Health Coaches, Health Data, and Their Interaction

Health Coaches, Health Data,
and Their Interaction

COLOPHON
This research presented in this dissertation was conducted at the Human-Technology Interaction group, School of Innovation Sciences, Eindhoven University of Technology. This research was supervised by prof. dr. Wijnand A. IJsselsteijn and dr. ir. Martijn C. Willemsen. A catalogue record is available from the Eindhoven University of Technology Library.

GRAPHIC DESIGN Bas Reijnen
ILLUSTRATIONS Michiel Rutjes
PRINTING GVO drukkers & vormgevers
ISBN 978-90-386-5364-8

Health Coaches, Health Data, and Their Interaction

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens,
voor een commissie aangewezen door het College voor Promoties,
in het openbaar te verdedigen op

dinsdag 12 oktober 2021 om 15:30 uur

door

Heleen Muijlwijk
(e/v Rutjes)

geboren te Noordoostpolder

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

Voorzitter	prof. dr. I.E.J. Heynderickx
Promotor	prof. dr. W.A. IJsselsteijn
Copromotor	dr. ir. M.C. Willemsen
Leden	prof. dr. G. Fitzpatrick (Vienna University of Technology)
	dr. S. Stumpf (City, University of London)
	dr. S. Munson (University of Washington)
	dr. B.E.R. De Ruyter (Philips Research Eindhoven)
	prof. dr. ir. A.C. Brombacher

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Voor Linda, mijn lieve zus,
die me de moed heeft gegeven om te leven

Summary

Health coaching can play a vital role in achieving and maintaining a healthy lifestyle. Health coaches, such as dieticians and personal trainers, are confronted with a vast growth of digital solutions for health coaching. Wearable tracking devices and mHealth smartphone applications enable continuous and in-situ collection of health data. Not only does this potentially empowers clients to manage their own health, also coaches may benefit from such data. It provides them additional information over a client's self-report, and models based on such data enable detection of trends and correlations. Still, these data are not widely used by coaches in practice, and the coach's needs regarding health data are still ill understood. For coaches, these data may indeed be inherently limited, as coaches have a much broader understanding of the client, including her motivations, daily context and personal values. Adding to current research on health data in relation to the client, for example in the field of Personal Informatics (PI), in this research we have a main focus on the coach. This dissertation aims at understanding health coaches' needs towards clients' health data, it compares strengths of both coaches and data, and explores means to facilitate effective collaboration between coaches and data that synthesize their complementary forms of knowledge.

Specifically, in *Chapter 2*, we interviewed health coaches on their current practices and their perspective on successful health coaching. This allowed us to identify opportunities for health technologies independent of possible prejudices towards technology. Thereafter, we more explicitly discussed the role of data and technology with health coaches in a focus group, and we observed coaches' hands-on practices with health data in a workshop. *Chapter 3* describes a field study where we observed coaching sessions enriched with clients' health data. We learned how these data influenced the health coaching process, both on content-level (e.g., the coaching advice) as well as on relationship-level (e.g., mutual understanding, conversation dynamics). *Chapter 4* reports on a longitudinal field study, in which we facilitated parents of newborns to customize their own data-practices to capture their baby's health issues and share this with their healthcare professional. By weekly interviews with healthcare professionals and parents, we gained understanding on how data-practices and underlying needs evolved over time, and how data-sharing influenced their communication and alignment.

Drawing from this qualitative work, we learned that coaches are generally willing to deploy tracking devices to collect data, but they want to keep control over *how*

these data are used. They emphasize that every client is unique, and they doubt whether data-driven models or algorithms can capture the right nuance to serve as effective input for health coaching. In response to these concerns, in *Chapter 5* we explored the effect of transparency of health recommendations on coaches' levels of trust and acceptance. Through an online survey we evaluated a data dashboard with coaches with various expertise levels. We found that laypeople, novices and experts have different needs regarding transparency. To encourage coaches to actively deploy their domain knowledge, in *Chapter 6* we developed an interactive support tool for running coaches to predict challenging yet realistic target finish times for their runners' next marathon. Multiple study iterations (i.e., pilot interviews, think-aloud sessions, and an online survey) allowed for developing a novel means of interaction in which coaches could meaningfully express themselves. We found that coaches were keen on deploying their knowledge to steer the model, especially when runners were familiar to them, and that this resulted in increased levels of trust and improved model performance.

Our results show that health data are rarely 'plug-and-play', and the effects of data on health coaching go beyond merely adding numerical information. For instance, clients' practices around collecting data and their reflections on the data showed to be highly informative to coaches. We observed that data disrupt roles and changes the dynamics of a conversation. Effective use of health data requires careful alignment of expectations and sufficient room to collaboratively reflect on the data. Our results show that coaches are not merely positive about the opportunities that health data provide, and indeed, coaches may add complementary and valuable knowledge. Coaches appreciate model transparency and interactivity, which allows them to be involved in these models and deploy their expertise. We also make a methodological contribution to the fields of explainable AI (XAI) and interactive Machine Learning (iML), by showing that different types of users may have substantial different needs towards support systems, highlighting the need for using knowledgeable participants working on realistic tasks when evaluating those systems. Starting at understanding coaches' practices and needs has shown to enhance the design of interactive interfaces, where the complementary strengths of health coaches and health data are effectively combined.

Samenvatting

Health coaching kan een belangrijke rol spelen bij het bereiken en behouden van een gezonde leefstijl. Health coaches, zoals diëtisten en personal trainers, worden geconfronteerd met een snelle groei aan digitale oplossingen voor health coaching. Draagbare technologie, zoals horloges en smartphone apps, maken het mogelijk om continu en overal gezondheidsdata te verzamelen. Dit stelt cliënten in staat om hun eigen gezondheid te managen, maar ook coaches kunnen hier hun voordeel mee doen. Het geeft hen extra informatie ten opzichte van het verhaal van een client, en modellen die gebaseerd zijn op deze data kunnen trends en correlaties detecteren. Toch gebruiken coaches deze data niet veel in de praktijk, en we weten nog weinig over de behoeftes van coaches ten opzichte van deze data. Voor coaches zijn deze data beperkt, omdat zij een veel breder begrip hebben van de cliënt, inclusief haar motivaties, dagelijkse context en persoonlijke waarden. In dit onderzoek focussen we op de coach, als aanvulling op bestaand onderzoek in bijvoorbeeld Personal Informatics, wat zich vooral richt op gezondheidsdata in relatie tot de cliënt. Dit proefschrift beoogt een beter begrip van de behoeftes van health coaches als het gaat om de gezondheidsdata van hun cliënten, het vergelijkt de kracht van zowel coaches als data, en het onderzoekt manieren om effectieve samenwerking tussen coaches en data te faciliteren, waarbij hun complementaire kennis samenkomt.

Meer specifiek, in *Hoofdstuk 2* hebben we health coaches geïnterviewd over hun huidige werkwijze en hun visie op succesvolle health coaching. Hiermee konden we kansen identificeren voor gezondheidstechnologie, zonder dat we daarbij afhankelijk waren van mogelijk vooroordelen ten opzichte van technologie. Daarna hebben we de rol van data en technologie meer expliciet besproken met coaches in een focusgroep, en hebben we coaches laten werken met gezondheidsdata in een workshop. *Hoofdstuk 3* beschrijft een veldstudie waarin we coaching sessies hebben geobserveerd die verrijkt waren met data. Daarin zagen we hoe deze data het coaching proces beïnvloedden, niet alleen op inhoudsniveau (bijvoorbeeld het coaching advies) maar ook op relatieniveau (bijvoorbeeld wederzijds begrip en gespreksdynamiek). In *Hoofdstuk 4* bespreken we een langer durende veldstudie, waarin we ouders van pasgeboren baby's tools hebben gegeven om hun eigen data te verzamelen rondom de zorgen over hun baby, en deze te delen met hun zorgverlener. Door middel van wekelijkse interviews met zorgverleners en ouders hebben we een beter begrip gekregen van hoe het gebruik van data en de achterliggende behoeftes veranderen over de tijd. Daarnaast gaf het inzicht in hoe de data de communicatie beïnvloed-

den, en hoe het ervoor kan zorgen dat zorgverleners en ouders al dan niet op één lijn zitten.

Deze kwalitatieve studies laten zien dat coaches in het algemeen bereid zijn om technologie in te zetten om data te verzamelen, maar dat ze controle willen houden over hoe deze data worden gebruikt. Coaches benadrukken dat elke cliënt uniek is, en ze betwijfelen of data-gedreven modellen of algoritmes de juiste nuance hebben om nuttig te kunnen zijn voor het health coaching proces. Als reactie hierop hebben we in *Hoofdstuk 5* gekeken naar het effect van transparantie van gezondheidsrecommandaties op het vertrouwen en de acceptatie van coaches. In een online studie hebben we een data dashboard geëvalueerd met coaches met verschillende expertise niveaus. De resultaten laten zien dat leken, beginners en experts verschillende behoeftes hebben als het gaat om transparantie. Om coaches aan te moedigen om hun domein kennis actief te benutten, hebben we in *Hoofdstuk 6* een interactieve tool ontwikkeld voor hardloopcoaches om een uitdagende maar realistische streeftijd te voorspellen voor een aankomende marathon van hardlopers. In meerdere studie iteraties (pilot interviews, hardop-denk-sessies, en een online studie) hebben we een nieuwe manier van interactie ontwikkeld waarin coaches zichzelf betekenisvol konden uitdrukken. De resultaten laten zien dat coaches graag hun kennis inzetten om het model te sturen, vooral als ze werken met data van hun eigen pupillen, en dat dit resulteert in meer vertrouwen in het model en verbeterde model prestaties.

Onze resultaten laten zien dat niet evident is hoe gezondheidsdata nuttig ingezet kunnen worden, en dat de effecten van data op health coaching verder gaan dan slechts het toevoegen van numerieke informatie. Het bleek bijvoorbeeld erg informatief voor coaches om te zien hoe cliënten hun data verzamelden en hoe zij hierop reflecteerden. We hebben gezien hoe data de gespreksdynamiek en de rollen van coaches en cliënten veranderden. Effectief gebruik van gezondheidsdata vereist een zorgvuldige afstemming van verwachtingen en voldoende ruimte om gezamenlijk op de data te kunnen reflecteren. Onze resultaten laten bovendien zien dat coaches niet alleen maar positief waren over de kansen die gezondheidsdata bieden, en dat coaches inderdaad veel complementaire en waardevolle informatie toe te voegen hadden. Coaches waardeerden het als modellen transparant en interactief waren, zodat ze betrokken werden en hun expertise konden aanwenden. Ons werk heeft daarnaast ook een methodologische bijdrage aan de velden explainable AI (XAI) en interactive Machine Learning (iML), omdat we hebben laten zien hoe verschillende soorten gebruikers op belangrijke manieren van elkaar verschillen in hoe ze reageren op ondersteunende systemen. Dit onderstreept het belang van het gebruik van ervaringsdeskundige en representatieve proefpersonen die werken aan realistische taken, om deze systemen goed te kunnen evalueren. Dit proefschrift illustreert hoe beginnen bij dagelijkse praktijk en behoeftes van gebruikers bevorderlijk is voor het ontwerpen van effectieve interfaces, waarin de complementaire kracht van health coaches en gezondheidsdata goed gecombineerd worden.

Data help me to ask

Coach

the right questions.

General Introduction

CHAPTER 1

General Introduction

What do data mean? This dissertation contains many data, for example, participants' quotes, questionnaire results, statistical relations and effect sizes. Yet, to make meaningful sense of these data, we need to put them in perspective, which we do in this dissertation by reflecting on them in the discussion sections. We elaborate on possible interpretations and we position them within the context of related work in the field. But if we would meet at a coffee machine, or at a conference, I would probably talk about these data at another level. I would explain to you what these data mean to me, and share the lived experiences associated with collecting them. I would talk about the ideas and beliefs from which the studies originate, how my practices were influenced by personal circumstances and context, and how the process of working with these data enhanced my skills, shaped my view of the world, and how I grew as a person. Thus, data as simple as effect sizes or p-values can be understood on many levels. Data are not merely numbers; they represent rich and meaningful information, depending on the person who assesses them.

The same principle applies to health data. Health data, typically collected by wearable devices or smartphone applications, may include step counts, heart rate data, sleep data, sedentary behavior, and nutrition intake, among other things. At first sight, these data represent behavior. For example, it may show that a person is typically active in the afternoons, consistently wakes up early, and that she has certain periods when she frequently eats chocolate, next to a generally healthy diet. Now imagine you would be a coach, and she would ask you to support her to attain a healthier lifestyle. If you would assess these data, what would that tell you? It would probably raise a number of questions. What does she do for a living? Does her environment differ between mornings and afternoons? Does she wake up by herself, her alarm, or her children? Does she buy the chocolate herself; is it a sign of weakness? A treat of her husband? A celebration of success? A deadline coming up? These easily emerging questions illustrate that merely clients' data are not sufficient to understand her. In order to be able to provide her appropriate support, these data should be understood beyond behavior, for instance in terms of the client's daily context and motivations.

This dissertation concerns the potential role and value of health data within a health coaching process. We will consider both human coaches' and data-driven models' strengths and limitations, and investigate how they may reinforce one another. In the remainder of this general introduction we will provide background on this research topic, and give an overview of the dissertation and its main contributions.

COACH, DATA, AND CLIENT

As the title of this dissertation suggests, this work focusses on health coaches, health data and their interaction. Yet, health coaches and health data are inevitably connected with a client. The client is, after all, the very matter of the coaching. Successful coaching is mainly determined by the extent to which the client is satisfied with the process and whether it supports her in reaching her health goals. Related fields, such as Personal Informatics (PI; Epstein et al., 2020) and Persuasive Technology (PT; Fogg, 2003; IJsselsteijn, De Kort, Midden, Eggen, & Van Den Hoven, 2006), largely focus on health technologies in light of the needs and perceptions of the client (Figure 1). There is, however, another important potential end-user of these data, whose needs are currently underexposed: the health coach. As tracking becomes more commonplace, coaches will encounter increasing numbers of clients who bring their data. Coaches may benefit from these data too; it potentially provides them additional information next to the clients' self-report. Coaches do have substantially different perspectives on these data compared to clients (Figueiredo, Su, & Chen, 2020; Pichon et al., 2020; Raj, Lee, Garrity, & Newman, 2019; Raj, Newman, Lee, & Ackerman, 2017). For one, they do not perceive the data from a first-person perspective and thus lack knowledge on the context, in turn, they typically have more domain knowledge on health than their clients.

It is worthwhile to take the coach into consideration when studying the role and value of data for health coaching. Coaches may bring in substantial domain knowledge on health, as well as knowledge on the client, that put the data in a new perspective. For example, coaches may add which motivational strategies are likely to resonate with the client, allowing for the use of data in effective and supporting ways. In turn, coaches' skills and interventions may be enhanced when using data, making coaching more effective. In order to exploit this mutual benefit, we need to understand the implications of introducing data in a health coaching process, if and how it changes the coach's role, and consequently, which new skills this requires of the coach and which different information needs the technology should meet. Therefore, in this dissertation we give the coach a central role, as we study the 'system' of coach, client and data (Figure 1). Understanding the coach's perspective enables effective collaboration between coaches and data, or data-driven models, that allow for better coaching practices, eventually improving clients' health.

THE IMPORTANCE OF HEALTH COACHING

A healthy lifestyle is of key importance for good health and wellbeing (Forouzanfar et al., 2016). The World Health Organization emphasizes the critical role of nutrition (World Health Organization, 2003) and physical activity (World Health Organization, 2020) for preventing chronic diseases and improving a broad range of health-out-

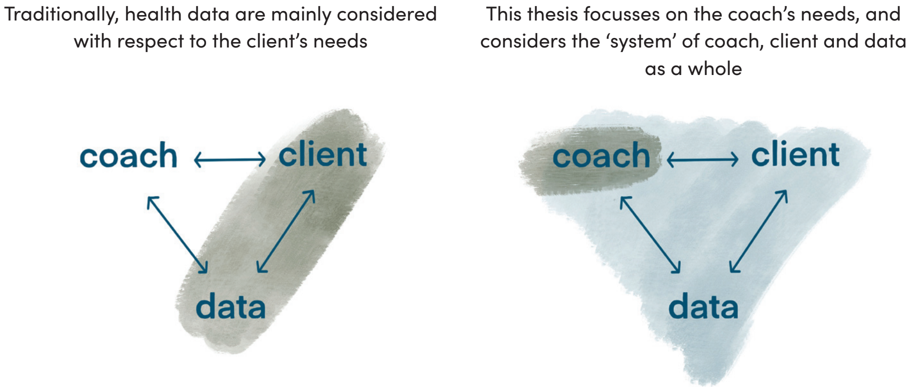

Figure 1 Triangle of coach – client – data; a conceptual model to frame the contribution of our work by having a central focus on the coach.

comes, such as improved cardio metabolic health and mental health. Behavior change is notoriously hard; health coaching can play a key role in achieving and maintaining a healthy lifestyle, for example using techniques such as goal-setting or motivational interviewing (Michie et al., 2011; Olsen & Nesbitt, 2010).

There are two main types of health coaching. Traditionally, health coaching refers to an interpersonal encounter between a coach and a client, while more recently, a broad range of digital solutions for health coaching (i.e., e-coaching) is emerging, driven by advances in wearable sensors and data processing. Both types of health coaching support clients to achieve their health goals, and if we are able to improve its effectiveness, this would result in improved population health. Potentially, human coaches may be more effective when taking advantage of these digital tools, and in turn, e-coaching applications may be more effective when they incorporate principles of interpersonal coaching. Hybrid forms of coaching that integrate human coaches and digital artefacts may be even more promising. Yet, to be able to capitalize this potential value, we need to improve our understanding of coaches' needs when interacting with health data.

HEALTH COACHES

Health coaches, in contrast to clinicians, do not diagnose illness nor do they provide clinical interventions (Wolever, Jordan, Lawson, & Moore, 2016). Instead, they support their clients by collaboratively working towards health goals and strengthening their clients' intrinsic motivation, for example using techniques such as motivational interviewing, goal setting, and content education (Wolever et al., 2013). Health coaches include professions such as dieticians, personal trainers, lifestyle coaches, and wellness coaches. Also nurses and physicians can take up the role of health coach, for example when supporting patients to manage their chronic disease. In this dissertation, we have mostly worked with health coaches working on wellbeing and health promotion. Though, in some chapters we have considered health coaches with a slightly different background and focus; that is, in *Chapter 4* we have also included general practitioners and pediatricians in a coaching role, and in *Chapter 6* we included running coaches.

Health coaching is an inherently interpersonal process. A key element of health coaching is its collaborative relation, where, in contrast to a hierarchical relation, the coach stands next to the client (Olsen, 2014) and facilitates her through for example active learning and self-discovery (Wolever et al., 2013). Thus, health coaches need to be sensitive to their clients' motivations, values, and everyday context. This process involves substantial situational awareness and mutual coordination. Literature highlights the key role of good communication for shared decision making in healthcare (Ong, de Haes, Hoos, & Lammes, 1995; Roter & Hall, 2006), and the importance to understand and incorporate patients' narrative (Murphy & Franz, 2016) and values (Berry et al., 2017) to provide appropriate care.

Drawing from the field of decision making, it is generally recognized that humans are strong in making fast and intuitive decisions based on tacit knowledge and common sense (Gigerenzer, 2008; Kahneman, 2011). Indeed, in health and sports coaching, despite increasing attention for data-driven approaches (c.f., Cardinale & Varley, 2017; Doherty et al., 2020; Sqalli & Al-Thani, 2020), coaches still largely rely on intuition (D. Collins, Collins, & Carson, 2016; Lyle, 2010). While some consider this as an efficient and adequate approach, other researchers have highlighted the limitations of intuitive decision making. They argue that humans are susceptible to a wide range of biases, especially in situations of high uncertainty (Kahneman, Slovic, & Tversky, 1982). Thus, it is worthwhile to consider if, and how, health data and data-driven models can complement coaches.

HEALTH DATA

Rapid advances in sensor technology and wearable computing have enabled a vast growth of commercially available wearable devices, such as smartwatches and activity trackers (Seneviratne et al., 2017). In addition, there is a large number of mHealth applications available for smartphones (Byambasuren, Sanders, Beller, & Glasziou, 2018). These wearable devices and mHealth apps allow for continuous and in-situ collection of behavioral data. For clients, these devices may support them with feedback on their daily routines, typically supplemented with timely and personalized recommendations (Sqalli & Al-Thani, 2020; Tong et al., 2021). A recent review of the Personal Informatics literature shows that these technologies are mostly used for changing behavior, raising awareness and managing chronic conditions (Epstein et al., 2020).

