d, the number of actions K and states S, and the total number of state switches into unique states $\tilde{L}$. Contextual or context-free cases then dominate the lower bound. We get for the lower bound.

Theorem 2. *For any context or context-free algorithm, the regret after T time steps in the CBCCAC setting, for any $K > 1$, minimum sub-optimally gap Δ_{min} for all actions, number of state changes L, number of states S and minimum detectable change $\varepsilon_{min} > \Delta_{min}$ in mean reward, context size d, with $(1-p_\nu)$ and $0 \leq p_\nu \leq 1$ and constant $\gamma > 0$, is lower bounded by*

$$\mathbb{E}[R(T)] \geq \min\left\{\gamma\sqrt{Td}, \sqrt{SKT} + \frac{\Delta_{min}\sqrt{KT\tilde{L}}}{\varepsilon_{min}}\right\}.$$

These results lead us to the conclusion that there exists either a contextual bandit or multi-armed bandit algorithm that solves the problem setting perfectly. While an interesting result, it also highlights that there might not be

a combination strategy that achieves better results than any *tuned* single algorithm, at least in expectation over an infinite time horizon. Nevertheless, from a practical point of view, it is still more desirable to have an algorithm that adjusts its strategy depending on the current state of the problem setting. More specifically, an algorithm that can effectively choose between different strategies that perform best over some period of time. The tricky part here is how to optimally corral several algorithms to achieve this goal, which is an open question in the bandit literature. In this thesis, we have explored several combination strategies with promising results, although a finite time horizon analysis is reserved for future work.

3.2.2 Regret Upper Bound of COMBINE-UCB

Now we turn to the regret our algorithm archives in the problem setting. Our proposed algorithm combines two algorithms using a gradient bandit strategy. The upper bound of COMBINE depends on the algorithms used. We analyze COMBINE for the UCB instantiation, that is, COMBINE-UCB. It uses a gradient bandit to select between the LinUCB algorithm and a nonstationary UCB bandit algorithm we call NUCB for the remainder of this section. We carry out the analysis using the D-UCB algorithm for NUCB. D-UCB computes the discounted average action reward as $\bar{R}_t(a) = \sum_{\tau=0}^{t} \mathbb{1}(J_\tau = a)\gamma^{t-\tau} r_\tau(a)$, where J is the sequence of action choices up to time t, $\mathbb{1}$ is the indicator function and $\gamma \in (0,1]$ is the discount factor. D-UCB also discounts the number of action plays computed as $N_t(a) = \sum_{\tau=0}^{t} \mathbb{1}(J_\tau = a)\gamma^{t-\tau}$ [74]. While the multi-armed bandit in our implementation does not discount the number of action plays, it is possible to modify the exploration parameter α_B to a time-dependent version to achieve the same behavior as D-UCB (see Paper III for details). With this modified exploration parameter, we can rely on previous results.

Theorem 3. *The regret of COMBINE-UCB, using exploration parameter $\alpha'_B(t)$, is upper bounded with probability $1-\delta$ by*

$$\mathbb{E}[R(T)] \leq \sqrt{(\tau_1) d \log^3(KT\log(\tau_1)/\delta)} + \mathbb{E}[\beta](T-\tau_1)^{(1+\eta)/2}\log(T-\tau_1),$$

with

$$\mathbb{E}[\beta] \leq K.$$

From this analysis, COMBINE using D-UCB matches the lower bound up to logarithmic factors when the gradient bandit settles on either *LinUCB* or *NUCB* eventually. Otherwise, if a dynamic equilibrium between the two policies is reached, we get an upper bound that matches the lower bound up to an additional $\sqrt{\mathbb{E}[\beta]}$ factor, being $\sqrt{K}$ in the worst case.

3.3 Exploiting Prior Collected Data: The Latent Bandit Problem

The problem with contextual bandits is the need for learning from scratch, albeit there exist some ad-hoc methods for policy initialization, such as optimistic value initialization for UCB-type algorithms or enforcing a prior distribution on Thompson Sampling algorithms. The framework does not explicitly account for the use of previously collected data to choose actions with higher rewards while avoiding noncompetitive actions deliberately. In mHealth, we would like to minimize the time it takes to choose from a set of "good enough" interventions while continuing to personalize and improve recommendations over time. To facilitate this vision, it is helpful to think about this problem from the perspective of utilizing available information to inform initial policy generation. Given some amount of offline data is available to construct suitable environmental and reward models, for example, from a user study with non-personalized interventions or simply by defining expert-driven rules that can be incorporated into a suitable reward model, the agent can use these models to more deliberately choose from a set of good interventions and avoid learning from scratch.

The problem can be more formally defined under the so-called *latent bandit* problem. An additional property of this setting is the dependence of the reward distributions on a hidden underlying state. An extension to the latent bandit setting considers the possibility of changing states studied under the *non-stationary latent bandits* introduced by [75]. It is quite common in practice that the latent state is subject to change, abrupt or gradual, as time progresses, e.g., a patient might be interested in a particular set of interventions for a shorter or longer period.

We formally defined the non-stationary latent bandit problem.

The (non-stationary) *latent bandit* [75] is an online learning problem with *bandit feedback*. That is, only the reward of the chosen action is revealed to the agent. The process goes as follows. At every time step $t = 1, 2, \ldots, n$:

1. the agent receives context $X_t \in \mathcal{X}$.

2. the agent chooses an action $A_t \in \mathcal{A}$ according to its *policy*, mapping *history* $\mathcal{H}_t = (X_1, A_1, R_1, \ldots, X_{t-1}, A_{t-1}, R_{t-1})$ and context X_t to actions in $\mathcal{A}$.

3. the environment reveals the reward $R^{A_t} \in \mathbb{R}$ according to the joint conditional reward distribution $P(\cdot|A, X, S; \theta)$, parameterized by $\theta \in \Theta$, where Θ is the space of plausible reward models.

The initial latent state $S_t \in \mathcal{S}$ is sampled according to a prior distribution $S_1 \sim P_1(s)$ and evolves according to parameterized transition kernel $P(S|S_{t-1};\phi)$, with $\phi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, that maps the current state to a distribution over next states. Unlike the full reinforcement learning setting, the agent does not affect the environment dynamics through its actions. Only the current state determines the distribution over the next latent state. We define the mean reward of action A_t in context X_t and latent state S_t under the model parameters θ^*, as $\mu(A_t, X_t, S_t; \theta^*) = \mathbb{E}_{R\sim P(R|\cdot)}[R]$. Since the reward model is provided in the latent bandit setting, there are no strong assumptions placed on the form of $P(\cdot|A, X, S; \theta)$, which can be an arbitrarily complex function of θ and contexts X can come from an arbitrary process. We only assume that the rewards for a particular A_t, X_t and S_t are Gaussian distributed, with mean $\mu(A_t, X_t, S_t; \theta^*)$ and variance proxy σ^2.

The performance of a bandit algorithm is measured through the regret an agent incurs by choosing a sub-optimal action at time step t. Given the latent state s^*, we define $A_t^* = \text{argmax}_{a\in\mathcal{A}}\, \mu(a, X_t, s^*; \theta^*)$ as the optimal action as a function of context X_t and parameters θ^*.

We define the regret non-stationary latent bandit problem as: For a fixed latent state sequence $s_{1:n} \in \mathcal{S}^n$, the expected n-time step regret is defined as

$$\mathcal{R}(n, \phi^*, S_{1:n}) = \mathbb{E}\left[\sum_{t=1}^{n} \mu(A_t^*, X_t, S_t; \theta^*) - \mu(A_t, X_t, S_t; \theta^*)\right]. \tag{3.1}$$

We consider the Bayes regret, computing the n-time step regret as an expectation over latent state randomness. The n-time step Bayes regret for the non-stationary setting is defined as

$$\mathcal{BR}(n; \theta^*; \phi^*) = \mathbb{E}_{S_{1:n}\sim\phi^*}[\mathcal{R}(n, \phi^*, S_{1:n})|\theta^*, \phi^*]. \tag{3.2}$$

Note that the Bayes regret is a much weaker benchmark than the dynamic regret. It does not consider instances where our developed algorithm might perform very poorly. Nonetheless, it is an often realistic benchmark to compare against since we are more interested in a particular solution that performs well on average. Thus, most users should be fairly satisfied with the recommendation performance.

