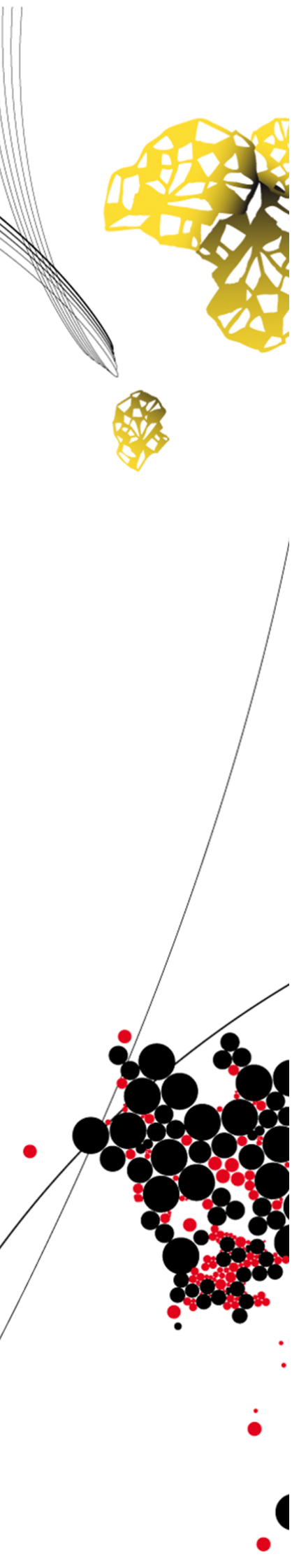


UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

User Interface Design Based on Human-Centered Explainable AI Methods

Zoe Zhang 2811960

M.Sc. Thesis, Interaction Technology (HCID)

October 2022

Supervisors:
dr. M. Theune
dr. A. Karahanoglu
dr. N. Strisciuglio

University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Acknowledgement

I would like to thank my supervisors Armagan Karahanoglu and Mariet Theune for their patience and detailed instructions.

I would like to thank the help from experts Shane T. Mueller, Robert Hoffman, Gary Klein, Martijn Redegeld, Nicolas Marini and Hans Nuijt.

I would like to thank all people who participated in the user research and user testing. This thesis cannot be done without your sincere and detailed feedback. Thank you for being kind and patient for so many times.

I would like to thank my best friends Sneha Ramesh, Sneha Borkar and Shyam Sundar who warm me with real friendship.

I would like to thank my parents for making me feel at home and lighting up my life.

Abstract

Artificial intelligence (AI) has demonstrated its considerable influence on every aspect of human life. However, algorithms are getting rather complex, and there are more black-box models as tasks that AI deals with increase and become more complex. Therefore, the eXplainable AI (XAI) attempts to solve this problem by making the algorithm understandable and trustworthy for human beings. Although numerous explanation methods are provided, most answer the Why question (why does the algorithm generate such decisions) from the technical experts' view. In contrast, the leading target group of explanations for AI is non-technical end users. As the calling for human-centered XAI gets stronger, researchers have proposed a set of requirements and design guidelines for human-centered XAI. However, these requirements merely stay at an abstract level and have not gone into a detailed design context. Only limited research provides clear guidance on how to implement and fulfil those instructions, nor does much research present the actual practice in design work. Moreover, the evaluation research for the current human-centered XAI still needs to be enriched. This thesis work cooperates with the company EatMyRide (EMR) which assists cycling enthusiasts in customizing and evaluating their nutrition plans. This research aims to make the working principles and algorithm of the application more understandable and trustworthy so that users will stick to this application. The main contribution of this work is that it will compensate for the deficiency of authentic practice and evaluation of human-centered XAI in an actual design context. There will be research on the current EMR application and interviews for its potential users to acquire more profound insights, especially regarding the aim of this thesis. Based on previous findings, the practice of human-centered XAI will be presented as new user interfaces in low-fidelity and high-fidelity prototypes, and user testing will be conducted to evaluate the effect of the design work. The design practice and evaluation based on the findings from the literature review and previous research are the main contributions to the current human-centered XAI field because it implements the XAI guidelines and evaluates the real effects of the actual practice. After that, the paper delivers discussions and conclusions regarding the research questions, the limitation of this thesis work, and insights and suggestions for future explorations.

Contents

List of acronyms

AI	artificial intelligence
XAI	explainable artificial intelligence
ML	machine learning
DARPA	Defense Advanced Research Projects Agency
FAT	Fair, Accountable, and Transparent algorithms
LIME	local interpretable model-agnostic explanations
SHAP	Shapley additive explanation
UI	user interface
UX	user experience
lo-fi	low-fidelity prototype
hi-fi	high-fidelity prototype
EMR	EatMyRide
RQ	research question
HR	heart rate
ATP	adenosine triphosphate
SD	standard deviation
ESS	Explanation Goodness Checklist
ESS	Explanation Satisfaction Scale
ETS	Explanation Trust Scale

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Artificial intelligence (AI) gradually shows its importance in numerous aspects of human society [1]–[8]. The need for explaining the machine learning (ML) models and algorithms has grown dramatically in past decades because people are reluctant to accept the result if there is a lack of transparency, interpretation and trust in those models [9]–[11]. Therefore, the concept and aim of eXplainable AI (XAI) have been proposed by numerous scientists, and the research on XAI has increased rapidly in past decades [12]–[17]. Many XAI methods such as local interpretable model-agnostic explanations (LIME) [18] and Shapley additive explanation (SHAP) [19] are trying to solve the Why question (why does the algorithm/system generate such decision) in an algorithmic and expert perspective by providing the importance and influence level of each feature input, which is not friendly and easy for non-technical lay users [10], [20], [21]. Therefore, the calling for human-centered XAI has developed fast in recent years [22]–[24]. Many principles, regulations and design guidelines are proposed to fill the gap and formulate the outline. However, these requirements merely stay at an abstract level and have not dived deep into the fundamental design context. For example, there is only limited research that provides detailed guidance on how to implement and fulfilled human-centered XAI requirements, nor does much research present the actual practice or evaluation of human-centered XAI in design work [14], [25]–[28]. It can be concluded that current human-centered XAI research still lacks actual case practice and evaluation. Therefore, the motivation of this thesis research falls into the scope of implementing the human-centered XAI instructions into a real design case with corresponding evaluation methods to measure its effectiveness and trustworthiness.

1.2 Goals of the Assignment

This thesis research aims to fill the gap between theory and practice, practically implementing the human-centered XAI guidelines into a tangible design context. It attempts to fulfill the requirements of human-centered XAI proposed in past studies to improve the overall user experience (UX), trust and understanding of a product. Besides, the research will evaluate the performance of human-centered XAI after finishing the design work. This thesis cooperates with the company EatMyRide (EMR) which assists cycling enthusiasts in personalizing and evaluating their nutrition plans. The requirement from the company is to design the user interface (UI) for the EMR application based on XAI principles so that users can understand the reason behind recommendations from the system. The logic is that by implementing XAI approaches, users will be more inclined to provide valuable and precise data to the EMR application, which will benefit both the company and users. On one side, cyclists will get more personalized and suitable recommendations to improve their cycling performance and health condition. Conversely, the company will get more detailed and valuable data to improve the algorithms and UX.

1.3 Research Question

How do we implement human-centered XAI methods in a practical design case to match users' needs, and improve their understanding and trust in the algorithm?

1. How to figure out users' perception of the current EMR application?
2. How to enhance or correct users' current understanding of the EMR application?
3. How to evaluate the results and effects after implementing the human-centered XAI?

1.4 Report Organization

The thesis report consists of six chapters and three appendixes. Chapter 2 provides a detailed description and analysis of the XAI with a conclusion of methods that will be used in the later research. Chapter 3 analyzes the current state of the EMR application, which includes research on the EMR application, and the expert, user, and target mental model analysis. This chapter also answers Research Question 1. Chapter 4 and 5 include the low-fidelity and high-fidelity prototype design and

evaluation, which answers Research Question 2 and 3. Chapter 4 focuses more on the human-centered XAI design, whereas chapter 5 addresses the evaluation. Both chapters include the design and evaluation results and analysis. Chapter 6 presents the overall answers to research questions and the analysis of the limitation of this thesis research. Lastly, Chapter 7 concludes the work and provides some suggestions for future research in this direction.

Chapter 2

Literature Review

This chapter presents a literature review on current XAI and its evaluation methods, which includes 1) the definition of explanation 2) the reason why explanation is needed 3) the reason why current AI applications lack explanation 4) the aims and principles of XAI 5) different XAI methods and the way to approach them. Besides, the literature review also includes an elaboration of human-centered XAI precisely. It contains the definition and methods to achieve human-centered XAI, followed by evaluation approaches and practical design guidelines. Finally, a conclusion will summarize design guidelines and evaluation methods that will be practiced in later research.

2.1 Background

AI has demonstrated its increasing importance in different fields such as healthcare [1], [2], finance [3], military [4], [5], legal [6], transportation [7], [8] by solving complex problems, which makes it indispensable for human society [29]. However, the sophistication of AI algorithms prevents users from figuring out their inside working mechanism, which causes security and trust problems in highly concerning areas mentioned above. As humans are reluctant to accept techniques that are not transparent, interpretable and trustworthy [10], the requirement of ethical AI is in growing demand and the concept of XAI is proposed to solve this problem [9]. However, only a few prototypes have implemented this theory during the past decades, while there are numerous proposals for making AI more transparent [30], [31]. The reason why current products lack the interest to put explanations into effect could be argued as follows [14]:

- It is hard to implement abstract-level XAI guidelines into variant and complex real-world scenarios

- Some XAI requirements are contradicted by real-world situations. For example, it is hard for mobile applications to merge many explanations within a limited space

- There is a lack of guidance about how to integrate explanations to already-existing UIs

- Many companies tend to use opaque algorithms such as deep learning

2.2 Explainability

There are various definitions of explainability and descriptions of what makes a good explanation from both algorithm and social aspects. Miller [32] summarizes that explanation is capable of enabling users to simplify and narrow down their observations and facilitating users to build a general model for future repetitive use. Abdul et al. [13] define explainability as transparency, interpretability, trust, fairness and accountability. However, the terms interpretability and explainability of AI are often misused in different literature [33]. Interpretability is more passive, meaning the AI Model is understandable for humans. In contrast, explainability is more active, denoting behaviors or actions executed by the model to explain to users what is happening. Moreover, Gilpin et al. [34] argue that explainability is broader than interpretability. They prompt that interpretability is the ability that systems can gain users' trust and provide users with the causes of some actions, whereas explainability gives users space to interact, for example, by answering specific questions and providing underlying reasons. It is not hard to find that an XAI system not only includes explainability, but also covers interpretability simultaneously, meaning that both active interactions and passive explanations exist in a single system.

To a more social-science and theoretical level, Miller [32] takes explanation as an answer to a WHY question, which fits the theory put up by Halpern and Pearl [35] that a good explanation should provide 1) users new knowledge 2) the possible reason of a specific result. Beyond this, Miller [32] also concludes that good explanations should 1) be contrastive (why doing X instead of Y) 2) have preferences (explanations should be selected because not all principles should be explained) 3) lean to socialization. Taking the second feature *preference* further, explanations that are too complex are improbable to be accepted by users, according to the research by Herlocker et al. [15]. Like Schaffer et al. [36] who conclude that numerous detailed explanations negatively influence users' confidence, patience and overall experience because users already have a certain level of trust for AI. The same opinion put up by Doshi-Velez and Kim [37] that explanation is only needed when

there is a mismatch between users' perception and results from the system, or the explanation lies in a new field and has a profound impact.

2.3 Goals and Requirements of XAI

The XAI need some clear goals in order to have basic principles of dos and don'ts. According to Dosilovi et al. [11], since there is a trade-off between the performance and transparency of AI models, which means decreasing the complexity will generally lower the accuracy of the AI, the XAI is prompted to leverage the current situation with two aims that are 1) improving the explainability of AI models while preserving their accuracy, and 2) making sure humans (especially non-technical users) understand how the model works and trust the results provided by algorithms.

From an abstract and algorithmic point of view, Samek et al. [38] point out that XAI should let people quickly figure out underlying working principles of the algorithm by extracting essential knowledge from AI under regulations from laws. It stresses that only necessary knowledge should be explained in an easy-to-understand way for users to learn the working principles of the system.

Other than goals and principles on the AI expert level, prior work simultaneously pays attention to the XAI from a social aspect. Work of Wachter et al. [39] describes the primary aims of XAI that are 1) "to inform and help the subject understand why a particular decision was reached" 2) "to provide grounds to contest adverse decisions" 3) "to understand what could be changed to receive a desired result in the future, based on the current decision-making model." Meanwhile, FAT(Fair, Accountable, and Transparent algorithms) academics also focus on the lay users of AI products. It gives the explainability in ML that "is to ensure that algorithmic decisions, as well as any data driving those decisions, can be explained to end-users and other stakeholders in non-technical terms." Same as the aim of the XAI given by D. Gunning is that "XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" [40]. The insightful point is that the XAI should be able to correct non-technical users bias and inform them how to get a better result in the future practice.

However, although some work notices the importance of XAI at the society level, current existing works still primarily focus on explanations generated by an algorithm to technical-background staff instead of applying a user-friendly approach that addresses usability, interpretability and understandability for lay users. Most of whom could have a less or non-technical background in the actual operational context, nor do they provide the proof and evaluation for fitting the real-world user tasks [10], [20], [21]. There is a growing demand to include more human factors in current XAI

research, which means separating lay users from AI researchers and domain experts [22]–[24]. It can be seen that a lack of research exists in developing, applying, and evaluating designs for the XAI domain at the same time [16], [41], [42].

2.4 Taxonomy

Based on different attributes of ML models, researchers conclude various XAI methods based on three algorithmic taxonomies [26]:

- The model itself is interpretable or not (which indicates black-box or white-box model)
- Interpretation based on the global or local scope
- Model-specific or model-agnostic

For the first taxonomy, if the model itself is interpretable, it has high transparency, such as Decision Trees, K-Nearest Neighbor, Rule-based Inference, Bayesian Models, and Linear Regression. Users themselves can easily understand these models, and post-hoc methods like explanation by simplification, explanation by feature relevance, and visual explanation can also be applied to explain in more detail. For the black-box model, the model itself is opaque and too complex to understand, such as Support Vector Machine, Convolutional Neural Network, Ensemble Method, Recurrent Neural Network, and only post-hoc methods can be used.

For the second taxonomy, global interpretability indicates the whole logic and reason of the model, which can be implemented with different results. Local interpretability facilitates understanding some outcomes generated locally, and it cannot be applied to explain other possible results acquired from different places. Global model interpretation via recursive partitioning called (GIRP) was proposed by Yang et al. [43], and Nguyen et al. [44] proposed an approach based on activation maximization. For local interpretability, the well-known one named local interpretable model-agnostic explanations (LIME) as a surrogate model is proposed by Ribeiro et al. [18]. Another counterfactual approach is made by T. Miller [32], which is an example-based explanation method.

The last taxonomy focuses on whether the XAI method could be implemented on different ML models or only on the specific one. For the model-agnostic method, explanation by simplification approach like rule-based learner mishra2017local can be applied as well as Shapley additive explanation (SHAP), which is a game-theoretic approach that uses Shapley value to measure the importance of different features of the data [19].

Another algorithmic taxonomy mentioned by Zhu et al. [10] categorizes different algorithms into two dimensions: reactive or deliberative (See Fig. 2.1). They claim that the XAI in the reactive process is to inject explainable notes which are created by the deliberative process, and it provides post-hoc analysis at the same time. The XAI in the deliberative process is to make the model easier for humans to control. "It is a reduction, reorganization, or reframing of the complex into something understandable that maintains the transparency and introspection of the model." Overall, this taxonomy is relatively straightforward and can be taken as a detailed explanation of the taxonomy "The model itself is interpretable or not (which indicates black-box or white-box model)." Compared to the work done by Zhu et al., a detailed graph (See Fig. 2.2) prompted by Wang et al. [45] indicates not only different kinds of human and XAI reasoning but also the links and connections between them, which provides a more logical and clear picture of how XAI works.

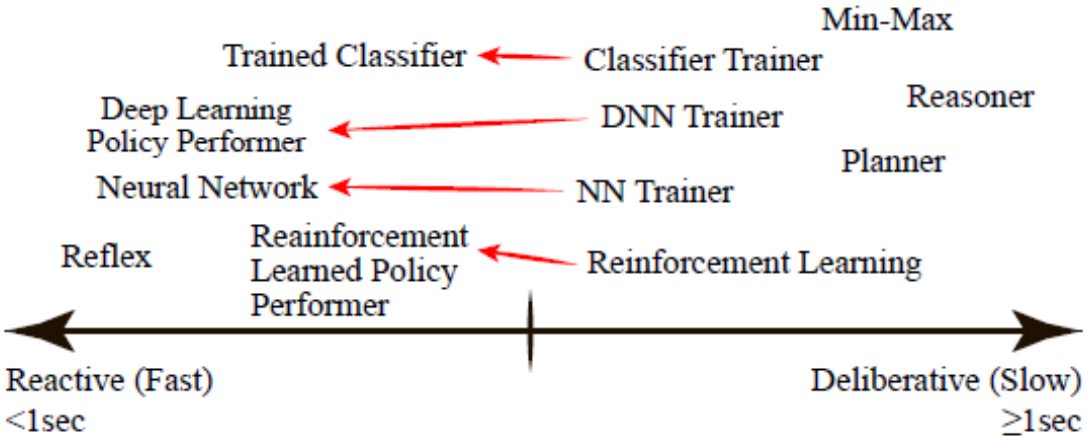


Figure 2.1: Mapping Deliberative AI/ML Techniques to Reactive Processes
Reproduced with permission [10] ©[2018] IEEE.

Apart from the taxonomy mentioned above, Miller's research [32] stresses the human factor and socialization in the XAI. The research work proposes three factors for explanations that 1) people usually ask why B does not happen instead of why A happens 2) explanations should focus on two or three causes instead of all the possible factors 3) there should be a balance between explanations generated by algorithms and users' mental perceptions. Miller's finding emphasizes the importance of the post-hoc explanation methods in XAI, which can be a reference for further human-centered XAI analysis and design practice.

Another researcher Lipton [17] also demonstrates that "To the extent that we might consider humans to be interpretable, it is post-hoc interpretability that applies." However, there should be a balance between explainability and cost effort. Bunt et

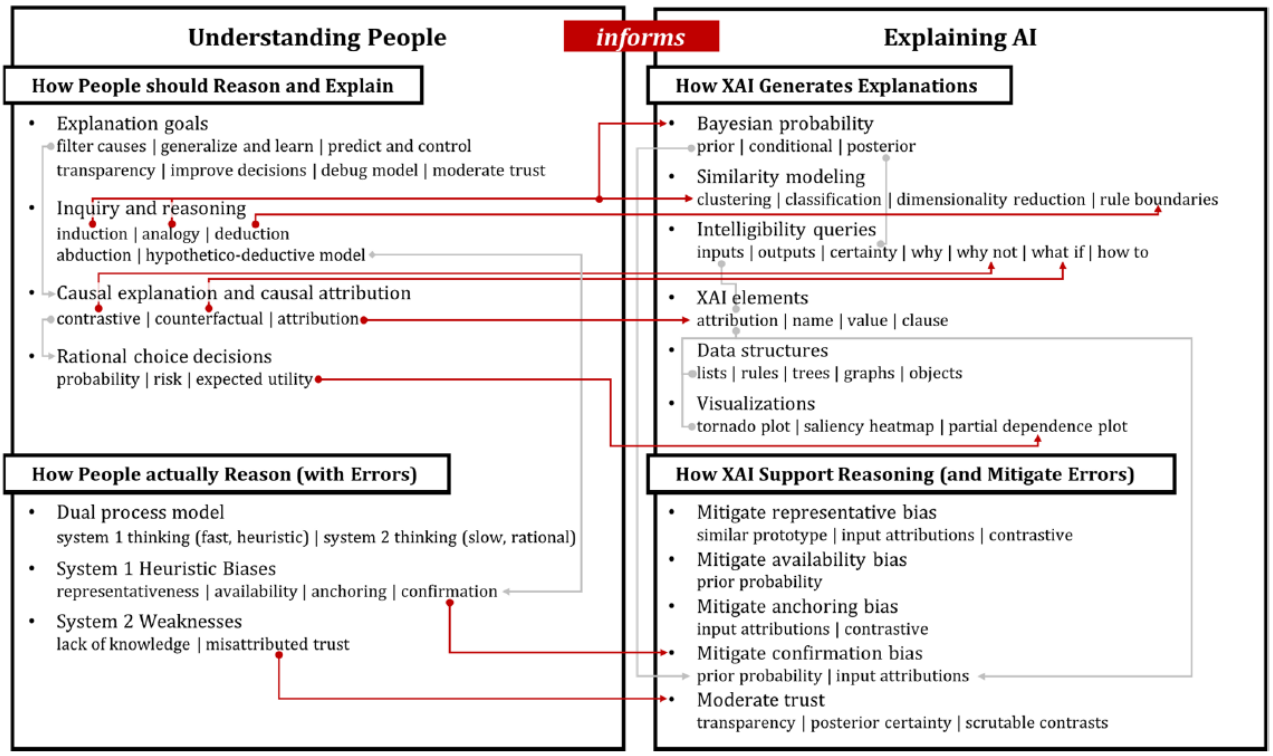


Figure 2.2: Conceptual Framework for Reasoned Explanations

Red arrows for how theories of human reasoning inform XAI features, and grey for inter-relations between different reasoning processes and associations between XAI features. Reproduced with permission [45] ©[2019] ACM.

al. [46] found that "While some users were interested in accessing more information, the dominant responses were that the applications were sufficiently transparent, or that the cost of viewing an explanation would outweigh the benefit," which doubts the necessity of applying XAI method to explain the low-cost or unimportant result. Work done by Kulesza et al. [47] also draws attention to the proper usage of XAI, "when soundness was very low, user experienced more mental demand and lost trust in the explanations, thereby reducing the likelihood that users will pay attention to such explanations at all." It can be concluded that the XAI for non-technical users should be highly concise and compact with only two or three critical factors explained. Otherwise, users will lose interest and trust in the explanations and the whole system, which is exactly the opposite aim of XAI.

2.5 Human-Centered XAI

The last section presents different XAI taxonomies and detailed methods to achieve explainability in AI. Since numerous XAI techniques focus on an algorithmic intuition rather than a user-centered view, the calling for human-centered XAI gets stronger [32]. This section will present content about the requirements and goals of human-centered XAI with approaches to implement it and evaluate the results in the real design context.

2.5.1 Definition and Goal

As mentioned before, most XAI work neglects users' perspectives and needs [48]–[51]. For example, methods like SHAP lists all different features used in the prediction and their corresponding importance and contribution levels [52]. However, it is worth questioning whether merely listing features with different values can satisfy the users' needs, which is to answer questions such as *why it does this* and *why does X instead of Y*.

In order to close the gap and bring more social and psychological elements to current XAI, it is necessary to bring human-centered approach and cross-field methods to XAI domain [45]. Although there are many open-source XAI toolkits online, it is difficult to bring them to real-world practice, which means to bridge the users' needs and theoretical guidance, and this mission often falls to the UX designers [25]. The work by Liao et al. [25] attempts to fill the gap according to Miller's definition of explanation, therefore, they propose an XAI question bank (See Fig. 2.3) specifically for lay users, which is valuable to build users' mental models and for future user research use. Similarly, Hoffman et al. [23] also draws a table (See Fig. 2.4) that links users' questions and learning aims together.

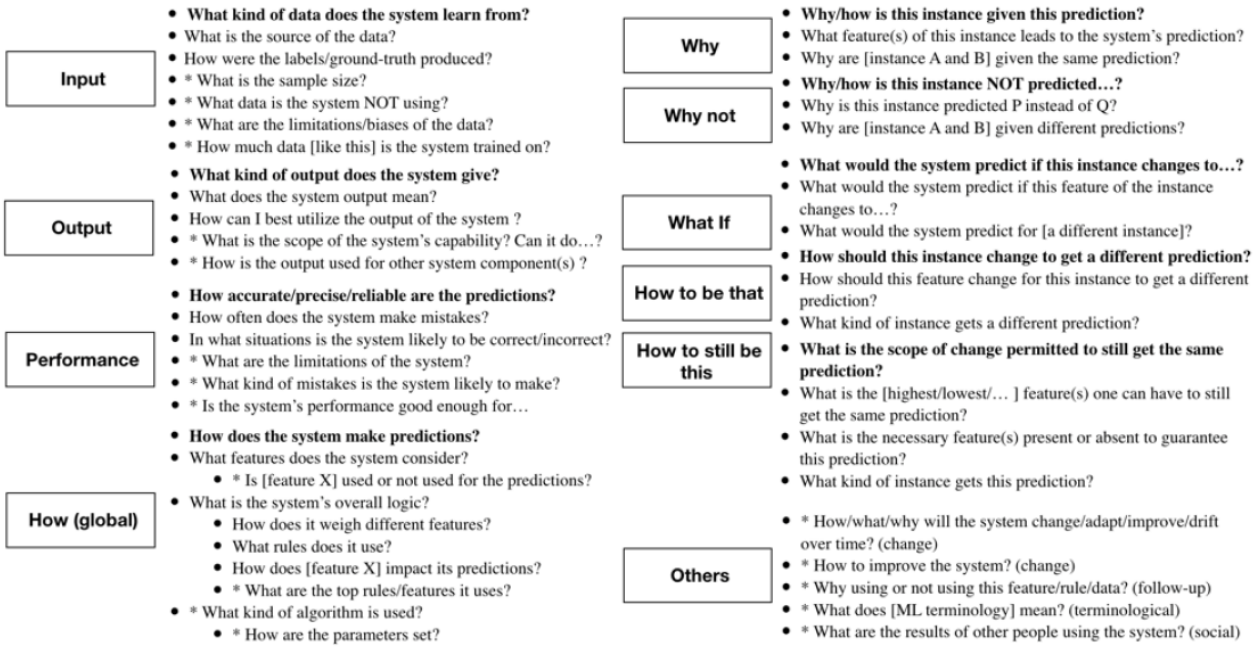


Figure 2.3: XAI Question Bank

With leading questions in bold, and new questions identified from the interviews with * Reproduced with permission [25] ©[2020] ACM.

TRIGGERS	USER/LEARNER'S GOAL
How do I use it?	Achieve the primary ask goals
How does it work?	Feeling of satisfaction at having achieved an understanding of the system, in general (global understanding)
What did it just do?	Feeling of satisfaction at having achieved a understanding of how the system made a particular decision (local understanding)
What does it achieve?	Understanding of the system's functions and uses
What will it do next?	Feeling of trust based on the observability and predictability of the system
How much effort will this take?	Feeling of effectiveness and achievement of the primary task goals
What do I do if it gets it wrong?	Desire to avoid mistakes
How do I avoid the failure modes?	Desire to mitigate errors
What would it have done if x were different?	Resolution of curiosity at having achieved an understanding of the system
Why didn't it do z?	Resolution of curiosity at having achieved an understanding of the local decision

Figure 2.4: Triggers and Users' Goal

Reproduced with permission [23] ©[2018] [Shane Mueller].