Advances in Artificial Intelligence (AI), Machine Learning (ML) and data science are reflected in how these health data are used. For example, some e-coaching systems are equipped with adaptive procedures, where they real-time and automatically adapt goals (Mohan, Venkatakrishnan, & Hartzler, 2020) and training schemes (Janssen et al., 2020; Vos, Janssen, Goudsmit, Lauwerijssen, & Brombacher, 2016) to a client's performance and physical workload. Machine Learning principles, combined with an expert system, can find patterns in health behavioral data, leading to actionable insights for clients (Mitchell et al., 2021). Also, it has been explored how e-coaches may tailor coaching strategies to clients' different types of motivations (Beinema, op den Akker, van Velsen, & Hermens, 2021), their preference for different communication styles (Niess, Diefenbach, & Woźniak, 2020), and their susceptibility to different persuasive strategies (Spelt, 2021). AI systems have also been used to predict the probability that a client will engage in healthy behaviors (Lewis, Liu, Groh, & Picard, 2021). Furthermore, AI models show to be well able to predict sports performance and injury risk, for example based on training load or past performances (Claudino et al., 2019; Doherty et al., 2020; Smyth & Cunningham, 2017). Lastly, novel means of feedback during sports training are explored, for example haptic feedback through electrical muscle stimulation for runners (Lu & Brombacher, 2020).

An important premise underlying tracking health data is that it potentially empowers clients. These devices and the data they comprise may indeed foster clients' engagement in their health (Pavel et al., 2013), and while it has been argued that this enables people to take control over their own health (Topol, 2015), clear evidence for patient empowerment is lacking (Alpay, Henkemans, Otten, Rövekamp, & Dumay, 2010; Storni, 2014). Not only can people collect and assess their personal data, they are also enabled to share and compare these data with others (e.g. on platforms such as Quantified Self[1] and Patients Like Me[2]). These initiatives facilitate self-care, and by its low-cost and remote nature they have the potential to reach people who would otherwise have no or limited access to healthcare (Bhatta, Aryal, & Ellingsen, 2015; Mileski, Kruse, Catalani, & Haderer, 2017). At the same time, this trend of digitalization of health is criticized for being too narrow and too judgmental, mainly because it revolves only around numbers and context is lacking (Ajana, 2018; Kersten – van Dijk, Westerink, Beute, & IJsselsteijn, 2017; Lupton, 2016a; Storni, 2011; Wu et al., 2018).

Another perspective on the value of health data derives from the main strength of computing; it allows for reliable detection of trends and correlations in large amounts of data. Algorithms have shown to often outperform human judgement on quantitative prediction tasks (Grove, Zald, Lebow, Snitz, & Nelson, 2000), and data-driven models allow the detection of otherwise hidden patterns. For instance,

1 www.quantifiedself.com
2 www.patientslikeme.com

a large study based on data of smartwatches has enabled the detection of cardiac arrhythmias (Turakhia et al., 2019). At the same time, others have argued that having access to ones' behavioral data does not necessarily imply that one is able to change it (Kersten – van Dijk et al., 2017; Patel, Asch, & Volpp, 2015). In general, when evaluating the performance of data-driven models, critics argue that we tend to neglect the value of tacit knowledge (Klein, Shneiderman, Hoffman, & Ford, 2017), and computational performance decreases in situations high in uncertainty (Cummings, 2014). Thus, while the promise of health data is clear, it is not evident how to effectively use these data in practice.

AND THEIR INTERACTION

In the previous sections, we have discussed how health coaches may benefit from health data; wearable devices allow for data collection anywhere anytime, adding considerable information to a client's self-report. In addition, models that deploy such data may help to overcome coaches' biases. At the same time, we have discussed that data by themselves are not sufficient. Data are hard to interpret as they inevitably lack personal and contextual background, and a human coach can easily add these interpretations by collaboratively reflecting on the data with the client. Also, by the interpersonal nature of health coaching, coaches can build an emphatic and trust relationship with the client, which enhances effective coaching. We argue that human coaches and health data may substantially reinforce each other, and we should seek to understand how to facilitate effective collaboration.

AN AUTOMATION PROBLEM PERSPECTIVE

We are studying how coaches and data can and should interact. One aspect of this problem is the division of tasks: what should a coach do, and what should technology do? For this, the concept of automation provides a helpful perspective, as it addresses precisely this problem of which tasks to automate, and which tasks to leave to humans. Wearable trackers and mHealth apps indeed automate and extend parts of coaching; they automatically collect behavioral data and they often automatically generate motivational messages or advice. Back in 1951, Paul Fitts posed the 'Fitts list' (Fitts, 1951), describing which tasks 'men are better at, and machines are better at' (sometimes abbreviated as MABA-MABA). For example, humans would be better at inductive reasoning, whereas machines would be better at deductive reasoning. Extending on this idea, Parasuraman, Sheridan and Wickens (2000) developed a broader framework of automation, where different tasks (e.g. information collection, decision making) can be automated on different levels. Such a way of thinking, where tasks can simply be distributed over humans and machines, has been criticized by others. Dekker and Woods (2002) have argued that task allocation is too simplistic and theoretically driven. In practice, for example, handing over tasks to a machine does not necessarily alleviate a human's working load, rather it transforms the task into something else (e.g., supervision) that is not necessarily easier (Bainbridge, 1983; Dekker & Woods, 2002).

AN INTERACTIVE SYSTEM PERSPECTIVE

Recently, the focus in research has shifted from task allocation to enhancing mutual understanding and collaboration between humans and machines. Having insight in and control over systems are key elements in this discussion. We have discussed examples of AI that enable e-coaching to be adaptive and self-learning, but e-coaching systems are rarely transparent on how they come to their recommendations (Eiband et al., 2018). Beyond coaching applications, there is a vastly growing body of work that concerns fair, accountable and transparent machine learning algorithms (FATML; Lepri, Oliver, Letouzé, Pentland, & Vinck, 2018; Zarsky, 2016) aiming at giving users better insight in the processes of machines. Besides insight, users have been given more control as well. Initiated by Fails and Olsen (2003), the field of Interactive Machine Learning (Dudley & Kristensson, 2018) has rapidly advanced, where humans are enabled to interrogate and control models. This idea, sometimes referred to as 'human-in-the-loop', is considered as highly important in domains where decisions are critical, such as healthcare (Holzinger, 2016). Thus, interactive systems that facilitate coaches to critically assess and adapt models that deploy health data are a promising direction to explore, potentially even mitigating coaches' trust issues towards data.

COACHES' TRUST AND ADOPTION

While the value of self-tracked health data for healthcare professionals is recognized (Dinh-Le, Chuang, Chokshi, & Mann, 2019; Figueiredo & Chen, 2020), current design of commercially available self-trackers does not always effectively support clients to share their data with doctors (Nunes, Andersen, & Fitzpatrick, 2019). Health coaches and other healthcare professionals appear reluctant to use health data and health technologies in their coaching practice, for example, the adoption of Patient Generated Health Data (PGHD; Demiris, Iribarren, Sward, Lee, & Yang, 2019) and telemedicine (Choi, Park, Choi, & Yang, 2019) is low. Factors influencing adoption show to be mostly of social and organizational nature rather than technical (Jacob, Sanchez-vazquez, & Ivory, 2020). Specifically, coaches report having limited time and skills to adopt these data in their working routines (Chung, Cook, Bales, Zia, & Munson, 2015; Gagnon, Ngangue, Payne-Gagnon, & Desmartis, 2016; Y. Kim et al., 2017; Lordon et al., 2020), they doubt the reliability of data, suspecting it to be incomplete, irrelevant, or lacking context (West, Kleek, Giordano, Weal, & Shadbolt, 2018), they have privacy and security concerns (Gagnon et al., 2016) and they fear that data will negatively impact their relationship with the client (Lordon et al., 2020). Health coaches' reluctance towards adopting health data may be encouraged by the fact that we typically refer to these devices as *e-coaches*. By doing so, we present them as social actors replacing coaches, rather than tools supporting them (c.f., Fogg, Cuellar, & Danielson, 2009).

To conclude, integrating the strengths of health coaches and health data is promising, as it may substantially improve health coaching practices serving population health. Yet, to date, we have not been able to capitalize on this opportunity. There are many open research questions considering the complex interplay between coaches, data and clients. It is our aim and hope that this dissertation yields insights that will contribute to improved health coaching practices, and that these insights will add to the general body of knowledge in the field of Human-Computer Interaction (HCI) too.

OVERVIEW OF THE DISSERTATION

This dissertation aims at understanding the value and impact of health data in a health coaching process, and the relative strengths of health coaches and health data, each on their own, as well as together. We explore ways to facilitate collaboration between coaches and data that effectively combine their complementary forms of knowledge. We start with a number of qualitative studies that provide insight in the nature of health coaching and the potential role, value and impact of data therein. Subsequently, we adopt more quantitatively oriented approaches to evaluate several means of interaction between coaches and data-driven models, including transparency and model interactivity.

More specifically, **Chapter 2** seeks to answer the research questions; What are coaches' understandings of successful health coaching? And what are their perceptions, attitudes and needs towards technology in this process? We interviewed health coaches (n=9) to explore their current practices and their view on successful health coaching. After, in a focus group (n=4), we gradually moved towards discussing the potential role of technology, and we concluded with a workshop (n=21 coaches, 2 clients) where we observed coaches' hands-on experiences with data. We deliberately took health coaching as our point of departure, and from there gradually shifted our focus to the role of data and health technologies. This allowed for understanding where technology would fit in current practices and meet needs, independent of possible prejudices or anxieties towards technology.

As a follow up in a more practical setting, **Chapter 3** revolves around understanding; How do data change the coaching process? The chapter presents a workshop (n=12 coaches, 3 clients) and a field study (n=5 coaches, 6 clients), where we provided clients with wearable health trackers, and observed coaching sessions where these data were discussed. We particularly focused on how these data changed the coaching process, both on the level of actual coaching content (e.g., conversation topics and coaching advice), as well as on the level of the coach-client relationship (e.g., conversation dynamics and mutual understanding). Our study design allowed to systematically contrast and compare the individual and combined value of a client's data, and a conversation with the client.

In **Chapter 4**, we describe a field study that aims to answer; What are coaches' and clients' experiences, needs and expectations when sharing and receiving data? How does this evolve over time? And, to what extent are they aligned? In this study, we facilitated parents of newborns (n=5) to customize their own tracking practices according to their babies' needs. These data were shared with their health professionals (n=7) over a period of 5 weeks, and tracking practices were regularly changed to meet evolving needs, either initiated by the parents themselves, or requested by the healthcare professionals. Enabling participants to customize their own data, and by following them over a longer period of time, allowed us to observe how they converged to desired practices around data sharing, and how data influenced their understanding of the problem and communication with each other. This study showed that not necessarily the data itself, but the parents' practices around these data, were highly informative for the healthcare professionals. It also emphasized the importance of aligning expectations regarding the data and each other.

These studies showed that, while coaches are generally open to use health data, they want to stay in control of *how* these data are used. They argue that every client is unique, and they doubt whether data-driven models or algorithms can capture the right nuance to serve as effective input for health coaching. In response to these concerns, **Chapter 5** investigates; How does transparency affect coaches' trust in and acceptance of data-driven support systems? Specifically, we designed a health data dashboard with various levels of transparency, and tested this with coaches (n=111) with different levels of expertise. Through an online study we found that laypeople, novice and expert coaches indeed have different needs regarding transparency. Specifically, the results suggest that novice coaches are more likely to change their opinion as result of medium transparency, whereas more experienced coaches are more likely to change their opinion as result of high transparency of data-driven recommendations.

To more actively engage coaches to deploy their domain knowledge, in **Chapter 6**, we build an interactive prediction tool for running coaches who help runners prepare for a marathon. This chapter considers; Does model interactivity improve coaches' levels of trust in the model? When they interact with the model, what exactly do they contribute, and is this improving the prediction accuracy? The chapter describes a set of user studies, i.e., pilot interviews (n=2), think-aloud sessions (n=7) and an online study (n=71), through which we developed a means of interaction that allowed coaches to meaningfully express themselves. Combining qualitative and quantitative methods yielded rich insight into what coaches do, need and value when interacting with support systems. Our findings indicate that coaches are keen on deploying their rich knowledge, and that being able to interact with the system results in increased levels of trust as well as increased performance of the system.

This dissertation finishes with a general discussion in **Chapter 7**, where we reflect on our findings and its implications for the design of wearable tracking tools and data-driven support systems for coaches.

MAIN CONTRIBUTIONS

This dissertation contributes to prior work in several ways. For one, as response to the large focus in current literature on the client, we focus on the coach's needs towards health data. This allows for understanding the unique strengths of coaches and data in the coaching process, and the according roles they may take on. We gain insight in the effect of data on the coaching process, which are not just adding information as one might expect, but also disrupting roles, as it enables both clients and coaches to communicate in different ways. Our work also explores practical ways where coaches and data-driven models may effectively collaborate, where novel means of interaction allows coaches to deploy their rich domain knowledge into those models. Furthermore, we learned that users' needs towards support systems may substantially differ, particularly across different levels of domain expertise.

In addition, our work makes several methodological contributions, mainly by its high ecological validity. We aimed to create realistic and relevant situations for our participants. In most of our studies, participating coaches were assessing data of their own clients, as opposed to often used fictitious tasks based on hypothetical data. This ensured the data were timely and relevant to them, enhancing their investment in the task and the outcome. Related to this, while it is notoriously hard to recruit domain experts – the health coaches – in empirical research efforts, as they are busy professionals, we managed to attain adequate sample sizes, supplemented with rich qualitative data. Lastly, a particular strength of our work is our multi-method approach. We have deployed interviews, focus groups, workshops, field studies, think-aloud sessions, and online experiments and surveys. This broad perspective enabled us to triangulate insights from both rich qualitative data as well as larger and focused quantitative data. In sum, this dissertation yields insights broadening the field of human-computer interaction in the health domain, by studying users' hands-on experiences with data. This informs the design of health technologies to effectively utilize health data in a health coaching process.

So I don't believe in
question is, what do
data, how do you
do you interpret the
you create insights
the client doesn't
it becomes

Coach

merely data. The
you do with the
give feedback, how
data, how do
from the data that
see, (...), then
interesting.

CHAPTER 2

Beyond Behavior: The Coach's Perspective on Technology in Health Coaching

In this chapter, we explore coaches' perspectives on their work, aiming at understanding what health coaching comprises and what role technology could possibly take in such a process. We present three consecutive qualitative studies, starting with interviews with health coaches on their working practices, followed by a focus group where we more explicitly discussed the role of health data and technology, finishing with a workshop where we observed coaches' hands-on experiences with clients' health data. Our results show that coaches are concerned that introducing technology in the coaching process puts too much emphasis on behavioral information, lowering the attention for the client's lived experience, while a good understanding of those experiences is key for successful coaching. We reflect on the value of ambiguity in data, as this potentially facilitates meaningful conversations. The work described in this chapter provides a starting point in our explorations on the possible role and value of health data in a coaching process.

This chapter is derived from:
Rutjes, H., Willemsen, M. C., & IJsselsteijn, W. A. (2019). Beyond Behavior: The Coach's Perspective on Technology in Health Coaching. *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI '19)*. Glasgow, Scotland UK.

Related publications:
Rutjes, H., Willemsen, M. C., Kersten – van Dijk, E. T., De Ruyter, B., & IJsselsteijn, W. A. (2017). Better Together: Opportunities for Technology in Health Coaching from the Coach's Perspective. In *Conference on E-Coaching for Health and Wellbeing*. Amsterdam.

INTRODUCTION

A healthy lifestyle has many benefits. Mortality is on average lower among physically active people than among their inactive peers, and an unhealthy diet is associated with leading causes of death, including coronary heart disease and stroke (Institute of Medicine, 2001). However, to attain a healthy lifestyle, established routines or habits have to be broken and deeply ingrained attitudes need to be changed (Webb & Sheeran, 2006). Health coaches may provide help in this difficult but beneficial process of behavior change.

We define health coaching as a client-centered process where a coach supports an individual client on achieving her goals related to health and wellbeing. The process of health coaching itself is an interpersonal process, where situational awareness, mutual coordination, and substantial knowledge about the personal characteristics and habits of the client are required. In this client-centered process, a good relationship between the client and the coach (O'Broin & Palmer, 2006) and good communication skills (Wolever et al., 2013) are of key importance. In related fields, the importance of a good relationship between caregiver and client is also emphasized, for example in psychotherapy (Norcross, 2002) and in medical settings (Bensing, 1991; Roter & Hall, 2006), in order to elicit patient's values (Berry et al., 2017), and to facilitate shared medical decision making (Ong et al., 1995).

Over recent years, various e-coaching systems have been introduced that offer some unique opportunities for behavior tracking and interventions which were hitherto unavailable, either to coaches or their clients. Specifically, technologies such as smartphones, activity-trackers and health watches are equipped with a broad set of sensors, which allow for higher resolution and potentially more objective tracking, over longer periods of time, than typically afforded through users' subjective self-assessment. Moreover, advances in (big) data processing enable increasingly personalized and contextualized behavioral recommendations and motivational feedback.

However, by and large, technology in health coaching to date has focused rather exclusively on the client's needs and technology affordances, typically arriving at fully automated, stand-alone e-coaching systems or apps. Neither the coach's information needs, nor the nature of the coaching process, have received sufficient consideration in the design and research of this technology. In the present chapter, we aim to address this research gap by taking the coaches' perspective as point of departure in understanding the coaching process, the role of supporting technology, and the design requirements of technology aimed at supporting the coaching process.

RELATED WORK

Personal Informatics and Patient-Generated Data

Over recent years, interest in Personal Informatics and Quantified Self has been growing, where technology and user-generated data (e.g., from wearable trackers) are employed to increase a user's self-awareness and provide her with actionable insights to support the attainment of a user's self-improvement and health goals (Epstein, Cordeiro, Bales, Fogarty, & Munson, 2014; Kersten – van Dijk et al., 2017; K. Lee & Hong, 2018; Li, Dey, & Forlizzi, 2011). This literature is predominantly client-oriented – that is, the end-user's needs and goals are the driving force in developing apps and trackers that allow for self-tracking. At the same time, self-tracking is increasingly construed as a social and collaborative activity, inevitably embedded in a social context (Epstein, Jacobson, Bales, McDonald, & Munson, 2015; Kersten – Van Dijk & IJsselsteijn, 2016; Maitland & Chalmers, 2011; Rooksby, Rost, Morrison, & Chalmers, 2014). This signals that health tracking frequently involves more stakeholders than the primary end-user alone, and that data may be shared. However, self-tracking systems to date have relatively limited functionality in supporting such sharing (Kersten – Van Dijk & IJsselsteijn, 2016).

In clinical settings, there is a growing interest in the value of personally tracked data to supplement existing clinical data, by providing more contextualized and continuous health information. Research on such patient-generated data (PGD) has a broader focus than the empowerment and needs of the patient; it also studies the clinicians' needs regarding PGD, and the extent to which PGD fits and impacts current healthcare practices and workflows (Chung et al., 2015; Gabriels & Moerenhout, 2018; Hong, Lakshmi, Olson, & Wilcox, 2018; Kelley, Lee, & Wilcox, 2017; Mentis et al., 2017; Raj et al., 2017; Rutjes, Willemsen, Kollenburg, Bogers, & IJsselsteijn, 2017; West, Giordano, Van Kleek, & Shadbolt, 2016; West et al., 2018; Zhu, Colgan, Reddy, & Choe, 2016). Even though the use of PGD is increasingly prevalent in chronic disease management, current PGD tools do not adequately support the effective collaboration and communication between caregivers and patients.

E-Coaching, Behavior Change- and Persuasive Technologies

In addition to health data being captured by people (patients, clients) using self-tracking technology, technology can also take a more active role in interpreting the data, and providing the end-user with relevant feedback, timely and personalized cues, and motivational rewards, all of which may support health behavior change. Such e-coaching technology, including Behavior Change Technology (Consolvo, McDonald, & Landay, 2009; Michie et al., 2011) and Persuasive Technology (Fogg, 2002; IJsselsteijn et al., 2006), implicitly or explicitly takes on a role of a health coach, (c.f., Adams, Costa, Jung, & Choudhury, 2015; Nahum-Shani, Hekler, & Spruijt-Metz, 2015; Purpura, Schwanda, Williams, Stubler, & Sengers, 2011). Health coaching is generally conceptualized as a client – or patient – centered process, supporting the health needs and goals of the client. In line with this, the systems in use to date, be they apps (e.g., Runkeeper[3], MyFitnessPal[4]) or wearable sensing devices (e.g., Apple Watch[5], Fitbit[6]), focus primarily on the end-user wearing the device, not on the social or professional context of use.

The Coach's Perspective on e-Coaching Systems

Designing systems that support the process of healthy behavior change, however, does not imply that technology assisting in this process should focus exclusively on the client. First, as e-coaching technology is, to an extent, emulating the role of a coach,

3 www.runkeeper.com
4 www.myfitnesspal.com
5 www.apple.com/watch
6 www.fitbit.com

a deep understanding of what constitutes a successful coaching process should be incorporated into the design of e-coaching systems. Second, as both personal informatics and e-coaching systems are frequently part of a larger ecosystem of behavior change agents which explicitly includes human professionals (e.g., health coaches, medical doctors), the design of such systems should also incorporate the perspective of these stakeholders, the dynamics of the interpersonal coaching relationship, and the requirements of a successful coaching process. Although the primary goal of the coach is to support the client, she has her own unique perceptions, information needs and attitudes towards technology that are fundamental to inform the design of e-coaching systems that will be of value to professionals as well as their clients, and not disruptive of the client-coach relationship nor the coaching process. In current literature, the coach's perspective is underrepresented.