3.3.1 Actively Choosing the Action that Maximized State Discriminablity

In this example, we aim to demonstrate the advantages of using information-gathering techniques to identify states quickly. Consider a scenario where two

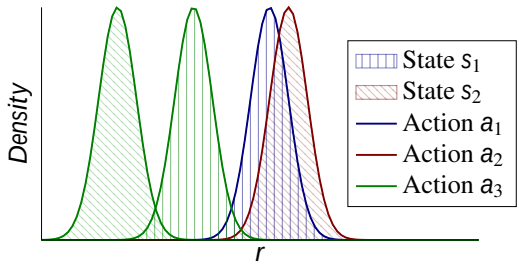


Figure 3.1: Qualitative example scenario where the highest-reward action is not ideal for quick identification of the state, leading to significant regret over longer periods of time.

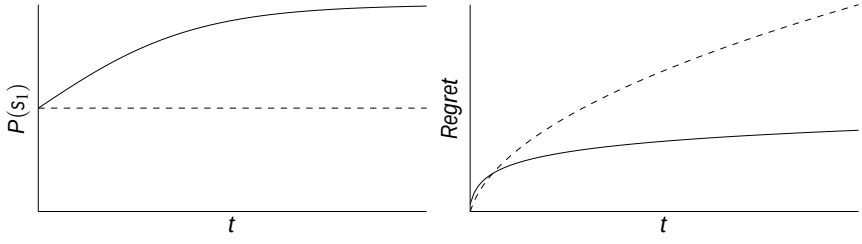


Figure 3.2: Belief state using cumulative regret from action a_1 and a_2 exclusively (dashed) and using action 3 occasionally (solid). Information-gathering helps to reduce regret through better state identification.

states determine the expected rewards for movies and the movies are categorized into three genres. One genre has high rewards in both states, while another genre has high rewards in one state and low rewards in the other. The agent operates under the assumption of a stationary latent bandit setting, where it knows the reward model but not the underlying state.

Previous approaches suggested considering movies that would score highly in both states. However, this strategy takes a long time to identify the true state despite minimizing regret in the short term. The agent faces difficulty in discerning the latent state from the received rewards. Uncertainty in the agent's belief state leads to sub-optimal movie recommendations given the true state, resulting in higher cumulative regret.

To illustrate this, we present reward distributions for the movie genres. Movies from genres 1 and 2 have similar rewards, while genre 3 exhibits significant differences between the two states shown in figure 3.1. From the agent's perspective, feedback from genres 1 and 2 could originate from either state. The agent's belief uncertainty often leads to continuously recommending sub-optimal movies from genre 2, despite the overall high cumulative reward shown in figure3.2.

Exploiting Structure: Resolving State Confusion

The principal idea of using *informative actions* to resolve state confusion can be attributed, at least in some part, to contemporary methods, particularly regarding *Information-directed sampling* in the bandit literature. Information-directed sampling (IDS) has been presented as an alternative approach to commonly used upper-confidence-bound and posterior sampling techniques to solve the exploration-exploitation trade-off in bandit problems [76]. The strategy involves minimizing the ratio between squared expected single-period regret and a measure of information gain. IDS has outperformed UCB and TS strategies in environments where knowledge about the reward gained from one action informs the reward of other actions. This general idea can be exploited to resolve state confusion in the following way. We compute a metric akin to the information ratio. When gathering information, the algorithm chooses the action that maximizes the ratio.

$$\Psi_{max} = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \frac{\bar{D}_{KL}(a)}{\overline{\Delta R}^2(a)} \tag{3.3}$$

For normally distributed action rewards, the KL-divergence $D_{KL}(P||Q)$ is defined as

$$D_{KL}(\mathcal{N}(\mu_1,\sigma_1)\,||\,\mathcal{N}(\mu_2,\sigma_2)) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2},$$

where μ_1 and μ_2 are the mean rewards and σ_1 and σ_2 are the reward standard deviations of the actions to be compared. Naturally, we can do pairwise comparisons between all actions and compute an average $\bar{D}_{KL}(a)$ for each action, encoding how "different" the action's reward distribution is compared to other actions.

Ψ_{max} encodes the intuition of the usefulness of an action for information-gathering. Only choosing the action that maximizes this ratio is insufficient since such a strategy is too myopic and may not work well for our problem setting. We would select an action with a sub-optimal reward too often, where it might not be necessary. The possible solution is to consider future time steps and ask: Am I better off with or without information-gathering? We can answer this question more precisely in the following way. Let a^e be the action that maximizes Ψ_{max} if a^e differs from the action a_t chosen by maximizing over the belief state. Then, compute the future reward of both a_t and a^e and choose the action that maximizes the reward over t' time steps.

This procedure is the core of our algorithm, which we call Active Greedy Exploration Model-Based Thompson Sampling (AGEmTS). AGEmTS extends the state-of-the-art posterior sampling algorithm mTS to take into account the relationship between reward distributions as well the transition model to choose action more deliberately and gain a significant performance advantage. The details of the algorithm can be found in Paper IV. How much we can expect our algorithm to generalize is discussed in the next section.

3.3.2 Theoretical analysis of AGEmTS

It is important to evaluate if the extension to the original mTS algorithm still retains sub-linear regret while at the same time improving regret significantly. The full proof is given in paper IV, but here we provide a shorter proof.

It is instructive to think about the problem from the stationary perspective before looking at the non-stationary setting. Under some conditions, AGEmTS may behave the same as mTS. We may not need to sample an information-gathering action if no significant state confusion exists. The two cases where information-gathering is unnecessary are

1. AGEmTS did not choose an information-gathering action. The states are challenging to distinguish. Over many samples, the agent has settled on the true latent state.

2. The reward distributions of actions are significantly different between states. Thus it is easy to distinguish them.

For the second case, AGEmTS will see no benefit in selecting an information-gathering action in most cases since the agent is never significantly confused. Suppose by an unlikely sequence of rewards. The agent cannot accurately determine what state the environment is in. In that case, it may not gather information if the potential loss in reward is not justifiable for the amount of entropy reduction in the belief state. Thus, the agent reverts to the behavior of mTS. Now, what about the event our estimation fails? AGEmTS may sample an information-gathering action unnecessarily, and we incur additional regret.

The idea here is to bound the probability of the estimation failing. While the information structure of the problem seems to be quite complex, previous work utilizes so-called martingale difference sequences that allow estimating the probability that the sum of random variables strays from its expectation by some number ε, provided that the random variables form a martingale difference sequence. More formally, this is stated in Proposition 1.

Proposition 1. *Let $(Y_t)_{t\in[n]}$ be a martingale difference sequence with respect to filtration $(\mathcal{F}_t)_{t\in[n]}$, that is $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0$ for any $t \in [n]$. Let $Y_t|\mathcal{F}_{t_1}$ be σ^2-sub-Gaussian for any $t \in [n]$. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\sum_{t=1}^{n} Y_t\right| \geq \varepsilon\right) \leq 2\exp\left[-\frac{\varepsilon^2}{2n\sigma^2}\right].$$

The task is to cleverly define the random variable $Y_t(s)$ so we have a martingale difference sequence and can apply **Proposition 1** for our analysis. This can easily be achieved by defining $Y_t(s) = \mu(A_t, X_t, s_I; \theta^*) - R_t$, where R_t is the realized reward at time step t. This definition does exactly what we need. We compare the sum of realized rewards to the sum of the expected reward, forming a martingale difference sequence. There is another technicality that is important for the proof. We need Y_t to be $\sigma^2-\text{sub}-\text{Gaussian}$. Intuitively, the reward follows a sub-gaussian distribution with a tail-decay at least as fast as a normal Gaussian. Note that according to this definition, a Gaussian is also a sub-Gaussian. This requirement of sub-Gaussianity is only a small obstacle, and in fact, it is commonly assumed in the literature that rewards are sub-Gaussian distributed.