However, these questions lack the illustration of the relation between lay users' and experts' mental models as the algorithms are developed by experts and follow the experts' cognition. A mental model can be regarded as one's understanding of the AI system the XAI context [23]. Question-driven framework steps from users' motivation for explanation and is used to develop the expert system [53], [54], there should be some analysis for expert mental model to sufficiently understand and close the gap between them. Therefore, human-computer interaction (HCI) endeavors more efforts in the XAI field to build a complete loop by building the models of all roles involved in different design stages [48], [55]. The work of Eiband et al. [14] maps five different stages with corresponding roles, questions, aims and approaches, which builds models for both experts and lay users to find a consensus:

- **What to Explain: Expert Mental Model** What happens to the best of our knowledge? What can be explained? What does an expert mental model of the system look like?

- **What to Explain: User Mental Model** How do users currently make sense of the system? What is the user mental model of the system based on its current UI? How does it differ from the expert mental model?

- **What to Explain: Synthesis – Target Mental Model** Which key components of the algorithm do users want to be made transparent in the UI? To what extent are users actually interested in the rationale behind the algorithm?

- **How to Explain: Iterative Prototyping** How can the target mental model be reached through UI design? How and where can transparency be integrated into the UI of the system?

- **How to Explain: Design Evaluation** How has the user mental model been developed? Has the target mental model been reached?

Moreover, Hoffman et al. build a more comprehensive model, including the whole reaction loop with users and XAI (See Fig. 2.5), which is valuable for analyzing the requirements and evaluation methods of XAI stepping from users' perspective. They claim that a good explanation should satisfy users needs so that it helps users build a good mental model and gain users trust in the AI by assisting users in understanding and operating the system. At a macro level, Tintarev [56] put up human-centered XAI goals should not only focus on the users' model but the overall picture, which are transparency, scrutability, trustworthiness, persuasiveness, effectiveness, efficiency and satisfaction. The following sections will elaborate on how to achieve human-centered XAI in a more precise and detailed manner.

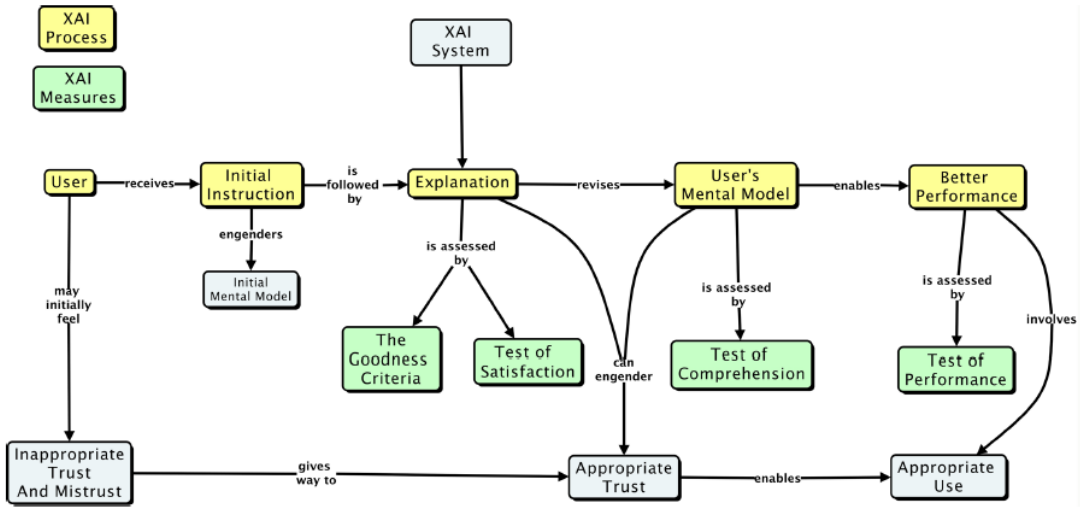


Figure 2.5: A conceptual model of the process of explaining in the XAI context Reproduced with permission [23] ©[2018] [Shane Mueller].

2.5.2 Requirement and Method

According to Nguyen et al. [44], the XAI requirements denote that the AI system should provide at least one of these functions 1) the understanding of its working mechanism 2) the visual explanation of its discrimination rules 3) or the possible causes that could disturb the model. Meanwhile, the work of Liao et al. [25] proposes more logical and specific requirements of what human-centered XAI should be able to achieve:

- Let users learn new knowledge of the system. Pay attention to separate things users already know and build new stuff on it
- Enhance users confidence or help users make hypothesizes of causes about following decisions of AI, also correct or mitigate users' own bias
- Involve more users' interaction and convince users to invest more in the AI system, such as letting users provide more information and feedback because "explanations as an integral part of a feedback loop to improve AI performance"
- Mimic how people explain things based on Miller's theory 1) explanation only focus on one or two reasons instead of all possible causes 2) explanation should be in a social and conversational manner
- Follow a progressive way, which means XAI should explain step by step in a question-driven approach

- Understand users' goals and implement XAI in a context-aware background

- "Help users understand the limitations of the AI, and make it actionable as to answer *Is the performance good enough for....*"

Other researchers' work has proved the above requirements and methods. According to suggestions from Herlocker et al. [15], users are not willing to accept over-complex explanations, which is also proven by Schaffer et al. [36] that comprehensive explanations of a recommendation system have side effects on users' trust and overall experience towards the AI.

For context-dependent scenarios, Bellotti and Edwards proposed that context-aware systems should be able to inform users about "what they know, how they know it, and what they are doing with that information" [57]. Explainability scenarios focus on what users need to understand instead of directly inserting explanations into the system centered on users' goals. According to Wolf's opinion [16], "Instead of asking what might an AI system be capable of explaining (a technology-first or solution-first orientation), a scenario perspective asks: what types of explanation might users need in the course of using AI systems?" Carroll [58] also proposes similar outlines that users themselves, their background, their goals, and the sequences of their actions are four elements of the design scenarios. Therefore, it can be concluded that XAI technologies should be implemented with specific user goals and aims in mind in order to fulfill their needs instead of merely explaining algorithms and data in detail.

As mentioned before, there is an issue in figuring out what type of explanations users might need. Tullio et al. [59] divide explanations into high-level and low-level explanations. High-level means explaining how information is related to each other, which is a more abstract and underlying principle approach, whereas low-level explanation is only about the input information and its calculation. High-level explanations can be taken as an abstraction of the low-level ones. For example, instead of explaining how algorithms do the exact calculation numerically, human-centered XAI should inform users of the underlying relationship of different factors and features, which can be regarded as the reason and logic of the system. Work of Borgman and Samek [38], [60] also proves that high-level explanations are much easier to modify users' mental models and previous beliefs, which is also verified by Liao et al. [25] that "the design challenge is to identify the appropriate level of details to explain the model globally."

2.5.3 Evaluation

An important issue that needs to be addressed is the evaluation of XAI, which is used to answer whether an explanation is proper and good enough among other explanations. Evaluating the XAI is of critical importance because it can not only pick up an optimized solution but also improves the personalization of the XAI such as interactive explanation because the effectiveness of an explanation is partly determined by its recipients and questions asked [13], [61]–[63]. Moreover, the same explanation may have different effects on different users or even on the same user in a different understanding process [23], [37]. The need for evaluating the XAI is in growing demand, however, the work on evaluating the XAI methods is considerably lower [26].

According to J.S van der Waa et al. [27], the overall user evaluation of XAI effectiveness should address three aspects 1) system understanding 2) persuasive power 3) task performance. Their work recommends the evaluation process should pay attention to the following three items:

- Constructs and relations: the aim and definition of the task should be clear, for example, the final aim of the work. After figuring out the motivation, the goal of the evaluation can be clear at the same time.

- Use case and experimental context: the selected use case can have a significant influence on the conclusion and evaluation, and that is the reason why quantitative data like age, education, and habits matter. Meanwhile, the environment can also influence the evaluation quality. Although online interviews can provide a large amount of information in a relatively short time, the results may not be that clear and insightful.

- Measurements: self-reported measures and behavior measure both matter. Self-reported measures are indicated as more subjective (e.g., users' perceived understanding of the whole XAI process). In contrast, behavior measures take a more objective view, which means observing the users' real behaviors and performances to learn their level of understanding.

The last item is given special notice because there may be a difference between users' perceived and real understanding. People sometimes overestimate how well they understand complex causal systems. This situation can be corrected by asking users to explain their understanding of how the system works [64]–[67]. Therefore, inquiring about users' thoughts and feelings is highly recommended. Hoffman et al. also mentioned this concept: "It is important that the method provide some sort of structure or "scaffolding" that supports the user in explaining their thoughts and

reasoning. One method is Cued Retrospection. Probe questions are presented to participants about their reasoning after the reasoning task has been performed" [23]. For instance, questions like *Can you describe your understanding of the...Can you describe the components or process of...* should be asked to collect information. This method works well with counterfactual reasoning, which means it can be used as a probe to evaluate whether the contrastive explanations are satisfying [68]. Besides, another method Diagramming is also used to measure users' real mental model. It effectively conveys their understanding to researchers, which helps analyze corresponding information in a workflow [69].

As the explanation process model shown in Figure 2.5, there are also other elements that matter in the process of evaluating the XAI, which are 1) Explanation goodness and satisfaction 2) Measuring mental model 3) Measuring curiosity and trust [23]. This multi-measurement method is also proven to be effective in Miller's research [32].

For the second item *Measuring mental model*, Hoffman et al. propose a hypothesis that measuring the performance of XAI is simultaneously measuring the soundness and goodness of the user mental model. Points from their work are similar to the suggestions from J.S van der Waa et al. mentioned before, which is to use Probe questions to let users describe the whole working process or some components and functions of the system instead of merely presenting subjective results from questionnaires evaluation or answers from questions like *How will you rate your understanding level of...* from users themselves. Meanwhile, work from Hoffman et al. also lists some casual links between users' mental model and performance. The following two sections will only explain the item one and three.

- Users' performance will improve after they receive good explanations from the system
- Users' performance is an externalization of their inner mental model
- Users' performance could be affected by their level of epistemic trust, in other words, the cognitive or deep level of trust.
- Users' performance will become more reasonable after they receive and understand the explanations properly

Explanation Goodness and Satisfaction

Based on the table of Triggers and goals (See Fig. 2.4), Hoffman et al. [23] propose a scale that is used to measure the goodness of the XAI (See Appendix C.1). This goodness checklist is especially for AI experts to evaluate the goodness of the

explanations of the system, therefore, its references are different properties of XAI. For explanation satisfaction (See Appendix C.2), Hoffman et al. conclude several attributes to measure satisfaction: understandability, feeling of satisfaction, the sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness. These attributes are determined after the review of other researchers' work. They develop a Likert scale after literature reviewing fields like cognitive psychology and philosophy of science. Attributes in that scale are designed to measure elements that can make a good explanation. The main difference between the Explanation Goodness Checklist and Explanation Satisfaction Scale is that the Goodness Checklist is for experts to measure the XAI from different aspects, while the Satisfaction Scale is for users to evaluate whether the explanation is good and satisfying enough.

Measuring Curiosity and Trust

Measuring curiosity is an essential factor because users' behavior of seeking explanations is motivated by their curiosity, while good explanations can also improve people's curiosity, which is proven by lots of cognitive and psychology research. When users realize a gap between their understanding and the system, they will actively seek help from explanations to close it. However, improper explanations could suppress people's curiosity in the following forms:

- Explanations that have too many details are overwhelming for people
- XAI system makes it difficult for users to ask questions
- Explanations that have too many uncontrollable or open variables
- Explanations make users feel uneasy because they are complicated or have too much deep knowledge that needs considerable effort to understand

Another indicator, trust measurement, is of critical importance in computer science and cognitive science fields [70], [71]. Researchers have found that the trust will automatically drop rapidly under time pressure, when systems have suspicious flaws, or when there are frequent alert alarms. It is rather challenging to rebuild users' trust once it has crashed down [72], [73]. In an ideal situation, users' trust will gradually grow up as time passes [74]. After reviewing numerous trust measurement scales, Miller et al. have made the Explanation Trust Scale (See Appendix C.3), which is built on some modifications to Cahour-Fourzy Scale and some questions merged from other scales. This scale explores whether users trust the system by measuring the predictability, reliability, efficiency and dependability directly, and it can also be used for individual testing.

2.5.4 Design Guidelines

Currently, there is a lack of design guidelines in the XAI field, and many instructions only stay at the theoretical level rather than dive deeper into real design practice [25]. According to Moore et al. [75], there are four elements a sound XAI system should equip with 1) naturalness (explain step by step in a conversational way) 2) responsiveness (users can put up questions in this system) 3) flexibility (variant explanation methods) 4) properness (explanations should consider users previous knowledge and interactions).

Research from Lipton and Gunning [17], [40] also shows that explanations in natural language with "analytic (didactic) statements that describe the elements and context that support a choice" has a noticeable impact on improving users' understanding of a system. Their work also claims that methods like counterfactual reasoning and explanation by example are more preferred than pure data or algorithm analysis explanations. This statement is also proven by Miller [12] that a user-friendly explanation should be contrastive and socialization (e.g., answering why A not B question).

The form of human-centered XAI should also be post-hoc and text or visual-based. According to the finding from Herlocker et al. [15], users care about the type of explanations such as text, graph, and video much more than the type of data and algorithm of a system. The work of Kouki et al. [76] has also proven this. They find that among other XAI methods, texts and images have a better performance. They also conclude that textual explanation is valuable and effective in step-by-step explanation and visualization performs better when the graphs are simple. However, they also mention that there is no preference between these two methods, which should be decided based on a specific context.

Besides the design form, there are some requirements relating to the contents. Systems could provide some support using example-based explanations, for instance, information about how other people make a choice when AI generates some recommendations. Another guideline related to content is that users seek high-level information rather than detailed explanations on how exactly the algorithm work, which is much more valuable for them to build an overview of certain system [59]. This statement is proven to be true in the research from Liao et al. [25] that participants prefer high-level post-hoc explanations to algorithmic methods like SHAP and LIME, and the former approach has indeed shown to be more effective. Similar suggestions from Bellotti et al. [57] that users need proper abstraction when informing them about the underlying calculations.

Another suggestion given by Liao's research is that the XAI should not only provide descriptive information on the outcomes of the algorithm, but also inform users about what they can do with the result, and how it relates to their goals. Their work

promotes the result from the algorithm to a higher level, which is how to maximize its utilization. Meanwhile, they also highly stress that when designing XAI systems and providing explanations, it is essential to discriminate between the knowledge that already exists in users' minds and new information that users do not know and needs to be explained [25].

2.5.5 Conclusion

The human-centered XAI should be implemented in a context-aware system [25], [57]. Researchers should figure out the user profile information such as users' background, goals, needs and action sequence, and take the users' aim as a navigator in design [16], [58]. Besides, XAI should be able to correct non-technical users' bias and inform them how to get a better result in the future practice [39]. It should also let users understand the working principles and discrimination rules of the algorithm, and possible causes of disturbing the model [44]. Good explanations can enhance the soundness of the users' mental model, and users' trust, satisfaction and curiosity of the product. It can also improve the users' performance as the soundness and goodness of the user mental model is improved [23], [27].

Gaps between the user mental model and expert mental model can be filled by building a target mental model [14]. The target mental model not only concerns about the knowledge gap, but also the users' goals and interest. For the content of the explanations, it should only include necessary knowledge for users in an easy-to-understand way [38]. The explanations should answer the *Why* questions (*why it does this*), or provide the content in a counterfactual way (*why does X instead of Y*) [32]. Moreover, explanations should provide knowledge and information that users have not acquired before, and inform users what they can do with the result, and how it relates to their goals [25].

The number of the causes in a single explanation should be limited to 2 or 3 [12]. Otherwise, numerous detailed explanations could negatively influence users' confidence, patience and overall experience [15], [36]. Moreover, there should be a balance between explainability and cost effort [46]. Post-hoc human-centered XAI methods such as textual and visual explanations have satisfying effect for non-technical users [15], [76]. The explanations should be interactively presented step-by-step using natural language with didactic statements [17], [25], [40], [75]. Besides, compared to explanations of how algorithm calculates, users prefer high-level explanations of the model's underlying principles, which is effective in re-building the user mental model [25], [38], [59].

The XAI evaluation is of critical importance because the effectiveness of an explanation is partly determined by its recipients and questions asked. Moreover, the

same explanation may have different effects on different users or even on the same user in a different understanding process [25], [37]. The effectiveness of the XAI can be evaluated by measuring the users' performance (simultaneously measuring the user mental model) and users' trust, satisfaction, and curiosity of the system [23], [27]. For building and measuring the user mental model, methods like Think-Aloud Task with Concurrent Question Answering, Cued Retrospection, and Diagramming allow researchers to know users' actual extent of command of the model [23], [27], [69]. During this process, probe questions like *Can you describe your understanding of the...Can you describe the components or process of...* are useful in collecting users real understanding of the system, and these questions performs well in evaluating the counterfactual reasoning [68]. For quantitative method, the Explanation Satisfaction Scale and Explanation Trust Scale are recommended by Hoffman et al. [23], which has satisfying effect in measuring the users' satisfaction and curiosity separately.

Chapter 3

Analysis of EMR

This section focuses on the research of the EMR application and its potential users. It aims to build a target mental model based on the gaps that need to be filled between expert and user mental models following the process proposed by Eiband et al. [14]. This chapter will also answer the Research Question 1 (RQ 1) and part of the RQ 2. Firstly, the expert mental model is built based on the analysis of the EMR application and interview results from one company's data specialist. Then, the user profile and mental model are set after the user research. The comparison between the expert and the user mental model will be presented. Finally, a target mental model is built on the user research, and requirements and guidelines for the XAI design. One thing needed to stress is that all mental models include two parts, which are 1) the overall working process and 2) perceptions of four function segments.

3.1 EMR Application Study

The EMR application is designed for cycling enthusiasts to assist them in customizing their daily nutrition plans based on their cycling schedule. The overall workflow is presented in Fig. 3.1, which shows how the EMR algorithm processes the data and generates the nutrition plan.

The user interface (UI) pages of the EMR application can be divided into four categories *Registration, Cycling and Nutrition Plan, Biological Explanation*, and *Result Renew and Comparison* according to their different functions (See Fig. 3.2, Fig. 3.3, Fig. 3.4, and Fig. 3.5).

The application has three modules **Plans**, **Learn**, and **Profile**. It first fetches users' basic information such as height, weight, and gender in the registration step. Then it will let users input their cycling data by selecting a course profile or importing routes they have cycled before from other cycling applications such as Strava and

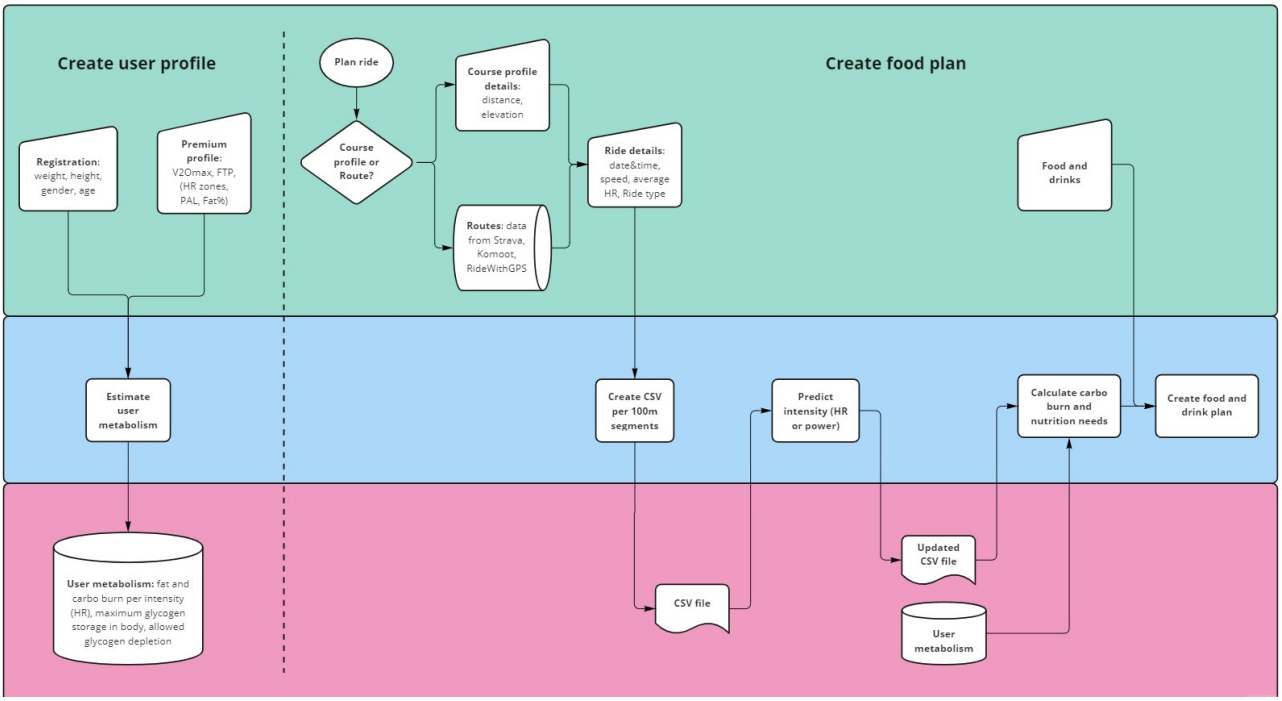


Figure 3.1: Workflow of EMR application

The figure explains how the EMR algorithm processes users' data and generates the final nutrition plan. The green section indicates users input, the blue section shows the algorithm behaviour, and the red section displays the data storage.

Komoot.

Selecting from a course profile requires users to choose the road type, distance, total elevation, and duration. Then, the application requires users to choose the ride type from recovery, endurance, interval training, and race. Besides, it will also ask for users' intensity which can also be taken as heart rate or average power.

After completing the basic personal profile and cycling data, the EMR application will require users to fill in their liquid and nutrition intake, such as what kind of drink and food they plan to have and its amount. Based on all these data, the algorithm will generate a personalized food plan, which contains the detailed recommendation of 1) how much and when to eat and drink 2) an evaluation of the food plan, such as its quality and amount 3) the glycogen usage of the body.

After filling out the current nutrition plan, the system will automatically pop up other food plans such as breakfast, lunch, dinner, and snacks for users to fill up. On the main page, the application also displays graphs about the number of carbohydrates, proteins, and fats that users should take, and some rough explanations of the calculation of the total energy intake.

The algorithm of the EMR application is decision tree regression, which builds models in the structure of a tree and generates the continuous output. The model is transparent and has the highest rank of interpretability compared to other algorithms. [26], [77]. However, it is worth noticing that the transparency is only for AI experts rather than target users, as the algorithm is not presented or explained in the EMR application UIs.

Diving deep to the algorithmic perspective, features chosen to build the baseline dataset are: ['weight', 'age', 'length', 'totaltimemove', 'total time stop', 'total distance (km)', 'average speed (km/h)', 'uphill/downhill hys=0 (m)', 'uphill0end', 'average temperature', 'm/v', 'elevationPerDistanceRelation', 'avgHeartRatePerUser', 'speedHeartRateRelation', 'timeRelation']. The EMR model is first trained with the baseline dataset and again with the data excluding the last four features. It turns out that the last four features used to enrich the dataset highly improve the algorithm's performance. Moreover, features 'age', 'totaltimemove', and 'timeRelation' show a P-value less than 0.05, meaning these three features influence the model significantly.

3.2 Expert Mental Model

As previously mentioned, a mental model can be regarded as one's understanding of the AI system in the XAI context [23]. The expert mental model of the EMR overall working process has been presented in the EMR Application Study section (See Fig. 3.1). This section will focus on the expert mental model divided into four categories:

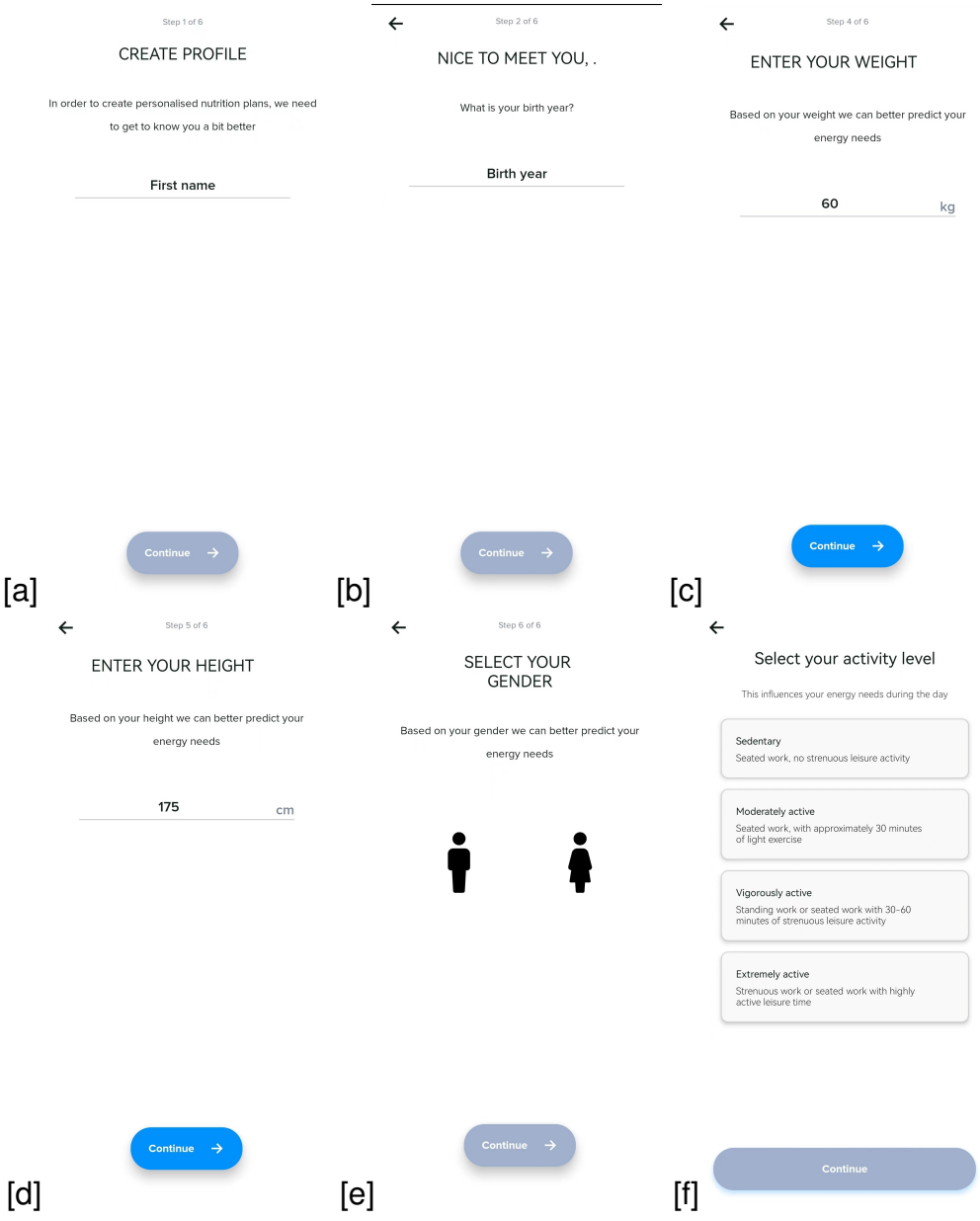


Figure 3.2: UIs for Registration

UIs belong to Registration category. Notice (f) is from user *Profile* module, and it is not presented when users fill in personal data at the registration step.