FOCUS OF THE PRESENT WORK

This chapter aims at understanding the health coach's perspective on the role and requirements of technology in health coaching. Medical treatment as part of the health coaching process are beyond our scope, since the process of diagnosing and treating diseases are often associated with more strict guidelines and standard procedures, and thus bring different dynamics into the process. We will take the health coach's perspective on the coaching process as a point of departure and will progressively zoom in on the potential role and impact of technology as part of that process. We will address the following research questions:

1. What defines and influences successful health coaching?
2. What are health coaches' perceptions, attitudes, and needs towards technology in their coaching practice?
3. What do these results imply for design of technology in health coaching, in order to fully utilize the potential of both human coaches and technology?

To answer these research questions, we conducted three consecutive qualitative studies. In Study 1, through semi-structured interviews, we explored coaches' reflections on the process of health coaching, addressing what defines successful coaching, as well as common barriers and success factors that coaches encounter in their day-to-day practice. In Study 2, using a focus group, we explicitly considered the potential role of technology in health coaching. We explored the extent to which potential technology interventions resonate with current coaching practices and focused on the coaches' perceptions and attitudes towards the use of technology. In Study 3, we observed coaches interacting with clients as well as health-data in a hands-on workshop, allowing for a deeper reflection on the potential role of technology, based on this experience.

This study was approved by the local ethics committee at Eindhoven University of Technology, Human-Technology Interaction group.

In the remainder of this chapter, we will discuss each of these studies in detail, and will reflect on the potential role and impact of technology in the coaching process. These insights will be used to formulate design considerations for future technology solutions supporting the coaching process.

STUDY 1: EXPLORATION OF THE COACHING PROCESS

METHODS

We recruited nine Dutch health coaches (seven women and two men, average age of 37 years, ranging from 25 to 56 years) from our personal network, who volunteered to participate in the interviews. On average, they had 10 years working experience, ranging from 2 to 28 years. The coaches were all working on individual basis with healthy clients having health- and wellbeing-related goals, by focusing on diet, physical activity, or both. We interviewed three dietitians, four personal trainers, one coach providing online coaching and one teacher/researcher in coaching, who is also a sports coach.

The interviews were semi-structured, and conducted by one researcher in a face-to-face setting. The duration varied between 30 and 60 minutes. The interview questions were:

1. Can you give examples of things you do and recommend as a coach, and how you motivate your clients? Does this change from person to person?
2. When is coaching successful in your opinion? What contributes to that?
3. Who sets the goal? Can you explain this process?
4. How do you translate a long-term goal into short term activities?
5. Which barriers do you see often with your clients?
6. If you could be a fly on the wall with your clients, what information would you focus on, and what would you do with this information?

As we aimed for a broad understanding of health coaching, we asked about many aspects of coaching (questions 1-5). With the sixth question, we hinted at opportunities for technology, without being overly explicit.

Saturation was used as a stopping criterion: data collection ended when for three consecutive interviews no major new insights emerged. Two researchers agreed that the data was saturated after nine interviews. All recordings were transcribed verbatim, and a thematic analysis was performed following established guidelines (Boyatzis, 1998), using a hybrid approach. That is, we followed the stages of the inductive, data-driven approach, with the exception of Stage II, Step 3, where we also drew from our theoretical background for the articulation of meaningful themes (see Boyatzis, 1998, page 52 for this specific hybrid approach). The final coding of the data was executed by two researchers. The inter-rater reliability (IRR), expressed in percentage of agreement per interview per theme, was on average 98.8%, with a minimum of 88.0%. All disagreements were resolved by discussion.

RESULTS

Four major themes emerged from the thematic analysis, which are summarized in Table 1 along with their subthemes. Below, all themes and subthemes are discussed.

Themes	Subthemes
I: Success goes beyond Achieving Goals	It is about the experience instead of the numbers There is often a more profound issue underlying a stated goal Success is also about learning
II: The Value of a Personal (Human) Approach	The relationship between coach and client is of key importance Social support is a major succetss factor Tailoring advice and coaching style is important and implicit
III: Adapt the Advice to Situational Characteristics	Fit the advice to daily life of client Contextual information of the client is informative to the coach Consider stress & personal barriers Not all information is shared
IV: Motivation is Important	Behavior change is hard – reasonable expectations are important Intrinsic motivation is essential Suitable and specific short-term goals help to motivate

Table 1 List of themes and subthemes of study 1.

Theme I: Success goes beyond Achieving Goals

In the interviews, coaches indicate that coaching is about much more than helping clients achieve their goals. First, many coaches report that success is not always measurable, and often related to experience instead of numbers. For example, Coach #8 often observes that clients who are only halfway their initial weight loss goal may be already satisfied, because they have more energy, sleep better and feel better. In line with this, Coach #1 states: *"Don't look at the scale, look in the mirror. [...] It's about how you feel."*

Coaches also note that often a more profound issue is underlying the explicit goals the client initially presents. For example, bad self-esteem might be underlying a weight loss goal. As Coach #7 explains: *"The first impression is that someone wants to lose weight, but in my experience, it's never about that. It's really about somebody fighting something within themselves [...] So it is my role as a coach to understand which emotions are there, and to be sensitive."*

Furthermore, the coaches indicate that client's awareness of their personal health and behavior is considered as a valuable benefit of coaching. Coach #5: *"In my opinion, if a client didn't achieve the weight loss, it can still be successful, because something changed in their awareness."* Awareness about the impact of a certain diet or behavior on health is important to help clients make deliberate choices. Coach #3 strongly argues that coaching can still be successful when clients are choosing the unhealthy option sometimes, if they are aware of the impact and if their choice is deliberate.

When talking about monitoring progress, some coaches are concerned about clients being obsessed with the numbers. For example, Coach #1 states that long term perseverance is more likely when clients enjoy what they do, instead of focusing on, for example, speed or burned calories. On the other hand, quantitative measures can also positively influence the experience. Coach #7 explains: *"I have people who have such a bad self-esteem, or lost touch with reality of their bodies, that unless I can show them on paper, 'look, you're making progress!' they won't believe me."* Coaches indicate that the effect of monitoring, either automatically or by keeping manual diaries, on their behavior and motivation, varies from client to client. For example, Coach #2 explains: *"Some people, when they see 8,000 steps on their activity tracker and know they have to reach 10,000, they will go for an extra walk to achieve their goal."*

Theme II: The Value of a Personal (Human) Approach

All coaches emphasize the value of the relationship between client and coach in order to be successful. For example, Coach #3 says: *"It is about the trust relationship between client and coach, which implies certain skills a coach should have: standing by someone, being open, not judgmental, providing safety, guiding someone, listening, being empathic."* Many coaches report that the first consultation is aimed purely at building a relationship with the client and that this is a prerequisite to start working towards goals.

The role of social support of friends and family is also mentioned as a success factor. Coach #1 and #8 stress the value of a supporting spouse, and Coach #5 explains how effective it is when clients share their health goals with their colleagues and friends, to help them stay motivated.

Throughout all interviews, a personally tailored approach emerges as being important. All coaches state that what they do depends on the client. For example, Coach #2 explains: *"I'm always looking for the things that a client needs. What motivates, helps, triggers him, or maybe just reassures him at this moment?"* Notably, almost all coaches report they tailor intuitively and that the process of tailoring is hard to explain explicitly. However, further probing does reveal certain personal and situational characteristics they use in their tailoring. Personal characteristics include personal goals or problems, the client's need for empathy (e.g., a strict approach versus 'hand holding'), how motivated clients are, their base level (*"With some of my clients, I'm already happy if they'd eat one piece of fruit a day"* (Coach #9)), potential physical injuries or limitations, gender, age, profession and their place of residence (rural or urban). Situational characteristics are even more commonly mentioned

I have people who have such a bad self-esteem, or lost touch with reality of their bodies, that unless I can show them on paper, ‘look, you’re making progress!’ they won’t believe me.

Coach

by the coaches as tailoring aspects and are discussed separately in the next section (Theme III).

Theme III: Adapt the Advice to Situational Characteristics

All coaches stress the importance of adapting the advice to situational characteristics of the client. They try to make the advice very specific and fit it into the daily life of the client. Practical constraints like working night shifts, truck drivers who are on the road, or different cultural backgrounds are extensively discussed in consultation meetings to find feasible solutions. Even back-up plans are made, as Coach #5 illustrates: *"If a client plans to go for a run, but does not want to run in the rain, I suggest installing a weather-app and we discuss a back-up plan."*

It may not always be easy to gain access to reliable information about a client's daily life and behavior. The coaches mention common problem with clients withholding information – simply because they don't know it is relevant, or to avoid shame.

Some coaches indicate that the client's context may often reveal highly relevant information about the client. For example, Coach #3 reports: *"consultations where a partner or parent joins gives me much more information. [...] Also, home visits are a very important source of information, seeing the kitchen and the fridge tells me a lot."* Coach #9 asks the clients to make a food journal: *"It says it all. Some people forget it, then you know they're not motivated, [...]. Some write very sloppy, others very tidy, including the times, others bring food to the consultation or make pictures. It's not only the information itself, but also the way it is presented, which is very informative."* Thus, examining the context of the client offers a rich source of information and helps to tailor the coaching process to the client's needs.

Not only external, practical barriers but also internal, personal barriers play a role in coaching. For example, the coaches report they must be sensitive to their client's stress levels in order to gauge their readiness to change behavior. Coach #2 explains: *"When there are big stressors like divorce or change of jobs, it is very hard to change behavior"*. Some coaches talk about 'the right moment' to make the change. Knowing barriers that impede adherence to the coaching plan helps to better understand and assist the client.

Theme IV: Motivation Is Important

All coaches report that for long-term behavior change, motivation is very important. First, reasonable expectations should be elicited; it makes a difference if clients realize that change is hard. Coach #8 and #9 report that they have had clients expecting them to be a magician and that just visiting them will initiate a change, not realizing that they need to change themselves. Second, being intrinsically motivated is indicated as key for success by all coaches. Some coaches indicate that clients who visit them because their doctor referred them (e.g., because of diabetes) are the most difficult to coach, because they do not come on their own initiative.

Although overall success is more than reaching explicit goals (see Theme I), all coaches emphasize that setting specific, measurable short-term goals is essential in providing success experiences, and thus motivation. Coach #7 compares the short-term goal "losing some body fat" with "losing 0.5% body fat". She described the power of making the goal very specific: *"[...] Next week, 0.5% done, high five, everybody happy, check off that goal. You see what I mean? [...] To the mental side, it makes a world of difference."* Not only measurable, also achievable goals are important, as they increase self-efficacy. It is common practice for coaches to make gradual, step-by-step changes.

In order to keep clients motivated, it is very important to focus on (small) successes, and divert the attention away from inevitable stagnations or lapses in adherence. Coach #2 explains that this is especially important when dealing with clients who have low self-esteem, for example those who are binge-eating. Coach #6 often creates a personalized progress report for clients to make the success over the longer term visible.

CONCLUSION: COACHES' REFLECTIONS ON SUCCESSFUL COACHING

The interviews demonstrate coaches' reflections on successful health coaching and the factors that may impact that process. Summarizing, health coaching is an interpersonal process that goes beyond measurable goals and activities. Goals and corresponding successes are often related to the client's experience rather than measurable behavior, and a good relation between client and coach is a prerequisite for successful coaching. The interviews also illustrate the importance of fitting the advice to situational characteristics: The more specific and tailored the advice, the more likely clients are to adhere to it. Tailoring shows to be an important aspect of successful coaching, yet coaches find it hard to formalize this process. Intrinsic motivation of the client is considered as a key success factor, and specific and achievable short-term goals help clients to stay motivated and increase self-efficacy.

STUDY 2: POTENTIAL ROLE OF TECHNOLOGY IN HEALTH COACHING

Purposely, we did not mention the role of technology explicitly in the interviews in Study 1. In elucidating the coaching process and its success factors, we did not want the results to be biased by coaches' perceptions on the capabilities of technology, or possible resistance against technology. In Study 2, as a follow up, we explored the potential role of technology in health coaching more explicitly. We used a focus group to explore coaches' perceptions and attitudes towards the use of technology as part of the coaching process. This method provides an additional advantage over interviews, in that the interaction between coaches may spark richer discussions and provide new insights.

METHODS

We invited all coaches of Study 1 for a follow up in a focus group; four out of the nine health coaches agreed to participate. The group consisted of a dietitian, two personal trainers and an online coach. The session lasted 2 hours and was facilitated by the same researcher who conducted the interviews. The coaches received a small financial compensation for their participation.

The focus group started with a short 'warming up': a car navigation system was discussed as a metaphor for the potential role of technology. This was framed by the levels of automation per sub task as proposed by Parasuraman, Sheridan, and Wickens (2000): information acquisition and analysis (e.g., road network, traffic congestion information), decision making (e.g., the driver is provided with route information while driving) and action (e.g., self-driving car, cruise control). The coaches were asked to apply the idea of such levels of automation on subtasks in health coaching. First, the coaches individually wrote down their thoughts for 10 minutes, after which a group discussion started. The group discussion was driven by the following questions, which were visible on a screen during the session. The facilitator refocused the attention to one these questions whenever the discussion was going off topic.

1. What can technology do for you? What do you need?
2. How would technology fit in your workflow?
3. What drawbacks of technology do you see?
4. On the allocation of tasks between you and technology: What do you prefer to do yourself, and which tasks do you prefer to outsource to technology?
5. Can you describe your new role, when assisted by technology?
6. Do you feel you have enough skills to work with technology?

The session was video-recorded and transcribed verbatim. Segments were selected when relevant to one of the themes 'potential role of technology' and 'perception of, or attitude towards technology'. These segments were clustered in topics using thematic analysis (Boyatzis, 1998), using a similar approach as in Study 1, by two researchers.

RESULTS

After the introduction a vivid discussion sparked easily among the coaches. The facilitator did not need to intervene often; all four coaches were open and willing to share their opinions and experiences. The facilitator only interrupted to give turn to other coaches, or to refocus the attention to one of the main questions. The emerging themes represent the coaches' view on the potential value of technology, as well as coaches' concerns around technology. See Table 2 for an overview. All themes are explained in the sections below.

Themes	
Potential value of technology	• More and better quality information • Being present 24/7 and motivating clients • Support with administrative tasks
Coaches' perceptions and attitudes towards technology	• Interaction with the client • Analysis of the client's data • Limitations of data and negative psychological effects

Table 2 List of themes of study 2.

More and Better Quality Information

The coaches report a clear value of technology in providing them with more and better quality information. For example, Coach #B explains *"At the start I want to know many things, related to physical activity, nutrition,* (...) *and I only get half of that. (...) that would be something that technology can support me with."* Having access to objective information is helpful, as often there is a mismatch between reality and what people report. They do notice a need for tracking the *right* type of information; they doubt if this is always feasible. For example, often information on a client's mental state and experiences is of interest, but not trivial to track.

Being Present 24/7 and Motivating Clients

The coaches indicate that technology can be helpful in being present 24/7 in order to motivate the client. Coach #C describes that with one of her clients: *"I don't have the time to always be by his side. But then he would always be aware that he is actually doing well."* Being present and sending motivational messages throughout the day can help clients stay motivated, according to the coaches, especially in cases when intrinsic motivation is lacking.

Support with Administrative Tasks

Another benefit of technology emerging from the focus group is support with administrative tasks, e.g., by automatically making notes during consultation meetings, or keep detailed track of exercises during workouts. Coach #D reports *"I try to avoid making notes as much as possible, because I always find it annoying* (...) *I think it would be valuable if notes could be made automatically, based on speech recognition."* When the coaches are supported with their administrative tasks, they can focus more on the main role that they see for themselves: being there for the client, interacting with her, and through that, understanding her goals and experiences.

Interaction with the Client

Supporting with administrative tasks, technology could play a more proactive role in the interaction with the client, for example by detecting emotions, non-verbal cues and change talk. Coach #C explains: *"As a coach it is very important to notice change* *talk, for example 'maybe I can' or 'I would like to' and intervene immediately. Maybe an algorithm can help to detect change talk, for example my watch would vibrate, so I don't miss these opportunities."*

According to the coaches, interpreting and reflecting on the story of a client is a very important element of successful coaching, and not something that technology can, or should, do. For example, they report that it is very effective to connect a client's goals to her values (e.g., being fit so you can play with your children), and as Coach #C explains: *"A computer couldn't interpret this, and reflect, like: 'I understand you'".* Even in the case technology would be able to do this, the coaches think it will be more effective for a human coach to take this role. All coaches agreed that a fully automated coaching solution would not sufficiently motivate clients. Coach #B reports: *"That personal touch is crucial, because only controlled by a computer… I myself would also be like: whatever. But if someone is looking, it's different."* Furthermore, the coaches express concern about losing skills by too much interference of technology in the process, e.g., being sensitive to cues of the client, and being alert and ready. Coach #C questions: *"What makes us humans unique? And what makes us sensitive to each other? And if we outsource that to technology… we lose a lot of human strength. That idea scares me."*

Analysis of the Client's Data

The coaches indicate that a structured use of a client's data can help in the coaching process. Coach #C illustrates this with an example of a client who was often feeling faint during exercises. By structurally keeping a record she could pinpoint this to a certain type of exercise and was able to tailor the training sessions such that the problem was resolved. Also, the coaches see a benefit of using data for calculating realistic short-term goals based on the client's history.

The coaches do not see much added value of relating the data of single clients to those of others, such as using data of a large group of clients to explore preferences and new opportunities for clients. When this topic emerged, Coach #C reported: *"Ok, maybe a database would be a good addition, but in the end of the day… You know, everyone is unique."* She explains that, although a coach is expert on health, the client is expert on herself. With that perspective, coaches consider it their task to tune the coaching program to the client's needs and possibilities and facilitate her in her own process. They do not feel technology could be sensitive enough to fulfill this serving and humble role, nor that additional information of other clients would improve this process. Another objection of the coaches against the idea of using data of other clients, is that this might lead to recommendations reinforcing current (potential unhealthy) behavior, instead of pushing a client slightly out of her comfort zone. They explain the fine line between engaging the client with activities that she enjoys, at the same time challenging her a little bit. Finding and keeping the right balance is a subtle and socially interactive process, and they do not trust the capabilities of technology in this regard.

A computer couldn’t interpret this, and reflect, like: ‘I understand you’.

Coach

Limitations of Data and Negative Psychological Effects

The coaches report that the use of (too much) technology may increase the risk that clients become obsessed by the data instead of listening to their bodies and enjoying the activity. The coaches state that tracking data encourages competition, whereas they feel that persistence of healthy behavior is related to joy of the activity. They consider themselves to have an important role in both reminding the client to listen to her body, and in interpreting and making sense of the data. Coach #D explains: *"So I don't believe in merely data. The question is, what do you* do *with the data, how do you give feedback, how do you interpret the data, how do you create insights from the data that the clients don't see, (...), then it becomes interesting."* The coaches stated that the effect of the data can be very different for different people, so it is the coaches' task to carefully use the data and show it at the right moment and in the right way. Furthermore, they are skeptical about the existence of a ground truth on what constitutes healthy behavior, so having the data does not necessarily imply a straightforward interpretation.

Importantly, the coaches note an additional responsibility when confronted with (more) data. They are worried about their liability when serious issues remain unnoticed. With this perspective, more data is not always better; instead, it provides them with additional responsibilities, higher work load and potential stress.

CONCLUSION: COACHES' PERSPECTIVES ON TECHNOLOGY

The focus group results illustrate that coaches see the added value of technology mainly in having access to more reliable information about their clients' health behavior, as well as in the opportunity to be a supporting presence in their clients' lives. Furthermore, the coaches see benefits of technology to support them with several tasks in their practice, ranging from exercise logging to recognizing a client's 'change talk' and calculating appropriate short-term goals.

The coaches also have a number of substantial reservations about the added value of technology on the coaching process. For one, coaches expect that an emphasis on data may foster obsessive behavior. Also, they think technology may not be sufficiently advanced to appropriately tailor its interventions to the individual client. Thus, they are skeptical about automatically generated recommendations or feedback, and feel that automated interventions that lack the involvement of a human coach will not be very effective. Furthermore, they worry about additional responsibilities on their part that may arise from being provided with (more) data.

STUDY 3: COACHES INTERACTING WITH DATA AND CLIENTS

Study 1 and Study 2 illustrate the complexity and social nature of the coaching process, and the challenges that may arise when introducing technology in the mix. As the first two studies did not allow coaches to actually gain some firsthand experience with technology or data, we organized a workshop where a group of coaches interacted with two clients who brought their self-tracked health data. Interacting with possible artefacts could potentially reveal value that remains hidden when just talking about technology in hypothetical terms. For the purpose of comparison, half of the coaches were invited to interact directly with the clients, while the other half were asked to consult the self-tracked data only. Both groups independently formulated an advice. Afterwards, in a group discussion, the results were compared and the value of both sources of information was discussed.

METHODS

Twenty-one coaches (mostly personal trainers) volunteered to participate in a 1.5 hour workshop, embedded in a seminar targeted at employees of all Dutch student sports centers. Two clients, recruited through the authors' personal network, introduced themselves briefly (3 minutes each), explaining their goal and question to the coaches. Client #1 aimed to be stronger, and wondered why she remained stable, despite her fitness training sessions and healthy diet. Client #2 had a hard time to keep up with her running training sessions, while she aspired to run the half marathon.