Without going further into technicalities, the next steps are the following. Defining Y_t allows us to estimate the probability that the estimated total cumulative reward deviates from the empirical sum by some ε. This tool quantifies

the probability of making a mistake and AGEmTS commits to information gathering. As soon as we have this probability, we can judge if estimation failure will have a significant impact on the regret bound of our algorithm.

To continue the intuitive proof, let's look at the event that Y_t is close to 0 with some confidence radius δ and define all time steps where this is true as E_t. We can similarly define all time steps where this is not true as event $\bar{E}_t$, according to Proposition 3 by [75], The probability of $\bar{E}_t$ is $P(\bar{E}) \leq 2|\mathcal{S}|n^{-1}$. This probability is already relatively low and decays even more as the number of time steps increases. Further, note that AGEmTS does two reward estimations. We make an error if the confidence intervals of both expected reward estimates overlap, and we cannot be confident that the expected improvement over vanilla mTS is achieved with high probability. Thus, we bound the probability that both estimations exceed their confidence radius of δ. Both of these estimations are independent of each other. Thus, we have the probability that both events will occur:

$$\mathbb{P}(\bar{E}_{mTS} \cup \bar{E}_{AGEmTS}) = \mathbb{P}(\bar{E}_{mTS})\mathbb{P}(\bar{E}_{AGEmTS}) \leq 4|\mathcal{S}|^2 n^{-2}.$$

This probability is exceedingly low and does not contribute significantly to the regret for a reasonable time horizon n. There is something to be said about the confidence δ needing to increase as n increases to ensure that the results of **Proposition 3** hold. In practice, we have noticed that a dynamic confidence radius leads to very conservative behavior where AGEmTS rarely does information gathering, so constant thresholds did result in better regret - A small cautionary tale of the difference between theory and practice. The next step is to estimate the potential gain when using information gathering, essentially estimating the problem-dependent constant that allows AGEmTS to outperform mTS. The proof is quite technical, but it boils down to the following. The roll-out procedure has an upper bound on how often an information-gathering action is selected. AGEmTS chooses information-gathering if

1. The entropy of the belief state is $H(P_t) \geq 1$.
2. The expected regret gain $R^{ig} - R^{ps} > R_u$.
3. Information gathering is simulated during each roll-out to compute R^{ig} if and only if $R_t^{ig} - R_t^{ps} > R_u$ at current time step t, which further limits information gathering.

Event 1. is the most important for our analysis since AGEmTS may never select an information-gathering action if the condition is not met, thus determining the maximum number of information-gathering action selections. Using the posterior over the belief state, we can estimate the expected change in the direction of the true latent state via a differential equation.

$$\frac{\delta P_t(s_0,a_t)}{\delta t} = \frac{P_t(s_0)P(r_{a_t}|a_t,x_t,s_0;\theta^*)}{P_t(s_0)P(r_{a_t}|a_t,x_t,s_0;\theta^*)+(1-P_t(s_0))P(r_{a_t}|a_t,x_t,s_1;\theta^*)} - P_t(s_0). \tag{3.4}$$

Conversely $\frac{\delta P_t(s_1,a_t)}{\delta t} = -\frac{\delta P_t(s_1,a_t)}{\delta t}$ for state s_1.

Now, this equation is a difficult-to-solve nonlinear first-order differential equation. The idea here is to linearly approximate it, which results in a much easier-to-solve linear first-order differential equation. Due to approximations, we get a slightly worse problem-dependent constant than the algorithm might actually exhibit, a caveat we must live with until an improved analysis.

In the end, the regret-upper bound for the stationary setting, including the problem-dependent constants, can be estimated as

$$\int_{t=0}^{n-1} 1-\tilde{P}(s_0,a_t)\,dt = \int_{t=0}^{n-1} \frac{e^{-t\tanh(d/2)}}{e^c+1}\,dt \tag{3.5}$$

$$= \frac{\coth(d/2)(1-e^{-(n-1)\tanh(d/2)})}{1+e^c}. \tag{3.6}$$

The next step is to analyze the more interesting, non-stationary setting. The analysis shifts slightly toward the scenario that we have to deal with several stationary segments instead of only one segment. From the perspective of figuring out the regret of AGEmTS, it is interesting to consider the scenario where we sample an information-gathering action one time step before the environment switches to another state. Thus, the information we have gathered is completely outdated, and we may need to wait until we reach the uniform belief state again to sample another information-gathering action. Naturally, this is the worst-case scenario, where we have invested resources to gather information but essentially have no payout, and even worse, be very confident in our current belief state and choose detrimental actions. In this worst-case scenario, we would like to know how much additional regret we can expect when making such a mistake using AGEmTS. We can perform a similar analysis as in the stationary case using the linearly approximated derivative of belief state posterior, computing the number of time steps we select the wrong state as

$$N(s_0) = \int_{t=0}^{-\frac{\log(\frac{1}{2\varepsilon})+e^d\log(\frac{1}{2\varepsilon})}{1-e^d}} 1-\tilde{P}_t(s_0)\,dt = \frac{1}{2}\coth(d/2)\Big(2\varepsilon+2\log\big((2\varepsilon)^{-1}\big)-1\Big). \tag{3.7}$$

As an example, if we let $\varepsilon = P_0(s_0) = 0.001$, $\Delta\mu_{s_1} = \Delta\mu_{s_0} = 0.05$, $\sigma = 0.5$, we have $t \approx 1140$. The time to reach a uniform belief state can be several times longer than the expected length of stationary segments. These results shake our

belief in the algorithm, but remember: This happens if we are very unlucky with our timing regarding the information-gathering action. More typically, AGEmTS should resolve confusion before the state switches, and we can gain an advantage. To prove this, we can now include the influence of the transition matrix, which was omitted due to the assumption of stationarity. Again, we use the derivative of the state belief to estimate the number of time steps we need to reach the uniform belief state. To allow a closed-form solution, we turn again towards approximation. Since the transition matrix is assumed to be fixed, we can approximate the derivative of the belief state in the following: $\delta P_t^m(s_0)/\delta t = P_t(s_0)P(s_0|s_0) - P_t(s_0)$. This is an interesting approximation since it would give us the upper bound on the expected number of time steps before a uniform belief state is reached as $t = -\log(2)/\log(p_{s0})$, which is $\geq \log(2)$ as much as the expected time steps between transitions. So far so good, but when does this approximation hold? Using a convexity argument and letting $P(r_t|a_t, x_t, s_0; \theta^*) = q$ and $P(r_t|a_t, x_t, s_1; \theta^*) = p$, we can compute:

$$\frac{\frac{1}{2}(P(s_0|s_0)q + P(s_0|s_1)p)}{\frac{1}{2}(q+p)} - P_t(s_0) = \frac{1}{2}P(s_0|s_0) - P_t(s_0) \tag{3.8}$$

$$\frac{2(P(s_0|s_0)q + P(s_0|s_1)p)}{(q+p)} = P(s_0|s_0) \tag{3.9}$$

$$\Rightarrow \frac{q}{p} = 3 - \frac{2}{P(s_0|s_0)},\ p > 0 \wedge P(s_0|s_0) \geq \frac{2}{3}. \tag{3.10}$$

Given that the relatively mild conditions in equation 3.10 are satisfied, $\delta P_t^m(s_0)/\delta t$ is an upper bound in the interval $(0.5, 1]$. This result shows that the number of time steps needed before an information-gathering action may be considered is bounded by $-\log(2)/\log(p_{s_0})$. Thus, given that the state changes are not too fast, i.e., $P(s_0|s_0) \geq \frac{2}{3}$, the algorithm selects an information-gathering action quite some time before we expect a switch in the state, and thus, benefits from better state identification for action selection.