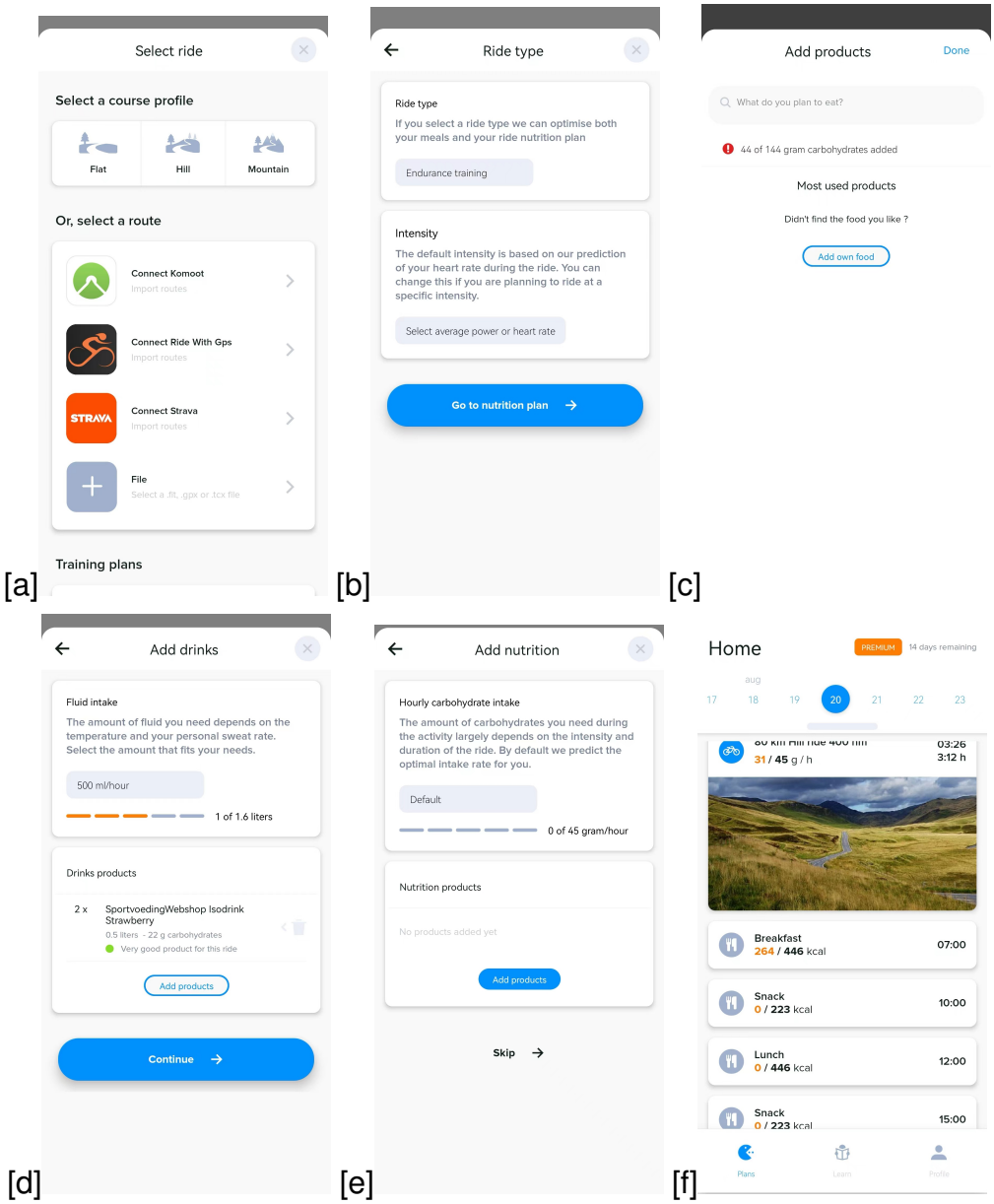


Figure 3.3: UIs for Cycling and Nutrition Plan

UIs belong to Cycling and Nutrition Plan category. Food plans in (f) pop up after users finish the cycling nutrition plan.

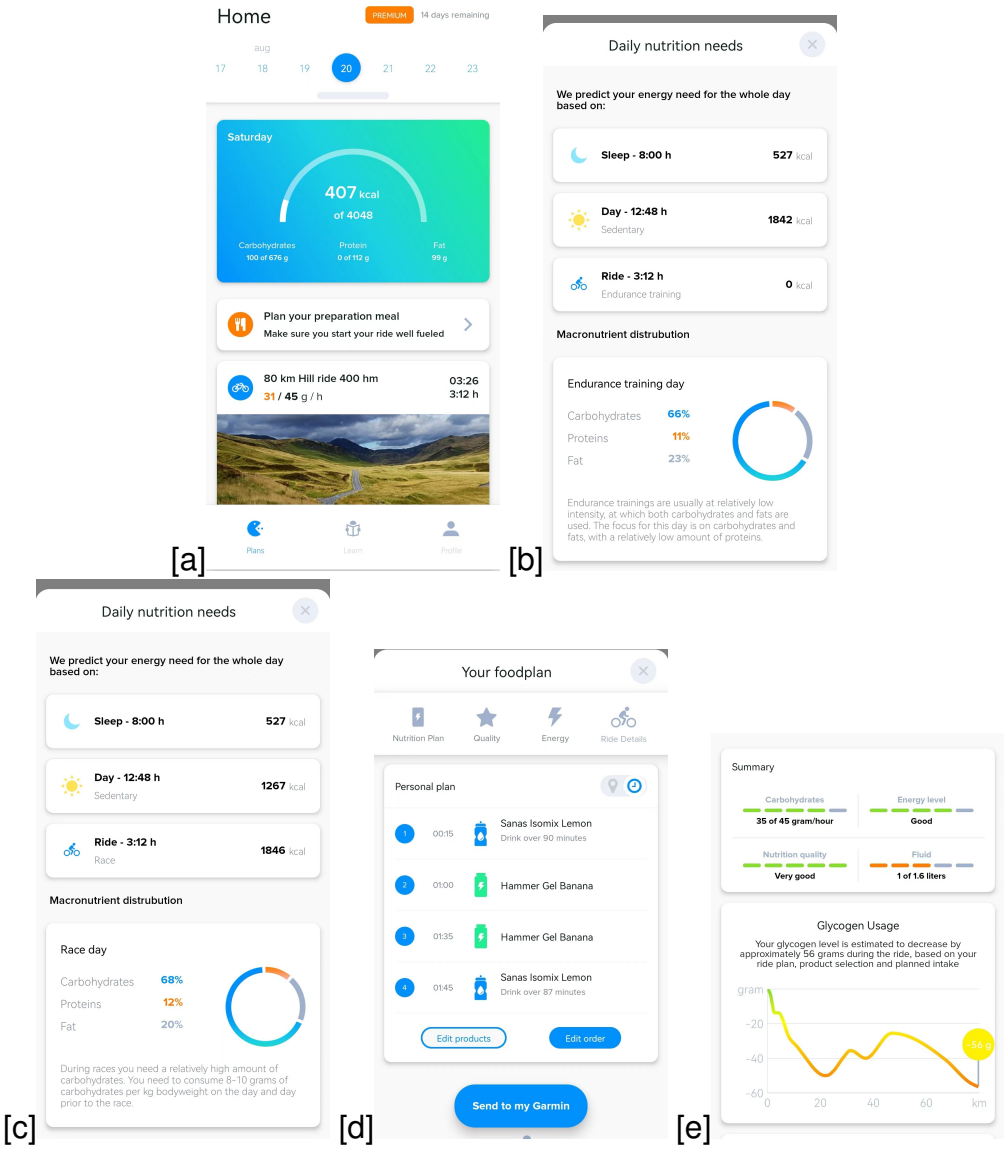


Figure 3.4: UIs for Biological Explanation

UIs belong to Biological Explanation category. These UIs are mainly responsible for the explanation work.

Figure 3.5: UI for Result Renew and Comparison
The UI belongs to Renew and Comparison category.

Registration, Cycling and Nutrition Plan, Biological Explanation, and *Result Renew and Comparison*. According to Eiband et al., the expert mental model should be able to answer questions like *What happens to the best of our knowledge? What can be explained? What does an expert mental model of the system look like?*

- **Participant** One data analyst from the EMR company
- **Method and Measure** Online semi-structured interview using ZOOM. Notes are taken down during the interview.
- **Procedure** Display the process of using the EMR application from registration until the end. Questions such as *What does this step stand for? Why do we need this step? What do you want to inform users? Does this have a significant influence on the algorithm and outcome? What is the relation between... ?* are promoted for the expert to answer.

For Registration, information such as weight, height, gender, and age are used in predicting the metabolism, which influences later energy calculation, whereas country and measurement do not. The Activity Level also affects the algorithm in predicting the daily energy consumption.

For Cycling and Nutrition Plan, *Select a route* provides more data, such as weather and detailed road information to the model, leading to better predictions than *Select a course profile*. Predicting the heart rate (HR) is the essential function of the EMR algorithm because the predicted HR data will be used multiple times in

calculating energy consumption and intake influenced by HR and carbo burns. The real HR data will update the predicted HR data after the actual cycling.

For Biological Explanation, the expert explains the fundamental reason for the differences in energy prediction from the biological aspect. Low-level intensity training, like endurance training, belongs to aerobic exercise, therefore, both carbo and fats are used. High-intensity training is anaerobic exercise, and most riding energy comes from the carbo. This primary principle determines the basic logic of the algorithm. Consequently, the calculation formulas of the model are different, and the prediction results are various. The same level of underlying reason also applies in other situations. For instance, if users plan to have a race day, the algorithm will assume that their bodies are almost full of carbo, and the energy needed mainly comes from their body storage instead of the food they have on race day. Therefore, the algorithm recommends that cyclists take much energy on the pre-race day. For the glycogen level graph, the data analyst indicates that it is used to determine whether the body has enough energy to repair the muscle. Users' muscles will be impaired if the line turns red.

One thing that needs to be stressed is that the information and reasons mentioned in the Biological Explanation are not displayed in the UIs (See Fig. 3.4 for reference), while the expert expressed in the interview that the knowledge is important for both the application and users' cycling performance.

There is only one UI in the Result Renew and Comparison category. The algorithm will re-calculate cyclists' energy consumption using their real HR data during cycling instead of the predicted HR. The previous and renewed results will be presented in one graph for comparison. The algorithm makes progress and becomes accurate by closing the disparity between its predicted HR value and users' real HR data. However, this information, recognized as the essential reason the algorithm keeps improving, is not presented in the UI.

3.3 User Profile and Mental Model

This section presents the user profile and user mental model. According to the research [25], [57], [58], human-centered XAI should be implemented in the real-world design context, and it is essential to figure out the user profile, which includes users themselves, their background, goals, and the sequences of actions. The user research contains two parts, which build the user profile and mental model separately. The user mental model includes users' understanding of the overall working process of the EMR application and its four different functional modules. The consent form and information brochure for participants can be found in Appendix A.1.

3.3.1 User Profile

After interviewing each participant, the user profile is built by organizing and integrating the representative answers.

- **Participant** Eight participants. Two are professional cyclists from The Union Cycliste Internationale (UCI). One is from the cycling club of the University of Twente. Another one is a cycling enthusiast in the Netherlands. The rest four are enthusiasts from a Chinese cycling club. All participants are fluent in reading English and have used English applications.

- **Method and Measure** Online semi-structured interview using ZOOM. Notes are taken down during the interview. Table 3.1 presents some questions for building the user profile.

Do you check and record your cycling data, why, and how?
Do you pay attention or make food plans for cycling? If yes, how? e.g. calculate the calories or make cycling meal
Have you ever used some nutrition plans application? Are you still using them? Why?
To what extent will you follow the suggestions from the application? Why?
How much do you think the application's nutrition plan helps you or makes a difference for cycling?
Will you adjust your diet (nutrition plans), especially for cycling? If yes, to what extent will you change them?

Table 3.1: Some questions asked in the semi-structured interview
Answers to these questions are used to build the user profile.

A user profile description of potential EMR application users is presented. The aim of cycling is to challenge themselves and prove their ability. All participants mentioned that cycling gives them a sense of control and achievement in real life. It also benefits their physical and mental well-being because of the sense of belonging from cycling clubs and friendships with other cycling enthusiasts.

Participants are interested in the nutrition plan and expressed that they believe scientific food plans positively affect cycling performance. However, the majority (six of eight) believe that the help from the food plan is limited, and dieting does not have much impact and improvement on cycling. Therefore, they consider nutrition plans less important in cycling.

Seven participants clearly expressed that they believe in their body's feelings much more than the prediction of the AI model. The reason is that there are numerous items that the algorithm cannot detect or predict precisely. For instance, it is hard to describe the condition and feelings of their body to the application, and information like that is complicated to standardize to a quantified data format to be used in the model. The situation varies everytime, therefore, the algorithm cannot provide a precise prediction. Moreover, except for two professional cyclists, participants value freedom more than being constrained by the food plan. They would like to take the plans from nutrition applications as a reference or suggestion instead of guidance they need to follow accurately.

Half participants used some similar nutrition applications before. However, they all gave up after weeks or months. The reasons mentioned by participants are 1) they have built a model in their mind after a period of learning from an application, meaning that they can roughly do the calculation and evaluation the same as algorithms 2) they quickly feel bored and lose patience after many times filling up the same kind of data (e.g., what and when do they eat) 3) the application does not provide anything new, and the core function can be done by users themselves after repetitive using.

When asked about their knowledge of nutrition from a biological aspect, except for one professional cyclist and two cycling enthusiasts who read nutrition books and papers, others do not know much about the underlying principles, such as how our body consumes energy from adenosine triphosphate (ATP), and the reason why types nutrition types needed varies with different cycling types. Those two cycling enthusiasts mentioned that getting help from reading scientific content is much more helpful than a nutrition application because the application's fundamentals also come from biological knowledge, except the calculation from algorithms is more accurate than personal estimation. However, participants considered this difference does not significantly influence their cycling performance.

3.3.2 User Mental model

This section aims to build a user mental model of the overall working process and four function categories of the EMR application. Think-Aloud Task with Concurrent Question Answering is widely used in building user mental models according to research from Hoffman et al. [23]. Their research also indicates that Cued Retrospection is an effective method to evaluate users' real understanding instead of their perceptive understanding of a system. Probe questions such as *Could you describe the process of... What is this data used for? Why does the system...* should be put up to make sure the user mental model is authentic and objective [23], [27]. More-

over, Diagramming is also used in visualizing and analyzing the mental model of a system's workflow [69]. Finally, there will be a conclusion summarizing the critical features and issues in the user mental model.

Method

Think-Aloud Task with Concurrent Question Answering, Cued Retrospection, and Diagramming. Only notes are taken down, no audio or video recording during the online interview.

Procedure

Participants were required to experience the usage of the EMR application from the registration step. The task was performed with the Think-Aloud Task with Concurrent Question Answering method. The Triggers and Users' Goal table (See Fig. 2.4) is provided as a reference in case users have no idea what to describe or evaluate. Some questions related to participants' feelings and performance were prompted to participants (e.g., *Why do you think it is... Why do you take this step?*). After that, the Cued Retrospection method was applied to acquire users' real understanding of the application. The information collected from Think-Aloud Task with Concurrent Question Answering and Cued Retrospection will be used to build the user mental model, where the workflow graph is made using Diagramming.

Result

The user mental model includes users' perceived overall workflow and functions. The application workflow is concluded using Diagramming (See Fig. 3.6).

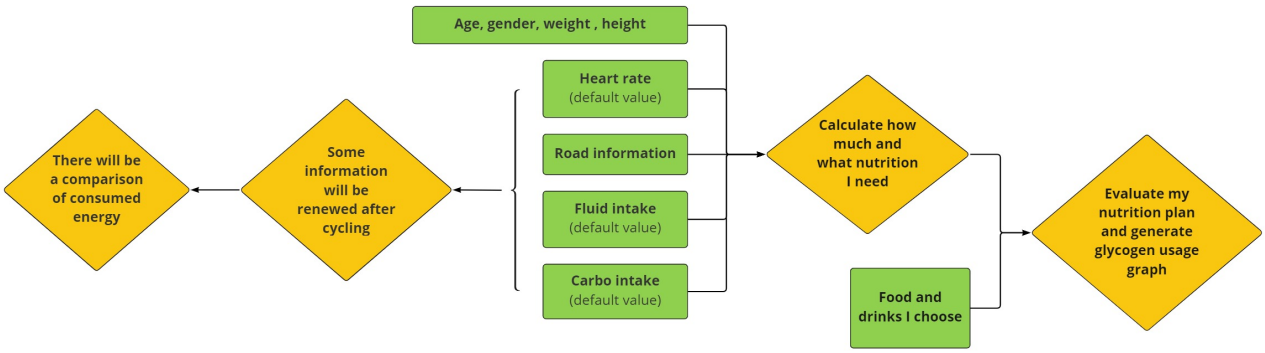


Figure 3.6: User mental model of the EMR workflow
The green color represents users and yellow is for the application. Square and diamond mean input and output separately.

For Registration, all participants noticed that the data explanations are repetitive. They understand that features *Age, Gender, Height, Weight* are used in the prediction model. However, participants also mentioned that the text does not provide helpful information about how and where their data will be used (See Fig. 3.2). Participants gradually lost interest in continuing reading it. They agreed that explanations in the registration part are knowledge they already know.

For Cycling and Nutrition Plan, participants had a rough understanding that road types, and fluid and food intake influence the energy consumption prediction. Common issues reported are 1) repetitive text and explanations 2) no new knowledge 3) the model workflow is logically confused. Moreover, two participants thought that the prediction results from *Select a course profile* and *Select a route* are approximately accurate. Others thought that *Select a course profile* provides more accurate results only because of detailed road information (See Fig. 3.3 [a]). Participants understood that selecting different ride types (See Fig. 3.3 [b]) influences energy consumption. However, they were unsure to what extent this choice would make a difference in the model.

When seeing the explanation *Default intensity is based on our prediction of your heart rate* (See Fig. 3.3 [b]), all participants mentioned that this application had not informed them of how it predicts their HR data. Therefore, they estimated that the HR prediction is based on their personal profile or riding types. Six participants questioned the accuracy of the intensity data because 1) the application uses predicted HR instead of real HR that can be measured by cycling equipment 2) the real HR data dynamically changes with time, whereas the predicted HR is probably not.

Besides, half of the participants indicated that the explanation of *Fluid intake* (See Fig. 3.3 [d]) has a counter effect on their understanding and trustworthiness of the algorithm. The reason is that the application does not ask for their sweat rate data, whereas the prediction is based on this data. Moreover, there is no clear sign of whether the system can get weather information. Participants mentioned that the course profile does not require any weather data input.

Out of the UX consideration, all participants reported confusion as there is no preset information when they were required to fill out the nutrition item, which led them to wonder whether they needed to manually input all the detailed information for each item, such as calories, fats, and proteins. Another point is that the application will automatically pop up four to five different food plans without notification after users finish the cycling nutrition plan. Most participants expressed frustration and loss of confidence when they needed to input their every meal plan into the EMR application.

In conclusion, the user mental about this category is unclear and logically confused. From the expert mental model analysis, it can be found that important infor-

mation is left out or misunderstood. Participants are not clear about the underlying workflow of the model, nor do they know how and where their data is used. Some explanations are redundant and provided without concerning the real-world context. Users do not understand the purpose of some explanations, and they consider those explanations not helpful or useful for their cycling. From the user experience (UX) aspect, all participants firstly doubted whether they needed to manually add the drinks and nutrition information until they kept inputting products' information before actually seeing them (See Fig. 3.3 [c]). Besides, new food plans suddenly appear without any notification after completing the cycling nutrition plan (See Fig. 3.3 [e]), which lowers the participants' interest and patience with this application significantly.

Participants described their primary understanding for the Biological Explanation pages: nutrition needs vary with riding types, and the algorithm will evaluate and rate their nutrition plans. However, the same opinions were put forward that users cannot figure out the purpose of some explanations. For example, explanations in Figure 3.4 [e] only provide a text version of the graph, which basically does not provide any new information, nor explain the graph's meaning. Participants are not sure what this graph tries to explain and inform. Moreover, the majority of participants mentioned that pages containing detailed numbers and calculations look the same, except some numbers and text are changed (See Fig. 3.3 [b][c]). Over half of the participants indicated that they will not pay attention to the explanations in the *Macronutrient distribution* after figuring out that the content does not change as long as the riding type is the same, even if the detailed cycling data is different.

When asked about the fundamental principle of the model (biological aspect), except for participants who have read nutrition articles and books before, others cannot tell the underlying reason, which indicates that there is a lack of a high-level explanation of this model. For example, when participants were asked about what the *Glycogen Usage* graph means, they had no clear answers but mainly speculations such as *The red color mains great amount of energy consumption*, nor do they understand what it means to their cycling performance.

For Result Renew and Comparison (See Fig. 3.5), participants reported that they got the idea that the model will compare the actual and planned energy consumption. However, they were confused about what data the model uses to update the prediction and how the algorithm performs this task. Because there is no prediction result of the total energy consumption presented in the application before actual riding. Similarly, because no graph or text shows the trend of real and planned energy consumption, participants cannot tell whether the algorithm is progressing and to what extent the accuracy is improved.

3.3.3 Conclusion

In conclusion, the user mental model of the working process is ambiguous, and participants are not clearly informed of the algorithm's use of their data. Participants who have not previously read nutrition books are unaware of the meaning of some textual and graphic explanations and how their actions will influence the body and cycling performance, especially from the biological aspect. All participants reported a large amount of repetitive textual information in the explanations, and the majority of explanations do not provide any new knowledge of the model.

Moreover, some explanations reduce users' interest and trust in the algorithm because of numerous or improper causes provided in a single explanation, which also lowers the UX. There is a lack of logical process and context-aware design in the explanations for the algorithm, which prevents users from believing the algorithm's progress and knowing the information addressed by the expert. For example, using predicted or real HR data is the essential reason for the difference between predicted and real energy consumption, and the accuracy of HR prediction decides the algorithm's accuracy to an enormous extent. Besides, the primary principle of the model is the different calculation methods for aerobic and anaerobic exercise, which belong to high-level explanations.

3.4 Target Mental Model

Figure 3.7 compares the expert mental model and user mental model from five aspects *Registration, Cycling and Nutrition Plan, Biological Explanation, Result Renew and Comparison*, and *Algorithm and Data*. Compared to purely algorithmic explanations, users are interested in high-level explanations that relate to the system's working principle instead of focusing on the algorithm, for example, how algorithms perform the calculation [25], [59]. Therefore, the target mental model should leave out the information in the last item *Algorithm and Data*. Besides, the repetitive information and knowledge that users already know should be discarded according to the research of Liao et al. [25]. The overall XAI design should imply users' goals and needs in a context-aware system [16], [25], [57].

For the Registration, the information from the expert mental model should be kept because it informs users how their data will be used [25]. It first introduces the predicted HR data, which is the essential component of the algorithm and the workflow. This aim is settled based on the user profile and mental model analysis. Previous research also proves that XAI should let users quickly figure out the underlying working principles of the algorithm and help users understand how a particular decision was reached [38], [39].

For the Cycling and Nutrition Plan, the comparison between *Select routes* and *Select from a course profile* will be presented to users because XAI should correct users' bias and facilitates them to get more desirable results in the future practice [39].

For the Biological Explanation, the knowledge from the expert mental model will be kept because it relates to users' performance improvement, which belongs to the users' aim. Moreover, high-level explanations such as the primary logic of the model are helpful in re-building users' mental model, and it also fits users' interest [25], [59].

The information in the Result Renew and Comparison relates to the principle of how the algorithm makes progress, and it also links with the expert's aim. The information on the relation between predicted and real HR data is essential to the model's workflow. Therefore, the target mental model should reserve knowledge in this category.

The overall explanations should be presented step-by-step using the logic of having a dialogue [25]. The number of causes in each explanation should be limited to 2 or 3 [12]. For the form of the explanations, post-hoc XAI methods such as textual and visual explanations are suitable for non-technical users [15], [76]. Besides, The content of the explanations should be presented in natural language with didactic statements [17], [40].

New table

	Expert	User
Registration	• Personal profile data (weight, height, gender, and age) is used to predict the metabolism (HR, energy storage, etc), which influences the prediction of energy consumption and needs. • The Activity Level also influences the daily energy consumption.	• Personal profile data (maybe weight, height, gender, and age) is used in predicting the energy consumtion and needs
Cycling and Nutrition Plan	• *Select routes* provides more data such as weather, road information and past cycling HR compared to *Select from course profiles,* which leads to better predictions of the cycling HR, and energy consumption and needs	• *Select routes* probaly provides more data than *Select from course profiles*
Biological Explanation	• Low intensity cycling (recovery training, endurance training) belongs to aerobic exercise. Both carbo and fats are needed • High intensity cycling (interval training, race) belongs to anaerobic exercise. The energy mostly comes from carbo • The energy for the race mainly comes from the energy storage on the pre-race day (90% energy comes from carbo), instead of the food intake on the race day. The algorithm assumes cyclists' body is full of carbo on the race day. Therefore, the food plan on the pre-race day is more important. • Red line of the *Glycogen level* means the body does not contain enough glycogen and the muscle wil be impaired., whereas green line means the opposite.	• Different training types have different nutrition needs and consumption • Foodplans for the pre-race day and race day are related together • Red line of the *Glycogen level* means the glycogen in the body decreases, whereas the green line means the opposite.
Result Renew and Comparison	• After cycling, users real HR data and road information will be uploaded from their device to the mdel to calculate the real energy consumption • The algorithm makes progress by closing the gap between the predicted and real HR data.	• There will be a coparison between the predicted and real energy consumption.
Algorithm and Data	• The application uses decision tree regression model. • The data used and prediction results such as HR, road information, energy consumtion and needs are stored per 100m segments in a CSV format. • Features *'age'*, *'totaltimemove'*, and '*timeRelation*' show a P-value less than 0.05, meaning these three features influence the model significantly.	