After this introduction, 4 groups were formed. Five coaches were interacting with client #1, 5 coaches jointly consulted the data of client #1, 5 coaches were in interacting with client #2 and 6 coaches jointly consulted the data of client #2. The coaches who were interacting with the clients had no access to data. In these parallel sessions (30 minutes each) the coaches were asked to formulate an advice for their client. Afterwards, a group discussion was initiated (40 minutes), where every group shared their insights and advice, and the differences between the groups and the added value of the tracking data and the interpersonal coach-client interactions were discussed.

Prior to the workshop, both clients wore a Jawbone UP3 for 2.5 weeks, which tracked their sleep, steps and resting heart rate. Additionally, client #1 wrote down her daily food intake. Client #2 brought the data of her TomTom Sports GPS watch, which contained all of her running and cycling training sessions of over a year. This included training duration, distance, GPS information and heart rate. The coaches could assess the data via the standard interfaces of Jawbone and TomTom Sports, using provided laptops and tablets.

The parallel sessions and the group discussion were audio or video recorded and transcribed verbatim. In the analysis, segments were selected from the transcripts that related to either the value of human coaches, or the value of technology and data. The segments were clustered by two researchers using thematic analysis (Boyatzis, 1998), with a similar approach as in Study 1 and 2.

RESULTS

All parallel sessions showed vivid conversations, either with the client or about the data. Also the group discussion was lively, with active contributions from participants of all groups. In the next sections, the emerging topics will be described. See Table 3 for an overview.

Themes
• Emotional background of health goals • Understanding beliefs underlying the behavior is important • Data and conversation bring different views on actual behavior

Table 3 List of themes of study 3.

Emotional Background of Health Goals

Client #2 has had negative sport experiences when she was a child in school. She explains to the coaches: *"the education was like a push, you* have *to... and you're graded for it."* Her goal of running the half marathon originates from this background. However, she doesn't like running, she enjoys cycling much more. For the coaches in dialog with the client, this is clear right away.

In the other parallel session, where coaches consult her data, the coaches only know from the introduction that she had a hard time keeping up her running training sessions and aimed for the half marathon. Still, they interpret her underlying motives based on the data: *"I think she is looking for a thrill, because she goes into a very high heart rate zone and stays there for 20 or 25 minutes, and then suddenly it is over."* Not only during the training session, but also from her daily routine, the coaches deduce something about her character: *"She's just not in balance. She's not moving there, that day only 5000 steps, and then the next day suddenly 18000 steps."* And later: *"It fits with what she wants, she wants the endorphins, it seems like she is punishing herself for a lazy day, to compensate directly."* Thus, the behavioral data and a sparse self-report in the introduction provide hints about the client's background, motives and character.

When the individual groups share their findings in the shared group discussion, both groups report that the information of the other group is complementary to their own information. The coaches with data state that the story explains why the data is like this, and the coaches interacting with clients directly report it is interesting to see these emotional issues expressed in terms of behavior.

Understanding Beliefs Underlying the Behavior is Important

Client #1 has strong beliefs on what is healthy, and reports she bases her ideas about health on scientific literature she studies extensively. The coaches in dialog with the client are quickly aware of her beliefs and determination. They also talk about persuasive strategies that are likely to be effective: *"If we provide her scientific articles supporting our advice, then we will probably be much more likely to convince her."* Similarly, in the other parallel session, the coaches consulting the data of client #1

She is just very structured. (...) She doesn't change things in her diet nor in her training sessions.

Coach

deduce from the data that she must have a determined and headstrong character. They report: *"She is just very structured. (...) She doesn't change things in her diet nor in her training sessions."* They advise the client to bring variation in her nutrition as well as in her training sessions, and to be a bit less persistent in her attitude towards a healthy lifestyle.

Data and Conversation Bring Different Views on Actual Behavior

Access to practical facts of the client's daily life is important to the coaches. Some are more easily accessible using data, others appear clearer for the coaches in dialog with the client. For example, client #1 went on holiday for one week in the period when she was tracking her behavior. This is obviously noted by the coaches consulting her data, but they draw the wrong conclusion: she might not work as hard as she reports. On the other hand, the coaches with data have access to much more detail on client #1's diet. Therefore, they advise her to take the carbohydrate-rich meal *before* her training session instead of after, to increase the effectiveness of the training. The timing of this meal remains hidden for the coaches in dialog.

Additionally, client #2 has limited knowledge on training schemes; she declared to the coaches: *"I started running immediately 4 or 5 kilometers. At once."* The coaches try to explain her about training zones based on heart rate, but that conversation stagnates because they lack detailed and accurate information of the training sessions. In the parallel session, the coaches consult the training-data of client #2 to understand her training style. Based on that, they advise a specific training scheme, even embedding it in the daily context of the client by inspecting a typical running track and adapting it on a very practical level: *"Here is a bridge, right? Then she can take a break there, and then go back."*

CONCLUSION: COACHES INTERACTING WITH TECHNOLOGY

In the workshop, we set up two extreme situations: having access to a client's data only, or interaction with the client only. This setup allows us to compare and contrast the added value of each source of information, separately and together. In line with insights from Study 1, the results of the workshop support the notion that goals can be ambiguous, and that there can be more profound issues underlying a client's goal, which are required for the coaches to know in order to give suitable advice. The results show that direct interaction with a client is an effective way to unravel these issues in qualitative and experiential terms, through self-report on how clients feel and what they believe. On the other hand, the coaches consulting the data discern most of these issues too, only in quantitative and behavioral terms. Having detailed information provides a good understanding of the actual behavior, and results in more specific advice, situated in the day-to-day context of the client.

Concluding, information is both hidden *and* revealed through behavioral data. Stand-alone data are not sufficient for effective coaching, yet at the same time, data provide additional insights that improves the understanding of the client. In the words of one of the coaches in the group discussion: "*Data help the coach to ask the right questions*."

DISCUSSION

In health coaching, many technological tools (apps, trackers) are being developed and used to support end-users in meeting their health goals. In developing such tools, the coach's role and perspective are rarely fully understood or well-represented. We believe that in order to develop successful e-coaching applications, we need to have a deeper appreciation of what constitutes a successful coaching process, which includes both the client and coach perspectives. In the current chapter we aim to extend our understanding of the coach's perspective on e-coaching technology, and the potential value that technology and (tracked) data may have in a successful coaching process and a productive coaching relationship.

Based on the interviews with health coaches in our first study, we conclude that health coaching is about understanding the client – both in terms of behavior and experience – and effectively anticipating on that. Inspired by the multi-channel telecommunication model (Zimmermann, 1980), we depict this process as two communication channels between client and coach, one focused on behavioral data and one on the client's lived experience, i.e., sharing stories, background and daily experiences (see Figure 2, left). Our second and third study reveal that coaches see that technology has the potential to improve the understanding of the behavioral aspects of the client, i.e., by tracking and sharing data. At the same time, coaches emphasize the importance of having information about the client's experience, and they believe that this is too subtle and ambiguous for technology to 'understand'. As a result, coaches fear that incorporating technology into their coaching practice results in overemphasizing 'objective' and numerical information on measurable behavior, thereby discounting subjective experience and personal context, which would be detrimental to the coaching process (see Figure 2, middle). We will discuss these results in more detail below, and propose a cross-over that bridges the behavior and experience channels (see Figure 2, right).

TECHNOLOGY PROVIDES MORE INFORMATION ON BEHAVIOR

In health coaching, information on health-related behaviors (e.g., nutrition, physical activity) is key for successful coaching. The coaches clearly recognize the value of technology giving them access to better and more reliable information about the training sessions and the client's daily behavior. This resonates with insights from the sports psychology domain where monitoring technology for athletes provides metrics that are helpful to plan and optimize training programs (Cardinale & Varley, 2017; Halson, 2014). Also the workshop (Study 3) shows that interacting with detailed behavioral data

Figure 2 Conceptual model of impact of technology on coach-client communication.
Left: Current situation: exchange of behavioral and experiential information.
Middle: Expected situation when technology is incorporated: predominance of behavioral information.
Right: Desired situation: behavioral information is represented in meaningful terms, facilitating the exchange of experiences.

is useful to get a richer and more precise view on the client's actual behaviors and routines, which may be less prone to errors than using a client's self-reported information.

Thus, technology can provide coaches with behavioral data, which is a valuable source of information to optimize the coaching process. Yet, the coaches are skeptical about automated normative interpretation of the client's behavioral data, because ground truth metrics on what is healthy behavior are lacking. Health data are inherently ambiguous; what might be good care for one patient under certain circumstances, can be sub-optimal for another (Kersten – Van Dijk, IJsselsteijn, & Westerink, 2016; Raad voor Volksgezondheid en Samenleving, 2017).

The availability of, and focus on, behavioral data might impact the client and coaching process, in particular regarding motivation. The coaches state that progress feedback or automated motivational messages in between consultation meetings can potentially be helpful, for example in making small progress become visible. There is general support in literature that feedback from technology on behavior can be an effective way to change behavior (Hermsen, Frost, Renes, & Kerkhof, 2016; Orji et al., 2018). At the same time, coaches in our study are aware of the potential drawbacks of (real time) feedback on clients' behavior and motivation. For example, they state that when clients are more focused on the feedback than on the joy of the activity itself, it easily undermines intrinsic motivation. As illustrated in Figure 2 middle, behavioral data get overemphasized, leaving the subjective experience underexposed. These insights resonate with controlled experiments demonstrating detrimental effects of external rewards (Deci, Ryan, & Koestner, 1999) or even just the presence of self-tracked information (Etkin, 2016) on intrinsic motivation.

Coaches in our study also express concern that their clients would become obsessed by the data. One's own health data is shown to inherently have emotional connotations and provoke value judgements, including shame and obsession (Ancker et al., 2015; Cordeiro et al., 2015; Orji et al., 2018). This means that behavioral data, in addition to inadequately representing experiential qualities, may even negatively influence that experience. Thus, as the effects of feedback on motivation are mixed, it requires a deeper understanding of the client to use it effectively, arguing for systems with a 'coach-in-the-loop'.

COACHING BEYOND BEHAVIOR

For successful coaching, coaches highlight the importance of understanding the client's experiences, guiding the interpretation and use of behavioral data. The importance of experiential and relational aspects of the coaching process was emphasized time and again, serving a variety of goals, including understanding a client's implicit motivations, providing adequate social support, and helping someone overcome personal barriers.

Several studies in the field of Quantified Self share the idea that behavioral data are just a proxy of underlying experiences. The real value of self-tracking is often pointed out as a contextualized, subjective and social process (Elsden, Kirk, & Durrant, 2016; Karapanos, Gouveia, Hassenzahl, & Forlizzi, 2016; Kersten – Van Dijk & IJsselsteijn, 2016; Rooksby et al., 2014), nicely illustrated by Rooksby, Rost, Morrison, and Chalmers who conceptualized the use of activity trackers as 'lived informatics': "Tracking was explained in terms of people's lives, worries, hopes, interests, careers and so on" (p. 1171, Rooksby et al., 2014). At the same time, Lupton (2016b) highlights a tension, also present in our results, between quantifiable data being perceived as more reliable on the one hand, whereas on the other hand "numbers alone tell us nothing, it is the context in which the numbers [...] are created that [is] important" (p. 6). Also for meaningful interpretation of health data among peers, context is essential for understanding (Agapie et al., 2018; Puussaar, Clear, & Wright, 2017).

Our results resonate with the related field of sports psychology, where coaching is frequently characterized as a complex and ambiguous process that is hard to capture through behavioral information alone (Bowes & Jones, 2006; Jones & Wallace, 2005). To infer experiential information from behavioral information, substantial contextual information as well as social intelligence is required. Despite rapid advances in artificial intelligence and context-aware computing, creating socially intelligent systems is still one of the major challenges in the field (Frey & Osborne, 2017; Green & De Ruyter, 2010). Because the client's behavior is relatively easy to measure, compared to the client's experience, it is likely that behavioral information will predominate a coaching process supported by technology. This effect has also been observed in healthcare, where incorporating information technology involves the risk of focusing too much on measurable and predictable workflows, not acknowledging the flexible and fluid nature of the work (Ash, Berg, & Coiera, 2004; Marcu, Dey, & Kiesler, 2014), failing to capture emotional aspects (Mentis, 2010), and sometimes even resulting in work-arounds to overcome constraints introduced by technology use (S. A. Collins, Fred, Wilcox, & Vawdrey, 2012).

SYNERGY BETWEEN BEHAVIORAL AND EXPERIENTIAL INFORMATION

Our workshop (Study 3) demonstrated that having access to behavioral data potentially facilitates the exchange of experiential information and enhances the relation between client and coach. The coaches reflect on the strength of combining data and personal interaction: *"Data help to ask the right questions."* It provides access to the client's context and experiences which otherwise might have been overlooked by the coaches. In this way, data can be a cue for interaction, potentially triggering new topics of conversation, or directing attention to underexplored aspects of behavior. This adds to studies in medical settings, where similarly the main value of patient-generated data is found to be a facilitator of collaborative reflection, supporting communication, mutual understanding and shared decision making (Chung et al., 2015; Hong et al., 2018; Marcu et al., 2014; Mentis et al., 2017; West et al., 2018).

To facilitate this synergy between behavior and experience in the coach-client interaction, it is important that information from the two channels is easily connected. Ideally, experiences should emerge intuitively from representations of behaviors. Therefore, we propose to facilitate a cross-between both channels (see Figure 2, right), by presenting behavioral data on a meaningful level. So, in addition to showing behavior on a low level, for example, 1000 steps, 7 hours of sleep, 500 kcal, data are aggregated and represented in a meaningful way, shifting it to information rather than data. For example, depending on temporal, locational, and other contextual information, 1000 steps could be construed as a 'lunch walk' or a 'hospital visit', or, 500 kcal as a 'dinner with friends' or a 'dinner at home later than usual'. This behavioral, yet contextualized and meaningful chunks of information will more easily trigger interpretations at an experiential level in interaction with the client, e.g., 'relaxing lunch walk with a colleague' or 'stressful hospital visit with my child', providing the coach with a rich set of pointers accessing the client's daily life, context, values and needs.

THE VALUE OF AMBIGUITY

An important question is, to what extent can and should technology interpret behavioral information, in order to facilitate the coaching process optimally? Context-aware computing makes a helpful distinction between contextual information of a particular behavior (the *who*'s, *where*'s, *when*'s and *what*'s) and the intention of that behavior (*why* the behavior has occurred) (Dey & Abowd, 1999). The contextual information is often not ambiguous. However, the *why* of the behavior, that is, the intention and meaning of the behavior to the client, remains ambiguous. A lunch walk might imply that the client was relaxed, but it can very well mean the opposite – in the case where she desperately needed a break on a stressful day. Thus, behavioral information remains ambiguous in terms of experience, and is therefore problematic to interpret automatically.

Ambiguity is not necessarily problematic. Gaver, Beaver, and Benford (2003) show that it may have value when things are left open for interpretation, as it reveals something about the identity, motivations and expectations of the interpreter. In the specific case of health data, it has been shown that the interpretation is colored by one's own beliefs (Kersten – Van Dijk et al., 2016; Lupton, 2016b). This does not have to be amiss; the ambiguity might actually comprise the value of using those data, as long as it can be used as a conversation topic in coaching (Rutjes et al., 2017). Ambiguity, or stronger, inconsistency, has also been valued in *triangulation* (Mathison, 1988), a methodological concept in social sciences. Triangulation describes the process where multiple sources of evidence are used to validate a claim, for example, using multiple methods, data sources or researchers. Mathison (Mathison, 1988) states that, traditionally, we tend to strive for converging evidence, where all sources of evidence are leading to a single claim. However, often evidence is inconsistent or contradictory, and this can actually be a valuable opportunity to learn. It invites the researcher to make sense of the differences, ending up with a more holistic view of the subject of interest and more valid claims (Mathison, 1988).

CONCLUSION: TECHNOLOGY SUPPORTING THE COLLABORATIVE COACHING PROCESS

In the complex and interpersonal process of health coaching, interpreting the client's behavior, i.e., talking about intentions and meanings of the behavior to the client, is a highly valuable activity in itself. Therefore, ambiguity make behavioral data useful. It is not possible, not required, and even stronger, it would be an unfortunate loss, to automate this task of interpretation. It would impede collaborative reflection and the enhancement of the coach-client relationship that is critical to success. Gaps and irregularities in the data, as well as situations where behavioral and experiential information are contradicting, are valuable starting points for effective coaching.

At the same time, presenting behavior in a meaningful way, easily triggering the recall of experiences, is an important facilitator of the collaborative coaching process. Along the lines of context-aware computing, the who's, where's, when's and what's provided by technology are used to determine why the behavior occurs (Dey & Abowd, 1999). Low-level data is often too distant from experiences to facilitate effective communication. The cross-over between the two channels (see Figure 2, right) aims at balancing between interpreting behavioral data on such a level that it optimally supports collaborative reflection and sharing experiences, at the same time, not restraining from the value of ambiguity of the behavioral data in this process.

Coaching is a social process, vigorously engaging both client and coach. As such, our focus on the coach's perspective provides only a partial view on the role and value of technology in health coaching. On the other hand, as argued in the beginning of this chapter, previous literature mainly focusses on the client's perspective on technology in health coaching, leaving the coach's perspective underrepresented. Our results show that health coaching is all about the interaction between the client and the coach, and therefore, future work should focus on how technology influences and may effectively support these coach-client interactions.

Our results show that the coaches value support of technology in terms of having access to meaningful facts and figures on the client's behavior. At the same time, the coaches clearly emphasize that they want to be in the lead when it comes to understanding the client and shaping the coaching process accordingly. Technologies in health coaching are often called *e-coaching* systems, giving the impression that these systems are aiming for replacing the coach, rather than supporting her. We conclude that in the coaching process, being a dynamic, contextualized and social activity, there is a unique and important role for the human coach. Technology potentially provides a valuable contribution, by informing and facilitating this process. By bringing meaningful information, yet accounting for the complexity and dynamics of the health coaching process, technology potentially promotes a better informed and more effective dialog, closer to the client's experience, and enhancing the relationship between the client and the coach.

Sunday, uh
going on this... You
20th. 13222 steps.

Coach

Monday? What was were working? The

The Influence of Personal Health Data on the Health Coaching Process

CHAPTER 3

The Influence of Personal Health Data on the Health Coaching Process

In Study 3 of *Chapter 2*, we have seen how both a client's data and her self-report reveal different types of information and provide different views on the client, including on her behavior, beliefs, and potentially effective coaching strategies. In this chapter, we further investigate the value of both sources of information, and observe how they merge in practice. We report on a workshop and a field study where we enriched coach-client interactions with a client's self-tracked data. We included both familiar and unfamiliar coach-client pairs, as well as alternating the timing of the data presented (i.e., at the beginning, or halfway through the session), which allowed for acquiring a variety of data-driven coaching interactions. Analyzing these through a mixture of qualitative and quantitative methods reveals that data are not 'plug-and-play'. There is an extensive process of interpreting and contextualizing data, in which the client has a key role, that is essential to gain relevant and actionable insights from the data useful to the coaching process. Our results also show how data shifts roles, and how both coaches and clients are motivated to put the data in the right perspective when talking about them. Thus, this chapter illustrates the influence health data can have on a coaching process, including on content level (e.g., coaching advice) as well as on relationship level (e.g., conversation dynamics).

This chapter is derived from:
Rutjes, H., Willemsen, M. C., Feijt, M. A., & IJsselsteijn, W. A. (*Under review*). The Influence of Personal Health Data on the Health Coaching Process.

INTRODUCTION

Tracking health-related behaviors is becoming increasingly commonplace, through the ubiquitous availability of consumer tracking technology including wearable devices and smartphone applications. Such technology enables a wide range of measurements, from simple step counts to more advanced measures such as sleep stages and heartrate measurements. Visualizations of these data provide users with a level of insight in, and potentially control over, their own health. These devices are typically presented as e-coaches, thus, helping users to achieve their goals rather than merely presenting information. This may include sending motivational messages, recommendations, and comparing the behavior with health standards (e.g., 10k steps a day) or user-set goals. Thus, when facing health issues or setting health goals, it is assumed that tracking one's data and interacting with e-coaches provide a helpful solution.

The increasing use of tracking technology is inevitably transforming health coaching and the practices of health coaches, such as personal trainers or dieticians. When people visit a health coach, bringing one's self-tracked data will be increasingly commonplace. When people have tracked their health data before meeting a coach, health coaches are facing clients who are potentially better informed and more engaged with their health. Also, the data provide coaches an additional source of information that is essentially different from traditional self-report. As wearable devices are carried along in the daily life of the client, they can provide continuous, high frequency and in-situ measurements, enabling a detailed overview of trends over time (Sqalli & Al-Thani, 2020) contextualized in the daily life of the client (Figueiredo & Chen, 2020). In addition, the nature of the data is different; it is initiated by the client herself rather than suggested by a coach or doctor, potentially better reflecting the client's perspective and needs. While these benefits are clear, data may not necessarily be informative to coaches or beneficial to the coaching process – it may as well be regarded as a distraction or be perceived as a threat challenging a coach's expertise. Thus, coaches may find themselves competing against rather than collaborating with the data. Indeed, current adoption of patient-generated data in healthcare contexts is low (Demiris et al., 2019). Health coaches have reported issues with data in their coaching practice, including disruption of their relationship with the client (see *Chapter 2*). By any means, wearable devices and the data they provide are changing the coaching process, which will affect coaches, clients and their relation in important ways. In the present work, we seek to understand this effect of data on the health coaching process.