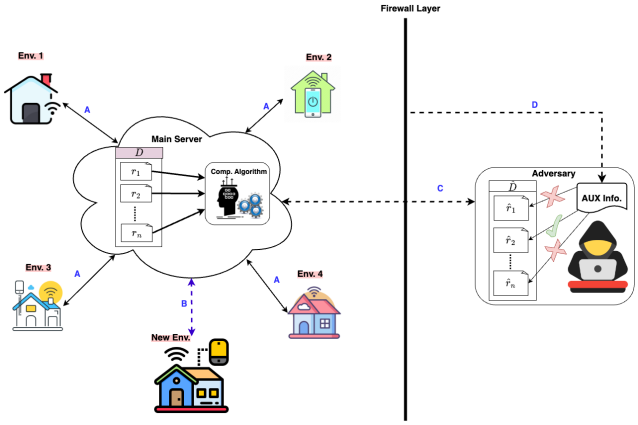


Figure 3.3: A high-level representation of the recommender system and adversary model.

3.4 Privacy aspects exemplified in the latent bandit setting

In this section, we outline the recommendation system and the linkage attack scenario. The system aims to provide personalized recommendations to users, such as music, movies, or mobile health interventions. Users receive recommendations from an agent, and the system adapts to their changing states.

Figure 3.3 demonstrates the information-sharing process among users, the learning setup, and the adversary. A centralized server handles the distribution of transition matrices, authenticating agent identities. Agents form a learning group and share anonymized transition matrices, forming dataset D (**A**). New agents initiate the learning process by requesting a suitable transition matrix from the server (**B**).

The adversary, pretending to be a new user, communicates with the server and obtains different transition matrices. This leads to the construction of an anonymized dataset $\hat{D}$ containing transition matrices of all users in the learning group (**C**). The adversary's goal is to match collected auxiliary information with the full transition matrices in $\hat{D}$ (**D**).

3.4.1 Data anonymization

Data anonymization employs various techniques to safeguard sensitive information while maintaining data utility. Common strategies that have been explored in the literature include randomization, K-anonymity [77], differential privacy [64], data masking and tokenization, generalization and suppression, and secure multi-party computation [78].

For the thesis, we are mainly concerned with strategies that manipulate

the data to ensure privacy. In particular: randomization; introducing noise or perturbation to the data, preventing the identification of individuals. Generalization and suppression: replacing specific values with more generalized or aggregated values, for example, replacing exact ages with age ranges or replacing precise geographic locations with broader regions. Specifically, we ask the question: given that an attacker has access to privatized data, how likely is it that an individual record is de-anonymized using publicly available data? This question is explored in the context of recommendation systems based on sharing some privatized data or models.

We implement the mechanisms of randomization by adding controlled noise to the data. Specifically, sensitive information is anonymized via a Laplace noise mechanism. More formally, we define the noisy version of the transition matrix as

$$\hat{\mathbf{r}} = \mathbf{r} + \mathbf{w} \cdot Z(\varepsilon), \forall \mathbf{r} \in D,$$

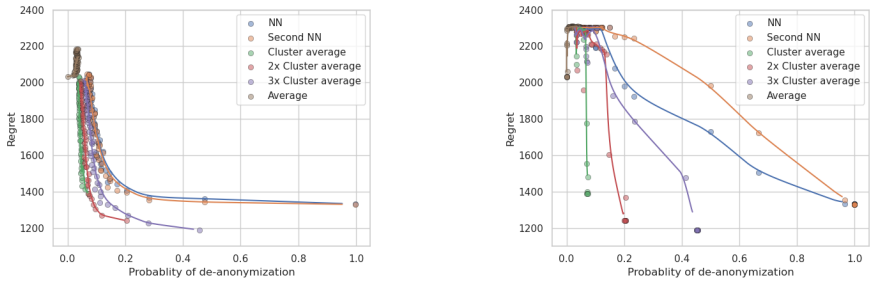
where Z is a random draw from a Laplace distribution with mean 0 and ε variance, and $\mathbf{w}$ represents the weight of the added noise to individual cells and can be chosen arbitrarily. Adding noise in such a fashion has the advantage of controlling the level of noise added to specific attributes. For example, if we define the weight $\mathbf{w}$ as

$$\mathbf{w} = (w_1, \ldots, w_d) = \big(1/H(X_i), \ldots, 1/H(X_d)\big),$$

with $w_i \in [0, \infty)$ and entropy $H(X_i) = \sum_{x_i \in \mathcal{X}_i} P(X_i = x_i) \log P(X_i = x_i)$, we can encode the concept of "uniqueness" regarding the cell's value distribution in the dataset. The more unique the cell value of a user's transition matrix, the more noise is added. This level of noise control is highly advantageous in finding an optimal trade-off between privacy and performance.

Another method for adding noise, albeit less controllable, is lossy compression. In our study, we look at truncated singular value decomposition (tSVD), widely used for matrix approximation. By considering only the top-k eigenvalues, tSVD compresses the information in the transition matrix while preserving essential information for approximate reconstruction. Specifically, tSVD solves the problem of approximating the transition matrix $\mathbf{r}$ with another matrix $\tilde{\mathbf{r}}$ by minimizing the Frobenius norm $||\mathbf{r} - \tilde{\mathbf{r}}||_F$ subject to the constraint that $\text{rank}(\tilde{\mathbf{r}})$ is equal to k [79]. Unfortunately, it is often not clear if sensitive information is effectively anatomized via lossy compression, leaving user data potentially open to re-identification attacks. Further, as demonstrated in our experiments, lossy compression often removes a significant amount of performance-relevant information, resulting in an unfavorable balance between privacy and regret.

Besides adding noise to individual records, we can employ methods of generalization and suppression via data aggregation. Averaging has the effect



(a) Laplace noise **(b)** tSVD transformation

Figure 3.4: De-anonymization vs. regret for the CASAS dataset. Legend: Nearest neighbor (NN), cluster average (Cluster average), global average (Average). Second nearest neighbor (Second NN). Double the number of clusters (2x Cluster average) and triple the number of clusters (3x Cluster average).

of hiding user-specific information since rare cell values are changed more toward the population average. Our results demonstrate that clustering offers a favorable balance between privacy and regret compared to nearest-neighbor approaches at a given privacy level. Clustering significantly enhances privacy, even if no noise is added to the cluster average, and in combination with noise, it outperforms all the other methods.

The results are shown in figure 3.4 for the CASAS dataset[80], where data was collected from several smart home apartments that contain ambient sensors. In general, no strategy we have investigated achieves the best privacy-performance trade-off over the complete range of de-anonymization probabilities. Thus, a combination of noise and aggregation proves necessary to assess for a given privacy level. The study's conclusion emphasized the need for a combination of aggregation strategies and noise to achieve the best privacy-performance trade-off for a given dataset. Our research emphasizes the importance of employing multiple privacy-preserving techniques in conjunction with data aggregation to enhance privacy guarantees in the era of big data.

4. Summary of Papers

4.1 Paper I: Pitfalls of medication adherence approximation through EHR and pharmacy records: Definitions, data and computation

Purpose and Conclusion We examined common pitfalls in refill adherence estimations using administrative databases from a theoretical and practical point of view, focusing on definitional and data-related aspects. We used data from a comprehensive EHR system that includes nearly complete prescribing and dispensing information on medication. Through appropriate experimentation, we show that slight changes in definition can lead to significant under or overestimation of refill adherence compared to the gold standard PDC measure. This has significant implications for, e.g., patient selection for interventions and may lead to false conclusions in real-world pharmacological studies. We analyzed different methodological and data-related issues, in particular for data-related issues. We investigated the effects of missing values, duplicate entries, and errors in data input, showing a small yet statistically significant effect on population averages and a large effect in individual cases.

4.2 Paper II: Prediction and pattern analysis of medication refill adherence through electronic health records and dispensation data

Purpose and Conclusion We investigated the predictability of refill mediation adherence under various data-splitting strategies using a comprehensive EHR system. Various machine learning algorithms were investigated under different predictive scenarios, with tree-based algorithms like Random Forest and Gradient Boosting Trees showing the highest performance. Predictive models have high discriminability on patients with high healthcare utilization, reaching AUCs of approximately 0.90 and 0.91 on the test sets using baseline predictors and baseline+history predictors, respectively. The models' lowest discriminability is observed in the most realistic scenario of forward-prediction of new prescriptions, with AUCs of 0.77 and 0.80 with baseline predictors and

baseline+history predictors, respectively. While model discriminability is relatively high, especially compared to previous studies, these models might still not be suitable for selecting patients for intervention since performance is quite low in "interesting" scenarios such as predicting a sudden change in adherence or predicting adherence of new patients that are at the beginning of treatment, where AUCs range between 0.56-0.65 are achieved on the test set.