Figure 3.7: Comparison between expert and user mental models

The figure shows the differences between user mental model and expert mental model

Chapter 4

Low-fidelity Design and Evaluation

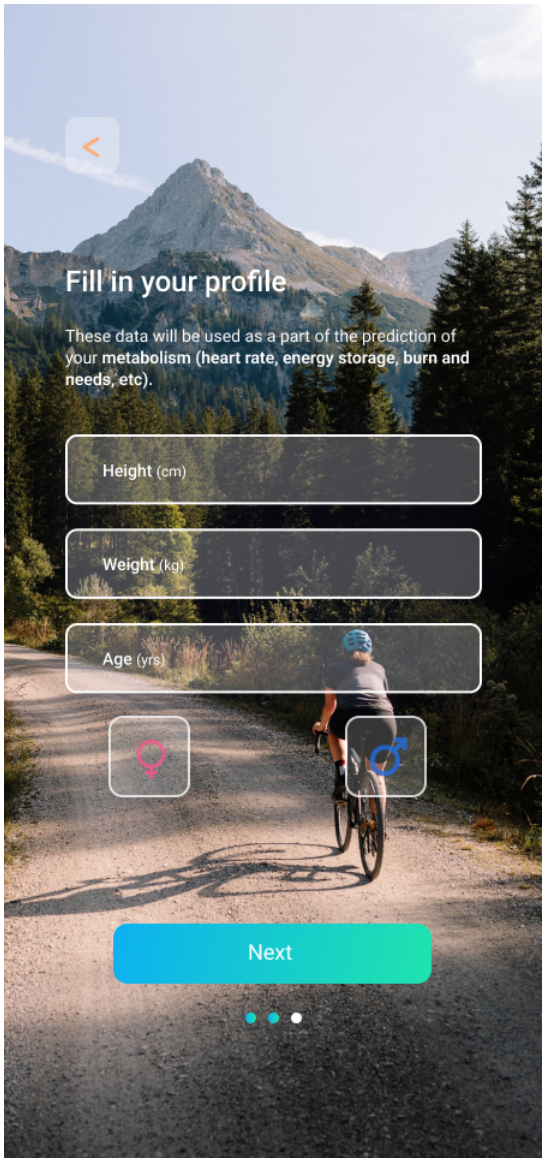
This chapter includes the process of improving the UI and UX of the current EMR application using human-centered XAI methods. The first section is the low-fidelity (lo-fi) prototype design, and the second is the human-centered XAI evaluation of the lo-fi pages for later improvement. Compared to the next chapter 5, this chapter focus more on the design part.

4.1 Low-fidelity Design

The UIs pages are presented in four categories as the original EMR application, which are *Registration, Cycling and Nutrition Plan, Biological Explanation*, and *Result Renew and Comparison*. Designs related to human-centered XAI will be presented and explained in each category. In order to provide a better UX, the lo-fi pages are designed in a high-fidelity (hi-fi) format, except they are non-interactive static pictures.

4.1.1 Registration

Compared to the original application, the new design removes all the repetitive explanations and the information users already know. It explains to users that their profile data, such as height, weight, age, gender and activity level, is used to predict the metabolism, which includes HR, energy storage, burn and needs. The *Activity Level* is moved to the registration step so that all the basic data is integrated, which enables users to quickly get a clear picture of what and how the algorithm uses their basic profile data (See Fig. 4.1).

[a] 

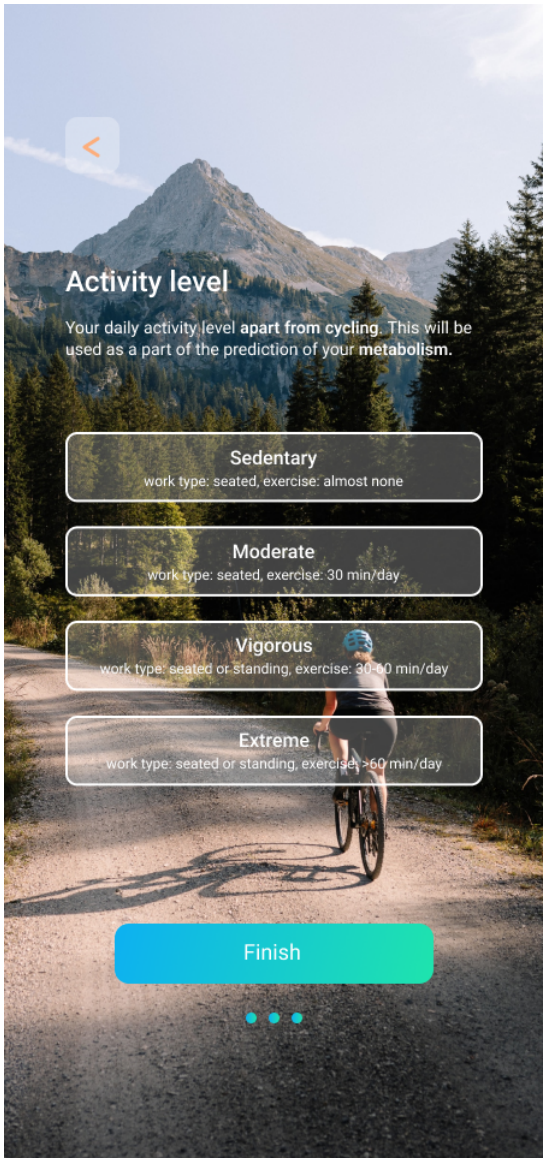
[b] 

Figure 4.1: Lo-fi Pages for Registration

[a] explains how users data will be used in the model. [b] shows the explanation for Activity Level.

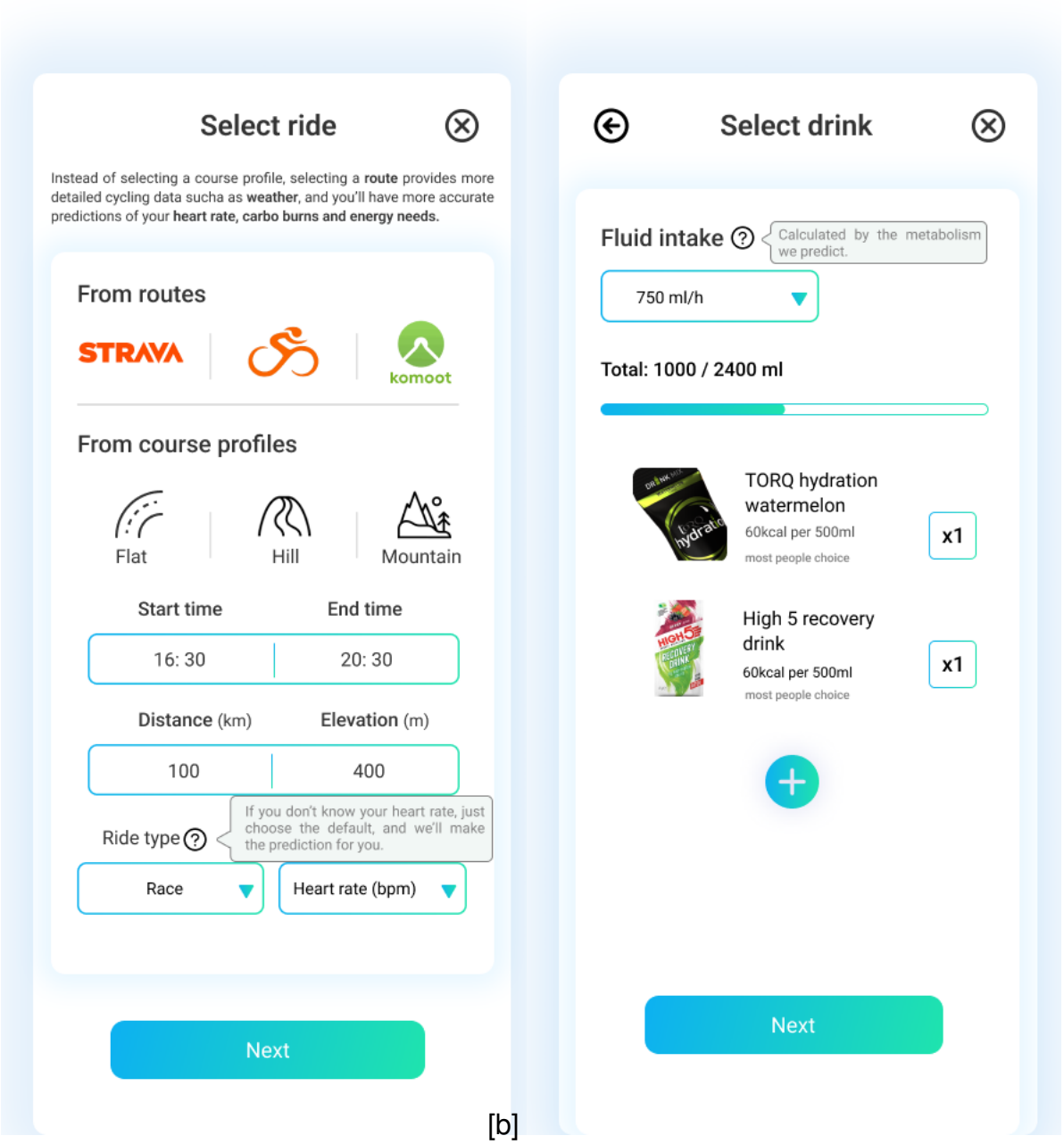
4.1.2 Cycling and Nutrition Plan

According to Wachter et al. [39], XAI should inform what users can do to improve the result of the system. Therefore, the counterfactual explanation (*Why X not Y*) aims to let users choose *From routes* instead of *From course profile* by presenting the advantages of *From routes* when users fill up the cycling data. This method is more acceptable and useful way for users in human-centered XAI aspect [32].

The explanation telling the difference between two choices is *Instead of selecting a course profile, selecting a route provides more detailed cycling data such as weather, and you'll have more accurate predictions of your heart rate, carbo burns and energy needs.* It stresses the weather information and HR prediction that users neglect or doubt what data the model can get and how the algorithm will use it (See Fig. 4.2[a]).

In the *Select drink* page, the temperature and personal sweat rate information is omitted from the previous explanation. The new explanation is *Calculated by the metabolism we predict* (See Fig. 4.2[b]). For food selection (See Fig. 4.3[d]), the explanation stresses the HR data by letting users know that their predicted cycling HR, road information, and metabolism data influence their carbo intake. Compared to the original EMR explanation *The amount of carbo you need during the activity largely depends on the intensity and duration of the ride. By default, we predict the optimal intake rate for you.*, the new explanation discards the knowledge users already know and limits the causes under three according to the human-centered XAI requirements [12], [25].

From UX aspect, *From routes* and *From course profile* are integrated to the same page. Therefore, the comparison is more intuitive, and fewer steps are required when filling up the cycling data (See Fig. 4.2[a]). Another improvement is the product list will be displayed directly before typing and searching to provide some information support for users (See Fig. 4.3[d]).



[a] [b]

Figure 4.2: Lo-fi pages for Cycling and Nutrition Plan

[a] explains the difference between two choice, it also improves the UX. [b] explains the Fluid intake.

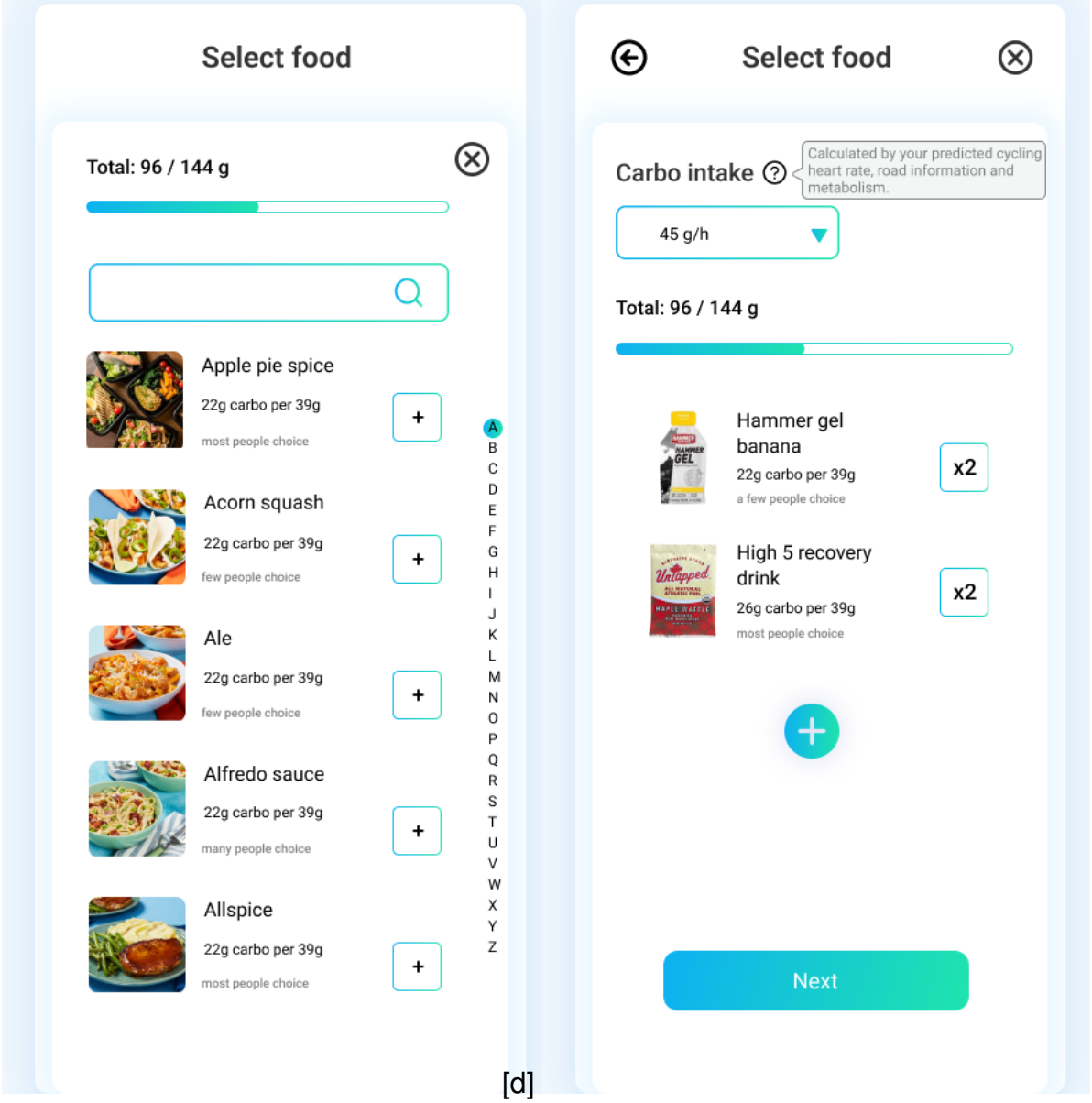


Figure 4.3: Lo-fi pages for Cycling and Nutrition Plan
[c] improves the UX. [d] explains the Carbo intake.

4.1.3 Biological Explanation

Compared to explanations of algorithm calculation, users are more interested in that of working principles [25]. High-level explanations are useful in re-building the user mental model [59]. This section contains the biological explanations of the primary principle of the application.

Firstly, each of the six different types of cycling: *No riding, Recovery training,*

Endurance training, Interval training, Race and *Pre-race day* are presented with different explanations (See Fig. 4.4 and Fig. 4.5). *Recovery training, Endurance training, Interval training*, and *Race* are divided into two categories: aerobic or anaerobic exercise based on the intensity. The *No riding* does not belong to training types, and it is displayed separately to provide an overview of the UI (See Fig. 4.4). Main differences of nutrition consumption and needs are presented in each explanation, and only the changed contents are presented (See Fig. 4.5[a][b][c][e]).

Besides, *Pre-race day* is addressed because it does not belong to a cycling type. However, it has a close link with race day, and the energy and nutrition intake on the pre-race day highly influences the performance of the race. Therefore, the model's calculation principle works differently from other cycling types (See Fig. 4.5[d]).

Moreover, the explanations of the *Glycogen level* are re-designed. Previous explanations do not provide helpful knowledge but only contextualize the graph. The conveyed information is the same except the numbers are changed. The new design explains how glycogen is calculated and what different glycogen levels mean to users' body condition4.6.

4.1.4 Result Renew and Comparison

The new design emphasizes the naturalness and reasonableness of the updated prediction results. It also highlights the updated data after cycling, such as the real HR and road information. Fig. 4.7[a] and [b] shows the comparison between before and after actual riding correspondingly. Part of the cycling information and food plan are renewed.

Fig. 4.8 displays the predicted energy consumption before riding and informs users that the actual burned energy will be uploaded after finishing cycling. The *Progress* module also presents the predicted and real energy consumption records and explains the differences between them. It stresses the change of HR data, which is the essential reason why and how the algorithm makes progress.

Fig. 4.9 shows the notification to users that they have a new record uploaded in the application and they can make the recovery meal after cycling. Fig. 4.10 presents the UI after users finish riding. The main change is the comparison in the *Burned energy* section.

4.1.5 Conclusion

The new design uses post-hoc human-centered XAI methods. Textual and visual explanations are provided because of their outstanding performance among other XAI methods for non-technical users [15], [76]. The overall reading load decreases.

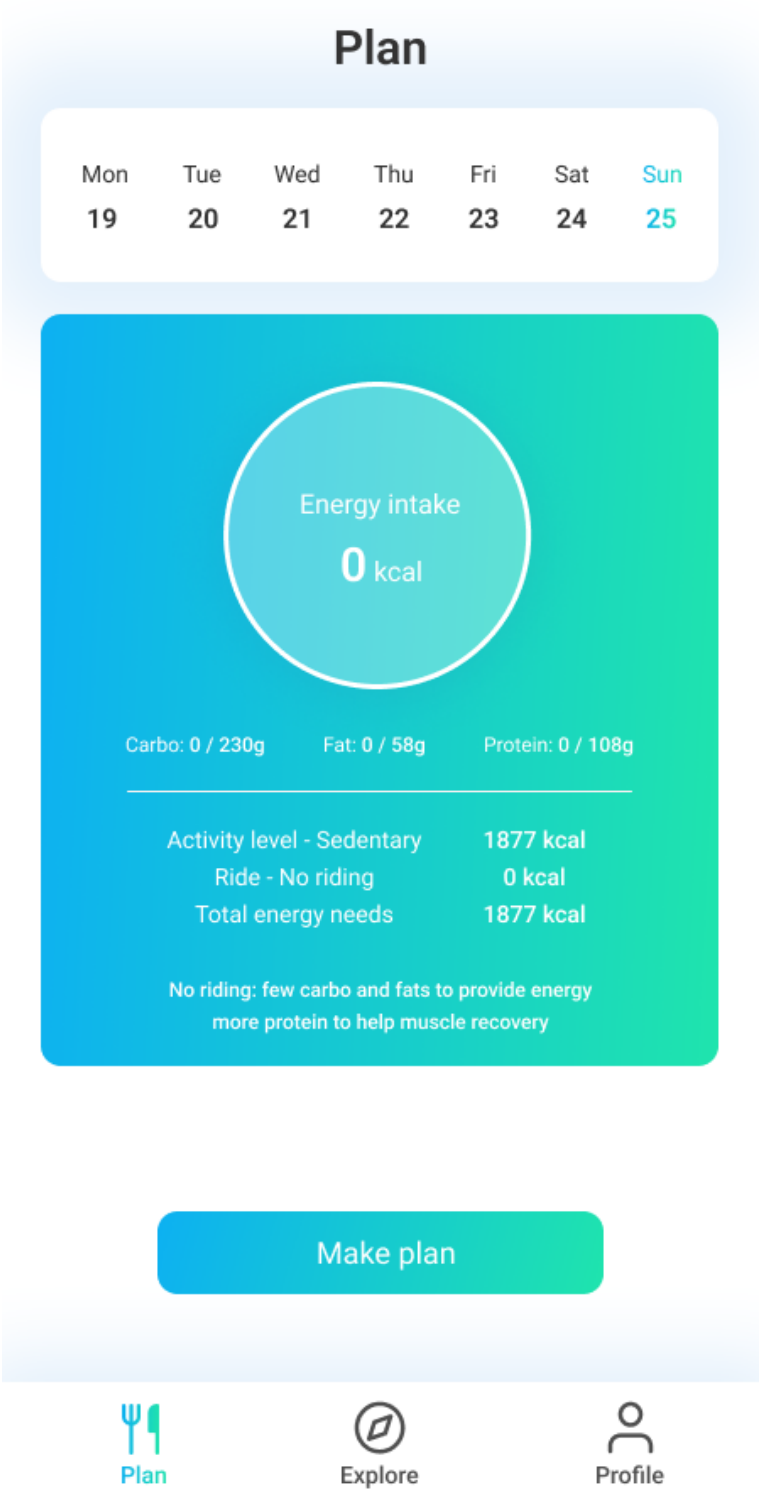
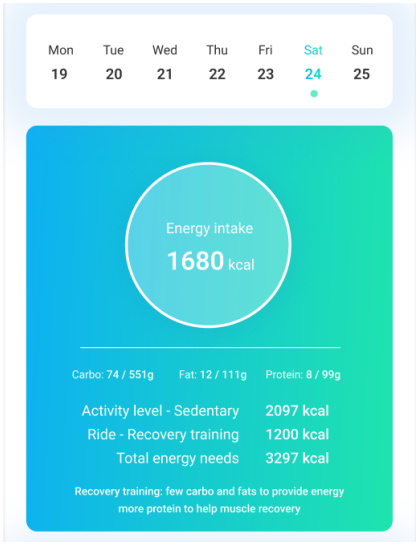
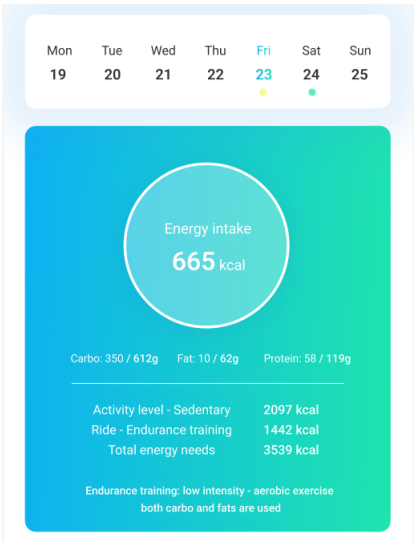


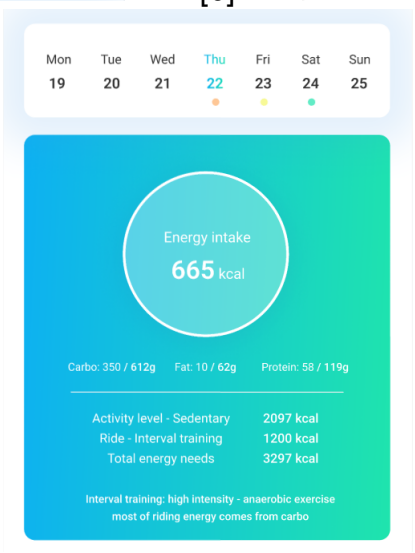
Figure 4.4: Lo-fi page for Biological Explanation
Explanation for No riding day. It shows an overall view of the main page.



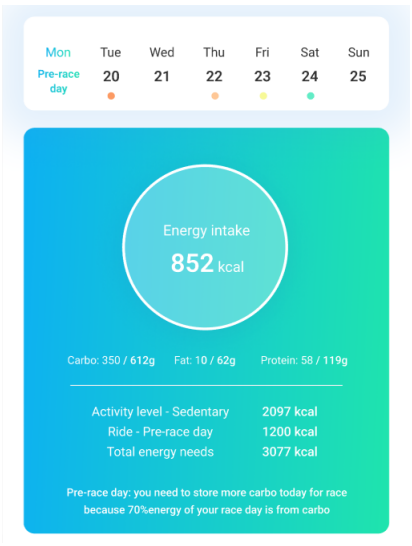
[a]



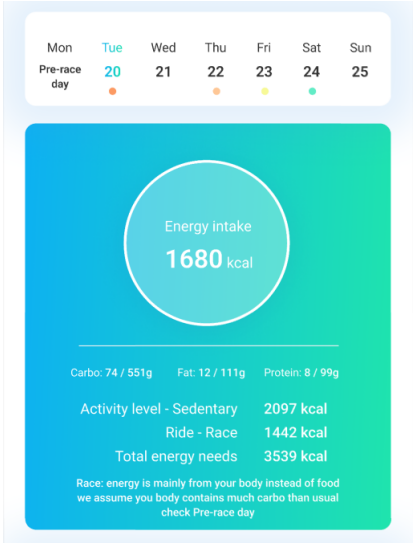
[b]



[c]



[d]



[e]

Figure 4.5: Lo-fi pages for Biological Explanation

[a][b][c][d][e] show the explanations for Recovery training, Endurance training, Interval training, Pre-race day, and Race separately.

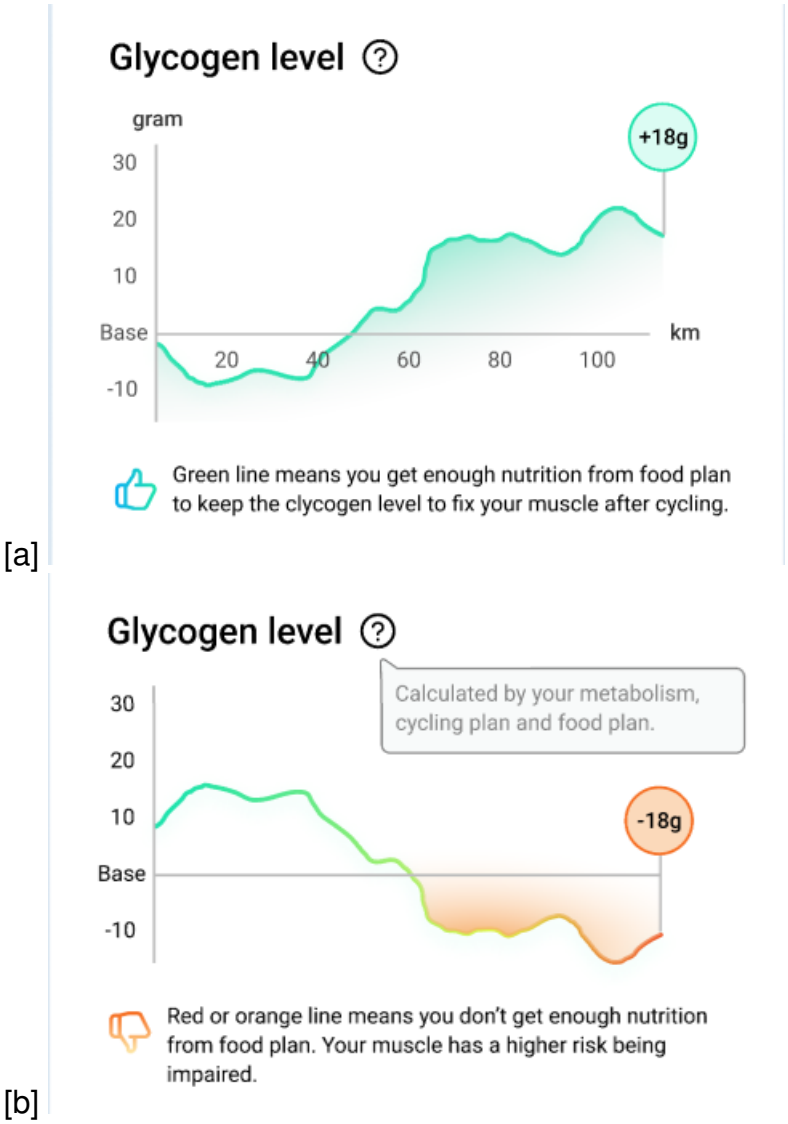


Figure 4.6: Lo-fi pages for Biological Explanation
[a] explains a good glycogen level while [b] presents the opposite result.