WHAT IS HEALTH COACHING?

Before discussing the possible effects of data on health coaching, we first consider the process of health coaching itself, independent of technology. A systematic literature review of Wolever and her colleagues (2013) reveals some key elements of health coaching. The authors find that health coaching is a patient-centered approach where a patient's goal is leading. Health coaching involves self-discovery, education or self-monitoring, all within the interpersonal relationship with the coach, who is guiding this process (Wolever et al., 2013). This is in line with general definitions of coaching, beyond health, which also consider coaching as an individualized and tailor-made approach, and based on a collaborative relationship rather than one based on authority (Ives, 2008).

Coaching is applied in many application areas related to health, varying from sports and wellbeing to clinical contexts. In sports, recreational and professional athletes are often supported by coaches (e.g., team coaches, personal trainers) to achieve their best sports performance (Cassidy et al., 2009). The term health coaching, on the other hand, is typically used in more clinically oriented contexts, including lifestyle and behavior change support for people with chronic diseases, obesity or hypertension (Olsen & Nesbitt, 2010; Sforzo et al., 2018). Wellbeing or wellness coaching, while typically taken together with health coaching (c.f., Sforzo et al., 2018; Wolever et al., 2013), often applies to healthy clients. These clients have no medical condition nor a specific sports goal, yet wish to increase their wellbeing or prevent illness. Sports, health and wellbeing coaching share a similar focus, including lifestyle, nutrition and physical activity. In the present paper, we focus on clients that are essentially healthy, yet, who have health-related goals. We exclude medical conditions such as chronic diseases, as that might narrow down the use of data towards those conditions. We also focus on coaching situations where it is plausible that a client brings her own data to a coach, tracked by a consumer device. Therefore, we exclude elite sports from our scope, as in these contexts, using data is already a common practice, and there are typically more advanced measuring devices available.

The effectiveness of health coaching is studied extensively, and shows to have mixed to positive results on improving health outcomes (Olsen & Nesbitt, 2010; Sforzo et al., 2018). The nature of health coaching itself, however, is less understood. What do coaches typically do, say and recommend? Which techniques do they apply? How long and frequently do coaches and clients meet, and what is the most effective format? What are typical dynamics of a coach-client conversation? Both the reviews of Olsen & Nesbitt (2010) and Wolever and her colleagues (2013) show that health coaching literature often lacks detailed reporting on these aspects, inhibiting systemic evaluation or elicitation of best practices. Nevertheless, literature provides some suggestions. For example, Olsen & Nesbitt's (2010) findings indicate that goal setting and motivational interviewing (MI) typically results in positive health outcomes. Both methods enhance self-awareness, accountability and confidence. Furthermore, we know from the clinical domain that effective doctor-patient communication is key to patient satisfaction and positive health outcomes (Ha et al., 2010). It has been argued that clinical conversations should go beyond biomedical topics, including the patient's narrative (Murphy & Franz, 2016) and the patient's values (Berry et al., 2017), to provide appropriate care. A good doctor-patient relationship is characterized by emotional connection and partnership (Dill & Gumpert, 2012), and this is likely to be the same in health coaching.

THE PROMISE OF DATA FOR HEALTH COACHING

For clients it can be beneficial to track one's health data, to gain insight in patterns and relations, possibly supplemented with e-coaching that contains personalized recommendations or motivational messages. It is generally recognized that clients' self-tracked health behaviors are also promising to serve as input for health coaches and healthcare professionals in general. First and foremost, self-tracked data potentially provide coaches with a more objective and reliable view on the client's behavior, compared to the more traditional information source, a client's self-report (Chung et al., 2015; Schroeder et al., 2017). These devices measure continuously, with high frequency, and are situated in the daily life of the client, thus allowing the collection of detailed information of trends in health (Sqalli & Al-Thani, 2020). This may lead to new or deeper insights about the client, and can facilitate personalized care (Figueiredo & Chen, 2020), tailored to a client's specific needs and experiences (see *Chapter 2*, and Sqalli & Al-Thani, 2020). Combining data of a large group of users allows for novel detection of health issues, which in turn can improve the algorithms in health coaching programs, (c.f., Turakhia et al., 2019). So, besides the frequently mentioned benefits self-tracking has for clients, (c.f., Epstein et al., 2020), there are substantial benefits for health coaches and their practices too.

Data may also serve as memory aid for clients (Figueiredo & Chen, 2020). It is notoriously hard to accurately recall day-to-day behaviors and experiences from memory (c.f., Kahneman & Riis, 2012). When clients report how they have been over the last days or weeks, they draw from their memory, which most likely gives coaches a biased representation. While it may be insightful for coaches to hear how clients remember certain events, to provide accurate support, coaches most likely also need to understand actual behavior and experience, including a client's actual food intake, training sessions, if they have been struggling, and how intense that felt in the moment. Bringing data to the coaching session may improve a client's memory retrieval, similar to Kahneman et al. (2004)'s 'day reconstruction method', where the reconstruction of a day through episodes has been proven to enhance reliable self-report.

These benefits, while important, only consider the individual gains for coaches and clients separately. Basically, coaches have more information, and clients are more engaged. Yet, when considering the coach-client relationship and interactions, new benefits emerge. It has been argued that data enhance effective communication between coach and client. Mentis and her colleagues (2017) observed clinical visits were patients' step-count data were discussed. They found that this process of co-interpretation, for example making sense of outliers and trends in a conversation, supports the re-construction of the patient's narrative. It shows how data serve as an opportunity for clients to share their lived experiences. *Chapter 2* shows that there potentially is a synergy between collaboratively reflecting on behavioral data and sharing lived experiences. It argues that particularly the ambiguity of behavioral data, i.e., a high step count can reflect intentionally healthy behavior or a broken car, provides relevant cues for meaningful coaching conversations. Figueiredo and her colleagues (2020) interviewed both healthcare providers and patients on their use of data in managing fertility issues and reveal that their data practices are essentially different. For patients, these are mainly driven by emotions, whereas for providers, this is a mainly rational process. The authors argue that in order to effectively utilize data, these different perspectives should be bridged, as both serve different purposes and add unique value. This is in line with several other studies (e.g., Chung et al., 2019; Pichon et al., 2020; Raj, Newman, Lee, & Ackerman, 2017) showing that both clients and healthcare professionals bring in their own expertise, resulting in complementary views on the data. Specifically, clients draw from their own lived experiences when reflecting on data, whereas healthcare professionals mostly rely on their medical expertise. In addition, Chung and her colleagues (2019) found that pre-visit notes by clients, based on their food diary data, guided explicit discussion on participants' goals, and thus increased alignment. To conclude, literature suggests that in addition to the data itself, it is the collaborative reflection on the data that adds value, and these collaborative reflections facilitate the alignment of goals, expectations, and perceptions on illness experiences.

BARRIERS TO EFFECTIVE USE OF DATA IN HEALTH COACHING

Despite these potential benefits of using data in health coaching, and the growing number of people that engage in self-tracking, the adoption of data in professional contexts is low. Demiris and his colleagues (2019) argue that the adoption of self-tracking tools in clinical practice are still in an 'early adopter' stage. Literature points to several barriers that could explain this slow uptake, both from a technical perspective, as well as from the healthcare professionals' point of view.

From a technical perspective, several challenges are identified that may inhibit leveraging these benefits. For example, measurements may be inaccurate (Mahajan et al., 2020; Sqalli & Al-Thani, 2020; West et al., 2017), and tracking devices and their underlying algorithms operate without expert guidance (Mahajan et al., 2020). Furthermore, these data typically comprise clients' health indicators, but lack contextual information that is needed to effectively serve as input for personalized health coaching programs (Sqalli & Al-Thani, 2020).

Health professionals have also expressed a range of perceived barriers withholding them to use data in their practice. This includes time constraints (Chung et al., 2015; Devaraj et al., 2014; Gagnon et al., 2016; West et al., 2018), privacy and security concerns (Gagnon et al., 2016; Watt et al., 2019), patients having unrealistic expectations about health professionals reviewing their data (Chung et al., 2015), patients misreading or over-monitoring their data which reinforces worries (Watt et al., 2019), lack of expertise to analyze the data (Chung et al., 2015), lack of familiarity with the technology (Gagnon et al., 2016), and data being incomplete, unreliable or irrelevant (West et al., 2018).

Besides these most practically oriented barriers, it becomes particularly clear that health professionals want to secure a good relationship with the client when introducing data. They want to avoid data disrupting the contact with their client (see *Chapter 2*, and Gagnon et al., 2016). For example, they fear that looking at a screen is misinterpreted as indifference for the client (Gagnon et al., 2016), and they want to prevent an overemphasis on numerical information distracting from the client's subjective experience (see *Chapter 2*). This resonates with other researchers' critical perspective on reducing health to numbers (Van Dijk et al., 2015; Lupton, 2016). When collaborating with hybrid eHealth technology, health professionals stress the need to establish and maintain an empathic relationship with their client (Brandt et al., 2018). And, interview studies with healthcare professionals and patients suggest that if collaborative reflection on data is not effectively supported, this may reinforce misunderstandings and unaligned expectations (Chung et al., 2016; Figueiredo et al., 2020; Pichon et al., 2020; Raj et al., 2017).

In sum, a large share of the expected benefits and barriers of data in coaching typically go beyond coaches and clients individually; they are situated within the coach-client relationship. To increase our understanding of the potential effect of data on relational aspects, we discuss this through the lens of the theories of distributed cognition (Hutchins, 2000) and communication theory (Watzlawick et al., 1967).

THROUGH THE LENS OF COGNITIVE AND COMMUNICATION THEORIES

We should note that the required knowledge for effective health coaching is distributed across the coach, the client and possibly the data from a tracker. This implies that for gaining a complete understanding that is needed to effectively coach, this distributed knowledge should be shared and coordinated. This process is well described in the distributed cognition paradigm (Hollan et al., 2000). Drawing from their observations in aviation (c.f., Hutchins & Klausen, 1996), Hollan and his colleagues (2000) argue that cognition needs a larger unit of analysis than just one individual; cognition is distributed across people and technological artefacts. They show how information is transmitted and transformed in such a sociotechnical system, and they argue that cognition is shaped by cultural expectations and social organization. Coaches and clients, like pilots in a cockpit, have expectations of each other in terms of what the other knows and how they are supposed to act based on the information at hand. In this light, coach-client communication can be seen as sharing knowledge representations and interpretations of the data on the one hand, and the client's situation and the expected outcome of coaching interventions on the other. This process includes contextualization of data, making predictions, and checking assumptions.

Furthermore, we have seen that in coach-client communication not only the subject matter of what is being discussed is important. It is key to also consider effective communication in terms of *how* things are discussed, and to situate what is being said within the relationship between the coach and client. This insight resonates with the communication theory of (Watzlawick et al., 1967), arguing that information transmission is always contextualized within a relationship between the sender and the receiver. Every instance of communication can be understood on a content level, i.e., the information that a message contains, as well as on a relationship level. The relationship level of communication comprises, among other things, the sender's expectations on how the message should be understood and what the recipient is expected to do with the information. It basically reveals how the communicators view one another. We have seen that indeed effective health coaching is strongly determined by the quality of the coach-client relation, in terms of mutual trust, respect, and investment. Thus, when a client brings her data to a coach, it might not only be the data per se that informs the coach, it may also be the act of initiating the tracking, the fact that she is willing to share, and the way she talks about the data, that is informative to the coach. This may signal needs and levels of motivation, dedication, or self-confidence, as well as need for acknowledgement, expectations of the coaching, or trust in the coach.

CONTRIBUTION AND RESEARCH QUESTIONS

Prior literature gives insight in how self-tracking data potentially influence the health coaching process. The current study aims to expand current understandings in two important ways. First, we contribute by exploring the value of data across various conditions. That is, we add data both in the beginning and halfway through the coaching sessions, and we let coaches assess the data both in presence and in absence of the client. This setup allows for understanding the value of data and a conversation, individually and collectively, and comparing those in both qualitative and quantitative ways. It also allows for contrasting coaching sessions that started with data, or started with a conversation, and compare the results when either one is taken as point of departure. This approach adds to prior work that typically draws insights from sessions where data were available from the start (e.g., Mentis et al., 2017; Raj et al., 2017), resembling our 'end-situation' where data and a client conversation come together. In some other studies (e.g., Figueiredo et al., 2020; Pichon et al., 2020), clients and healthcare professionals are only interviewed individually on their needs and experiences. There have also been studies that observed patients' and healthcare professionals' individual and collaborative interactions with data (e.g, Chung et al., 2019; Schroeder et al., 2017). Yet, we add to this by an explicit comparison of the coaching sessions with data-only, conversation-only, and data and client conversation together.

We want to particularly highlight the value of the condition where coaches assess the client's data in absence of the client, as we expect this may yield interesting results. In this condition, we ask the coach to formulate an advice based on the data solely, not knowing the client other than reading her goal or question. Essentially, this mimics an e-coaching situation through a Wizard of Oz like approach, where a coach, be it a human or an artificial one, generates advice merely based on a client's data and goals. Following this up with a conversation with the client, after which the coaching advise is updated, gives insight in not only the value of data, but also its possible limitations and information gaps. It allows to identify the additional information that a client conversation yields, and it sheds light on the feasibility of e-coaching and specific design considerations for such applications.

Second, we contribute to prior literature with a focus on essentially healthy clients who wish to improve their wellbeing and fitness or prevent illness. Prior work on the effects of data on health coaches has mostly focused on medical contexts, for example working with chronically ill people with Irritable Bowel Syndrome (IBS; Chung et al., 2019; Schroeder et al., 2017), diabetes (Raj et al., 2017) or Parkinson's disease (Mentis et al., 2017). The study of Chung and her colleagues (2019) used both healthy participants as well as chronically ill participants (IBS patients), and they found that the use of data differed across these cases. For the IBS patients, the focus was mostly on identifying and managing symptom triggers, whereas for healthy participants there was more time spend on discussing potential goals and possibilities (Chung et al., 2019). In the current study, we particularly focus on how data affect the health coaching process for healthy clients, where goals are more open-ended compared to health coaching in medical contexts.

We have described how data are likely to have effect beyond merely bringing in additional information, as they also influence relational aspects of coaching. Therefore, we aim to address the effect of data in terms of content and relation separately. Furthermore, while we acknowledge the client's perspective, we will have a main focus on the coach's perspective in our analysis. We seek to broaden the understanding of collaborative use of data, and we believe that this benefits in the first place from exploring how data challenge the roles and working practices of health coaches and how data meets their information needs. Of course, we will also recognize clients' perspectives and needs, and study how data affects the collaborative process as a whole. Accordingly, we aim to answer the main research question

How do data change the health coaching process?

by answering the following sub-questions:

a. How do coaches and clients relate to the data, i.e., how do they interpret it and utilize it in a coaching session?
b. Can a client's data already be informative to coaches in the absence of explanation or contextualization from the client?
c. How do data change the coaching at the level of coaching content, i.e., topics that are discussed, insights that are gained, and advice that is given?
d. How do data change the coaching at the relationship level, i.e., the roles of the coach and client in the coaching process, and their relation?

METHODS

In a workshop and a field study, we let coaches interact with clients and their data in various ways. That is, some started with a conversation with the client (*conversa-*

tion-first condition), after which the client's data were introduced and discussed. Other coaches started with assessing the client's data (*data-first* condition) in absence of the client, after which the client came in and a conversation started. We both worked with coach-client pairs familiar to each other, to understand the effect of adding data into an existing coaching relationship, as well as coach-client pairs unfamiliar to each other. The latter was representative for a coaching 'intake' situation and allowed us to study the value of data in isolation, with no background information of the client. Altogether, this resulted in a broad range of setups, covering several phases of the coaching (intake sessions or further progressed), enabling us to compare and contrast across more data-driven and more conversation-driven sessions. In this section, we will describe our approach in more detail.

This study was approved by the local ethics committee at Eindhoven University of Technology, Human-Technology Interaction group.

PARTICIPANTS AND STUDY PROCEDURE

Workshop

The workshop was organized as one of the parallel sessions at a symposium organized by the sports coaching academy at an applied university, organized for teachers and practitioners in sports coaching. In total, twelve coaches voluntarily participated, with various backgrounds and professions, such as teachers at the sports coaching academy, sports-related community workers, and physiotherapists. There were two workshop rounds; four coaches participated in the first round, and eight coaches participated in the second round. The coaches were split up in small groups of two to three participants and randomly assigned to one client (see Figure 3). The duration of each workshop round was 45 minutes and was facilitated by two researchers.

We recruited three clients from our personal network who had a health-related issue or question, although not indicating severe health issues. We invited them to join the sessions as if they were clients visiting a coach and asked them to bring any relevant self-tracked data of any type. All clients wrote down their question as input for the sessions. See Table 4 for an overview of their question and the data they brought. Note that all coaches were unfamiliar to the clients, as we brought in 'stand-in clients' ourselves. All participants in the workshop, including coaches and clients, participated on a voluntary basis.

Field Study

For the field study, we recruited five coaches from our personal network, of which three personal trainers in a university sports center (A, L and I), and two dieticians (M and K). Their experience as a coach ranged from 4 to 20 years (*median* = 5), and their average age was 31 years (*SD* = 5.7). Aiming for a realistic setting, we asked the coaches to join the study with one of their own clients that they were currently coaching. This guaranteed that we included coach-client pairs where tracking data

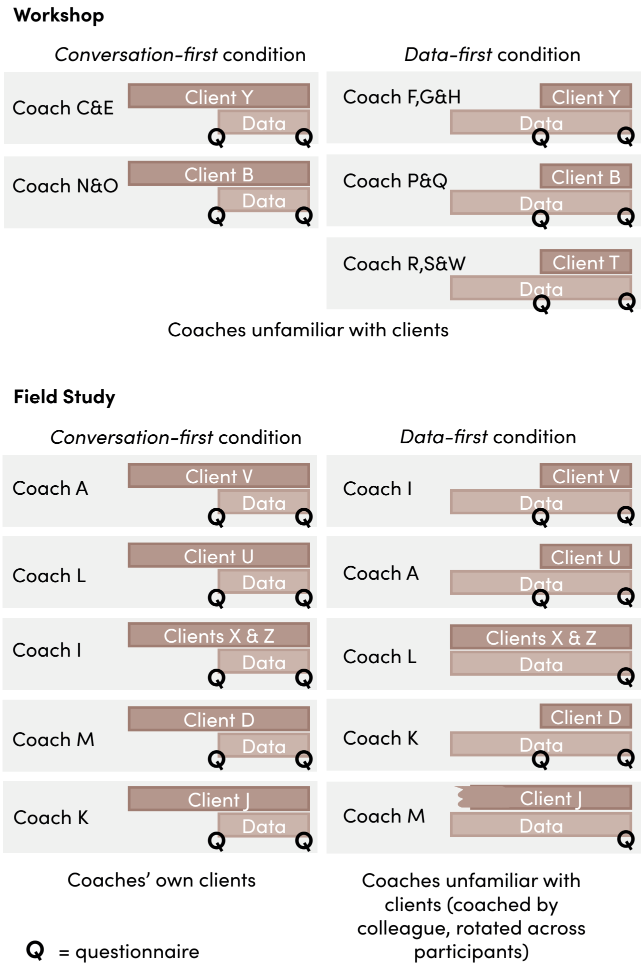

Figure 3 Overview of coaching sessions in the workshop and field study, split by conversation-first and data-first condition.

was relevant, at least from the coach's perspective. The resulting coach-client pairs had been working together for minimum a month up to a year, with a frequency varying from once a week to once a month. None of the pairs had been using self-tracking data before in their coaching sessions, although one of the coaches did recently start exploring the use of self-trackers with other clients. We provided the clients with a health watch to track their behavior, i.e., the Samsung Gear Fit 2 Pro, which they used for approximately two weeks. Again, aiming for a realistic situation, we let clients themselves decide how they would use the tracker. It would automatically track step count, floors, calories burned and physical activity level, but they were free to take the watch off at any time. Additionally, clients could decide to switch on the heart rate measurement, track sleep (i.e., keep the watch on at night), and manually track specific sports training sessions, nutrition, water intake and coffee intake. By letting the coaches choose the client, and letting the clients choose how to use the tracker, we aimed to create situations where tracking meets realistic needs.

Data may not only be added during existing coaching processes, it may also be brought by clients at the start. To mimic this situation, where data is available during an intake session where coaches and clients are yet unfamiliar with each other, we rotated the participating clients across the participating coaches, see Figure 3. We asked the clients to write down their goals, for the sessions with the unfamiliar coaches. The clients' coaching goals are listed in Table 4. The sessions lasted approximately 30 minutes and were guided by one or two researchers. We compensated the coaches with a €15 voucher, and the clients with a €5 voucher.