We discovered common patterns of refill adherence for patients who start their anti-hypertensive treatment, that is, receive medication for the first time. We observe distinct patterns of refilling and, notably, the non-unique dependence of the second-year adherence trajectory on adherence in the first year. While patients are more likely to continue treatment after the first year, there is a non-negligible number of patients who discontinue treatment. While some patterns are more likely than others, this highlights the necessity to support patients early, even if they are refill adherent in the first year. We simulated medication taking and correlated the results with refilling patterns. Our simulations show that certain pathological medication consumption patterns are incompatible with the observed pattern of refill adherent patients, implying that nonadherence can occur suddenly without prior warning signs in refilling patterns. Furthermore, certain medication consumption patterns result in similar pickup patterns, obfuscating potential pathological patterns of medication-taking that can be relevant for intervention.

4.3 Paper III: A New Bandit Setting Balancing Information from State Evolution and Corrupted Context

Purpose and Conclusion We define and investigate a novel variant of the contextual bandit problem with corrupted context motivated by mHealth applications. mHealth is a challenging environment for reinforcement learning agents to perform well in, given that the information content provided by users is often missing, incomplete, and unreliable. Furthermore, keeping user engagement is paramount for the success of interventions, and therefore, it is vital to provide relevant recommendations on time. Additionally, users might transition through different treatment stages that require more targeted action selection approaches. The purpose of this study was to formulate a problem setting that would exhibit the aforementioned properties and give an algorithm that can learn and act more effectively than simpler solutions.

We develop a meta-algorithm called COMBINE, which uses a "referee" that dynamically combines the policies of a contextual bandit (CB), which uses a "context" or feature vector to make decisions, and a multi-armed bandit (MAB) which aims to find the best action irrespective of context. The multi-

armed bandit selects actions through a simple correlation mechanism that captures action-to-action transition probabilities, effectively learning a dynamics model of the environment and allowing for more efficient exploration of time-correlated actions than standard bandit algorithms such as LinUCB or LinTS. We empirically evaluate the performance of the developed algorithm on simulated and real-world data.

In most settings where the performance of the combined algorithms differ significantly, the COMBINE approach outperforms single methods, where adjusting to one policy over the other in the short term can result in a reduction in regret compared to using CB and MAB algorithms individually. Given that user might vary in their responses to surveys over time, using COMBINE will provide a significant advantage over simpler solutions. On simulated data, we observe that for the extreme of high action fluctuations and low context corruption or low action fluctuations and high context corruption, using a simple agent approach, i.e., either a CB or MAB, might minimize incurred regret. This highlights the necessity to find more efficient exploration and exploitation schemes when combing multiple bandit algorithms such that the overall regret is not significantly larger than the regret of the best single algorithm.

4.4 Paper IV: Information-gathering in Latent Bandits

Purpose and Conclusion We introduce information gathering in the latent bandit problem. In the latent bandit problem, the learner aims to identify latent states and make optimal choices based on reward distributions. Previous solutions focused on choosing high-reward arms without considering information-gathering arms. We propose a new method that strategically selects arms to improve state estimation and reduce cumulative regret. The method retains sub-linear regret rates and outperforms state-of-the-art algorithms in synthetic and real-world scenarios. The timing of selecting information-gathering arms is crucial, striking a balance between gaining the most benefit and avoiding unnecessary regret.

The proposed method addresses the latent bandit problem by incorporating information-gathering arms. These arms may not offer immediate high rewards but provide long-term benefits in state discrimination. By carefully selecting arms based on reward structures and transition matrices, the method improves state estimation and reduces regret. Synthetic experiments demonstrate the importance of timing in selecting information-gathering arms, as gathering information too early or too late can result in limited gains or unnecessary regret. The algorithm's performance is validated in various synthetic environments and real-world datasets, showcasing its superiority over existing methods.

The interplay between reward distribution and transition matrix is highlighted, indicating the potential gains achievable through information-gathering arms. In reducible Markov chains, knowing the current state can lead to significant advantages in selecting future states. Overall, the proposed method offers a promising solution for the latent bandit problem, providing a better understanding of the role of information-gathering arms and their impact on cumulative regret.

4.5 Paper V: Beyond Random Noise: Insights on Anonymization Strategies from a Latent Bandit Study

Purpose and Conclusion We explore privacy concerns in a learning scenario where users share knowledge for recommendation tasks. It emphasizes the need for tailored privacy techniques that address specific attack patterns instead of relying on generic solutions. The study utilizes the latent bandit setting to assess the privacy-recommender performance trade-off using different aggregation strategies, including averaging, nearest neighbor, and clustering with noise injection. We simulate a linkage attack scenario where the adversary uses publicly available auxiliary information.

The findings reveal that adding random noise to individual user data records, specifically using the Laplace mechanism, is ineffective as it results in high regret relative to de-anonymization probability. Instead, we suggest combining noise with appropriate aggregation strategies, such as utilizing averages from clusters of varying sizes, to achieve a better balance between privacy and performance.

The research highlights that relying solely on random noise for privacy may not be satisfactory, and aggregations generally offer a more favorable trade-off between privacy and downstream task performance. It emphasizes that there is no single technique that universally optimizes the trade-off for a desired level of privacy. Additionally, the study highlights the discrepancy between privacy metrics like ADS-GAN and the probability of de-anonymization in real-world linkage attacks. It underscores the importance of evaluating privacy techniques in attack scenarios in the wild, considering the abundance of auxiliary information that can undermine such methods' effectiveness.

5. Conclusion and Future Work

This thesis is the culmination of ongoing work in the domain of mHealth and was driven in part by the iMedA project that aimed to improve medication adherence through personalized adaptive interventions.

We described methods to measure medication adherence using EHRs and identified common pitfalls that can significantly affect the accurate estimation of MA. Secondly, we developed prediction models for MA and discussed the implications of using data from EHRs and pharmacy databases to conduct analysis and patient selection for interventions. While an important first step, estimating adherence using these data sources is necessarily incomplete. The actual level of medication adherence is unobserved and unavailable in the records, and we must be content with approximate methods using dispensations. These limitations will not only have significant implications on the accuracy and veracity of adherence values in adherence research. They also more broadly implicate a need for alternative potential proxy metrics for mobile Health applications where adherence is primarily self-reported and thus potentially unreliable. The issue of appropriate proxy metrics for adherence is an open problem and an avenue for future research.

Furthermore, many factors that influence medication adherence are insufficiently recorded in EHRs. This highlights the need to monitor adherence outside the clinical or primary care setting to identify patterns and reasons for nonadherence early. Here, mobile Health may provide tools and methods that allow the collection of additional information about the patient's daily life that could inform decision-makers to design appropriate interventions to improve outcomes. It is not clear how data from EHRs and Mobile Health applications may be used in conjunction to design better interventions and algorithms for their delivery. EHRs are usually low-frequency and high data heterogeneity sources with missingness not at random. At the same time, mobile health applications record data with high frequency using standardized data collection methods and inputs. Generating a common representation that allows for the design of better algorithms and interventions is an open problem for future research. Specifically: How can we efficiently integrate EHR and mHealth data in the intervention process and How to efficiently integrate expert knowledge for better initial policies?

We present a problem setting in sequential decision-making that exhibits

proprieties of the mHealth domain, particularly context uncertainty and non-stationarity. For mobile Health applications, it is imperative to provide good intervention in a timely manner to ensure high patient engagement. We investigated strategies that can utilize existing data to speed up the provision of "good enough" initial policies in the latent bandit framework, where online learning remains a critical component to adapt to changing patient preferences. But these are just the first steps. A more complicated setting that includes all relevant properties to mHealth, such as context uncertainty, non-stationarity, good initial policy, and fast online learning capabilities, requires further study from a theoretical perspective in the context of bandits as well as how to align such a setting to the practical realities to avoid or recover from model mismatch. Specifically, one might ask: How do we construct an RL or Bandit setting that is more realistic for mHealth applications and, at the same time, amenable to theoretical analysis? How do we address the model mismatch between synthetic environments and real environments? How can we more effectively exploit other data sources to improve the exploration and exploitation trade-off for faster learning?