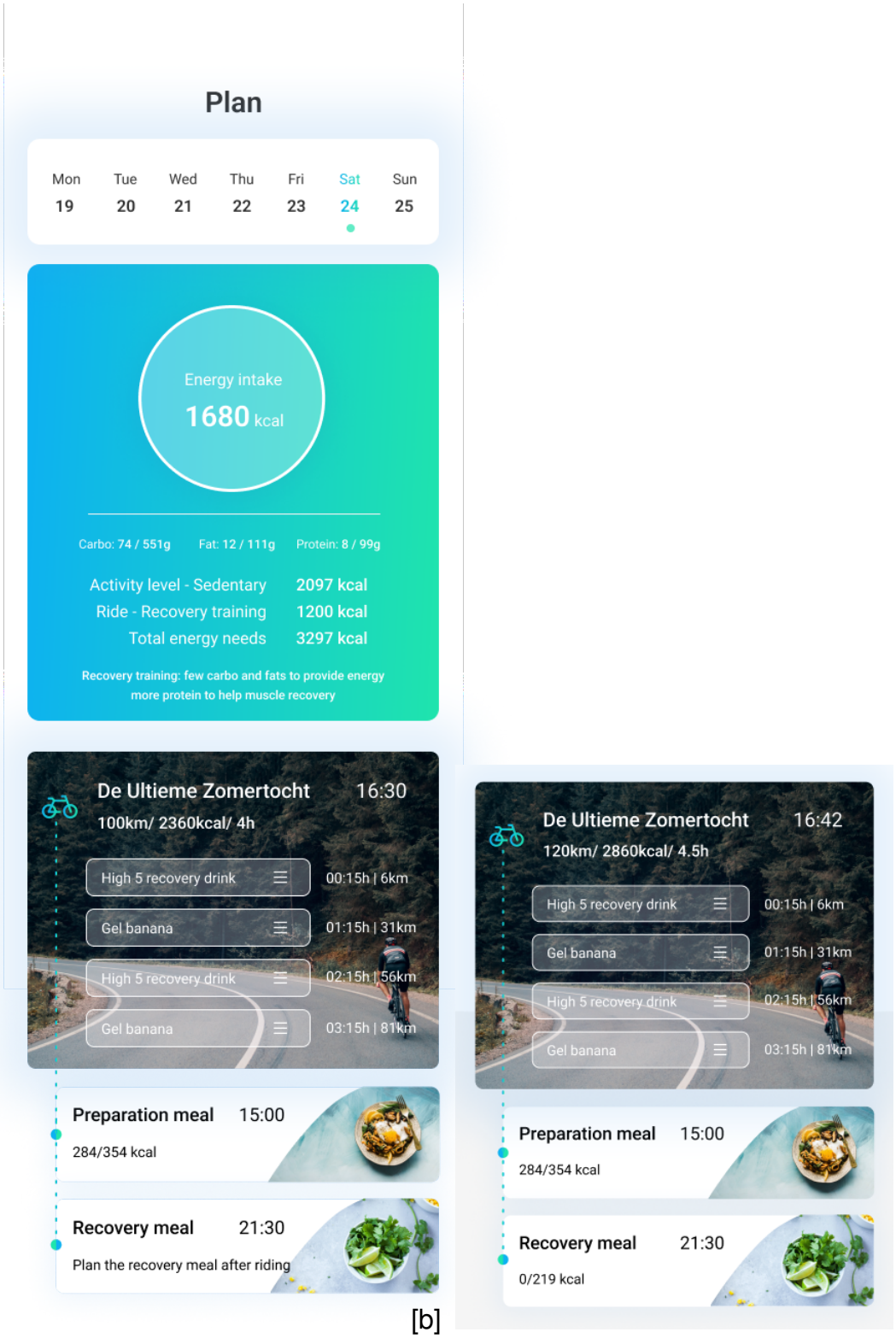


Figure 4.7: Lo-fi pages for Result Renew and Comparison
[a] dislays the information before riding. [b] shows what part of the information is changed after cycling.

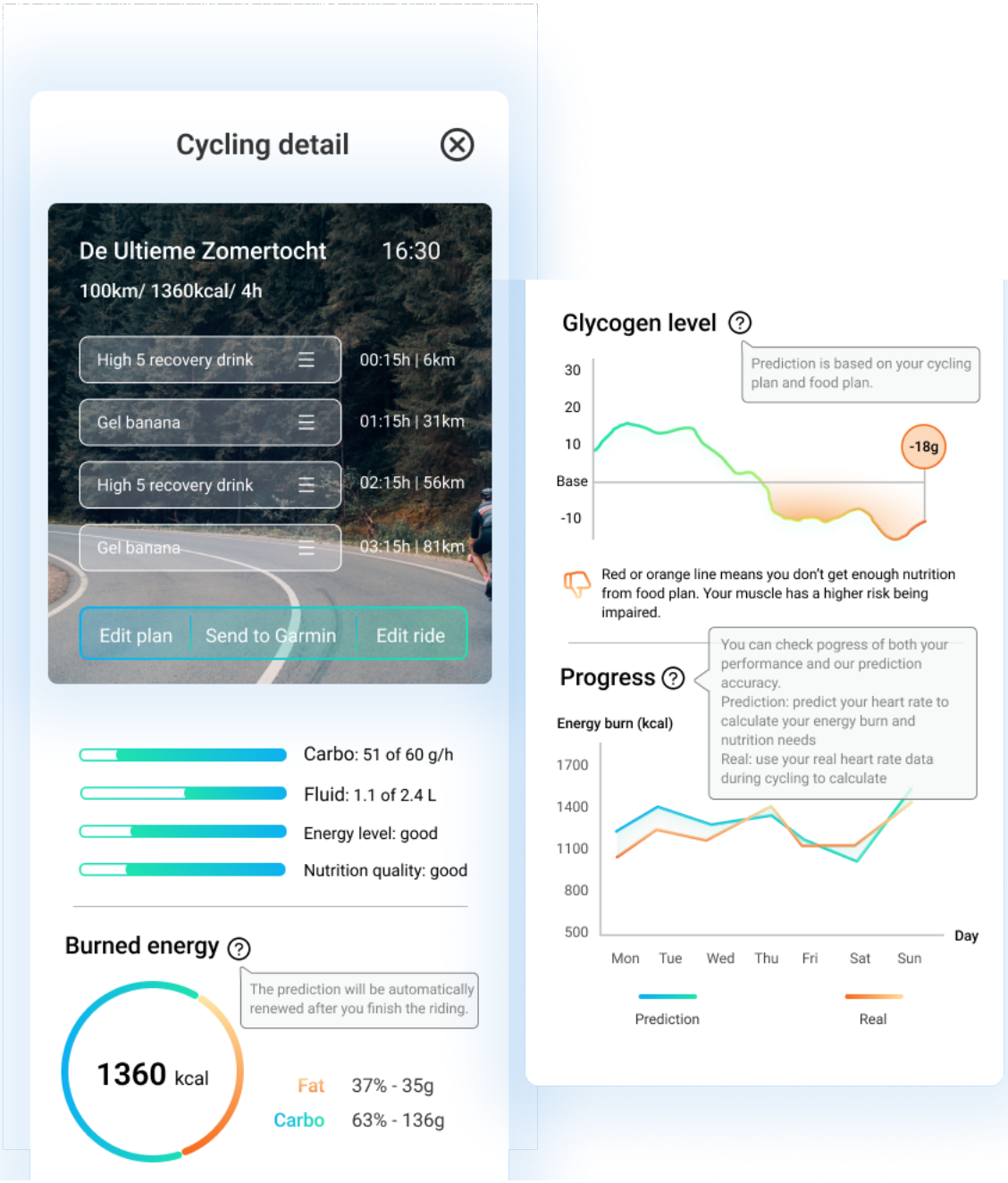


Figure 4.8: Lo-fi pages for Result Renew and Comparison
This figure contains the explanations for Burned energy and Progress.

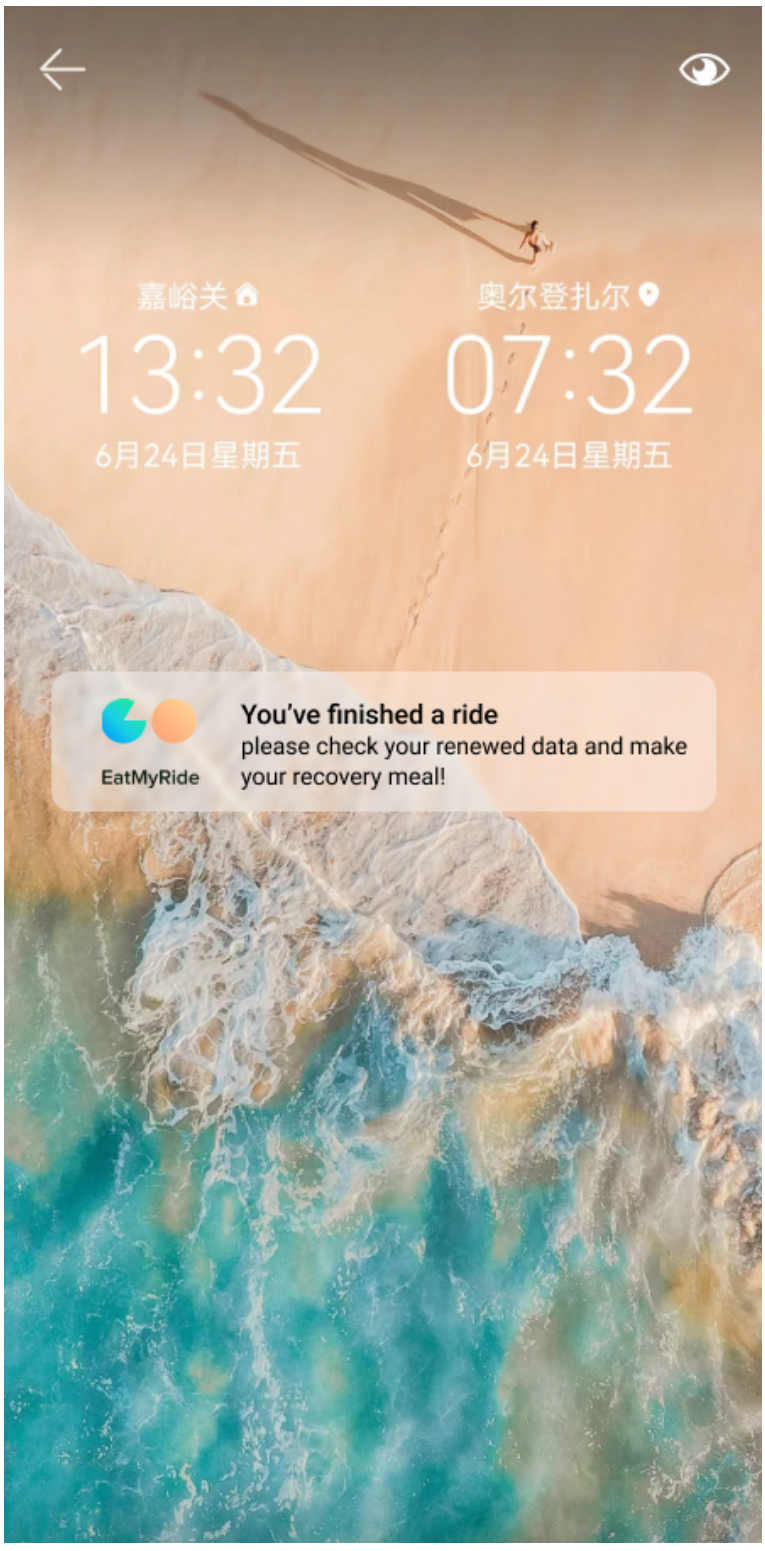


Figure 4.9: Lo-fi pages for Result Renew and Comparison
This page is especially designed for notification.

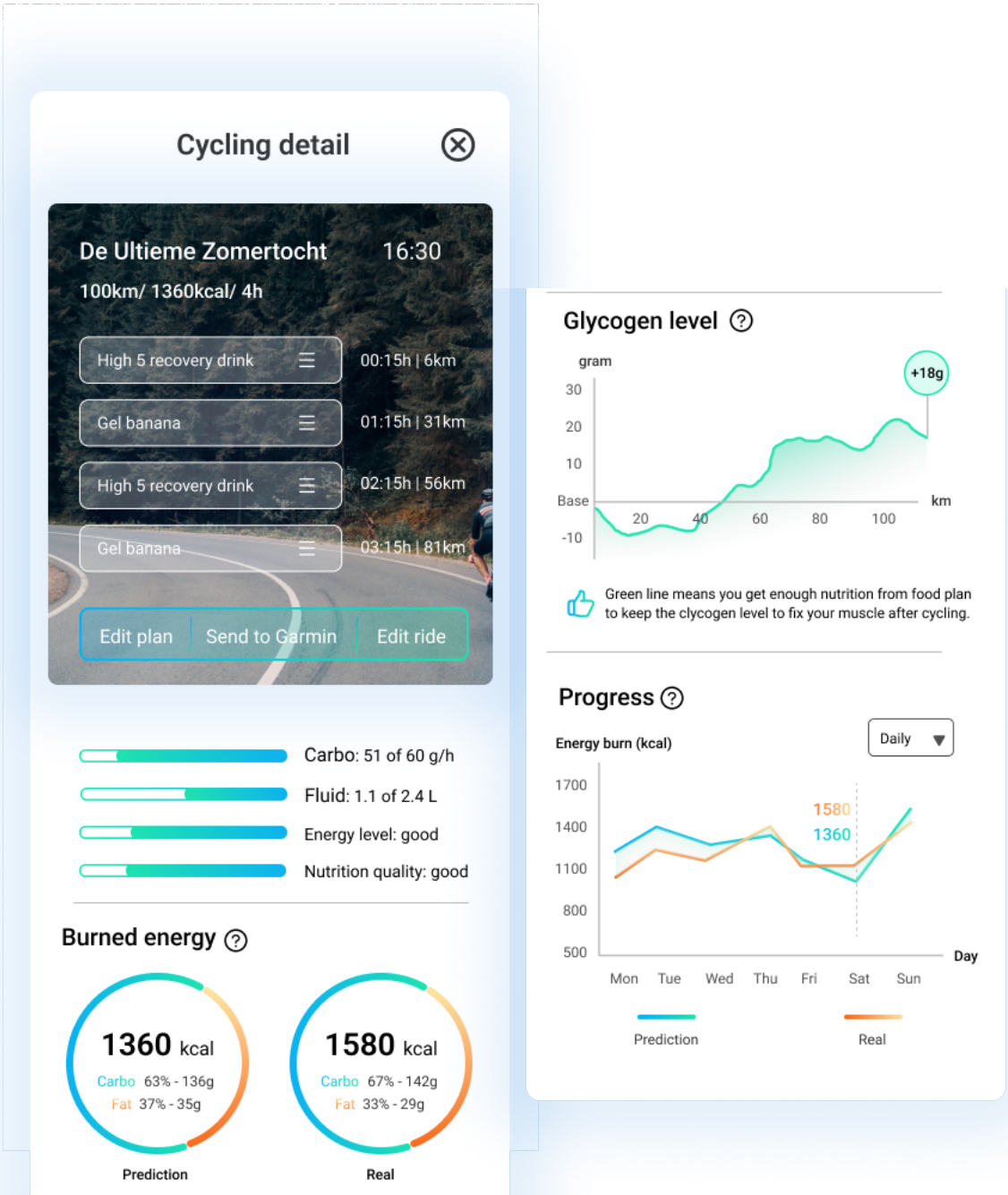


Figure 4.10: Lo-fi pages for Result Renew and Comparison
This figure shows the interface after actual cycling.

Repetitive information, the knowledge that users already know, and wordy explanations are discarded because over-explanation can cause significant side effects on users' overall feeling and attitude toward the algorithm [36], [46], [47]. All explanations are designed in a context-aware precondition and guided by users' aims, needs, and benefits instead of simply explaining how the algorithm works [25], [57].

New explanations introduce the workflow step by step according to the requirement of human-centered XAI [25], [75]. For example, new explanations firstly inform users that their profile data will be used to predict their metabolism (including HR), energy needs, and consumption. Then the carbo intake will be calculated based on the predicted cycling HR, road information, and metabolism data. The predicted energy consumption of cycling will be displayed in the UI first. After finishing the cycling, the real HR data will be uploaded automatically, and the prediction of energy consumption will be renewed. There will be a comparison and progress graphs showing the performance of AI, which is more intuitive. Important information like the HR data appears several times to show how the system processes the predicted and real HR data, which is the essential reason why the algorithm keeps making progress. Therefore, the causal link is clearer and more logical, and the final comparisons are more acceptable, understandable, and reasonable for users.

The new design also includes high-level explanations of biological knowledge, which are the fundamental working principles of the system. According to the previous research [25], [59], users are interested in high-level explanations of the model's underlying principles instead of pure algorithm calculations, and globally explaining the system is more effective and useful in re-building the user mental model. Those biological explanations are presented in natural language with didactic statements according to the human-centered XAI guidelines [17], [40].

4.2 Low-fidelity Evaluation

The lo-fi evaluation includes two parts: 1) the Explanation Goodness Checklist (EGC) 2) measuring the user mental model. The EGC (See Appendix C.1) has been mentioned in the chapter 2. It is for AI experts who have not participated in the model development process to evaluate the goodness of the XAI [23]. Measuring the user mental model is for understanding the users' perception of the application after implementing human-centered XAI, which uses the same method (Think-Aloud Task with Concurrent Question Answering, and Cued Retrospection) applied in the user research of the previous chapter 3. The consent form and information brochure can be found in Appendix A.2.

- **Participant** One AI specialist with a doctoral degree for rating the scale. Five

cycling enthusiasts from previous EMR user research for understanding the user mental model. According to Eiband et al. [14], it is feasible for researchers to use either the within-group design (involving the same users as in investigating mental model), or a between-groups design (involving different but comparable users). The lo-fi evaluation adopted the first option.

- **Method and Measure** The EGC is for evaluating the goodness of the XAI. The checklist is filled up in an electronic format. Think-Aloud Task with Concurrent Question Answering, and Cued Retrospection methods are used to measure the user mental model. Only notes are taken down, no audio or video recording during the online interview.

4.2.1 Explanation Goodness Checklist

The specialist was required to experience the EMR lo-fi prototype, then fill out the EGC. Opinions from the specialist were collected during and after the process. Fig. 4.11 shows the completed EGC. All seven questions were checked as *YES*. The specialist mentioned that answers for questions 1 and 5 mainly come from the algorithm's simplicity. The participant expressed that the model's workflow is straightforward and does not contain complex or abstract results (e.g., results from different layers of deep learning algorithms). The answers to questions 2, 3, and 4 come from the biological explanations and the predicted and real HR, which helps users get an overall picture of the model. Questions 6 and 7 are checked as YES primarily due to the explanations and display of the prediction results update and comparison.

4.2.2 User Mental Model

The process of measuring the user mental model of the lo-fi prototype is similar to that of the original EMR application. Evaluators mentioned that they have a relatively clear and logical perception of how this model works, which is confirmed when using probe questions to learn their objective understanding of the model.

Participants understood that their profile data is used to predict metabolism, energy consumption, and needs at the registration step. When filling up the cycling information and nutrition plan, they knew that selecting routes could provide weather information which is not predictable in selecting from course profiles. Participants also learned that their energy intake is decided by their predicted HR, metabolism, and cycling data. Besides, they mentioned that the new explanations are less confusing than those in the original EMR application because of the deleted sweat rate and previously explained weather information.

1. The explanation helps me **understand** how the [software, algorithm, tool] works.

YES	✓
NO	

2. The explanation of how the [software, algorithm, tool] works is **satisfying**.

YES	✓
NO	

3. The explanation of the [software, algorithm, tool] sufficiently **detailed.**

YES	✓
NO	

4. The explanation of how the [software, algorithm, tool] works is sufficiently **complete.**

YES	✓
NO	

5. The explanation is **actionable**, that is, it helps me know how to use the [software, algorithm, tool]

YES	✓
NO	

6. The explanation lets me know how **accurate or reliable** the [software, algorithm] is.

YES	✓
NO	

7. The explanation lets me know how **trustworthy** the [software, algorithm, tool] is.

YES	✓
NO	

Figure 4.11: This checklist is filled up by the AI specialist. It is used to measure the goodness of the XAI.

For high-level explanations, participants learned the different cycling and exercise types from a biological aspect and understood why the nutrition needs to vary with different cycling intensities. They knew the relation and nutrition calculation methods between race and pre-race day. Moreover, participants understood that the red line of glycogen level means they have not got enough nutrition from the food plan, therefore, the glycogen in their body is insufficient for repairing the muscle. Whereas the green line means the nutrition plan meets the requirement of the energy needs for cycling.

Participants understood that the real HR data will be automatically uploaded, and the energy consumption will be renewed after finishing cycling. The essential reason why the algorithm makes progress is by closing the gap between the predicted and real HR data.

For the overall UX, four participants appreciated the design of the help button with a text box because it enables them to look after the information and explanations when they need and want to know. Moreover, the UIs look more concise after hiding the constant textual explanations participants have read and learnt. All participants expressed that lowering the amount of meals they need to fill up (only preparation and recovery meals are kept) gives them a sense of control and freedom. Besides, they all agreed that the experience of selecting drinks and food becomes much more natural and comfortable.

However, participants also pointed out that there are redundant explanations for filling up the cycling data, such as the explanation for *Ride type* (See 4.2[a]). Besides, they hope the comparison between *From routes* and *From course profiles* can be stronger, and there will be more explanations for the *From routes* choice. For the *Result renew and comparison* UIs, participants clearly expressed that the link and relation between the previous and renewed predictions are not strong enough, and the explanations should highlight the difference and comparison.

4.2.3 Conclusion

The effectiveness of the explanations is not merely attributed to the human-centered XAI methods, but also the simplicity, understandability, and interpretability of the algorithm and the model itself. After applying human-centered XAI, users' overall understanding of the application becomes more precise and logical. Participants' mental model has met the requirements of the target mental model.

Compared to the original user mental model, participants have learnt new knowledge such as the core principle stressed by the EMR expert, which is how the algorithm keeps progressing. They have also understood the overall workflow of the model from the step-by-step global explanations. High-level explanations such as

biological knowledge help rebuild the user mental model, and they also assist users in knowing the underlying reasons why the model makes different predictions based on different situations. Moreover, designing in a context-aware situation enables the explanations to comply with users' aims and needs, which is also helpful in deciding how to construct the textual and visual explanations and which explanations should be kept or discarded. The explanations for *Select ride* and *Result renew and comparison* should be improved or re-designed.

Chapter 5

High-fidelity Design and Evaluation

This chapter includes 1) the hi-fi design based on the users' opinions from previous lo-fi testing 2) the evaluation of the hi-fi prototype. Compared to the previous lo-fi chapter, this chapter focuses more on evaluating the user mental model, satisfaction, and trust.

5.1 High-fidelity Design

The drawbacks of the lo-fi pages reported by participants in the evaluation phase are re-designed in the hi-fi prototype. After the adjustment, the hi-fi is added with interactive activities for the second time in user testing.

For the *Cycling and Nutrition Plan*, the comparison between *From routes* and *From course profiles* is stronger and more explanations are provided for the *From routes* (See Fig. 5.4). For example, existing routes also include past cycling HR and road information per segment, improving prediction accuracy.

For the *Result Renew and Comparison*, the explanations highlight the importance of renewed data which is the real HR and road information. The *Burned energy* module also explains how the model acquires and uses the updated real data. The explanations in the *Progress* section become more precise, which is expected to form a more intuitive and clear comparison between predicted and real results (See Fig. 5.2).