CONDITIONS

Conversation-first Condition

In this condition, we sought to observe how data would affect an ongoing coaching process. We instructed the coaches and clients to first have a regular coaching conversation, serving as a baseline. In the workshop, this would resemble an intake

	Client	Own Coach	Coaching goal or question to coach	Data Source
Workshop	Y	*n.a.*	I have really low energy after lunch and dinner. How can I overcome that?	iPhone Health app
	B	*n.a.*	I lost substantial weight over the last year, and now I want to maintain my current weight, while still building some strength. What is be a suitable food intake for me?	MyFitnessPal (food intake), Fitbit (physical activity, weight)
	T	*n.a.*	In periods when I do less sports, I lose weight relatively fast. What can I do to avoid this?	Google Fit app
Field Study	V	A	Fix knee problems to be able to play basketball again. Lose some weight.	Samsung Gear Fit 2 Pro (provided by the study)
	X & Z (couple)	I	X: Improve core / body condition / muscle strength, especially after suffering from a discal hernia (lower back). Z: Lose weight, tips to get fitter.	
	U	L	Lose weight and get toned to look good in wedding dress. Tips to control hunger pangs.	
	D	M	Lose weight, have a healthy BMI.	
	J	K	Lose weight and live without medicine for high cholesterol and diabetes. Tips for a healthy lifestyle.	

Table 4 Overview of participating clients' goals, questions and data sources.

situation, as coaches and clients were yet unfamiliar to each other. Those clients brought their written question or coaching goals as input for the session. In the field study, the coaches and clients had been working together for a while. They typically talked about how they had been since the last session, sometimes while stepping on a scale to measure weight in the meantime, or while walking on a treadmill. Coaches were asked to indicate when they had sufficient information to provide an advice, after which they filled in the first questionnaire, including their current advice for the client, and an evaluation of the information gained from the client's self-report. Subsequently, the client's data were introduced, and they continued their conversation, now supplemented with data. At the end of the session, the coaches filled in a questionnaire again, asking for any updates in their advice, and an evaluation of the information gained from the data.

Data-first Condition

In the data-first condition, the coaches were presented with the client's data at the start, in absence of the client. This enabled us to study the value of mere data, lacking contextualization in a conversation with the client, which essentially represents pure e-coaching. It also provided a baseline, to which we could compare the added value of a conversation. In these sessions, the coaches were assessing the data through the client's devices, i.e., their phone and/or watch interface, and during the workshop sometimes also through web interfaces. The client's question or coaching goal was always shared on paper, in some cases elucidated by clients before they left the room. The coaches were prompted to think aloud as much as possible, and when they had questions, the researcher helped them to find their way through the data. As soon as coaches reported they gained sufficient information from the data to give an advice, they filled in a questionnaire, including their advice to the client and an evaluation the information gained from the data. After, the client was asked to join the coach, and they were instructed to have a coaching session like they would normally have, though informed by the data. They could ask or discuss anything, with the aim of helping the client on her goal or question. At the end of the session, again they filled in a questionnaire, asking for any updates in their advice, and an evaluation of the information gained from the client's self-report.

We should note that condition (*data-first*, *conversation-first*) and coach-client familiarity are confounded, that is, the *data-first* condition is only applied to unfamiliar coach-client pairs. To understand the mere value of data to coaches, it was not feasible to apply the *data-first* condition to existing coach-client pairs, as coaches' interpretations would inevitably be colored by their background knowledge of the client. This did not limit our findings; our conditions did not serve the purpose of a balanced interventional experiment. Rather, we used our conditions to create a variety of realistic situations that allow us to understand the influence of data on health coaching in a broad sense.

As the study progressed, we found that the assessment of data in absence of the client was experienced as very uncomfortable by both coaches and clients. To avoid unnecessary tension, we loosened some constraints in the execution of the protocol, depending on people's responses in the moment. As a result, we allowed clients' presence, or even help, in some data-first sessions. As the goal of the conditions was to introduce variance in our data, rather than making a strict comparison across conditions, we could permit these deviations. More specifically, coach L was assessing the data of clients X & Z in their presence and with their help, client D was not sent out of the room when coach K was assessing here data (which limited the think-aloud, obviously), and client J was enthusiastically explaining the data to coach M from the beginning, which we deliberately let happen (see Figure 3).

All clients, except client T, participated in both the *conversation-first* as well as the *data-first* condition, with different coaches. The coaches in the field study also participated in both conditions, with different clients. Four out of five coaches had the *conversation-first* session before the *data-first* session. The coaches in the workshop were assigned to one condition, however, in the last workshop round, we ended with a short (5-minute) group discussion to reflect on the differences across the conditions.

MEASUREMENTS AND DATA ANALYSIS

We analyzed the sessions from the workshop and the field study together. All sessions were audio recorded and transcribed verbatim, allowing for a detailed analysis of the coaching sessions, including the dynamics of the coach-client conversation, their reflections on the data, and the coaches' questions and advice to the client. The transcripts of the sessions were analyzed through thematic analysis (Boyatzis, 1998), in the software package MaxQDA. In this process, we used a mostly inductive approach, comparing and contrasting across the four subsamples of coaching sessions, being data-only, conversation-only, data-after-conversation and conversation-after-data. This supported our goal of understanding the value of data and the value of a client conversation, individually and collectively. We expected that both would have their unique contributions to the coaches and the coaching process. Emerging themes that differentiated these subsamples were iteratively and systemically tested against the corpus of transcripts. Intermediate versions of themes were frequently shared and discussed with the research team to check their validity and relevance. One researcher coded all data, and when the final thematic codes were set, another researcher coded 20% of the data, resulting in an Inter-Rater Reliability (IRR) of 82%. All disagreements were resolved by discussion. Most disagreements were resulting from the different but consecutive codes 'understanding behavior' and 'understanding experience'. One may argue that these codes could be merged because they are very similar. When doing so, the IRR increased to 88%.

The coaches and clients filled in two questionnaires, enabling us again to systematically compare and contrast the value of data, conversation, and their combination.

Are you going to analyze me now?

Client

One halfway, i.e., right before the data or client was introduced, and one at the end of the session. The full questionnaires are provided in Appendix A. For our results, we only analyzed the coach-questionnaire, as the coach's perspective was most relevant. For the analysis we used multilevel models, with the coach as grouping variable (four measurements per coach in the field study, two measurement per coach in the workshop). The type of information (data or conversation) and timing of the measurement (halfway or end) were used as predictors. Our small sample, despite the repeated measurements, limited drawing conclusions on the questionnaire solely. Yet, we will interpret the results in combination with the qualitative findings resulting from the thematic analysis.

RESULTS

Coaches and clients were generally open and cooperative during the sessions. The sessions with familiar coach-client pairs showed seemingly natural coaching conversations, and in the sessions with unfamiliar coach-client pairs, they seemed motivated to get to know each other. The researchers were generally accepted as observers, though, often at some point the participants seemed to expect a specific task on what to do with the data, and we reiterated that we expected them to use it freely according to their own needs and interests. This was typically followed by lively discussions, driven by data to a more or lesser extent, which we will reflect on in more detail in the sections below.

The study setting seemed to be natural to the coaches and clients, with the notable exception of the first phase of the data-first condition, where the coach assessed the data in absence of the client. For clients it was awkward to give away their phone – a highly intimate an personal device (illustrated by a client stating with a mixture of being funny and being nervous: *"if there are messages coming in, don't answer them!"*,), and coaches felt put on the spot to assess data and come up with an advice without input from the client (*"this is completely against my principles!"*). The moment when the client entered the room again was often accompanied with ice-breaking statements like *"when will I die?"* or *"are you going to analyze me now?"*. Therefore, as the study progressed, we decided to lose this constraint, and allowed clients to be present or even help while coaches were assessing their data.

Figure 4 provides an overview of the results, both from the thematic analysis and the questionnaires. In the sections below, we will describe the incremental coaching activities in a session and the role and value of data therein, followed by a description on how data are affecting health coaching on content as well as on relationship level.

THE INCREMENTAL ACTIVITIES IN A COACHING SESSION

To understand the context wherein we introduced the data, we first provide a general description of the coaching sessions, in terms of the dynamics and activities.

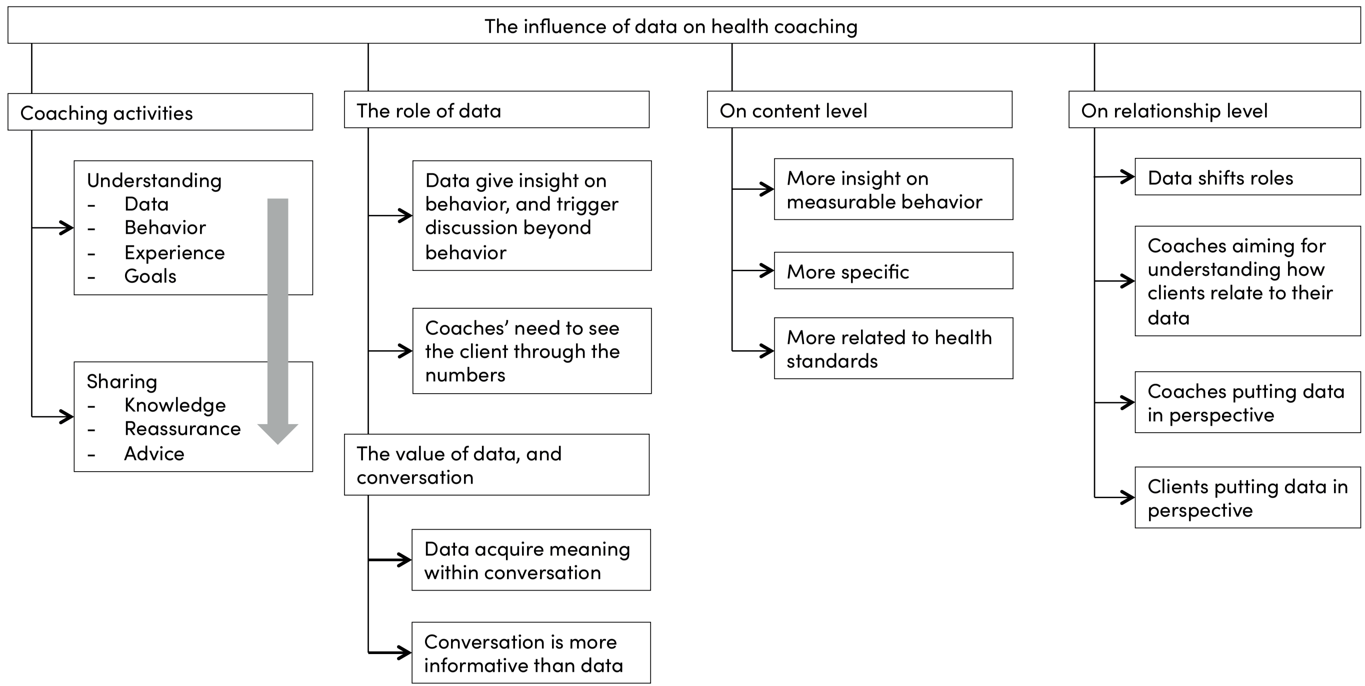

Figure 4 Overview of results in four main themes: (1) incremental coaching activities and (2) the role of value and data therein, (3) how data affects health coaching on content level and (4) on relationship level.

	Coaching activity	Example quotes
Understanding	Data itself and context wherein data were tracked	- For how long did you track? - Did you not move here, or were you just not wearing your watch? - Are these activities (walking, cycling) automatically tracked, or did you manually switch it on? - What does this blue line represent?
	Behavior (what-questions)	- How many times a week do you work out? - What kind of sports do you do? - What did you have for lunch?
	Experience (how- and why-questions) and daily life	- It was on the last week when your steps dropped; can I know why? - Do you wake up fresh? - Do you like to play tennis? - What kind of work do you do?
	Goals and current status	- Are you satisfied with you weight now? - I see that your activity levels are already quite good. Do you have a certain goal with that? - Did you have enough energy this week to do what you wanted to do?
Sharing	Knowledge, expertise	- Your lack of energy can be caused by so many factors, it may be your sugar intake, stress, or screen time. The impact can be different for everyone, so we need to explore what works for you. - If you make soup yourself, you could try to make it low in salt by using […]. - Let me explain you how it works with sleep cycles. - This heartrate is normal when you do an intense training.
	Reassurance and compliments	- I know it's hard, but you did it before, so I'm sure you can do it again. - Very good, the average is 6000 steps a day, well done! - Don't be too hard on yourself if you did not reach your goal for a day, look at what you've already achieved!
	Advice	- Add some higher intensity activities. - It's always good to work out together, other people can motivate you. - Try to walk a bit more. For example, at work, use your break to walk around the company. - If you see in your food tracker at the end of the day that you have some room left in the calories, first check whether you've had all the required nutrients, and avoid eating 'empty calories' like a cookie.

Table 5 Overview of coaching activities, including example quotes.

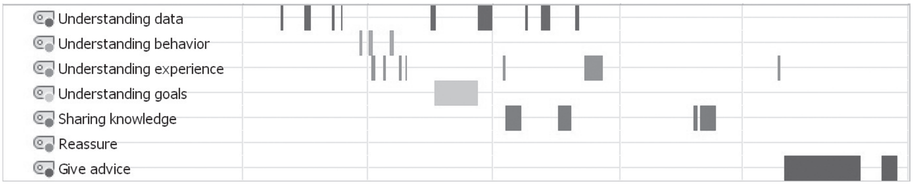

Figure 5 Typical blueprint of a coaching session, illustrating the incremental pattern from data-oriented to client-oriented, from understanding to giving advice. All codes referring to coaching activities are highlighted; x-axis represents time in the session. (Coaching session with Coach L and Clients X and Z, both data and clients present.)

From Understanding to Sharing

Throughout all coaching sessions, we identified two main type of activities. First, there were activities targeted at understanding. Here, the coach mainly asked questions and listened, trying to build up an image of the client's data (if available), her recent behaviors and experiences, current status and goals. As soon as there was a sufficient understanding on these aspects, the coach moved toward activities revolving around sharing. Here, the coach shared her knowledge and expertise, reassured the client and gave her compliments, and gave specific advice. An overview of these activities, including example quotes, can be found in Table 5. Typically, during 'sharing-activities', the coaches took a more leading role in the conversation compared to during 'understanding-activities'. Still, it also happened that they took a more facilitating role, trying to let clients themselves come up with actionable insights.

From Data to Client

The coaching activities showed to be typically incremental, which is illustrated by the code-line visualization of the session of Coach L and Clients X and Y in Figure 5. Activities aiming at understanding occurred mainly at the beginning of a session, and sharing mainly at the end, while it also happened that coaches were switching back and forth when new knowledge gaps emerged. Furthermore, we found incremental levels of understanding, gradually moving from data- or behavior-oriented to client-oriented (see also Table 5). A typical sequence started with coaches seeking after understanding the data itself (e.g., this number is your step count?), followed by understanding the clients' behavior (e.g., how often do you walk?) and then soliciting their experience (e.g., do you like to walk?). Finally, they aimed at understanding this information considering their current status and goals. E.g., does the particular behavior or experience disclose the client's struggles and challenges? Or, can it potentially contribute to the client's goals and wellbeing? Only after the coaches had a sufficient understanding on the how the clients were doing in light of their goals and challenges, they were ready to move to sharing activities, ultimately giving advice.

THE ROLE AND VALUE OF DATA

It is useful to consider the role of data within the dynamics of a coaching session. In the following sections, we will situate the role of data within the incremental activities of a coaching session, as well as reflect on the value of data, conversation and their combination.

Data are Considered beyond the Behavior they Represent

Most notably, data mostly led to insights on behavioral level. For example, coach L already concluded in the third minute of the session *"So basically, during the week, you're biking about 45 minutes to the office. And then in the weekends, you're really walking a lot more."* But such a straightforward understanding of the tracked behavior never appeared to be enough for the coaches. They were clearly seeking after clients' reasons for their behavior. For example, when Coach P noticed the client had an unhealthy snack, she solicited for her reasons: *"was that a moment of weakness, or perhaps you did not have a healthy alternative at that moment?"* after which the client reported *"I'm just exploring how to find a sustainable diet, I think a snack every now and then should be acceptable within a normal healthy lifestyle"*. This illustrates that the reasons for certain behavior were essential for coaches to accurately interpret it. Interestingly, coaches sometimes even started discussing seemingly straight-forward behavior to increase their understanding of the client. For example, while Coach G already knew the answer from the data, she still asked *"how often do you use the stairs instead of the elevator? And is that only in the morning, when you still have the energy, or do you do it in the afternoon too?"*. Asking the client these questions provided the coach with extra information on *how* the client answered them, and the client's answer added the context of a specific colleague always motivating her to take the stairs.

Thus, while data mostly added information in terms of the client's behavior, it also prompted conversations on the experience of those behaviors, the context wherein the behavior was performed, the triggers that motivated the client to execute the behavior, and their personal value judgement on the behavior. In this sense, the data provided input and tools for the coaches on all levels of understanding, from low-level behavior to higher-level lived experiences and goals. Yet, data were rarely self-explanatory. Higher level insights on the client were only gained through effective communication, where interpretations were shared, and data were collaboratively reflected on.

Coaches' Need to See the Client through the Numbers

Coaches showed to be keen on moving their focus from the data to the client as soon as possible. They showed little interest in the numbers per se; for example, they rarely engaged in efforts to analyze the data themselves. Rather, the coaches quickly shifted to what the data meant in terms of the client's narrative, by soliciting the client's

Was that a moment of weakness, or perhaps you did not have a healthy alternative at that moment?

Coach

experiences associated with these data. They easily disregarded data when there was no clear connection with the client's goals and experiences relevant for coaching. For example, in one coaching session the client's goal of losing weight clearly had a highly emotional connotation; her unhealthy food intake and low level of physical activity was due to feelings of depression and low self-esteem. These feelings may have been triggered by the fact that she recently moved abroad for her studies and did not yet feel at home in her new place. This made the step count data not only irrelevant but also very inappropriate to discuss. In this specific session, the data were barely discussed, other than an abstract discussion on how self-tracking at some point could be helpful as a motivation or to get more insight. Concluding, we observed coaches' urge to move away from the numbers to the client as soon as they reasonably could. In this light, the data did slow down the coaching process in some cases, being something that needed to be clarified before they were able to focus on the client in ways they considered more meaningful.

The Individual and Combined Value of Data and Conversation

Next to our qualitative results from the thematic analysis, the questionnaire allowed us to systematically contrast and compare the value of data versus conversation. In this questionnaire, which coaches filled in halfway and at the end of the sessions, coaches were asked to write down their advice (intermediate or final), after which we measured their perceived usability, objectivity, clarity and relevance of the information source at hand (i.e., data or conversation), on a 5-point Likert scale. Furthermore, we measured their perceived understanding of the clients' personal experience and daily life, and to what extent they felt the information (i.e. data or conversation) supported their effectiveness as coach, all on a 5-point Likert scale. Lastly, we asked them to reflect on the value of the data or conversation in an open question. This resulted in multiple measurements per coach (four measurements per coach in the field study, two measurements per coach in the workshop), thus, for the analysis we used multilevel models with the coach as grouping variable. As independent variables we used information source (i.e., data or conversation) and order (i.e., beginning or halfway through).

To be able to compare across the different information sources, the questionnaire was kept as constant as possible. Thus, all questions were applicable to evaluate both data and conversation. Note that, depending on the condition, the value of data was measured halfway (*data-first* condition) or at the end (*conversation-first* condition) of the session, and vice versa for the conversation. This allowed us to also test possible interactions, i.e., were data evaluated differently at the end, when coaches assessed the data within a conversation with the client, compared to halfway, when only assessed in absence of the client? In the results, we will focus on the main effect of information source (i.e., data or conversation) and possible interaction effects of information source and order, as those are most relevant with respect to our research questions.

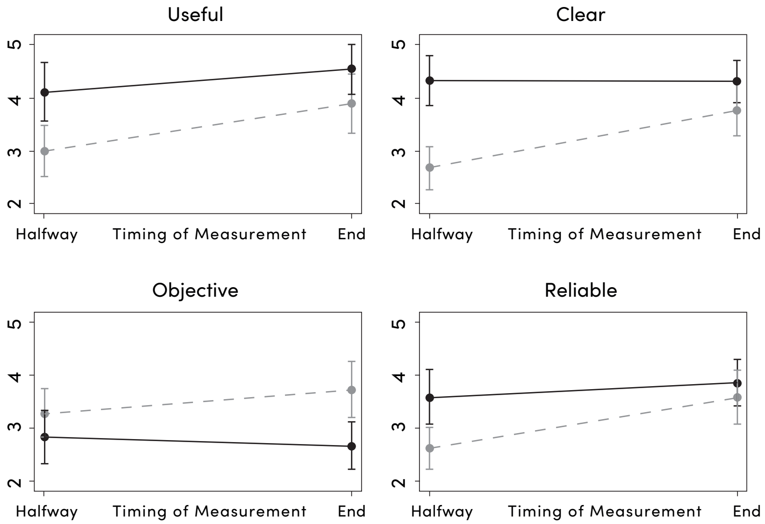

Figure 6 Results of multilevel models, showing coaches' evaluations of data (dashed line) and conversation (solid line), halfway and at the end of the coaching sessions.