Last but not least, we investigated important privacy concerns that will arise when such a recommender system is deployed in practice. Using a linkage-attack scenario, we have shown that contemporary privatization techniques might be ill-equipped to prevent an adversary from re-identifying an existing record using publicly available auxiliary information. Further, we exemplify the need for more sophisticated privatization strategies using the latent bandit setting to optimally solve the privacy-performance trade-off rather than simply adding noise to the records. It is not entirely clear how to optimally solve the privacy-performance trade-off for a given privacy guarantee in this setting. Future work may encompass more intelligent combinations of aggregation and noise strategies that adapt to more systematic changes in the environment, such as patients traversing different stages of behavior change or more tailored to individual variations within stage behavior. We may ask: Does an optimal strategy for the privacy-performance trade-off, given a certain privacy level, in the aforementioned recommender system exist? Can we devise an algorithm that matches or comes close to the regret lower bound? How can we address the issue of unreliable contexts and model mismatch and what are the fundamental performance limits?

References

[1] ANDREA M. BURDEN, J. MICHAEL PATERSON, ANDREA GRUNEIR, AND SUZANNE M. CADARETTE. **Adherence to osteoporosis pharmacotherapy is underestimated using days supply values in electronic pharmacy claims data**. *Pharmacoepidemiology and Drug Safety*, **24**:67–74, 2014. 2, 25

[2] A.V. PROCHAZKA J.F. STEINER. **The Assessment of Refill Compliance Using Pharmacy Records: Methods, Validity, and Applications**. *Journal of Clinical Epidemiology*, **50**:105–116, 1997. 2

[3] ALEXIS A. KRUMME, JESSICA M. FRANKLIN, DANIELLE L. ISAMAN, OLGA S. MATLIN, ANGELA Y. TONG, CLAIRE M. SPETTELL, TROYEN A. BRENNAN, WILLIAM H. SHRANK, AND NITEESH K. CHOUDHRY. **Predicting 1-Year Statin Adherence Among Prevalent Users: A Retrospective Cohort Study**. *Journal of Managed Care & Specialty Pharmacy*, **23**(4):494–502, 2017.

[4] J. M. FRANKLIN, W. H. SHRANK, J. LII, A. K. KRUMME, O. S. MATLIN, T. A. BRENNAN, AND N. K. CHOUDHRY. **Observing versus Predicting: Initial Patterns of Filling Predict Long-Term Adherence More Accurately Than High-Dimensional Modeling Techniques**. *Health Services Research*, **51**(1):220–239, Feb 2016.

[5] ALEXIS A KRUMME, GABRIEL SANFÉLIX-GIMENO, JESSICA M FRANKLIN, DANIELLE L ISAMAN, MUFADDAL MAHESRI, OLGA S MATLIN, WILLIAM H SHRANK, TROYEN A BRENNAN, GREGORY BRILL, AND NITEESH K CHOUDHRY. **Can purchasing information be used to predict adherence to cardiovascular medications? An analysis of linked retail pharmacy and insurance claims data**. *BMJ Open*, **6**(11):e011015, November 2016. 2

[6] F D R HOBBS. **Cardiovascular disease: different strategies for primary and secondary prevention?** *Heart*, **90**(10):1217–1223, 2004. 3, 13

[7] AMBUJ TEWARI AND SUSAN A. MURPHY. *From Ads to Interventions: Contextual Bandits in Mobile Health*. Springer International Publishing, Cham, 2017. 3, 13, 19, 20

[8] AWAIS ASHFAQ, STEFAN LÖNN, HÅKAN NILSSON, JONNY A ERIKSSON, JAPNEET KWATRA, ZAYED M YASIN, JONATHAN E SLUTZMAN, THOMAS WALLENFELDT, ZIAD OBERMEYER, PHILIP D ANDERSON, AND MARKUS LINGMAN. **Data resource profile: Regional healthcare information platform in Halland, Sweden, a dedicated environment for healthcare research**. *International Journal of Epidemiology*, 01 2020. 7

[9] P. A. GRADY AND L. L. GOUGH. **Self-management: a comprehensive approach to management of chronic conditions**. *American Journal of Public Health*, **104**(8):25–31, Aug 2014. 11

[10] C. DOWRICK, M. DIXON-WOODS, H. HOLMAN, AND J. WEINMAN. **What is chronic illness?** *Chronic Illness*, **1**(1):1–6, Mar 2005. 11

[11] S. S. MCMILLAN, E. KENDALL, A. SAV, M. A. KING, J. A. WHITTY, F. KELLY, AND A. J. WHEELER. **Patient-centered approaches to health care: a systematic review of randomized controlled trials**. *Medical Care Research and Review*, **70**(6):567–596, Dec 2013. 11

[12] E. BALINT. **The possibilities of patient-centered medicine**. *The Journal of the Royal College of General Practitioners*, **17**(82):269–276, May 1969. 11

[13] MARY E TINETTI AND TERRI FRIED. **The end of the disease era**. *The American Journal of Medicine*, **116**(3):179 – 185, 2004. 11

[14] R. M. EPSTEIN AND R. L. STREET. **The values and value of patient-centered care**. *Annals of Family Medicine*, **9**(2):100–103, 2011. 11

[15] G. BLUMROSEN, N. AVISDRIS, R. KUPFER, AND B. RUBINSKY. **C-SMART: Efficient seamless cellular phone based patient monitoring system**. In *2011 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, pages 1–6, 2011. 12

[16] A. MINUTOLO, G. SANNINO, M. ESPOSITO, AND G. DE PIETRO. **A rule-based mHealth system for cardiac monitoring**. In *2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 144–149, 2010. 12

[17] G. M. TURNER-MCGRIEVY, M. W. BEETS, J. B. MOORE, A. T. KACZYNSKI, D. J. BARR-ANDERSON, AND D. F. TATE. **Comparison of traditional versus mobile app self-monitoring of physical activity and dietary intake among overweight adults participating in an mHealth weight loss program**. *Journal of the American Medical Informatics Association*, **20**(3):513–518, May 2013. 12

[18] J. WEI, I. HOLLIN, AND S. KACHNOWSKI. **A review of the use of mobile phone text messaging in clinical and healthy behaviour interventions**. *Journal of Telemedicine and Telecare*, **17**(1):41–48, 2011. 12

[19] J. W. MCGILLICUDDY, A. K. WEILAND, R. M. FRENZEL, M. MUELLER, B. M. BRUNNER-JACKSON, D. J. TABER, P. K. BALIGA, AND F. A. TREIBER. **Patient attitudes toward mobile phone-based health monitoring: questionnaire study among kidney transplant recipients**. *Journal of Medical Internet Research*, **15**(1):e6, Jan 2013. 12

[20] PHILLIP OLLA AND CALEY SHIMSKEY. **mHealth taxonomy: a literature survey of mobile health applications**. *Health and Technology*, **4**(4):299–308, 2015. 12

[21] I. NAHUM-SHANI, S. N. SMITH, B. J. SPRING, L. M. COLLINS, K. WITKIEWITZ, A. TEWARI, AND S. A. MURPHY. **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, 05 2018. 12

[22] GILLIAN KING, MELISSA CURRIE, AND PATRICIA PETERSEN. **Child and parent engagement in the mental health intervention process: a motivational framework**. *Child and Adolescent Mental Health*, **19**(1):2–8, 2014. 12

[23] I. NAHUM-SHANI, E. B. HEKLER, AND D. SPRUIJT-METZ. **Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework**. *Health Psychology*, **34S**:1209–1219, Dec 2015. 12