5.2 High-fidelity Evaluation

XAI enables users to quickly figure out the underlying working principles of the algorithm by extracting essential knowledge from AI, which rebuilds the user mental model [23], [38]. Research also indicates that besides the mental model, trust, persuasiveness, curiosity, and satisfaction should also be considered as critical criteria

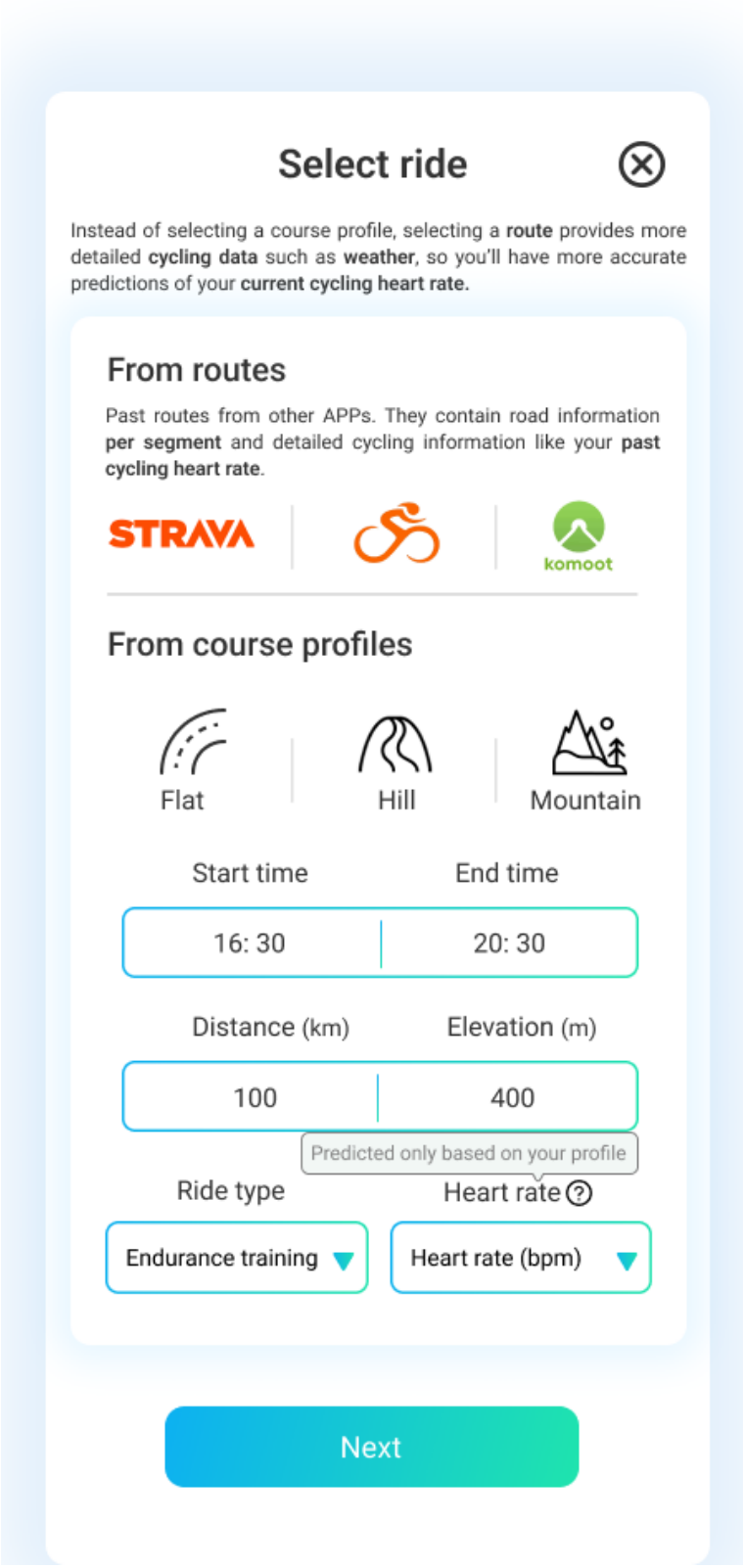


Figure 5.1: Hi-fi page for Cycling and Nutrition Plan
The figure provides more explanations for *From routes* to enhance the contrast.

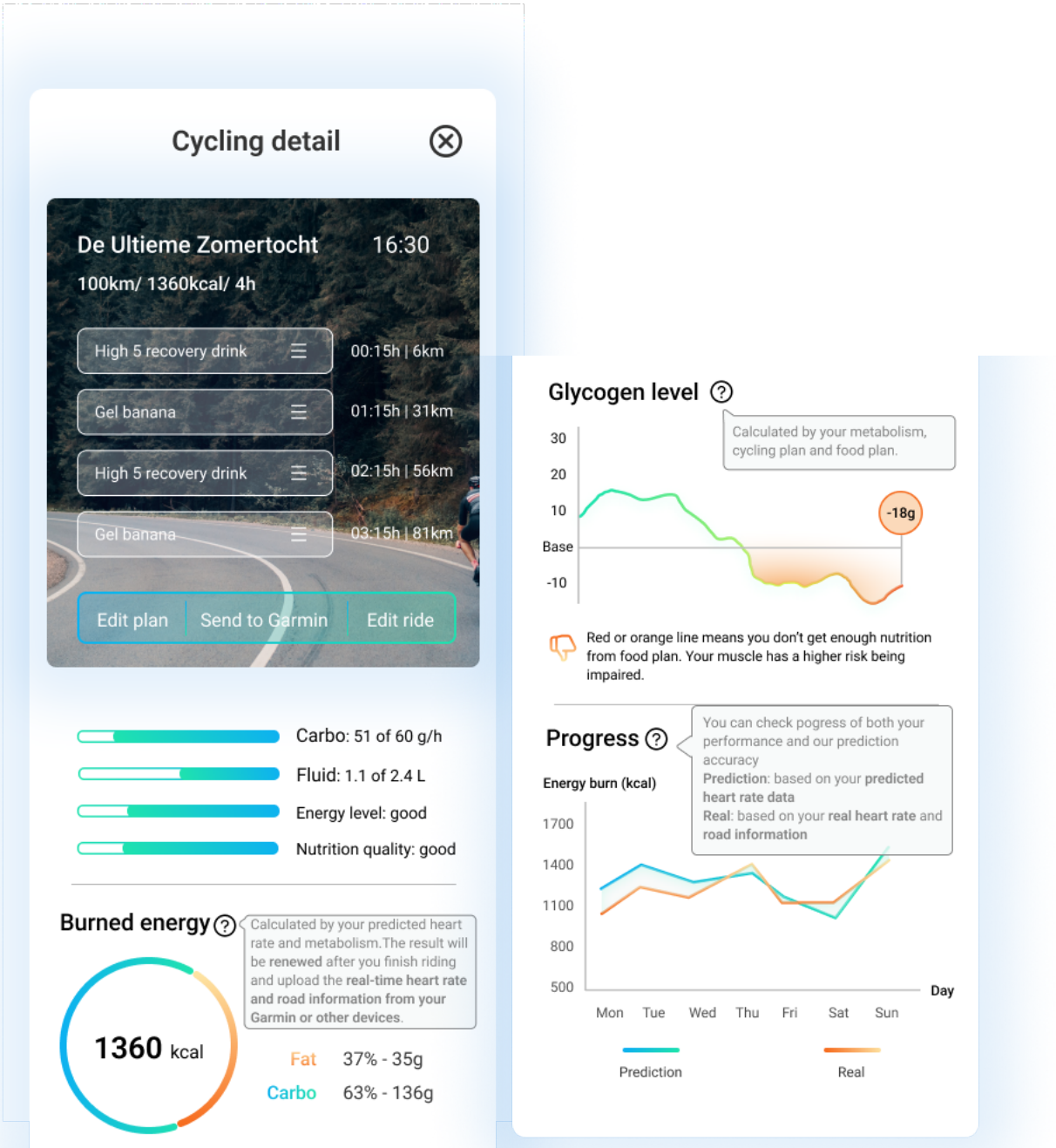


Figure 5.2: Hi-fi pages for Result Renew and Comparison

The explanations are much clear and precise, which also highlights the comparison between predicted and real energy consumption.

in human-centered XAI [11], [23], [56]. Due to the aim of the thesis research that is highly related to users' trust, besides the user mental model, the hi-fi evaluation will measure users' trust and satisfaction.

Participants were asked to perform the whole process from registration to reviewing the comparison on the hi-fi prototype. The observation is based on four stages which corresponding to the four different categories. Think-Aloud Task with Concurrent Question Answering, Cued Retrospection, and Diagramming will be used in measuring the user mental model and collecting users' feedback and opinions. According to the previous research, these methods are useful in measuring the effectiveness of counterfactual explanations from XAI [68].

Explanation Satisfaction Scale (ESS) and Explanation Trust Scale (ETS) (See Appendix C.2 and C.3) are for evaluating users' satisfaction and trust separately [23]. The consent form and information brochure can be found in Appendix A.3.

- **Participant** The hi-fi evaluation invited the same five cycling enthusiasts from lo-fi evaluation. The choice named within-group design is reasonable according to research from Eiband et al. [14].

- **Method and Measure** ESS and ETS for measuring users' satisfaction and trust. The scales are filled up in an electronic format. Think-Aloud Task with Concurrent Question Answering, Cued Retrospection, and Diagramming methods are used to measure the user mental model. Only notes are taken down, no audio or video recording during the online interview.

- **Procedure** Participants were required to experience the EMR hi-fi prototype with Think-Aloud Task with Concurrent Question Answering, and Cued Retrospection methods. Then the user mental was visualized by Diagramming. After that, the ESS and ETS were sent to users for rating. Participants' opinions and feedback were collected during and after these two sections.

5.2.1 User Mental Model

The user mental model of the hi-fi design is relatively the same as that of the lo-fi prototype. Participants mentioned that the explanations for *Ride type* present a stronger comparison between *From routes* and *From course profiles*, and there is more information on *From routes*. For instance, participants understood that selecting *From routes* can provide more accurate predictions of cycling HR not only by fetching the weather data but also from the past cycling HR data and road information per segment. The importance of cycling HR data is also outlined in this page.

For the *Result renew and comparison*, participants claimed that they had a better understanding of how the real cycling data such as HR and road information is uploaded to the model in the *Burned energy* section, which is absent in the explanations from the lo-fi. Moreover, participants also reported that the comparison in the *Progress* module is more concise and comparable.

The user mental model of the hi-fi prototype is made by collecting answers from probe questions related to the participants' objective understanding of the hi-fi workflow (See Fig. 5.3). Figure (**??**) presents a comparison between the original and rebuilt user mental model.

Compared to the original user mental model, the new model complements the information that the expert stressed as the essential principles of the EMR application from four categories *Registration, Cycling and Nutrition Plan, Biological Explanation*, and *Result Renew and Comparison*. Users not only understand what performance can improve the prediction result (e.g., choose the desirable cycling category), which aligns with Wachter et al. [39], but also figure out the underlying principles of how the algorithm makes progress, which is closing the gap between the predicted and real HR data. The logic of the current user mental model is much clearer than the original one, which contains numerous speculations and presumptions from participants themselves (See Fig. 3.6). Besides, participants have also known the difference in the algorithm makes different calculations and recommendations from a biological aspect.

From the overall UX side, users expressed that the process of using the hi-fi prototype is more smooth and more natural, and there are fewer steps required to finish tasks such as filling up the personal profile, cycling data, and nutrition plan. Besides, three participants straightforwardly expressed that they are willing to spend time uploading more data and reading the explanations due to the improvement of the UX.

5.2.2 Explanation Satisfaction Scale and Explanation Trust Scale

The ESS and ETS were sent to users after measuring the user mental model. The results of the scales contain mean value and standard deviation (SD) from five participants' ratings (See Fig. 5.5 and Fig. 5.6).

In the ESS, users held relatively different opinions on statements 6, 7, and 8. For item 6, participants generally expressed that the explanations help get more accurate predictions for the nutrition plan so that the cycling performance will be improved, and they can pay more attention to the nutrition for cycling because of the biological knowledge.

However, two participants also indicated that those explanations do not differ

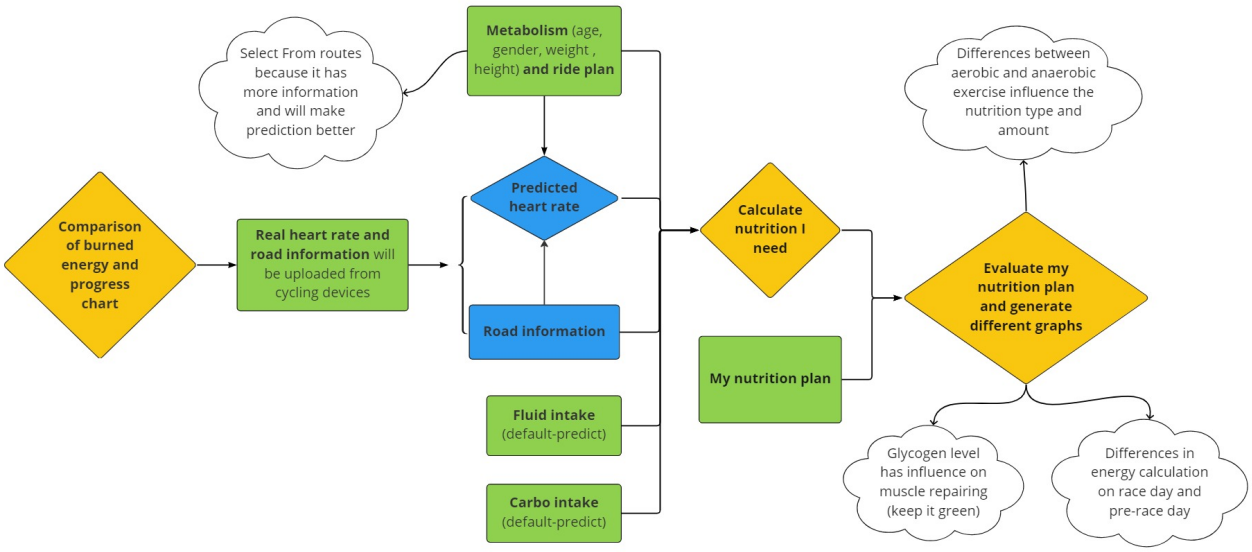


Figure 5.3: New user mental model after implementing XAI

The figure explains the user mental model of the EMR hi-fi prototype. The green color represents users and yellow is for the application. Square and diamond mean input and output separately. The blue color stands for important information that is newly acquired.

from the knowledge in the nutrition book, which can be learnt after several times. Because there are only limited types of textual explanations content, and no new knowledge comes in during the usage of the application, users will build a similar prediction model in their mind, except the calculation will not be as precise as the application's algorithm. Consequently, the application may be given up by users as the similar applications they previously had used.

One participant also mentioned the explanations for *Glycogen level* as an example. The participant expressed that *It is good to know what the different colors mean, but I want to know more about what I should do when the line turns red. For example, how much and what kind of nutrition should I take? Will there be a situation like the glycogen in my body is too much and what will happen?*

For statement 7, participants' opinions vary in the definition and display method of accuracy. Participants wondered whether showing the comparison between the predicted and real HR data and energy consumption can be regarded as a display of accuracy. Evaluators who hold disagreeing opinions think that the accuracy should be presented clearly and precisely in a percentage number.

Besides, all evaluators rated a high score (3, 4, or 5) for item 8 because the explanations do not contain information on when users should not trust the algorithm or when the prediction is inaccurate. However, two participants questioned this statement and why the algorithm should provide predictions or recommendations to users

New table

	Before	After
Registration	• Personal profile data (maybe weight, height, gender, and age) is used in predicting the energy consumtion and needs	• Personal profile data (weight, height, gender, and age) is used in predicting the metabolism, which will be used in predicting the HR and energy
Cycling and Nutrition Plan	• *Select routes* probaly provides more data than *Select from course profiles*	• *Select routes* is more desirable than *Select from course profiles.* Because *Select routes* provides more data than *Select from course profiles,* such as weather and road information (in segments), and past cycling HR data. Therefore, choosing *Select routes* can predict HR data (before cycling) more accurately.
Biological Explanation	• Different training types have different nutrition needs and consumption • Foodplans for the pre-race day and race day are related together • Red line of the *Glycogen level* means the glycogen in the body decreases, whereas the green line means the opposite.	• Low intensity cycling (recovery training, endurance training) belongs to aerobic exercise. Both carbo and fats are needed • High intensity cycling (interval training, race) belongs to anaerobic exercise. The energy mostly comes from carbo • The energy for the race mainly comes from the energy storage on the pre-race day instead of the food intake on the race day. Therefore, the food plan on the pre-race day is more important. • Red line of the *Glycogen level* means the body does not contain enough glycogen and the muscle wil be impaired, whereas green line means the opposite.
Result Renew and Comparison	• There will be a coparison between the predicted and real energy consumption.	• After cycling, users real HR data and road information will be uploaded from their device to the model to calculate the real energy consumption • The algorithm makes progress by closing the gap between the predicted and real HR data.

Figure 5.4: Comparison between the original and rebuilt user mental model
The comparison is based on four different types of functions.

when it knows the result is unreliable.

In the ETS, participants gave a high score of 4 to statement 3, and their reasons are relatively similar. Participants claimed that although they understand how the algorithm makes progress and can see it directly from the *Progress* module, they firmly believe that there is numerous data and situations that the model cannot get and predict. Three participants also mentioned that although the real HR data and road information will be used to update the result, calculations finished by the algorithm are still based on numerous predictions, such as metabolism from the personal profile data and other factors in counting energy consumption. The exact amount of energy consumption and nutrition needs are always generated from the predictions of some unknown information.

New table

Explanation Satisfaction Scale **1 to 5** is corresponding to **Strongly Agree** and **Strongly Disagree**		**SD**
1. From the explanation, I **understand** how the [software, algorithm, tool] works.	**1**	**0**
2. This explanation of how the [software, algorithm, tool] works is **satisfying**.	**1.2**	**0.16**
3. This explanation of how the [software, algorithm, tool] works has **sufficient detail.**	**1**	**0**
4. This explanation of how the [software, algorithm, tool] works seems **complete**.	**1**	**0**
5. This explanation of how the [software, algorithm, tool] works **tells me how to use** it.	**1.2**	**0.16**
6. This explanation of how the [software, algorithm, tool] works is **useful to my goals**.	**1.4**	**0.24**
7. This explanation of the [software, algorithm, tool] shows me how **accurate** the [software, algorithm, tool] is.	**1.8**	**0.56**
8. This explanation lets me judge when I should **trust and not trust** the [software, algorithm, tool]	**4**	**0.4**

Figure 5.5: The number beside each question is the mean value of all participants' answers.

New table

Explanation Trust Scale **1 to 5** is corresponding to **Strongly Agree** and **Strongly Disagree**		**SD**
1. I am confident in the [tool]. I feel that it works well.	**1.4**	**0.24**
2. The outputs of the [tool] are very predictable.	**1**	**0**
3. The tool is very reliable. I can count on it to be correct all the time.	**4**	**0.4**
4. I feel safe that when I rely on the [tool] I will get the right answers.	**1.2**	**0.16**
5. The [tool] is efficient in that it works very quickly.	**1**	**0**
6. I am wary of the [tool]. (adopted from the Jian, et al. Scale and the Wang, et al. Scale)	**4.4**	**0.24**
7. The [tool] can perform the task better than a novice human user. (adopted from the Schaefer Scale)	**1**	**0**
8. I like using the system for decision making.	**1.2**	**0.16**

Figure 5.6: The number beside each question is the mean value of all participants' answers.

5.2.3 Conclusion

In conclusion, the completeness of the user mental model after experiencing the hi-fi design improves considerably compared to the original user mental model. The new user mental model has met the requirements from the target mental model in Chapter 3. Participants have learnt what performance can improve the prediction results of the algorithm. However, the design has disadvantages considering the users' goals and trust.

Firstly, although users appreciate the biological explanations in the hi-fi, the limited and non-changing explanations still prevent users from keeping learning. Moreover, users are more concerned about what the explanations can contribute to their aims. Users understand how the algorithm progresses, and they believe that improving it leads to more accurate predictions. However, they doubt to what extent the algorithm's improvement can substantially enhance the nutrition plan and cycling performance, especially for the latter. Because participants believe that the food plan's influence is limited, they consider the system's recommendations as a reference rather than a decisive factor in improving riding performance.

Besides, participants stressed that since there is numerous information the model cannot sense or get as input feature, they trust their feeling of the body more than the prediction from the algorithm, which confirms the conclusion from Hoffman et al. [23] that users' performance is affected by their level of epistemic trust, in other words, the cognitive or deep level of trust. For the accuracy of the predictions, some participants believe that a desirable way to present it clearly and intuitively is to display the specific value of accuracy. Participants also reported that the explanations lack information on when they should and should not trust the algorithm, which is required by the work of Liao et al. [25] that the XAI should inform users about the limitation of the algorithms. However, some participants question its necessity, considering the algorithm knows that the predictions are not trustworthy and reliable.

Chapter 6

Discussion

This chapter includes the answer to the research question, the insightful findings during the research process, and the limitation of the thesis work.

6.1 Answer to the Research Question

How do we implement human-centered XAI methods in a practical design case to match users' needs, and improve their understanding and trust in the algorithm?

The prerequisite of applying the human-centered XAI in design practice is that the research should be in a context-aware situation. Researchers should investigate stakeholders' profiles, such as their backgrounds, goals, and action sequence, and the overall explanations should align with users' goals and requirements. After that, build the expert mental model and user mental model of the product, and collect their opinions and feedback using different methods. Based on the mental models and user profile, build the target mental model which closes the information gap between experts and users and fit users' aims and interest at the same time.

Therefore, the EMR application must ensure what its target users need and focus on those points instead of making its function much more extensive. For example, it includes every meal plan in a day, whereas it should highlight the primary function, assisting cyclists in customizing the cycling nutrition plan. Therefore, this research indicates that the EMR application should leave out all meal plans except the cycling nutrition plan, the preparation, and the recovery meal plan. Besides, the application should re-design the way of presenting the product information when users search the food and drinks when filling out the nutrition plan.

After setting the target mental model, the explanations should be presented using post-hoc methods to users, such as textual and visual explanations, and provide new knowledge instead of information that users already know. Moreover, explanations

should be presented in a step-by-step approach using natural language with didactic statements.

Therefore, this research uses a simple chatbox form to realize these requirements. It not only allows users to seek helpful information when needed, but also conceals the information that they already know after learning one to two times, leaving a certain level of choice to the users.

For the content of the explanations, counterfactual (why A not B) and high-level explanations are more effective in rebuilding the user mental model and correcting users' biases. The explanations should let users understand the product's workflow and what they can do to improve the algorithm's results. More importantly, the explanations and recommendations should not be complicated or overwhelming. The number of reasons in a single explanation should be limited to two or three, and the recommendations generated from the algorithm should leave users a certain extent of freedom. Otherwise, users will lose interest and curiosity to explore the product and explanations and stop trusting the system.

Therefore, this research suggests that the EMR application adjusts the current content of the explanations. Firstly, the application should leave out all the repeated explanations and only present the leading causes to the users. Secondly, the explanations should present the system's workflow step by step logically instead of directly showing all the possible causes, regardless of whether they are relevant to users. For example, the explanations for which users fill out their cycling and nutrition data lead to a counter-effect. Finally, the application should provide more explanations related to the biological aspect, which fits users' aims and improve users' trust, satisfaction, and curiosity about the product.

After implementing the XAI, corresponding evaluations are needed because the effect of XAI is different for each user. This research indicates that the EMR should conduct several user evaluations to assess the new user mental model and other traits like trust and satisfaction using professional scales like the Explanation Satisfaction Scale and Explanation Trust Scale in a quantitative way. Alternatively, it can use methods like Think-Aloud Task with Concurrent Question Answering, Cued Retrospection, and Diagramming to measure the effect of XAI qualitatively.

6.2 Insight

There are several insightful findings during the research process. Firstly, users' trust and satisfaction with the model are influenced by the application's explanations and overall UX. Higher quality UX can persuade users to invest more time in learning explanations and providing information and feedback, improving both the XAI's effectiveness and the algorithm's accuracy.

Contrarily, users' trust in the algorithm can significantly decrease when 1) the UX is unwell-designed and 2) the algorithm shows its complexity and intelligence on purpose. For example, participants straightforwardly expressed doubt when there is no provisioned product information in the *Add drinks* and *Add products* pages. Moreover, users complained that the system automatically pops up five new food plans which cannot be canceled, waiting to be filled out after they finished the cycling nutrition plan, which suppressed their confidence and curiosity in exploring the application because they felt constrained by the recommendations and lost their autonomy.

Besides, users' epistemic trust (cognitive level of trust) profoundly impacts the users' perspectives and attitudes toward the system. For example, most participants claimed that they trust their feeling about the body much more than the predictions from the algorithm because there is a quantity of data that the application cannot sense or get. Therefore, the model's deep level of trust and perceptions will probably only be influenced and changed after a long period of using the product in real life.

For the explanations themselves, users prefer that new knowledge is continuously provided. Instead of presenting some unchanged high-level explanations which can be learned several times, users would like to have various explanations that can change with different situations, which keeps them distinct from physical books. Besides, continued knowledge gathering can also enable users to stick to the application.

6.3 Limitation

There are several deficiencies if the thesis work. Firstly, the whole user research and evaluation process were performed online, whereas the physical experiment could provide better results according to the research of J.S van der Waa et al. [27]. Moreover, although it is feasible for researchers to use either the within-group design (involving the same users as in investigating mental model) according to Eiband et al. [14], the evaluation results probably become more inclusive if a between-groups design (involving different but comparable users) is applied.

All participants are the target users of the EMR application. They cycle more than four times a week or even daily, and the normal riding distance is approximately 50 kilometers. Some have participated in the four-day outdoor riding, which ranges up to 550 kilometers at high altitudes (more than 2000 meters). According to the Nielsen Norman Group, "Five participants will discover over 80 % of the problems" [78]. Research from Janet M. Six and Ritch Macefield has verified that the method from the Nielsen Norman Group has a 95 % confidence level and margin of error of ±18.5 %, which means there is a 95 % chance that a group of five participants will

find between 66.5 % and 100 % of the problems. However, While some groups of Nielsen's study found nearly all of the problems, one group found only 55 % of the problems [79]. Therefore, the results could be more accurate and comprehensive if the lo-fi and hi-fi prototype evaluation recruited more than 5 participants, especially for the feedback from the EGC, as only one AI specialist was invited to fill up the checklist.

Lastly, the ESS and ETS should have been sent to participants in the user research phase to quantitatively evaluate the explanations in the original application, which could provide a direct and clear comparison of the XAI between the initial application and the hi-fi prototype.

Chapter 7

Conclusion

The paper presents a design practice of carrying out the human-centered XAI in a context-aware situation. It contains the process of applying human-centered XAI methods to the EMR application design and evaluating the effectiveness after implementation.

The thesis work first walked through the past literature related to the XAI, then concluded the requirements and methods used in later research. After that, the paper analyzed the current state of the EMR application, including its UIs, algorithm, and workflow. The UIs were divided into four categories *Registration, Cycling and Nutrition Plan, Biological Explanation*, and *Result Renew and Comparison* according to the different functions.

The first-time user research was performed with an expert from the EMR company and target application users. The interview first built the expert mental model, then set the user profile, including users' goals, needs, and behaviors. Next, the user mental model was built using Think-Aloud Task with Concurrent Question Answering, Cued Retrospection, and Diagramming. Based on the user and expert mental model and the user profile, the target mental model is set, which provides guidelines for the explanations design in the lo-fi and hi-fi prototype.

The lo-fi prototype design is built on the previously mentioned four categories. The new explanations introduce the workflow step by step using post-hoc methods. Textual and visual explanations are provided because of their outstanding performance among other XAI approaches for non-technical users. The prototype discards the repetitive information, the knowledge that users already know, and wordy explanations. The design informs users of why the algorithm keeps progressing and compares predicted and real data. High-level explanations related to biological knowledge are provided in natural language with didactic statements. The second time lo-fi testing invited one AI specialist and the same target users for evaluation. The lo-fi prototype used the Explanation Goodness Checklist, and the user mental model was measured with the same methods from the previous user research. After

that, opinions and feedback were collected from the specialist and users.

The hi-fi prototype adjusts the lo-fi design based on the opinions of users. The third-time hi-fi testing was performed only with the same target users. Besides the mental model, the hi-fi evaluation used the Explanation Satisfaction Scale and Explanation Trust Scale to measure users' satisfaction and trust in a quantitative method. User feedback was collected during the evaluation process.