Due to the small sample, while somewhat mitigated by the repeated measurements per coach, it is not possible to draw strong conclusions from these analyses solely. At the same time, the effects we found are consistent over the multiple items in the questionnaire and are consistent with our qualitative findings. Therefore, we will nevertheless present the results, with the remark to interpret these results with caution and only in combination with the qualitative results.

Data Acquire Value When Situated Within a Conversation

We first inspect the main effect of conversation versus data for each of the dependent measures. As Figure 6 shows, coaches value conversation as more useful ($p=0.001$), clearer ($p<0.001$), and more reliable ($p=0.014$) than data, whereas data were valued as more objective ($p<0.001$). Furthermore, there was one significant interaction effect ($p=0.015$); data were valued as significantly clearer at the end of the study, when assessed within a conversation with the client, compared to halfway, when assessed in absence of the client. Thus, data apparently became clearer when contextualized in a conversation. Similar interaction effects with the other outcome variables were not significant.

Furthermore, we measured the extent to which the coaches felt they had a complete picture of the client's personal experience and their behavior in daily life, after assessing the data, the conversation with the client, or both. It shows that a conversation provides coaches with good insight (M=mean=4.11) on the client's personal experience, and so does the combination of data and conversation (M=4.00 for data-first condition and M=4.11 for conversation-first condition). The data in isolation, however, are providing significant less insight (M=1.69) on the client's personal experiences (interaction between order and information source is significant with $p<0.001$). The same holds for insight in the client's behavior in daily life. A conversation provides coaches with good insight (M=3.89) on the client's daily life, and so does the combination of data and conversation (M=3.69 for data-first and M=3.78 for conversation-first). Similar to the client's personal experience, also for the client's daily life, data in isolation provides significantly less insight (M=2.08, interaction significant with $p<0.001$). Thus, these results suggest that data only provide good insight in a client's personal experiences and her behavior in daily life when used within a conversation. Having a conversation only, without any data, results in an equal level of understanding compared to having both a conversation and data. This suggests that data do not substantially contribute to the coaches' picture of their client's experiences and daily life. This is generally in line with our qualitative findings presented in the other results sections.

Conversation is More Often Indicated as Sufficient and Supportive, Compared to Data

Halfway through the session, when the coaches had faced only one source of information (i.e., data or conversation), we measured the extent to which the coaches felt they had sufficient information to give appropriate advice. Coaches scored significantly higher on having enough information after solely a conversation (M=3.56) compared to after solely data (M=1.69, two-sided t-test[7], p=0.001). At the end of the sessions, we measured the extent to which coaches felt they had more information than in the first half of the session. Both a conversation (M= 4.54) and data (M=3.78) showed to add more information, and no significant difference was found between them (two-sided t-test, p=0.085). So, both conversation and data seem to supplement each other, and the conversation was valued as more informative by itself.

Lastly, we asked the coaches both halfway and at the end whether they agreed with the statement "The use of data (or: Having a conversation with the client) supports my effectiveness as a coach." In the coaches' responses, mean scores regarding conversation (M=4.22 halfway, M=4.69 at the end) were significantly higher (p=0.018, for the main effect) than those regarding data (M=3.62 halfway, M=4.33 at the end). The interaction was not significant. Again, this reveals the value of conversation over data, in this case being more supportive to coaches' effectiveness.

In conclusion, the analysis of the questionnaire, while based on a limited sample, shows a coherent message that is largely in line with our qualitative findings. It

7 It was not feasible to apply a multilevel model here, because the question halfway the session (i.e., whether they agreed with the statement "I have enough information to give the client appropriate advice.") was substantially different than the question at the end of the session (i.e., whether they agreed with the statement "Because of the data / the conversation with the client, I have more information than before."). Therefore, we applied a t-test to compare the group scores across the information sources data and conversation.

suggests that coaches generally gain more value from a conversation with the client, compared to assessing the client's data. The results show that data do acquire value when situated within a conversation, allowing for discussing the data and sharing interpretations. Indeed, when only the client's data were available to the coaches, in absence of the client, many coaches expressed difficulty to interpret the data and formulate an advice. For example, Coach K reflected: *"I have tons of questions, why this, why that"*. And when the client was introduced after the data assessment, the coached valued the conversation as *"very valuable"* (Coach L) in order to *"get a clear picture of their goals"* (Coach L) and to *"learn the reason of the data results"* (Coach I).

DATA CHANGING CONTENT-ASPECTS OF COACHING

We have described which role and value data may have in a coaching session. In the next sections, we will specifically focus on the influence of the data on the content-aspects of the coaching. Specifically, data put forward different topics to be discussed, leading to new insights. Mainly, data were adding insight in measurable behavior, making conversations more specific, and more driven by health standards.

Data Provide Insight into Measurable Behavior

Although we have seen that the mere value of data is limited, our results do reveal the value of data according to the coaches. The responses to the open question regarding the value of data showed data gave coaches *"an overall understanding of client's activity level"* (Coach A) or *"an indication of their basic health stats, such as rest heart rate and activity levels"* (Coach L). Indeed, advices based on solely data often included those 'basic health stats', such as *"try to aim for 10.000 steps a day"* (Coach 19). Interestingly, when this advice was updated after the client conversation, this sometimes showed to move away from the data, i.e., *"learn to trust your body and rely on your own intuition"* (Coach 13), while in other cases the focus on data was only strengthened, i.e., *"monitor other things that potentially explain weight, nutrition and sleep habits, and see if you can find relations there"* (Coach 19).

Data Provide More Specific Cues for Coaching

While we previously discussed the coaches' need to see the client through the numbers, at the same time, we observed that in some cases data showed to provide an additional lens on the client, revealing new information relevant to the coaching. When coaches and their clients were discussing the data, this sometimes resulted in topics which clearly would not have been discussed without the data. These topics were often very specific and highly contextualized in the daily life of the client. For example, when a coach was checking the food intake of her client, she asked: "*you eat Kung Pao? Did you make it yourself, or...?"* The client replied: *"No, it was from a Chinese restaurant. I was eating out (...)"* Coach: *"How often do you eat out?"* Client: *"It really depends, like, sometimes it can be once a week, sometimes it can be a month* *that I don't eat out at all."* Coach: *"Okay, so maximum once a week. That's okay."* The topic of cooking yourself or eating out, and the corresponding value judgement that it is okay to eat out as long as it is not more than once a week, would most likely not have emerged without the data. In another example, the data served as memory aid for the client. The coach asked *"Sunday, uh Monday? What was going on this... You were working? The 20th. 13222 steps."* Client: *"Uhm... I have to check what I did that day; I cannot remember. Ah, then I had a day off! I had a funeral and in the morning; I made long walk with my neighbor."* Such specific statements provide coaches with useful cues to deepen their understanding of the client's daily life, social environment and experiences. While triggered by data, this provides insights beyond data, on behavioral and even experiential level, and a starting point for meaningful coaching.

Data Trigger Comparisons with Health Standards

Also, we observed that data triggered conversations on standards and norms; on what is 'normal' for a person, or outside of a normal range. Particularly when coaches were checking the data without the client, thus lacking the background of the data, they were typically comparing the data with health standards. As coach F reflected: *"I need to ask many questions first. The only thing I take from the data now, is whether she meets the standards for physical activity."* Also, when the clients were present and collaboratively discussing the data, comparing the data with the standards was a common occurrence. Such standards regarded daily step counts, water intake, light and intense physical activity, sleeping time and sedentary time.

DATA CHANGING RELATIONSHIP-ASPECTS OF COACHING

Data did not only change the content of the coaching; it also affected the relation between the coaches and clients. Data shifted roles, and both coaches and clients were keen on understanding how the other would relate to the data. This motivated their own efforts to put the data in the right perspective.

Data Shift Roles, Typically Putting the Client More Central

In the *conversation-first* condition, where data were added within an ongoing session, clients often took a leading role in the conversation as soon as the data were presented. They felt ownership over it, because have been living and working with the data over the recent weeks, thus, they took their responsibility to explain it to the coach. This was not only driven by the clients, it also happened that coaches explicitly asked the client what they wanted to discuss regarding the data, such as Coach K: *"is that the most important for you to evaluate now, the food?"* Not only the data pushed the clients in a more leading role, it also happened that the data itself was leading in the conversation. That is, some coaches systematically 'checked off' the tabs in the menu of the tracker (i.e., *"Let's see, what else you have tracked. Ah, sleep, let's have look"*), following the data rather than their own agenda.

Coaches' Efforts to Understand How the Clients Relate to their Data

Coaches showed to be motivated to understand how the client's perceived their data and felt about their data. They frequently asked clients how they experienced the tracking, for example, how they used the tracker throughout their day, how often they checked the numbers on their watch or phone, and whether it motivated them or made them nervous. Furthermore, coaches derived information from observing how the clients engaged in the tracking. For example, one client brought large amounts of data (self-initiated) to the workshop, very detailed and over a long period of time. Based on this, the coaches drew the conclusion that this client was very persistent, and at the same time risking focusing too much on the numbers rather than on how she felt. And indeed, this became an important topic in the coaching session with the client. Another client forgot to bring her phone to both sessions, which forced the coaches to look at the data on the small interface on the watch. Coaches attributed this to a possible lack of engagement or fear of showing her data. Thus, how clients related to their data was informative to coaches.

Coaches Putting the Data in Perspective

Mostly as a response to clients' worries, coaches put the data in perspective. They typically reflected on how they understood the data, how the data related to their knowledge, and then giving their value judgement on the client's behavior. For example, when a client reported *"it shocked me to see that my natrium intake is apparently too high"*, the coach explained her knowledge on natrium, how it is different from salt, and what it meant in context of the high blood pressure of the client. Then, the coach challenged the threshold for natrium in the app. She recalculated it using her own formula, concluding that there was no reason for the client to worry. In another example, a client expressed *"when I eat out, the next day my weight increased with 1.5 kilograms. This cannot be all fat, can it? And I did not even take a desert! This really worries me. How is this possible?"* The coach guaranteed that this could indeed not be only fat. She shared some knowledge on how it could be due to salt-intake, but mostly, the coach was trying to draw the focus away from the weight measurement. She explained that weight may vary a lot on the short term, and that therefore it only makes sense to measure it with longer intervals. This pattern frequently happened across the coaching sessions. Data, supplemented with client's thoughts and feelings on it, provided the coaches the opportunity to reassure the clients, make compliments, or share additional knowledge. In some extreme cases, the coaches even recommended to stop tracking, to put their minds at ease and focus on the benefits their healthy lifestyle brings them.

Clients Putting the Data in Perspective

Also, clients showed to be motivated to put their data in the right perspective. First and foremost, clients frequently reflected on the reliability of the data and tried to guide coaches to interpretations that they found accurate. For example, when checking the number of stairs climbed, a client reflected *"it doesn't recognize this. I have two floors at home, and I go up and down so many times a day, but it only recognizes a few times."* Or, when a coach found high-calorie peanut candy in the nutrition list, a client responded surprised *"oh that is a mistake; that should be the healthy nut bar! Peanut candy, oh no, no I wouldn't eat that."* And, when a coach read out loud that the client drank 6 beers that week, the client responded *"now you making it sound like I had a beer every night!"* and explained that it was actually due to a party at work. Thus, clients showed to be highly invested in making coaches understand and interpret their data accurately and with the right nuance. They cared about their image that coaches would build from their data.

Lastly, clients showed to have expectations on how the data would be valued compared to their self-report. Coaches mostly focused on self-report as their main source of information, but in the rare cases where they had a stronger focus on the data, this was not always appreciated by clients. For example, when Coach A said while checking the data: *"You didn't eat much yesterday"*, the client replied rather frustrated: *"That's what I said!"* She seemed offended that coach did not take her word for it, and that the data apparently added information to her self-report.

DISCUSSION

Personal tracking data plays an increasingly important role in current healthcare practices. Healthcare professionals, sports coaches and lifestyle coaches are expected to benefit from the additional insights that the availability of data may bring. However, evidence is accruing that the mere insertion of more data into a health coaching practice does not linearly result in better outcomes, or indeed, a better process. The focus of the current chapter is to improve our understanding of the role of data in the health coaching processes. Specifically, we look at how clients' self-tracked data influence health coaching, both in terms of coaching content and the relationship between coaches and their clients. In a workshop and a field study, we observed coaching sessions where personal data were added in various ways; at the start and halfway through the session, in the presence and absence of the client whose data was being inspected, and within familiar coach-client relationships or in an intake situation where coaches and clients were unfamiliar to each other. Our real-world observations enabled us to situate our insights regarding the data within the dynamics of the health coaching process. In addition, we gained insight in the value of data and conversation, individually and collectively, by presenting the coaches with clients' data and client conversations in various order.

Throughout the study, the data-enriched coaching sessions demonstrated a pattern of incremental activities, moving from an initial need for low-level understanding of data and behavior, towards understanding higher-level client aspects such as the

Oh, that is a mistake; that should be the healthy nut bar! Peanut candy, oh no, no I wouldn’t eat that!

Client

context wherein the behavior was performed and how this relates to the client's goals and experiences. Only after the coaches gained sufficient understanding, they gradually moved to sharing knowledge and giving advice. Within this process, coaches and clients showed to be in a continuous process of negotiation on the meaning of the data, where they were motivated to put the data in the right perspective, for themselves and for the other. For example, coaches were seeking to connect the data to the client's goals and experiences, and clients were trying to make sure that the coach would build an accurate and nuanced picture of them based on the data. Furthermore, we observed that the presence of data could also bring up different topics. These topics were typically more specific, more related to health standards and more oriented to measurable behavior. Yet, data were rarely self-explanatory. Both our qualitative and quantitative analyses strongly show that collaborative reflection on the data, where interpretations are shared and data are contextualized within the clients' narrative, was required for data to be meaningful and useful in the coaching process.

DATA ARE NOT 'PLUG-AND-PLAY'

Wearable tracking devices and e-coaching applications are mostly presented as finished products or solutions. They are built on the premise that the personal tracked data provide an objective view on behavior, as opposed to subjective experience and biased self-report. Through a set of rather linear cause-effect relationships, data are expected to enable detection of trends and correlations, resulting in insight and ultimately effective coaching. This implies that such data must add value for health coaches as well; after all, more information is assumed to be better. Our results paint a more nuanced picture. While data do bring certain value to the coaching process, this value does not come from the data in and of themselves. Data are not plug-and-play, they need contextualization from the client to be meaningful in the coaching process. Specifically, merely presenting behavior does not reveal, among other things, why the behavior was performed, whether it was a pleasant experience for the client or a struggle, which belief or contextual situation triggered the behavior, and whether the behavior was beneficial at all in terms of the client's goal and narrative. We argue that the inherent value of data is very limited; data do not have value *because* they are objective, rather, data only acquire meaning when *seen through subjective perception* of the client, as part of a dynamic and collaborative process of meaning making, involving intrapersonal, interpersonal, and data-driven reflections and interactions.

We expected that the data would serve as memory aid for clients (Figueiredo & Chen, 2020), and indeed, our results show that clients recalled specific events and experiences when discussing data. Coaches, however, found it hard to gain actionable insights from the data. Both our qualitative and quantitative results clearly show that data is more informative to coaches when assessed in combination with a client conversation. Building on prior findings (see *Chapter 2*, and Figueiredo et al., 2020; Mentis et al., 2017; Pichon et al., 2020) that emphasize the value of collaborative reflection on data, our results show that data provide useful conversation starters and facilitate sharing lived experiences. Our results additionally show that data disconnected from interpersonal exchange typically result in more questions than answers. We expect this effect may be amplified by the character of health coaching for healthy clients, where the goal, and thus the use of data, is more open-ended compared to more medical contexts.

E-HEALTH TECHNOLOGY SHOULD NOT MERELY FOCUS ON TRANSFERRING INFORMATION

When designing self-tracking devices and e-health technology for collaborative use, our results show that it is key to facilitate broader collaboration than merely sharing data. To be able to effectively use and interpret data, we should allow these data to acquire meaning within a coach-client conversation. In this conversation, we have to acknowledge that coaches and clients are not only sharing information; at the same time they are establishing and maintaining a relationship (c.f., Watzlawick et al., 1967). Data are added to a dynamic interplay between a coach and a client that is subject to trust, expectations, empathy and investment. This calls for a broader view on self-tracking devices than merely a computational system. Drawing from distributed cognition theory (Hollan et al., 2000), we may consider the coach, client and tracking device as a sociotechnical system, wherein it is important that all agents are enabled to effectively share and utilize their unique knowledge representations of the data and the status and needs of the client. Thus, these technologies do not provide one-on-one solutions and data do not provide answers, rather, these technologies and the data they bring forward are enablers of a good coach-client relationship and effective communication, together, resulting in effective coaching. Health coaching is, after all, a client-centered process based on a collaborative relation (Wolever et al., 2013).

Thus, data visualizations and dashboards for clients and their coaches will need to support the coaching process and the coach-client relation with giving the right cues. Specifically, our results show that information that is very specific and well-contextualized (e.g., specific food or exercises; where the client was and with whom) yielded useful coaching conversations. Our results also show that such specific information alone is not enough; even seemingly self-evident behavior was still frequently questioned by the coaches and discussing this led to deeper insights concerning the client. Furthermore, presenting this information is only helpful when it is meaningful in terms of the client's status and goals. For example, when a client's struggles are rather emotional, presenting simple behavior such as step counts can turn out to be very inappropriate.

Prior literature typically points to low-resolution, incomplete or unreliable data as main barriers for data to effectively serve as input for health coaching (Mahajan et al., 2020; Sqalli & Al-Thani, 2020; West et al., 2017). Yet, our results extend these findings in that for data to be useful, it is not only a matter of measuring more consistently and more accurately. In addition, it is important to measure those things that are relevant to a client's goals and struggles, and to enable her to explain these experiences through the data. We have seen that meaningful coaching advice is typically based on information at the level of the client's experience rather than her data. Thus, collaborative reflection on the data allow coaches to understand the data through their client's eyes, which is needed to be able to provide appropriate support. A coaching process is, after all, an inherently social process that goes beyond an optimization problem based on data.

IMPLICATIONS FOR E-COACHING

It is interesting to consider the implications of our results for e-coaching applications, for example based on artificial intelligence principles. While our findings highlight the value of a coach-client conversation on the data, not everyone may have access to a human coach. Thus, when designing stand-alone e-coaching we may try to implement some of these beneficial elements of a conversation with a human coach in other ways.

Across our coaching sessions, data were mostly used as a tool to explore. Specifically, data facilitated talking about, and thus thinking about, what goals a client would have, what wellbeing would mean for her, and possibilities to achieve her goals that would fit her daily life. It is interesting to consider whether a fully automated e-coach could potentially also trigger such a process, for a client by herself. Our *data-only* condition, where coaches assessed the client's data in absence of the client, reveals coaches' unmet information needs that represent the gap that needs to be bridged between the data and appropriate coaching advice. Specifically, coaches were seeking to understand, among other things, how the client's health data connect to her goals, the particular challenges she would face while trying to achieve her goals, and the social context of subjective experience of certain activities. Ideally, we would support clients to go through such a process themselves. A study by Choe, Lee, Zhu, Riche and Baur (2017) reveals that this may be challenging. They show that people tend to have low levels of reflection on their self-tracking data, for example descriptive reflection, and that higher levels of reflection are rarer, for example transformative and critical reflection (c.f., Fleck & Fitzpatrick, 2010). They argue that these higher levels of reflection are not easy to foster through visual data exploration tools (Choe, Lee, et al., 2017), while this may be exactly what is needed to make health data effective in terms of coaching. Kocielnik, Xiao, Avrahami and Hsieh (2018) offer an interesting and practical solution-path to this problem. They designed an application that prompt users, through a conversational agent, to reflect on their data, for example by asking what happened during peeks or low points in the data, or by asking about goals, motivations, or contexts (Kocielnik et al., 2018).

Such use of reflective prompts is promising given the results of our study, specifically because it leaves the interpretation up to the client, it acknowledges that goals are dynamic, and it avoids value judgements based on data.

THE USE AND EXPECTATIONS OF DATA WILL BE EVOLVING

We observed that the role of data, and coaches' and clients' expectations of each other and the data, was not yet settled in the coaching sessions. For most coaches and clients, it was the first time they used data in such a way. This is a limitation of our study design, as coaches' unfamiliarity with the interface and the wearables may have amplified their concerns on the usefulness of the data and inhibited effective use of the data. Still, our study setup represents a realistic scenario, where a client buys a tracker, uses it for a while, visits a coach and brings her data. So, while by design of the study our focus was mostly on the early phases of data sharing, it still gives valid insights in what happens as soon as data are introduced to a coaching process.

It is interesting, though, to consider how this 'configuration' of coach, client and data, including their roles and expectations, will possibly evolve over time. Our results show that coaches and clients were largely attentive to how *the other* related to the data. They were interested in what the data would mean to them, and tried to understand the others' intentions and expectations on how to use the data in the coaching session. It is likely that coaches' and clients' common ground on these aspects will grow over time, when data have been used throughout several coaching sessions. Furthermore, coaches' and clients' data literacy may grow by having more experience in handling and interpreting the data, and this is likely to increase their self-efficacy and feeling of control. When coaches learn about the possibilities and limitations of data, and experience that clients still care about their opinions on top of what the data are representing, this might make coaches more comfortable and willing to use data. As a result, our observations regarding coaches' tendency to refer to health standards or to give the clients a leading role when exploring the data, may decrease over time when coaches acquire strategies to effectively utilize the data themselves.