[24] KAREN CHURCH, DENZIL FERREIRA, NIKOLA BANOVIC, AND KENT LYONS. **Understanding the Challenges of Mobile Phone Usage Data**. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '15, page 504–514, New York, NY, USA, 2015. Association for Computing Machinery. 12

[25] R. L. CUTLER, F. FERNANDEZ-LLIMOS, M. FROMMER, C. BENRIMOJ, AND V. GARCIA-CARDENAS. **Economic impact of medication non-adherence by disease groups: a systematic review**. *BMJ Open*, **8**(1):e016982, 01 2018. 13

[26] HAYDEN B. BOSWORTH, BRADI B. GRANGER, PHIL MENDYS, RALPH BRINDIS, REBECCA BURKHOLDER, SUSAN M. CZAJKOWSKI, JODI G. DANIEL, INGER EKMAN, MICHAEL HO, MIMI JOHNSON, STEPHEN E. KIMMEL, LARRY Z. LIU, JOHN MUSAUS, WILLIAM H. SHRANK, ELIZABETH WHALLEY BUONO, KAREN WEISS, AND CHRISTOPHER B. GRANGER. **Medication adherence: A call for action**. *American Heart Journal*, **162**(3):412–424, 2011. 13

[27] EDUARDO SABATE´ AND WOLRD HEALTH ORGANIZATION. **Adherence to long-term therapies: Evidence for action**, 2003. 13

[28] VINAY KINI AND P. MICHAEL HO. **Interventions to Improve Medication Adherence: A Review**. *JAMA*, **320**(23):2461–2473, 12 2018. 13

[29] SUSAN E. ANDRADE, KRISTIJAN H. KAHLER, FERIDE FRECH, AND K. ARNOLD CHAN. **Methods for evaluation of medication adherence and persistence using automated databases.** *Pharmacoepidemiology and drug safety*, **15**:565–574, 2006. 13

[30] WONG L CADARETTE SM. **An Introduction to Health Care Administrative Data**. *The Canadian journal of hospital pharmacy*, **68**:232–237, 2015. 13, 25

[31] P. MICHAEL HO, JOHN S. RUMSFELD, FREDERICK A. MASOUDI, DAVID L. MCCLURE, MARY E. PLOMONDON, JOHN F. STEINER, AND DAVID J. MAGID. **Effect of Medication Nonadherence on Hospitalization and Mortality Among Patients With Diabetes Mellitus**. *Archives of internal medicine*, pages 1836–1841, 2006. 14

[32] JASON YEAW, JOSHUA S. BENNER, JOHN G. WALT, SERGEY SIAN, AND DANIEL B. SMITH. **Comparing Adherence and Persistence Across 6 Chronic Medication Classes**. *Journal of Managed Care Pharmacy*, **15**:728–740, 2009. 14

[33] RICHARD BELLMAN. *Dynamic Programming*. Dover Publications, 1957. 15

[34] D. PRAKASH. **Target organ damage in newly detected hypertensive patients**. *Journal of Family Medicine and Primary Care*, **8**(6):2042–2046, Jun 2019. 15

[35] BEFEKADU ABEBE T. ABEGAZ TM, TEFERA YG. **Target Organ Damage and the Long Term Effect of Nonadherence to Clinical Practice Guidelines in Patients with Hypertension: A Retrospective Cohort Study**. *International Journal of Hypertension*, **2090-0384**(10):749, 2017. 15

[36] PASCAL C. BAUMGARTNER, R. BRIAN HAYNES, KURT E. HERSBERGER, AND ISABELLE ARNET. **A Systematic Review of Medication Adherence Thresholds Dependent of Clinical Outcomes**. *Frontiers in Pharmacology*, **9**:1290, 2018. 16

[37] MICHEL BURNIER. **Is There a Threshold for Medication Adherence? Lessons Learnt From Electronic Monitoring of Drug Adherence**. *Frontiers in pharmacology*, **9**, Jan 2019. 16

[38] DAVID SILVER, AJA HUANG, CHRIS J. MADDISON, ARTHUR GUEZ, LAURENT SIFRE, GEORGE VAN DEN DRIESSCHE, JULIAN SCHRITTWIESER, IOANNIS ANTONOGLOU, VEDA PANNEERSHELVAM, MARC LANCTOT, SANDER DIELEMAN, DOMINIK GREWE, JOHN NHAM, NAL KALCHBRENNER, ILYA SUTSKEVER, TIMOTHY LILLICRAP, MADELEINE LEACH, KORAY KAVUKCUOGLU, THORE GRAEPEL, AND DEMIS HASSABIS. **Mastering the Game of Go with Deep Neural Networks and Tree Search**. *Nature*, **529**(7587):484–489, jan 2016. 17

[39] ORIOL VINYALS, IGOR BABUSCHKIN, JUNYOUNG CHUNG, MICHAEL MATHIEU, MAX JADERBERG, WOJTEK CZARNECKI, ANDREW DUDZIK, AJA HUANG, PETKO GEORGIEV, RICHARD POWELL, TIMO EWALDS, DAN HORGAN, MANUEL KROISS, IVO DANIHELKA, JOHN AGAPIOU, JUNHYUK OH, VALENTIN DALIBARD, DAVID CHOI, LAURENT SIFRE, YURY SULSKY, SASHA VEZHNEVETS, JAMES MOLLOY, TREVOR CAI, DAVID BUDDEN, TOM PAINE, CAGLAR GULCEHRE, ZIYU WANG, TOBIAS PFAFF, TOBY POHLEN, DANI YOGATAMA, JULIA COHEN, KATRINA MCKINNEY, OLIVER SMITH, TOM SCHAUL, TIMOTHY LILLICRAP, CHRIS APPS, KORAY KAVUKCUOGLU, DEMIS HASSABIS, AND DAVID SILVER. **AlphaStar: Mastering the Real-Time Strategy Game StarCraft II**. `https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/`, 2019. 17

[40] OpenAI, :, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. **Dota 2 with Large Scale Deep Reinforcement Learning**, 2019. 17

[41] Thomas Dean, Leslie Pack Kaelbling, Jak Kirman, and Ann Nicholson. **Planning under time constraints in stochastic domains**. *Artificial Intelligence*, **76**(1):35 – 74, 1995. Planning and Scheduling. 18

[42] Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. **Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity**. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, **4**(1), March 2020. 19

[43] S. S. Villar, J. Bowden, and J. Wason. **Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges**. *Statistical Science*, **30**(2):199–215, 2015. 20

[44] Hamsa Bastani and Mohsen Bayati. **Online Decision Making with High-Dimensional Covariates**. *Operations Research*, **68**(1):276–294, 2020. 20

[45] Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. **Portfolio Choices with Orthogonal Bandit Learning**. In *International Conference on Artificial Intelligence*, IJCAI'15, page 974–980. AAAI Press, 2015. 20

[46] Xiaoguang Huo and Feng Fu. **Risk-aware multi-armed bandit problem with application to portfolio selection**. *Royal Society Open Science*, **4**(11):171377, 2017. 20

[47] Stefano Boldrini, Luca De Nardis, Giuseppe Caso, Mai Le, Jocelyn Fiorina, and Maria-Gabriella Di Benedetto. **muMAB: A Multi-Armed Bandit Model for Wireless Network Selection**. *Algorithms*, **11**(2):13, Jan 2018. 20

[48] R. Kerkouche, R. Alami, R. Féraud, N. Varsier, and P. Maillé. **Node-based optimization of LoRa transmissions with Multi-Armed Bandit algorithms**. In *2018 25th International Conference on Telecommunications (ICT)*, pages 521–526, 2018. 20

[49] Zheng Wen, Branislav Kveton, Michal Valko, and Sharan Vaswani. **Online Influence Maximization under Independent Cascade Model with Semi-Bandit Feedback**. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, **30**, pages 3022–3032. Curran Associates, Inc., 2017. 20

[50] Eric Schwartz, Eric Bradlow, and Peter Fader. **Customer Acquisition via Display Advertising Using Multi-Armed Bandit Experiments**. *Marketing Science*, **36**, 04 2017. 20