After the prototype design and evaluation phase, the paper answers the research question and provides some insightful findings during the research process, which probably indicate some future research suggestions. Firstly, the need for XAI to inform users when the results from the algorithm are not trustworthy is worth reconsidering as different situations and products. Besides, there is no specific guideline for informing users about the unreliable results or the limitation of the algorithm using the XAI methods, especially for the UI design. Moreover, users' epistemic trust (cognitive level of trust) profoundly impacts both users' behavior and perspective of a product. Therefore, future research could also focus on what kind of XAI methods can influence the users' epistemic trust effectively and efficiently and the detailed practice guidelines on how to implement it in a design case.

Bibliography

[1] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," ser. KDD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1721–1730. [Online]. Available: https://doi.org/10.1145/2783258.2788613

[2] E. Soares, P. Angelov, S. Biaso, M. H. Froes, and D. K. Abe, "Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification," *medRxiv*, 2020.

[3] C. Howell, "A framework for addressing fairness in consequential machine learning," 2018.

[4] E. Krotkov and J. Blitch, "The defense advanced research projects agency (darpa) tactical mobile robotics program," *The International Journal of Robotics Research*, vol. 18, no. 7, pp. 769–776, 1999.

[5] A. Henelius, K. Puolamäki, and A. Ukkonen, "Interpreting classifiers through attribute interactions in datasets," *arXiv preprint arXiv:1707.07576*, 2017.

[6] S. Tan, R. Caruana, G. Hooker, and Y. Lou, "Detecting bias in black-box models using transparent model distillation," *arXiv preprint arXiv:1710.06169*, 2017.

[7] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[8] J. Haspiel, N. Du, J. Meyerson, L. P. Robert Jr., D. Tilbury, X. J. Yang, and A. K. Pradhan, "Explanations and expectations: Trust building in automated vehicles," ser. HRI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 119–120. [Online]. Available: https://doi.org/10.1145/3173386.3177057

[9] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[10] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, "Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation," in *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, 2018, pp. 1–8.

[11] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 0210–0215.

[12] T. Miller, P. Howe, and L. Sonenberg, "Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences," *arXiv preprint arXiv:1712.00547*, 2017.

[13] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–18.

[14] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann, "Bringing transparency design into practice," in *23rd international conference on intelligent user interfaces*, 2018, pp. 211–223.

[15] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 2000, pp. 241–250.

[16] C. T. Wolf, "Explainability scenarios: towards scenario-based xai design," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 252–257.

[17] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[20] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088.

[21] T. A. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. Van Den Bosch, "Human-centered xai: Developing design patterns for explanations of clinical decision support systems," *International Journal of Human-Computer Studies*, vol. 154, p. 102684, 2021.

[22] M. Caro-Martinez, G. Jimenez-Diaz, and J. A. Recio-Garcia, "A theoretical model of explanations in recommender systems," in *ICCBR*, vol. 52, 2018, p. 2018.

[23] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[24] M. Ribera and A. Lapedriza, "Can we do better explanations? a proposal of user-centered explainable ai." in *IUI Workshops*, vol. 2327, 2019, p. 38.

[25] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–15.

[26] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[27] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating xai: A comparison of rule-based and example-based explanations," *Artificial Intelligence*, vol. 291, p. 103404, 2021.

[28] M. Chromik and A. Butz, "Human-xai interaction: a review and design principles for explanation user interfaces," in *IFIP Conference on Human-Computer Interaction*. Springer, 2021, pp. 619–640.

[29] D. M. West, *The Future of Work: Robots, AI, and Automation*. Brookings Institution Press, 2018.

[30] B. Y. Lim and A. K. Dey, "Design of an intelligible mobile context-aware application," in *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, 2011, pp. 157–166.

[31] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th international conference on intelligent user interfaces*, 2015, pp. 126–137.

[32] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[33] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[34] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[35] J. Y. Halpern and J. Pearl, "Causes and explanations: A structural-model approach. part ii: Explanations," *The British journal for the philosophy of science*, 2020.

[36] J. Schaffer, P. Giridhar, D. Jones, T. Höllerer, T. Abdelzaher, and J. O'donovan, "Getting the message? a study of explanation interfaces for microblog data analysis," in *Proceedings of the 20th international conference on intelligent user interfaces*, 2015, pp. 345–356.

[37] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[38] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[39] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[40] D. Gunning, "Explainable artificial intelligence (xai)," *Defense advanced research projects agency (DARPA), nd Web*, vol. 2, no. 2, p. 1, 2017.

[41] B. Shneiderman, "Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 4, pp. 1–31, 2020.

[42] A. Springer and S. Whittaker, "Progressive disclosure: empirically motivated approaches to designing effective transparency," in *Proceedings of the 24th international conference on intelligent user interfaces*, 2019, pp. 107–120.

[43] C. Yang, A. Rangarajan, and S. Ranka, "Global model interpretation via recursive partitioning," in *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 2018, pp. 1563–1570.

[44] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," *Advances in neural information processing systems*, vol. 29, 2016.

[45] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–15.

[46] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[47] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? ways explanations impact end users' mental models," in *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 2013, pp. 3–10.

[48] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu, "Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–12.

[49] R. Kocielnik, S. Amershi, and P. N. Bennett, "Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[50] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 29–38.

[51] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Pro-*

ceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1–52.

[52] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[53] W. R. Swartout and S. W. Smoliar, "On making expert systems more like experts," *Expert Systems*, vol. 4, no. 3, pp. 196–208, 1987.

[54] S. Gregor and I. Benbasat, "Explanations from intelligent systems: Theoretical foundations and implications for practice," *MIS quarterly*, pp. 497–530, 1999.

[55] C. J. Cai, J. Jongejan, and J. Holbrook, "The effects of example-based explanations in a machine learning interface," in *Proceedings of the 24th international conference on intelligent user interfaces*, 2019, pp. 258–262.

[56] N. Tintarev, "Explanations of recommendations," in *Proceedings of the 2007 ACM conference on Recommender systems*, 2007, pp. 203–206.

[57] V. Bellotti and K. Edwards, "Intelligibility and accountability: human considerations in context-aware systems," *Human–Computer Interaction*, vol. 16, no. 2-4, pp. 193–212, 2001.

[58] J. M. Carroll, "Becoming social: expanding scenario-based approaches in hci," *Behaviour & Information Technology*, vol. 15, no. 4, pp. 266–275, 1996.

[59] J. Tullio, A. K. Dey, J. Chalecki, and J. Fogarty, "How it works: a field study of non-technical users interacting with an intelligent system," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007, pp. 31–40.

[60] C. L. Borgman, "The user's mental model of an information retrieval system: An experiment on a prototype online catalog," *International Journal of man-machine studies*, vol. 24, no. 1, pp. 47–64, 1986.

[61] J. Schneider and J. Handali, "Personalized explanation in machine learning: A conceptualization," 2019. [Online]. Available: https://arxiv.org/abs/1901.00770

[62] D. S. Weld and G. Bansal, "The challenge of crafting intelligible intelligence," *Communications of the ACM*, vol. 62, no. 6, pp. 70–79, 2019.

[63] S. Bromberger, *On What We Know We Don?T Know*. Chicago and London / Stanford: University of Chicago Press / CSLI, 1992.

[64] R. A. Bjork, J. Dunlosky, and N. Kornell, "Self-regulated learning: Beliefs, techniques, and illusions," *Annual review of psychology*, vol. 64, pp. 417–444, 2013.

[65] M. T. Chi, N. De Leeuw, M.-H. Chiu, and C. LaVancher, "Eliciting self-explanations improves understanding," *Cognitive science*, vol. 18, no. 3, pp. 439–477, 1994.

[66] L. Rozenblit and F. Keil, "The misunderstood limits of folk science: An illusion of explanatory depth," *Cognitive science*, vol. 26, no. 5, pp. 521–562, 2002.

[67] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proceedings of the national conference on artificial intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004, pp. 900–907.

[68] R. M. Byrne, "Counterfactual thinking: From logic to morality," *Current Directions in Psychological Science*, vol. 26, no. 4, pp. 314–322, 2017.

[69] A. J. Cañas, J. W. Coffey, M.-J. Carnot, P. Feltovich, R. R. Hoffman, J. Feltovich, and J. D. Novak, "A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support," *Report to the Chief of Naval Education and Training*, pp. 1–108, 2003.

[70] E. T. Chancey, J. P. Bliss, A. B. Proaps, and P. Madhavan, "The role of trust as a mediator between system characteristics and response behaviors," *Human factors*, vol. 57, no. 6, pp. 947–958, 2015.

[71] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human factors*, vol. 57, no. 3, pp. 407–434, 2015.

[72] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International journal of human-computer studies*, vol. 58, no. 6, pp. 697–718, 2003.

[73] P. Madhavan and D. A. Wiegmann, "Effects of information source, pedigree, and reliability on operator interaction with decision support systems," *Human factors*, vol. 49, no. 5, pp. 773–785, 2007.

[74] V. Riley, "Operator reliance on automation: Theory and data," in *Automation and human performance: Theory and applications*. CRC Press, 2018, pp. 19–35.

[75] J. Moore and C. Paris, "Requirements for an expert system explanation facility," *Computational Intelligence*, vol. 7, pp. 367 – 370, 04 2007.

[76] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor, "User preferences for hybrid explanations," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017, pp. 84–88.

[77] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens, "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models," *Decision Support Systems*, vol. 51, no. 1, pp. 141–154, 2011.

[78] J. Nielsen. (2000) Why you only need to test with 5 users @ONLINE. [Online]. Available: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

[79] R. M. Janet M. Six. (2016) How to determine the right number of participants for usability studies @ONLINE. [Online]. Available: https://www.uxmatters.com/mt/archives/2016/01/how-to-determine-the-right-number-of-participants-for-usability-studies.php/

Appendix A

Information brochure and consent form

A.1 First time user interview

A.2 Second time lo-fi testing

A.3 Third time hi-fi testing

Information brochure & Consent Form

Our research is collaborating with EatMyRide (EMR), a company that dedicates to helping fanatic cyclists to improve their cycling performance by assisting them in customizing their nutrition plan using the EMR app. By implementing Explainable AI (XAI) (a method that makes AI more transparent and clear, and let things behind algorithms more understandable and trustable for users), we want to improve user experience (UX), give users more reasonable recommendations, and finally let users trust and stick to this app.

You as a participant will go through (go through means to experience the app from the registration part, and imagine you are going to use this app to make nutrition plans for the riding, it contains the whole process) our current EMR APP design first, and give us your valuable feedback and opinions based on your feelings about this APP, especially about its planning system. Afterward, there will be a few questions related to your cycling and eating habits such as how you prepare your nutrition plans for cycling. This will be an online interview last about 30min. There will be no recording but only some notes were taken down, which will be completely anonymous and will only be used in this thesis research. You can quit the interview at any time you want. Also, you can have your data deleted completely.

"I hereby declare that I have been informed in a manner that is clear to me about the nature and method of the research. My questions have been answered to my satisfaction and I agree with my own free will to participate in this research. I reserve the right to withdraw this consent without the need to give any reason and I am aware that I can quit this research at any time. I understand that the research results will be used in scientific publications or the EatMyRide company in a completely anonymous manner. My personal data will not be disclosed to any third parties without my permission. If I request further information about the research now or in the future, I can contact Zoe Zhang (j.zhang-15@student.utwente.nl) or Ethics Committee (ethicscommittee-cis@utwente.nl)."

☐ I give permission for the data collected in this research will be used for the analysis and possibly presented to the board members of the master thesis defense anonymously.

☐ I give permission for the answers to be quoted and published in the research output anonymously.

Signature:

Date:

If you have any complaints regarding this research, please contact the Ethics Committee of the Faculty of Electrical Engineering, Mathematics and Computer Science at the University of Twente, P.O. Box 217, 7500 AE Enschede (NL), email: ethicscommittee-cis@utwente.nl

Figure A.1: First time user interview Information brochure and Consent form

Information brochure & Consent Form

Our research is collaborating with EatMyRide (EMR), a company that dedicates to helping fanatic cyclists to improve their cycling performance by assisting them in customizing their nutrition plan using the EMR app. By implementing Explainable AI (XAI) (a method that makes AI more transparent and clear, and let things behind algorithms more understandable and trustable for users), we want to improve user experience (UX), give users more reasonable recommendations, and finally let users trust and stick to this app.

You as a participant will experience the design EMR APP lo-fi design and give us your valuable suggestions and feedback especially on its planning system. You will receive a web link containing the lo-fi design. You are required to go through the lo-fi design (go through means experience the app from the registration part, and imagine you are going to use this app to make nutrition plans for the riding, it contains the whole process). At the same time, you are encouraged to think out loud when you experience the lo-fi, and you will be asked some questions during and after this process. The whole process lasts about 30min. There will be no recording but only some notes were taken down, which will be completely anonymous and will only be used in this thesis research. You can quit the interview at any time you want. Also, you can have your data deleted completely.

"I hereby declare that I have been informed in a manner that is clear to me about the nature and method of the research. My questions have been answered to my satisfaction and I agree with my own free will to participate in this research. I reserve the right to withdraw this consent without the need to give any reason and I am aware that I can quit this research at any time. I understand that the research results will be used in scientific publications or the EatMyRide company in a completely anonymous manner. My personal data will not be disclosed to any third parties without my permission. If I request further information about the research now or in the future, I can contact Zoe Zhang (j.zhang-15@student.utwente.nl) or Ethics Committee (ethicscommittee-cis@utwente.nl)."

☐ I give permission for the data collected in this research will be used for the analysis and possibly presented to the board members of the master thesis defense anonymously.

☐ I give permission for the answers to be quoted and published in the research output anonymously.

Signature:

Date:

If you have any complaints regarding this research, please contact the Ethics Committee of the Faculty of Electrical Engineering, Mathematics and Computer Science at the University of Twente, P.O. Box 217, 7500 AE Enschede (NL), email: ethicscommittee-cis@utwente.nl

Figure A.2: Second time lo-fi testing Information brochure and Consent form

Information Brochure

Research Topic

User Interface Design through Human-Centered Explainable AI Method

Researcher

Zoe Zhang
+31684341835
j.zhang-15@student.utwente.nl
Electrical Engineering Mathematics and Computer Science, University of Twente
Enschede,

Supervisor

Armağan Karahanoğlu
a.karahanoglu@utwente.nl
Engineering Technology, University of Twente

Mariët Theune
m.theune@utwente.nl
Human-Media Interaction group, University of Twente

Ethics Committee
ethicscommittee-cis@utwente.nl

Description

Our research is collaborating with EatMyRide (EMR), a company that is dedicated to helping fanatic cyclists to improve their cycling performance by assisting them in customizing their nutrition plans using the EMR app. By implementing Explainable AI (XAI) (a method that makes AI more transparent and clear, and makes things behind algorithms more understandable and trustable for users), we want to improve user experience (UX), give users more reasonable recommendations, and finally let users trust and stick to this app.

You as our research participants can experience how and what a scientific nutrition plan could bring to you and really improve our current design. This research requires no pre-knowledge to AI, and you don't need to have previous experience using EMR or other similar applications. The whole process lasts for about 40 min.

Procedure and Data Collected

- **Familiar with original EMR app (5min)**

  You will experience the original EMR app design first. You are required to start from the registration part and finish the whole process by imagining you are going to take a cycle and are here to make your nutrition plan.

- **Think-Aloud Task with Concurrent Question Answering (30min)**

  You will experience the new design (Hi-Fi). There is only one task: Imagine that you are going to take a cycle, and you will make your nutrition plan from the User Register step and explore the EMR app (it won't involve any of your private data, or you can just fill in the fake data, it won't influence the result). You will perform this task and think out loud about the Hi-Fi prototype. During this process, you may also be asked some simple questions such as how and why you feel this way during and after the testing.

  You will participate online, and there is no recording only notes were taken down. The Hi-Fi document will be sent to you and you just need to open the webpage to experience it.

- **The Explanation Satisfaction scale and the Trust scale (5 min)**

  You will be asked to fill out two scales in an online form. These are simple scales that let you measure the satisfaction and trust level of the explanation.

Data storage

Data will be securely stored according to the GDPR guidelines with correct encryption in place and will be anonymized as early as possible. Data is only stored until the end of the thesis research, according to VSNU guidelines.

Data Usage

Your data will be anonymized, and will only be used for this master thesis research. All your information will be completely anonymous. After the research, all your raw data (video and audio data) will be deleted completely. If you want to withdraw from the research, all your data collected will be deleted directly. But the observation notes we taking down may be used (quoted) in this thesis paper anonymously.

Contact

If you have any questions after the research, please contact me via j.zhang-15@student.utwente.nl or the Ethics Committee via ethicscommittee-cis@utwente.nl

Consent Form

"I hereby declare that I have been informed in a manner that is clear to me about the nature and method of the research as described in the information brochure. My questions have been answered to my satisfaction and I agree with my own free will to participate in this research. I reserve the right to withdraw this consent without the need to give any reason and I am aware that I can quit this research at any time. I understand that the research results will be used in scientific publications or the EatMyRide company in a completely anonymous manner. My personal data will not be disclosed to any third parties without my permission. If I request further information about the research now or in the future, I can contact Zoe Zhang (j.zhang-15@student.utwente.nl) or Ethics Committee (ethicscommittee-cis@utwente.nl)."

☐ I give permission for the data collected in this research will be used for the analysis and possibly presented to the board members of the master thesis defense anonymously.

☐ I give permission for the answers to be quoted and published in the research output anonymously.

Signature:

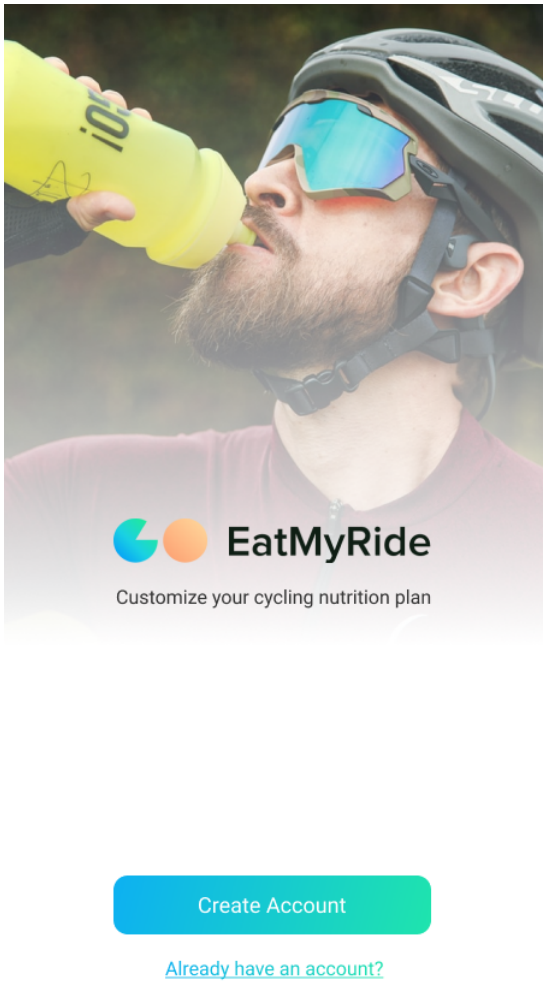
Date:

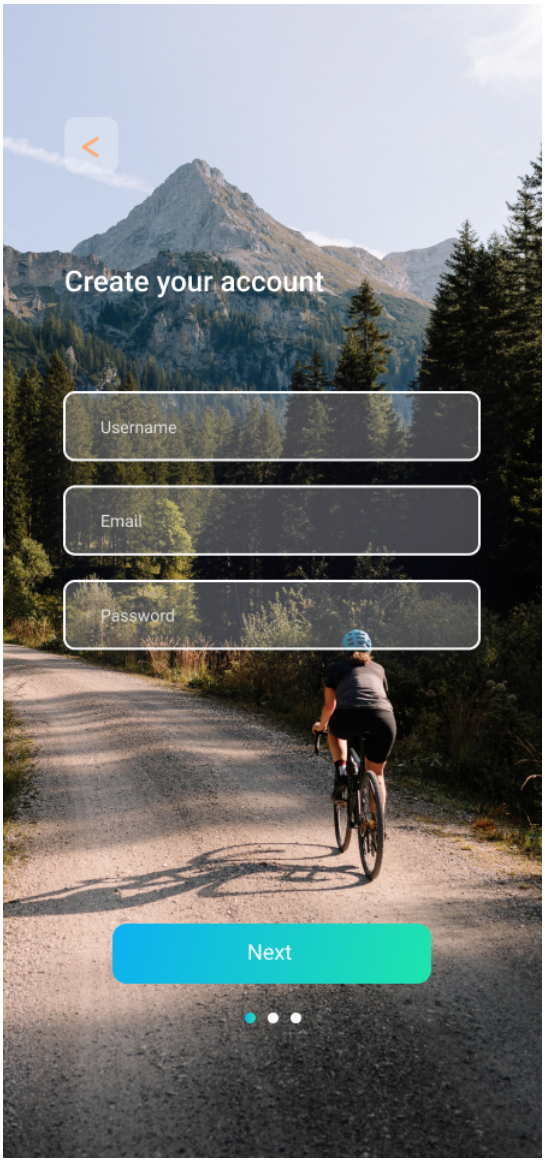
If you have any complaints regarding this research, please contact the Ethics Committee of the Faculty of Electrical Engineering, Mathematics and Computer Science at the University of Twente, P.O. Box 217, 7500 AE Enschede (NL), email: ethicscommittee-cis@utwente.nl

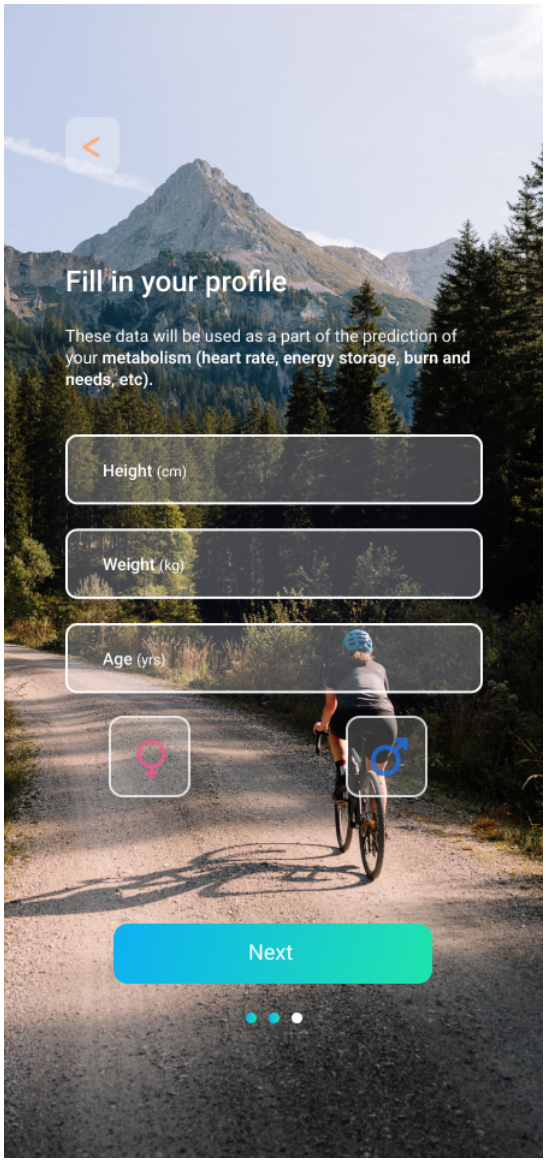
Figure A.3: Third time hi-fi testing Information brochure and Consent form

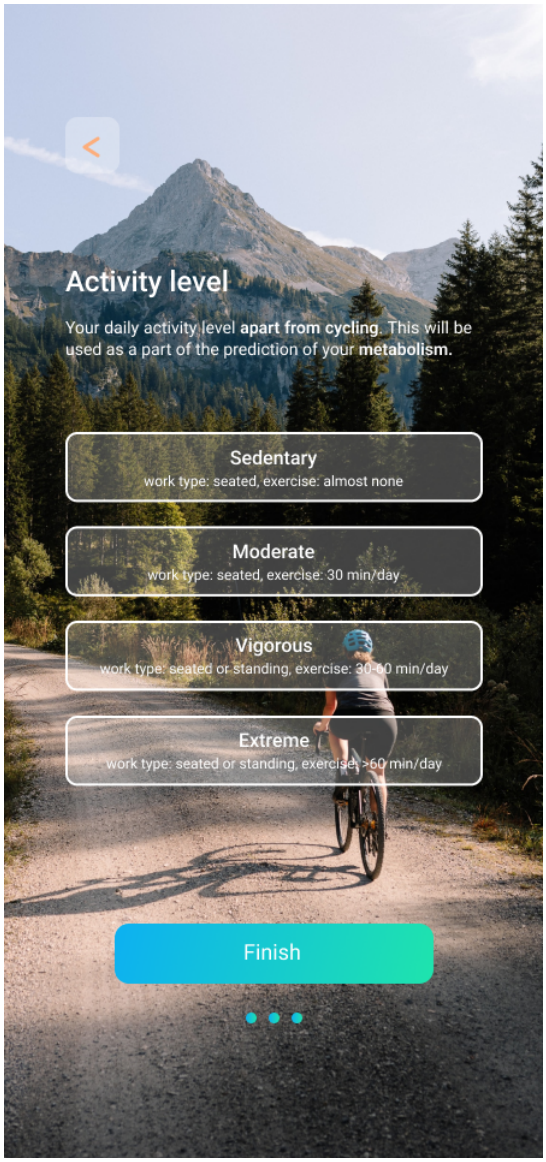
Appendix B

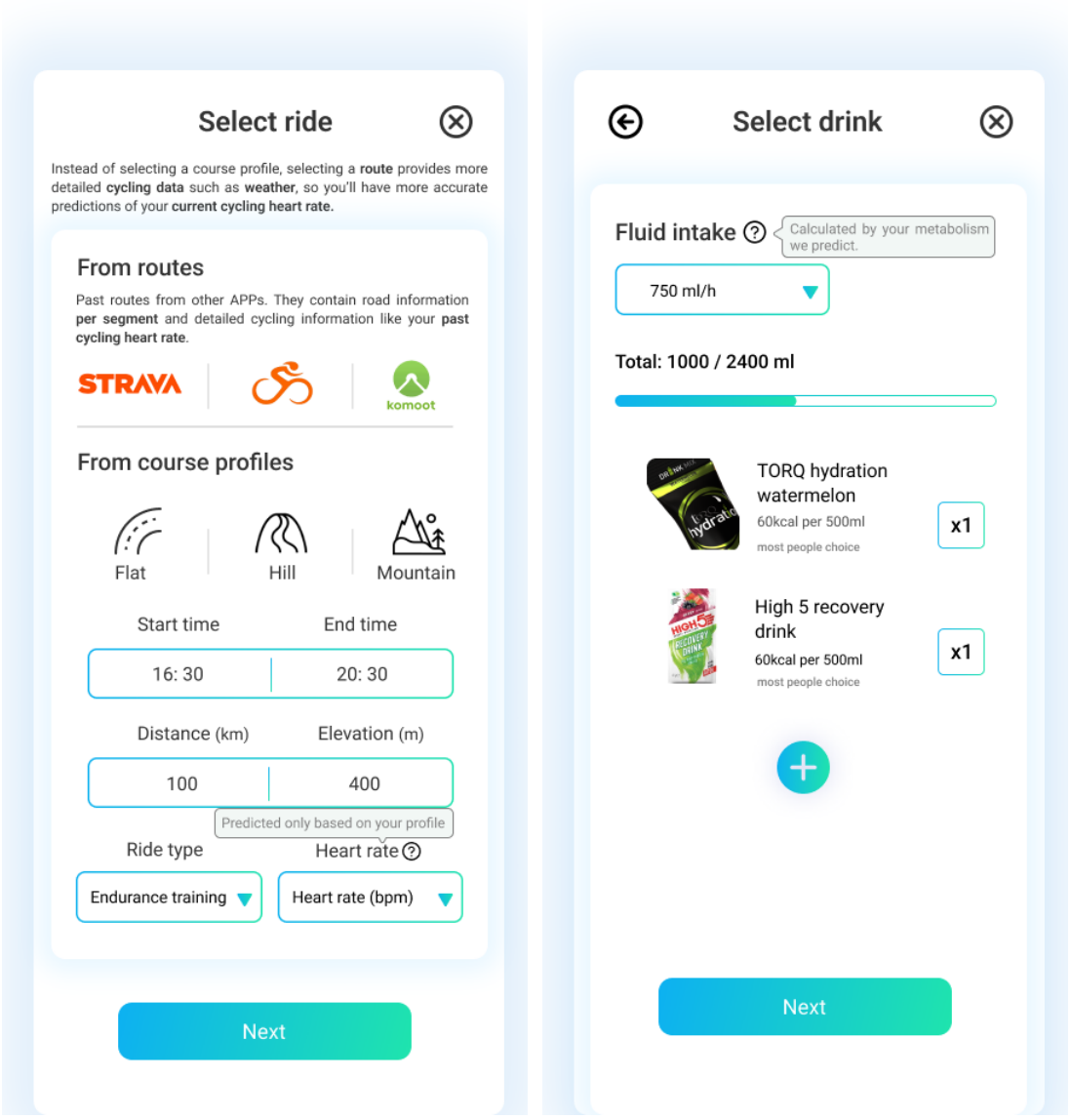
Hi-fi prototype pages

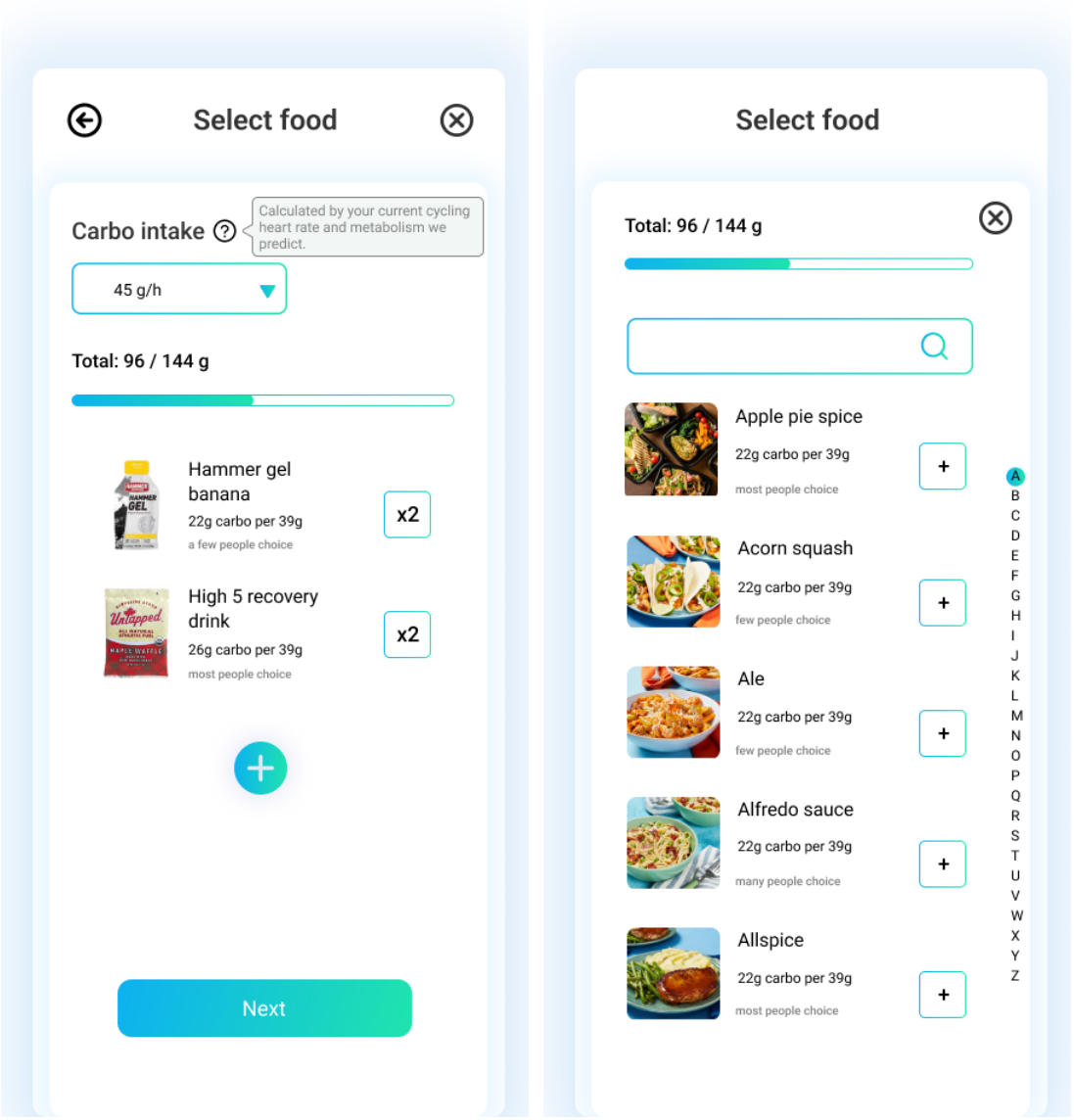


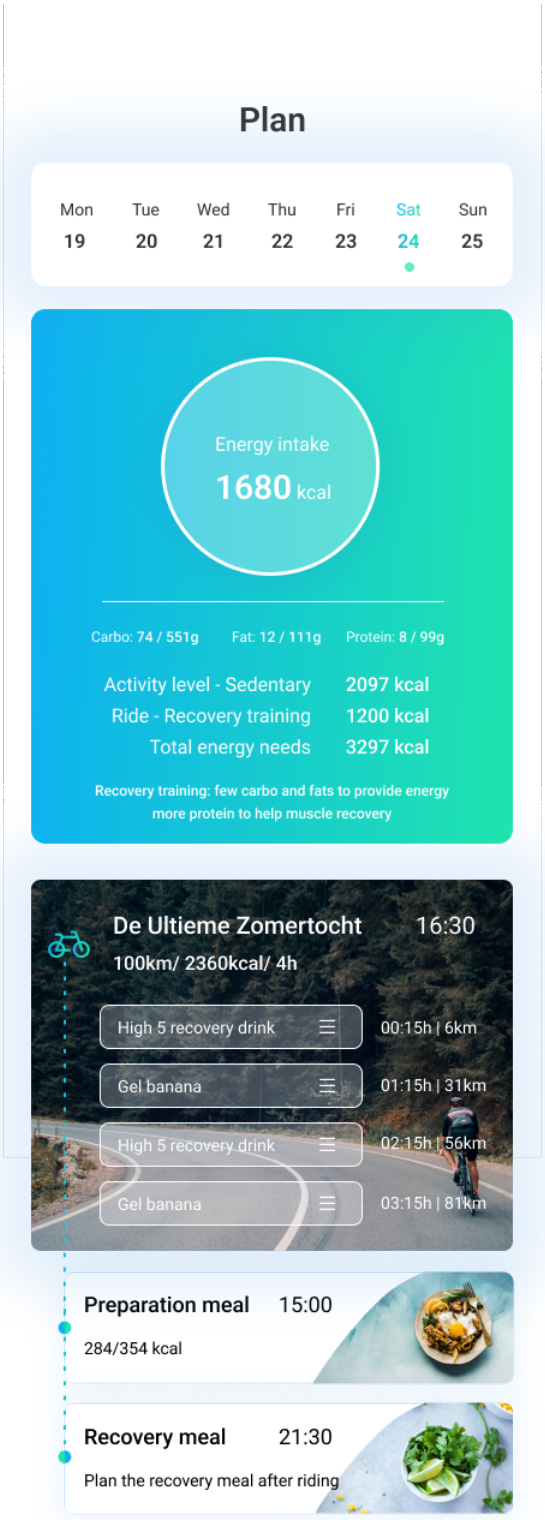


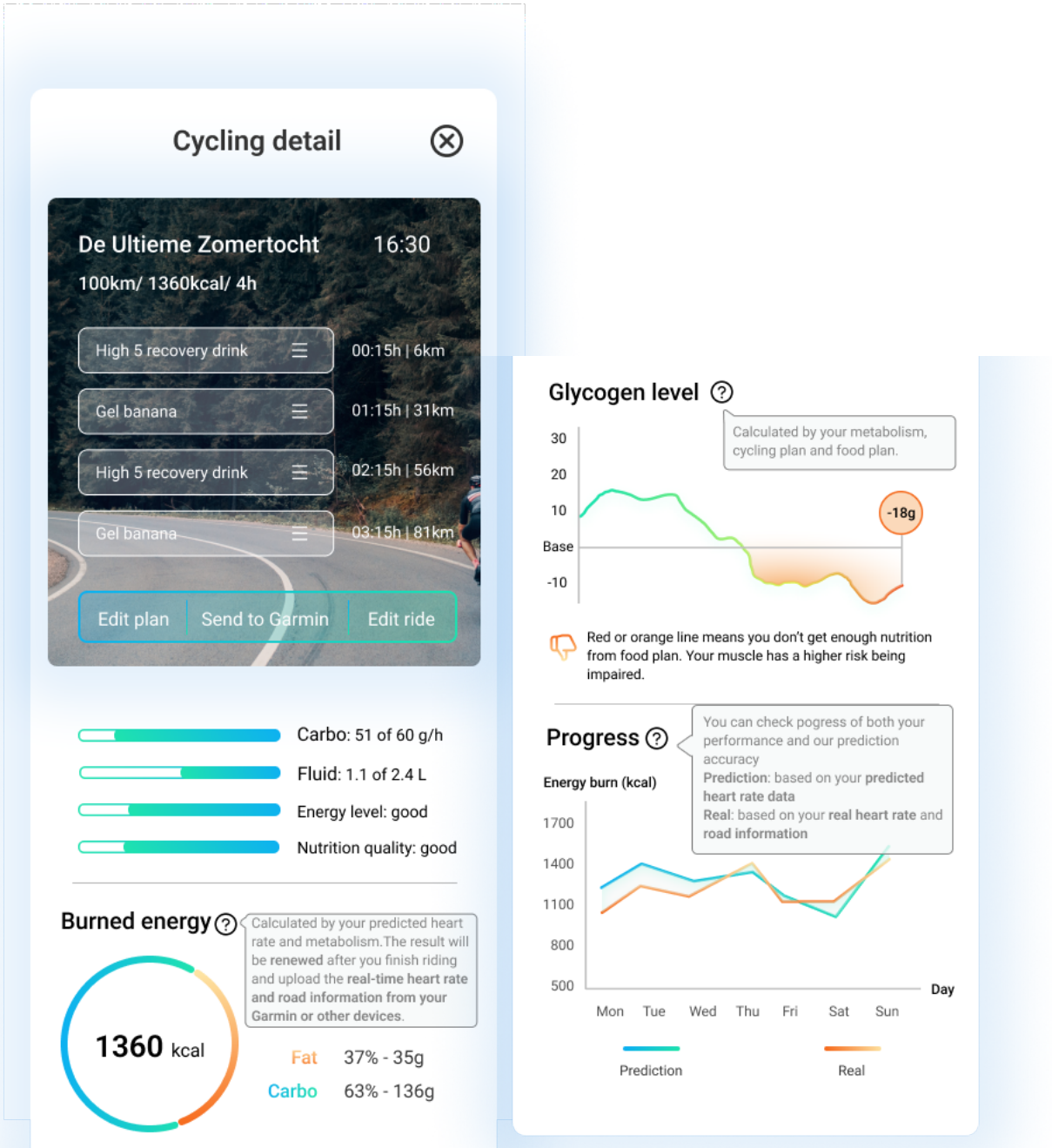

Fill in your profile
These data will be used as a part of the prediction of your metabolism (heart rate, energy storage, burn and needs, etc).
Height (cm)
Weight (kg)
Age (yrs)
Next

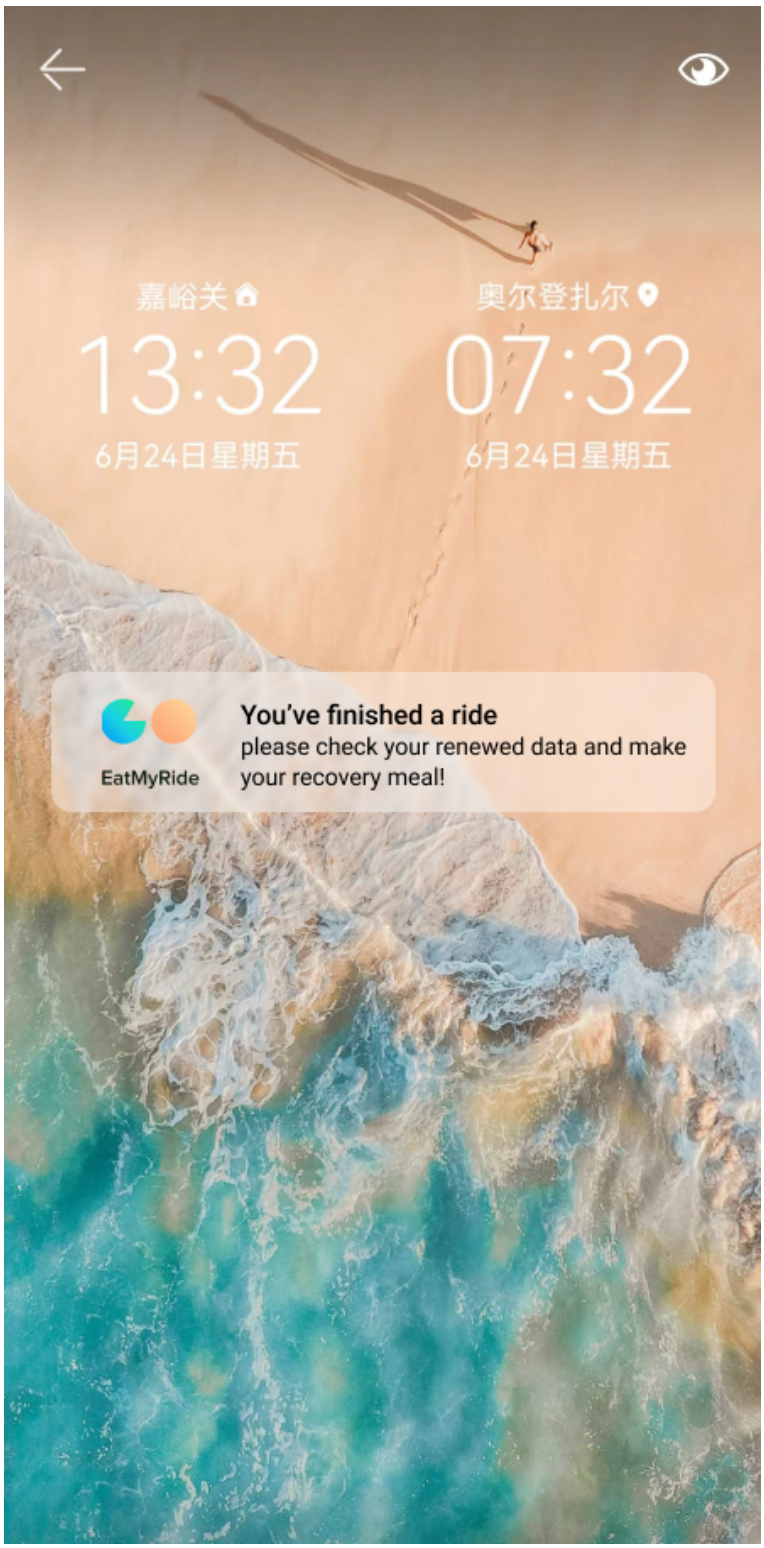

Activity level
Your daily activity level apart from cycling. This will be used as a part of the prediction of your metabolism.
Sedentary
work type: seated, exercise: almost none
Moderate
work type: seated, exercise: 30 min/day
Vigorous
work type: seated or standing, exercise: 30-60 min/day
Extreme
work type: seated or standing, exercise: >60 min/day
Finish


Select ride
Instead of selecting a course profile, selecting a route provides more detailed cycling data such as weather, so you'll have more accurate predictions of your current cycling heart rate.
From routes
Past routes from other APPs. They contain road information per segment and detailed cycling information like your past cycling heart rate.
STRAVA
komoot
From course profiles
Flat
Hill
Mountain
Start time
End time
16: 30
20: 30
Distance (km)
Elevation (m)
100
400
Predicted only based on your profile
Ride type
Heart rate
Endurance training
Heart rate (bpm)
Next
Select drink
Fluid intake
Calculated by your metabolism we predict.
750 ml/h
Total: 1000 / 2400 ml
TORQ hydration watermelon
60kcal per 500ml
most people choice
x1
High 5 recovery drink
60kcal per 500ml
most people choice
x1
Next


Select food
Carbo intake
Calculated by your current cycling heart rate and metabolism we predict.
45 g/h
Total: 96 / 144 g
Hammer gel banana
22g carbo per 39g
a few people choice
x2
High 5 recovery drink
26g carbo per 39g
most people choice
x2
Next
Select food
Total: 96 / 144 g
Apple pie spice
22g carbo per 39g
most people choice
Acorn squash
22g carbo per 39g
few people choice
Ale
22g carbo per 39g
few people choice
Alfredo sauce
22g carbo per 39g
many people choice
Allspice
22g carbo per 39g
most people choice
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z


Plan
Mon 19
Tue 20
Wed 21
Thu 22
Fri 23
Sat 24
Sun 25
Energy intake
1680 kcal
Carbo: 74 / 551g
Fat: 12 / 111g
Protein: 8 / 99g
Activity level - Sedentary 2097 kcal
Ride - Recovery training 1200 kcal
Total energy needs 3297 kcal
Recovery training: few carbo and fats to provide energy more protein to help muscle recovery
De Ultieme Zomertocht 16:30
100km/ 2360kcal/ 4h
High 5 recovery drink 00:15h | 6km
Gel banana 01:15h | 31km
High 5 recovery drink 02:15h | 56km
Gel banana 03:15h | 81km
Preparation meal 15:00
284/354 kcal
Recovery meal 21:30
Plan the recovery meal after riding


Cycling detail
De Ultieme Zomertocht
16:30
100km/ 1360kcal/ 4h
High 5 recovery drink
00:15h | 6km
Gel banana
01:15h | 31km
High 5 recovery drink
02:15h | 56km
Gel banana
03:15h | 81km
Edit plan
Send to Garmin
Edit ride
Carbo: 51 of 60 g/h
Fluid: 1.1 of 2.4 L
Energy level: good
Nutrition quality: good
Burned energy
Calculated by your predicted heart rate and metabolism.The result will be renewed after you finish riding and upload the real-time heart rate and road information from your Garmin or other devices.
1360 kcal
Fat 37% - 35g
Carbo 63% - 136g
Glycogen level
Calculated by your metabolism, cycling plan and food plan.
30
20
10
Base
-10
-18g
Red or orange line means you don't get enough nutrition from food plan. Your muscle has a higher risk being impaired.
Progress
You can check pogress of both your performance and our prediction accuracy
Prediction: based on your predicted heart rate data
Real: based on your real heart rate and road information
Energy burn (kcal)
1700
1400
1100
800
500
Day
Mon Tue Wed Thu Fri Sat Sun
Prediction
Real


嘉峪关
13:32
6月24日星期五
奥尔登扎尔
07:32
6月24日星期五
EatMyRide
You've finished a ride
please check your renewed data and make your recovery meal!

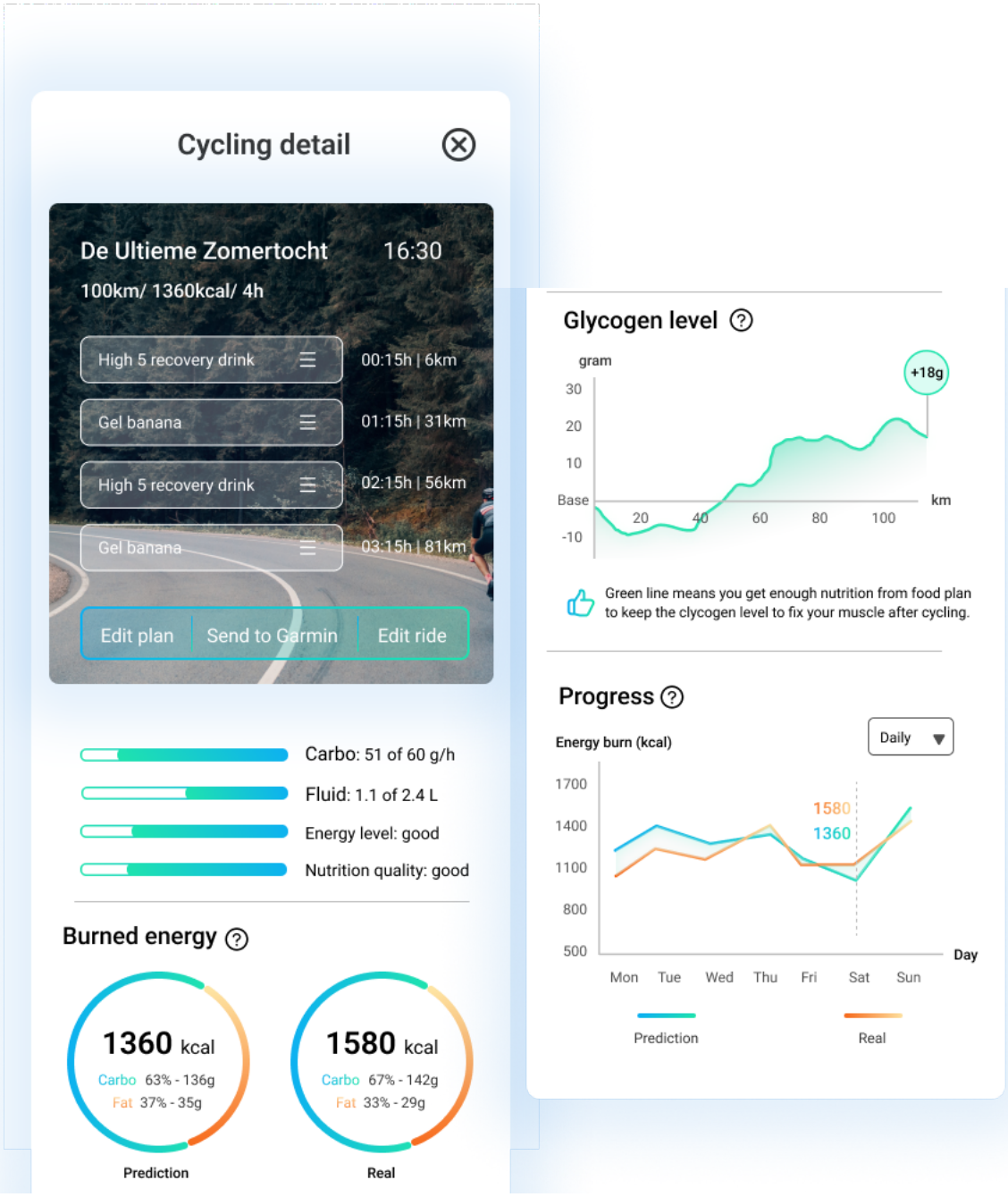


Figure B.1: Hi-fi prototype pages

Appendix C

Different measurement scales

C.1 Explanation Goodness Checklist

C.2 Explanation Satisfaction Scale

C.3 Explanation Trust Scale

The explanation helps me **understand** how the [software, algorithm, tool] works.

YES	
NO	

The explanation of how the [software, algorithm, tool] works is **satisfying**.

YES	
NO	

The explanation of the [software, algorithm, tool] sufficiently **detailed.**

YES	
NO	

The explanation of how the [software, algorithm, tool] works is sufficiently **complete.**

YES	
NO	

The explanation is **actionable**, that is, it helps me know how to use the [software, algorithm, tool]

YES	
NO	

The explanation lets me know how **accurate or reliable** the [software, algorithm] is.

YES	
NO	

The explanation lets me know how **trustworthy** the [software, algorithm, tool] is.

YES	
NO	

Figure C.1: Explanation Goodness Checklist

[23]

1. From the explanation, I **understand** how the [software, algorithm, tool] works.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

2. This explanation of how the [software, algorithm, tool] works is **satisfying**.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

3. This explanation of how the [software, algorithm, tool] works has **sufficient detail.**

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

4. This explanation of how the [software, algorithm, tool] works seems **complete**.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

5. This explanation of how the [software, algorithm, tool] works **tells me how to use** it.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

6. This explanation of how the [software, algorithm, tool] works is **useful to my goals**.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

7. This explanation of the [software, algorithm, tool] shows me how **accurate** the [software, algorithm, tool] is.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

8. This explanation lets me judge when I should **trust and not trust** the [software, algorithm, tool]

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Figure C.2: Explanation Satisfaction Scale

1. I am confident in the [tool]. I feel that it works well.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

2. The outputs of the [tool] are very predictable.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

3. The tool is very reliable. I can count on it to be correct all the time.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

4. I feel safe that when I rely on the [tool] I will get the right answers.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

5. The [tool] is efficient in that it works very quickly.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

6. I am wary of the [tool]. (adopted from the Jian, et al. Scale and the Wang, et al. Scale)

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

7. The [tool] can perform the task better than a novice human user. (adopted from the Schaefer Scale)

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

8. I like using the system for decision making.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

Figure C.3: Explanation Trust Scale

[23]