[51] Q. Wang, C. Zeng, W. Zhou, T. Li, S. S. Iyengar, L. Shwartz, and G. Y. Grabarnik. **Online Interactive Collaborative Filtering Using Multi-Armed Bandit with Dependent Arms**. *IEEE Transactions on Knowledge and Data Engineering*, **31**(8):1569–1580, 2019. 20

[52] Linas Baltrunas, Karen Church, Alexandros Karatzoglou, and Nuria Oliver. **Frappe: Understanding the Usage and Perception of Mobile App Recommendations In-The-Wild**. *CoRR*, **abs/1505.03014**, 2015. 20

[53] Huitian Lei, Ambuj Tewari, and Susan A. Murphy. **An Actor-Critic Contextual Bandit Algorithm for Personalized Mobile Health Interventions**, 2017. 21

[54] T.L Lai and Herbert Robbins. **Asymptotically efficient adaptive allocation rules**. *Advances in Applied Mathematics*, **6**(1):4 – 22, 1985. 21

[55] PETER AUER, NICOLO CESA-BIANCHI, AND PAUL FISCHER. **Finite-time Analysis of the Multiarmed Bandit Problem**. *Machine Learning*, **47**:235–256, 05 2002. 21

[56] WILLIAM R. THOMPSON. **On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples**. *Biometrika*, **25**(3/4):285–294, 1933. 21

[57] OLIVIER CHAPELLE AND LIHONG LI. **An Empirical Evaluation of Thompson Sampling**. In J. SHAWE-TAYLOR, R. S. ZEMEL, P. L. BARTLETT, F. PEREIRA, AND K. Q. WEINBERGER, editors, *Neural Information Processing*, pages 2249–2257. Curran Associates, Inc., 2011. 21

[58] RICHARD S. SUTTON AND ANDREW G. BARTO. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. 21

[59] AUER-D. HOFER D. MOHAMED, A.K.Y.S. AND J. KÜNG. **A systematic literature review for authorization and access control: definitions, strategies and models**. *International Journal of Web Information Systems*, **18**, 2020. 22

[60] ADIL HUSSAIN SEH, MOHAMMAD ZAROUR, MAMDOUH ALENEZI, AMAL KRISHNA SARKAR, ALKA AGRAWAL, RAJEEV KUMAR, AND RAEES AHMAD KHAN. **Healthcare Data Breaches: Insights and Implications**. *Healthcare*, **8**(2), 2020. 22

[61] KHALED EL EMAM, ELIZABETH JONKER, LUK ARBUCKLE, AND BRADLEY MALIN. **A Systematic Review of Re-Identification Attacks on Health Data**. *PLOS ONE*, **6**(12):1–12, 12 2011. 22

[62] VLADLENA BENSON, GEORGE SARIDAKIS, AND HEMAMAALI TENNAKOON. **Information disclosure of social media users**. *Information Technology & People*, **28**(3):426–441, Jan 2015. 22

[63] LATANYA SWEENEY. **Weaving Technology and Policy Together to Maintain Confidentiality**. *The Journal of Law, Medicine & Ethics*, **25**(2-3):98–110, 1997. PMID: 11066504. 23

[64] CYNTHIA DWORK, AARON ROTH, ET AL. **The algorithmic foundations of differential privacy**. *Foundations and Trends® in Theoretical Computer Science*, **9**(3–4):211–407, 2014. 23, 41

[65] CHUANYIN WANG, CUNQING MA, MIN LI, NENG GAO, YIFEI ZHANG, AND ZHUOXIANG SHEN. **Protecting data privacy in federated learning combining differential privacy and weak encryption**. In *Science of Cyber Security: Third International Conference, SciSec 2021, Virtual Event, August 13–15, 2021, Revised Selected Papers 4*, pages 95–109. Springer, 2021. 23

[66] L. M. HESS, M. A. RAEBEL, D. A. CONNER, AND D. C. MALONE. **Measurement of Adherence in Pharmacy Administrative Databases: A Proposal for Standard Definitions and Preferred Measures**. *Annals of Pharmacotherapy*, **40**:1280–1288. 25

[67] W. M. VOLLMER, M. XU, A. FELDSTEIN, D. SMITH, A. WATERBURY, AND C. RAND. **Comparison of pharmacy-based measure of medication adherence**. *BMC health services research*, **12**:155, 2012. 25

[68] KATHLEEN A. FAIRMAN AND BRENDA MOTHERAL. **Evaluating Medication Adherence: Which Measure Is Right for Your Program?** *Journal of Managed Care Pharmacy*, **6**:499–506, 2000. 26

[69] REDA ABOUSERIE. **Sources and Levels of Stress in Relation to Locus of Control and Self Esteem in University Students**. *Educational Psychology*, **14**(3):323–330, 1994. 27

[70] B. GARDNER, P. LALLY, AND J. WARDLE. **Making health habitual: the psychology of 'habit-formation' and general practice**. *Br J Gen Pract*, **62**(605):664–666, Dec 2012. 27

[71] DJALLEL BOUNEFFOUF. **Online learning with Corrupted context: Corrupted Contextual Bandits**. 2020. 28, 30

[72] ALEXANDER GALOZY, SLAWOMIR NOWACZYK, AND MATTIAS OHLSSON. **Corrupted Contextual Bandits with Action Order Constraints**, 2020. 28

[73] JAMES O PROCHASKA, COLLEEN A REDDING, KERRY E EVERS, ET AL. **The transtheoretical model and stages of change**. *Health behavior: Theory, research, and practice*, **97**, 2015. 29

[74] LEVENTE KOCSIS AND CSABA SZEPESVÁRI. **Discounted UCB**. *2nd PASCAL Challenges Workshop*, 2006. 31

[75] JOEY HONG, BRANISLAV KVETON, MANZIL ZAHEER, YINLAM CHOW, AMR AHMED, MOHAMMAD GHAVAMZADEH, AND CRAIG BOUTILIER. **Non-Stationary Latent Bandits**. *CoRR*, **abs/2012.00386**, 2020. 32, 38

[76] DANIEL RUSSO AND BENJAMIN VAN ROY. **Learning to Optimize via Information-Directed Sampling**. In Z. GHAHRAMANI, M. WELLING, C. CORTES, N. LAWRENCE, AND K.Q. WEINBERGER, editors, *Advances in Neural Information Processing Systems*, **27**. Curran Associates, Inc., 2014. 35

[77] PIERANGELA SAMARATI AND LATANYA SWEENEY. **Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression**. 1998. 41

[78] CHUAN ZHAO, SHENGNAN ZHAO, MINGHAO ZHAO, ZHENXIANG CHEN, CHONG-ZHI GAO, HONGWEI LI, AND YU AN TAN. **Secure Multi-Party Computation: Theory, practice and applications**. *Information Sciences*, **476**:357–372, 2019. 41

[79] IVAN MARKOVSKY. **Structured low-rank approximation and its applications**. *Automatica*, **44**(4):891–909, 2008. 42

[80] DIANE J COOK. **Learning setting-generalized activity models for smart spaces**. *IEEE intelligent systems*, **27**(1):32, 2012. 43

Paper I

Pitfalls of medication adherence approximation through EHR and pharmacy records: Definitions, data and computation

Alexander Galozy, Sławomir Nowaczyk, Anita Sant'Anna, Mattias Ohlsson, Markus Lingman

International Journal of Medical Informatics 2020.

Paper II

Prediction and pattern analysis of medication refill adherence through electronic health records and dispensation data

Alexander Galozy, Sławomir Nowaczyk

Journal of Biomedical Informatics 2020.

Paper III

A New Bandit Setting Balancing Information from State Evolution and Corrupted Context

Alexander Galozy, Sławomir Nowaczyk, Mattias Ohlsson

Submitted 2022.

Paper IV

Information-gathering in Latent Bandits

Alexander Galozy, Sławomir Nowaczyk

Knowledge-Based Systems 2023.

Paper V

Beyond Random Noise: Insights on Anonymization Strategies from a Latent Bandit Study

Alexander Galozy, Sadi Alawadi, Victor R. Kebande, Sławomir Nowaczyk

Submitted 2023.