Additionally, Watzlawick and his colleagues (1967) argue drawing from their experience with couples psychotherapy, that the 'healthier' a relationship, the more the relational aspects of communication move to the background, allowing a more dominant role for the subject matter itself. In contrast, with malfunctioning relationships, there is hardly any room for the content, as people are constantly struggling about the nature of the relationship. So, it is expected that when a good coach-client relationship is maintained and secured, more room is available for discussing the information itself that the data comprise. Indeed, our results show that discussing specific trends in data or drawing actionable insights was not always relevant or appropriate, yet, this might change over time. Future research is needed to validate these effects.

CONCLUSION

We report on a workshop and field study where we analyzed a variety of health coaching sessions enriched with a client's self-tracked health data. Observing the role and value of data within a realistic setting enabled us to situate our findings within a broader perspective, including the dynamics of a coaching session. Our results highlight the importance of considering the coach, client and the data as a whole, when evaluating the value of personal tracking data for coaching, or when designing tools that may support this process. Self-trackers and e-coaching applications are not independent computational systems, yet, they are embedded in a broader context of health coaching. This constitutes a process where coaches and clients are constantly involved in negotiating interpretations and aligning expectations as they collaboratively work towards health goals. Within this context, self-tracking devices should not be presented as solutions, rather, as helpful tools to support this process.

Now I possibly
with irrelevant data,
a specific question
the other hand, by
routines, she might
or suboptimal
recommend me to
that I would not

Mother

burden the nurse
unless I have
about it. (...) On
checking my
indicate anomalies
routines, and
do things differently
think of.

Evolving Practices and Value of Sharing Customized Home-Collected Data with Healthcare Professionals in Newborn Care: A Field Study

CHAPTER 4

Evolving Practices and Value of Sharing Customized Home-Collected Data with Healthcare Professionals in Newborn Care: A Field Study

This chapter reports on a field study where we followed participants over a period of five weeks, in which we enabled parents of newborns to collect data that they ought relevant to share with their healthcare professional. We observed their tracking and sharing practices, and investigated parents' and healthcare professionals' needs and reflections on these data through weekly interviews. Similar to *Chapter 3*, this study provides insight in the influence of data on the coaching process in practice, with two notable differences. First, in this chapter we worked with an open-ended tracking tool, enabling parents (whether or not in collaboration with their healthcare professional) to customize their own tracking, for example by creating their own labels. Second, we followed participants over a longer period of time, which allowed us to investigate how data-practices, -needs and shared understanding would evolve over time. Our results show how some parents and healthcare professionals were converging, and others were diverging, over time. To converge to a shared understanding of the data and the problem, aligned expectations showed to be of key importance. Furthermore, we observed how data played different roles in the coaching process; sometimes they helped to unravel problems, other times they helped to solve problems.

The participants in this chapter slightly deviate from the other chapters; we studied healthcare professionals (i.e., nurses in preventative care, general practitioners and pediatricians) with more clinically oriented roles than regular health coaches. They did, however, all engage in coaching activities for this particular context, as they were supporting parents to care for their babies. We excluded severe medical issues, and focused solely on cases where parents had worries or questions, or were in general need for support.

This chapter is derived from:
Rutjes, H., Willemsen, M. C., Bogers, S., Kollenburg, J. Van, & IJsselsteijn, W. A. (*Under review*). Evolving Practices and Value of Sharing Customized Home-Collected Data with Healthcare Professionals in Newborn Care: A Field Study.

Related publications:
Rutjes, H., Willemsen, M. C., Kollenburg, J. Van, Bogers, S., & IJsselsteijn, W. A. (2017). Benefits and Costs of Patient Generated Data, From the Clinician's and Patient's Perspective. *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '17)*, 436–439.

Kollenburg, J. Van, Bogers, S., Rutjes, H., Deckers, E., Frens, J., & Hummels, C. (2018). Exploring the Value of Parent-Tracked Baby Data in Interactions with Healthcare Professionals: A Data-Enabled Design Exploration. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '18). Montréal, Canada.

INTRODUCTION

The increasing availability of wearable sensor technologies on the consumer market, such as health watches and mHealth smartphone apps, is democratizing healthcare by empowering patients to take increasing control over their own health (Topol, 2015). Inevitably, healthcare professionals will encounter a growing number of patients who bring self-collected data to the clinical visit. This trend may shift roles in healthcare. Healthcare professionals are facing potentially more engaged and better informed patients, and patients' self-collected data provide an extra source of information regarding patients' health issues, additional to more traditional sources such as anamnesis or physical examination.

Healthcare professionals' and patients' expectations regarding data-sharing are not necessarily aligned. Patients have expressed various motives to track and share their data with healthcare professionals, including a need to explain oneself in more detail and with more objective information, and seeking acknowledgement for personal efforts or emotional support (Chung et al., 2016). By tracking, patients take increasing ownership over their health and the care process. While healthcare professionals have welcomed this favorable effect of tracking, they also have expressed concerns regarding this trend (see *Chapter 2* and *Chapter 3*, and Jacob et al., 2020; Lavallee et al., 2020; Lordon et al., 2020). For example, lack of standardization of tracking tools is potentially putting high burden on healthcare professionals to adopt these data in their current workflows (Lavallee et al., 2020; Lordon et al., 2020). Furthermore, the patient's self-tracked data may be incomplete or unreliable (West et al., 2018), or inherently limited (see *Chapter 3*), making it unsuitable to use, and healthcare professionals worry about the unintended negative implications of these data on patients themselves (see *Chapter 2*, and Lavallee et al., 2020). It is yet unclear what constitutes effective and satisfying data collection, sharing and reviewing practices.

The current chapter presents a study where parents of newborns customize and share their home-collected data with healthcare professionals, through a custom-made and open-ended toolkit. We follow five families and their healthcare professionals over several weeks, and observe how their data-collection, -sharing, and -review practices evolve over time, and how this influences the communication between parents and healthcare professionals. We believe that the newborn-context is relevant for studying the value and practices of sharing data in a healthcare setting, for several reasons. First, tracking is increasingly commonplace in this patient group. For example, baby-wearables allow for tracking a baby's breathing motion[8], oxygen-levels and heart rate[9] (J. Wang, O'Kane, Newhouse, Sethu-Jones, & Barbaro, 2017), and smart phone applications facilitate manual tracking of the baby's behavior[10], including feedings, diapers, sleep and developmental milestones. Furthermore, parents of newborns are generally motivated to track and share their baby's data with healthcare professionals. Parents are highly engaged with the wellbeing of their child (Lupton, 2011), and high physical and emotional impact of becoming a parent raises parents' need for reassurance and acknowledgement (Barclay, Everitt, Rogan, Schmied, & Wyllie, 1997). For healthcare professionals, home-collected data potentially add value to current sources of information such as parental report and physical examination, particularly when the problematic behavior of the baby does not show during the clinical visit. Lastly, healthcare for newborns goes typically beyond the health of the baby; it also considers the family as a complex and inter-related system, where parents' experiences and wellbeing are important to consider. For instance, excessive crying of a baby is not only important in light of the baby's health, but it is also essential to understand how the parents experience this behavior, and provide them support if necessary. Overall, the context of parents tracking and sharing data of their newborns provides a relevant case to learn about the value and practices of tracking and sharing home-collected data.

In the current study we investigate healthcare professionals' and parents' experiences, needs and expectations towards data sharing. We will explore how data-sharing transforms traditional roles and communication in healthcare settings, and the extent to which parents' and healthcare professionals' expectations of data are aligned. In the remainder of this section, we will discuss related work, highlight our contribution, and present our research questions.

RELATED WORK

Tracking and Children

Prior work has identified opportunities of tracking technology for different phases of a child's life. For preterm infants who were just discharged from the hospital, tracking the infant's and mother's health and mood showed to be a helpful tool for parents to cope with emotional challenges and gain insight in relations between the infant's and mother's health (Hayes et al., 2014). For young children, record-keeping of developmental milestones showed to be promising identify developmental delays and share concerns with healthcare providers (Kientz et al., 2007). Also for teens self-tracking can be beneficial, as it provides them insight in

8 www.snuza.com

9 www.owletcare.com/products/owlet-smart-sock

10 Examples include: www.philips-digital.com/baby-new, www.oviahealth.com, www.thewonderweeks.com/about-the-wonder-week-app

their behaviors and potentially supports their wellbeing (Potapov, Vasalou, Lee, & Marshall, 2021).

Tracking technology for children deviates from tracking technology for adults, especially because the tracking is embedded in families where parents, siblings, and other care-givers play an important role too. Pina and her colleagues (2017) argue that *Family Informatics* asks for a different perspective than Personal Informatics. For example, there may be important ripple-effects at play, where the health of one family member is influenced by another (i.e., children's sleep affecting parent's sleep and mood), and this can be important to capture and visualize. Furthermore, tools should not only facilitate *self*-tracking but also *second-hand* tracking, where parents track their children's health (Pina et al., 2017). When parents' tracking is combined with children's own tracking, it is important to be sensitive to different motivations that children and parents may have (Oygür, Su, Epstein, & Chen, 2021). For tracking technology to meet needs on family level, it is key to enable parents to *support* rather than monitor their children (Zehrung, Huang, Lee, & Choe, 2021).

Another important consideration for tracking technology in the context of children, is the implications for parental stress and wellbeing. Tracking babies' behaviors or physiology may increase parents' anxiety, because it brings forward new metrics that may worry parents (Lupton, 2020; J. Wang et al., 2017). Furthermore, tracking potentially inhibits parental intuitions (Gaunt, Nacsa, & Penz, 2014), and it may enhance the tendency among parents of judgmental comparison of maternal behaviors (Lupton, 2011). On the other hand, when parents self-track their own behaviors and experiences, this may potentially mitigate stress (Jo, Toombs, Gray, & Hong, 2020).

Healthcare professionals in childcare also potentially benefit from home-collected data. Indeed, it has been found that these data may support interactions with pediatricians, by giving insight in whether and when the child has completed developmental milestones (Kientz et al., 2007). Kientz, Arriaga and Abowd (2009) showed that parents and pediatricians were generally more satisfied with their communication when more detailed data were shared. At the same time, parents and pediatricians were not always aligned; in some cases parents were disappointed as they expected larger involvement of their pediatricians in their data, while these pediatricians themselves were positive about the interactions (Kientz et al., 2009). Besides some notable exceptions (Kientz et al., 2009, 2007), healthcare professionals' needs regarding data sharing are not well studied in childcare contexts; most literature is considering the families' needs. Therefore, we broaden our discussion of related work beyond childcare to a more general discussion on patient-generated data and their value for healthcare professionals.

Patients' and Healthcare Professionals' Expectations of Data

Self-tracking technology facilitates patients to track their (health) behaviors and experiences and share these data with their healthcare professionals (Nittas, Lun, Ehrler, Puhan, & Mütsch, 2019). These so-called patient-generated data (PGD) potentially contribute to an increased understanding of the patients' daily life and experiences (Lordon et al., 2020). For patients, the main benefit of sharing PGD is that it provides them with additional ways to express themselves. Effective communication between patients and healthcare professionals, where patients are actively involved, has shown to be hard, yet of key importance for the quality of care (Ha et al., 2010; Robinson, 2003). Tracking and sharing data may enhance this communication. For example, referring to data helps patients to explain their symptoms (Baos et al., 2005), and it supports patients to share their values in life, facilitating shared decision making and prioritizing treatment (Berry et al., 2019). Patients expect that sharing PGD helps their healthcare professionals to have a more complete understanding of their daily life, potentially improving personalized care, acknowledgement of their personal efforts, and emotional support (Chung et al., 2016). They also expect that healthcare professionals gain additional insights from their data, based on their medical expertise (Chung et al., 2016).

Healthcare professionals also indicate a number of benefits of using PGD in their practice. It may facilitate them to observe longitudinal changes, trends and correlations in behavior and wellbeing, that supports diagnosis and understanding the impact of treatment (Chung et al., 2015; Raj et al., 2017; Yoo & Choudhury, 2019). Furthermore, PGD potentially provide access to a patient's lived experiences, enhancing understanding how symptoms affect the quality of life, which is often overlooked in traditional clinical visits without data (Hong et al., 2018). Healthcare professionals often highlight the value of collaborative reflection on the data with the patient (see *Chapter 2* and *Chapter 3*, and Chung et al., 2015; Lordon et al., 2020; Mentis et al., 2017; Raj et al., 2017; West et al., 2018). More specifically, PGD may provide helpful starting points for meaningful conversations (see *Chapter 2* and *Chapter 3*), and data may be used to motivate and educate patients (Chung et al., 2015).

Patients and healthcare professionals' viewpoints regarding the use of PGD are not always aligned; they perceive the data differently, and have different expectations of its use. When reviewing PGD, healthcare professionals base their perspectives on their medical background, whereas patients use their lived experiences to make sense of the data (Raj et al., 2017). These different backgrounds potentially enhance misunderstanding and disagreement on the use and interpretation of PGD (Raj et al., 2017). At the same time, other researchers have framed these different backgrounds as an opportunity, to leverage the value of PGD by exploiting each other's knowledge (Marcu et al., 2014). Similar to the findings of Kientz and her colleagues (2009) in the childcare context, also in other domains patients typically expect higher levels of engagement of their healthcare professional with their data (Chung et al., 2016; Lavallee et al., 2020; Lordon et al., 2020; Zhu et al., 2016). Healthcare professionals often do not intend to extensively review the data, even when they initiate the data-collection themselves. Rather, they mostly use it to increase the patient's self-awareness and engagement with their own health (Lordon et al., 2020). This may lead to disappointment with patients, deviant tracking behaviors not compliant to the healthcare professionals advice, and even termination of data collection or sharing (Chung et al., 2016; Lavallee et al., 2020; Zhu et al., 2016).

Tracking Tools Designed for Efficiency, Flexibility and Collaborative Use

For healthcare professionals to be able to use data effectively, prior literature points to a critical balance in design of tracking tools between flexibility and efficiency. Efficiency is important, as healthcare professionals indicate time pressure as an important barrier to adopt PGD in their practice (Lordon et al., 2020). They worry that they will be overwhelmed by data when it is not sufficiently processed (Zhu et al., 2016), not standardized (Lavallee et al., 2020), or not relevant (West et al., 2016). There is a strong need for clear and efficient data presentations, for example through comprehensible visualizations (Schroeder et al., 2017), or by highlighting specific data that are considered as most relevant (Yoo & Choudhury, 2019). Standardized representations can support healthcare professionals to make sense of the data; it helps recognizing patterns and identification of missing or inaccurate information (West et al., 2016). While these examples of automatic data processing are clearly lowering the burden on healthcare professionals, in turn, this comes with the cost of losing flexibility (Chung et al., 2015). Tools presenting PGD through standardized structures risk hindering the often informal and unstructured communication during a clinical visit, which is not desirable (Marcu et al., 2014). Furthermore, flexible data collection, allowing for a broad range of potentially relevant behaviors and experiences, is often required to meet specific information needs across different situations and patients (Luo, Liu, & Choe, 2019; Schroeder et al., 2017). It is not straight-forward how to combine the favorable aspects of both flexibility and standardization in the design of tools that facilitate sharing PGD.

Current tracking tools do typically not adequately support collaborative use. First, tools often do not facilitate sensemaking or analysis of data, for users themselves (Choe, Lee, Lee, Pratt, & Kientz, 2014), nor in a family setting (Pina et al., 2017; Yamashita et al., 2017), nor with healthcare professionals (Nunes et al., 2019; Raj et al., 2017) or health coaches (see *Chapter 2* and *Chapter 3*). Collaborative use of data in healthcare settings may be particularly challenging, because patients and healthcare professionals have different needs regarding data representations (Berry et al., 2019; Raj et al., 2017; Schroeder et al., 2017; Vizer et al., 2019). Furthermore, current tools lack the flexibility that may be required for patients to express themselves in meaningful ways. The very act of tracking is a highly dynamic and personal process (Rooksby et al., 2014), and flexibility and adaptability may be required for patients to shape and use the tracking tools according to their own needs and situation (Ayobi, Sonne, Marshall, & Cox, 2018; Choe et al., 2014; Y.-H. Kim, Jeon, Lee, Choe, & Seo, 2017; Nunes et al., 2015; Storni, 2011). Active involvement in data collection also enhances accountability and engagement (Choe, Abdullah, et al., 2017), and may be a meaningful and positive learning process in itself (Ayobi et al., 2018).

FOCUS OF THE PRESENT WORK

We report on a study where we follow several parents as they collected data on their newborns and share and discuss these with their healthcare professionals. We repeatedly interview parents as well as healthcare professionals on their experiences and needs. Prior work provides a great starting point for understanding patients' and professionals' needs when collecting, sharing and discussing home-collected data. Yet, these studies are typically based on *one-off* interviews or observations, where healthcare professionals and patients are reflecting on the value of data in a specific moment. We contribute to prior work in multiple ways. For one, we took an iterative approach, where we followed evolvement of health issues, parents' questions, data-sharing practices and needs over time. Not only did this enable us to observe development over time, it also allowed for continuously responding to emerging needs with design interventions to deepen our understanding. Furthermore, it guaranteed that the data, and the health issues that they represent, were timely and relevant in a particular moment, making retrospective reflections unnecessary. Lastly, the data-sharing was situated in a realistic setting, where parents and healthcare professionals knew each other and had recently met in a clinical visit. Working with real data, from real people with real issues, within real relationships, makes this study highly ecologically valid. Altogether, the current chapter adds to prior studies by taking an iterative, timely, situated and realistic approach.

Since self-tracking is a relatively new trend, there is not yet consensus on the ideal use of self-tracked health data in clinical contexts. Healthcare professionals and patients typically have not yet found a common ground on how to use these data, and what to expect from each other. Chung and her colleagues (2016) argue that using such data nevertheless, is facilitating negotiation of roles, responsibilities, desired practices and meaning of data. In the present study, we aim to maximally facilitate this process by providing parents and healthcare professionals with open-ended tools to collect, share and discuss data, and observe to which desired practices they converge. We purposely discounted off-the-shelf devices and data, instead, we custom-build an open-ended toolkit that allowed for customizable tracking not constraining what would be tracked, how it was tracked, for how long, and who initiated it. It is important to note that our toolkit was used as a means rather than an end. We were not interested in evaluating the tools per se, rather, we sought to explore the evolving and ultimate use of such an open-ended product. We believe that following and facilitating this hands-on process allows for a deeper understanding of participants' needs and desired practices.

Our study was set up along the lines of data-enabled design (Bogers, Frens, Kollenburg, Deckers, & Hummels, 2016), where data are used to inspire and guide the design process, rather than a solution in themselves. Our earlier work already reported on this study (Kollenburg et al., 2018), but focused solely on the methodological aspects and value of data-enabled design. In this work, we revisit the data (both from interviews, observations and focus groups, as well as the home-collected baby-data) and analyze it more thoroughly using thematic analysis, aiming for a deeper understanding of the user needs and experiences when sharing home-collected data in a healthcare setting.

This study aims to answer the following research questions:

1. What are healthcare professionals' and parents' experiences, needs and expectations when sharing/receiving customized home-collected data?
2. How do data-sharing practices evolve over time, and how is this process influenced by, and influencing, communication between parents and healthcare professionals?
3. To what extent are parents' and healthcare professionals' expectations regarding home-collected data aligned?

METHODS

PARTICIPANTS AND RECRUITMENT

Seven healthcare professionals participated in the study (Figure 7). We included two nurses working in preventative care offices for children, two pediatricians, two general practitioners (GPs), and one children's daycare officer. The healthcare professionals were recruited either via the network of the researchers team, or via an official recruitment agency. They were given a financial compensation per hour spent on the study, equal to an average salary for those professions. Healthcare professionals with conflicting interest were excluded.

Three healthcare professionals were involved in the main track of our study (i.e., one nurse, one pediatrician and one GP, see Figure 7). They recruited one or two families from their own practice to participate in the study (Table 6). The inclusion criteria were: the baby is younger than 6 months, there is a question or worry about the feeding of the baby (however, not indicating severe health issues) and the parents take care of at least 50% of the feeds of the baby. The parents were also given a financial compensation. The pediatrician recruited only one family (Table 6), as she had difficulties finding suitable families. However, during the study, the GP referred one of her families to the pediatrician, so eventually, also the pediatrician could reflect on two cases. The four other healthcare professionals were involved in a *reflective* track, consisting of two focus groups, one interview, and two work observations (Figure 7). We used this reflective track to validate our insights from the main track with a broader group of healthcare professionals.

The study consisted of a home-data collection phase, which lasted approximately five weeks per family (Figure 7). During this phase, families collected and shared

	Family 1	Family 2	Family 3	Family 4	Family 5
Age baby at start of the study	5 weeks	3 weeks	5 months	3 months	6 weeks
Health issues	No issues	Born underweight	Multiple past hospitalizations	Excessive crying	Reflux

Table 6 Connections of families with healthcare professionals, and family demographics.

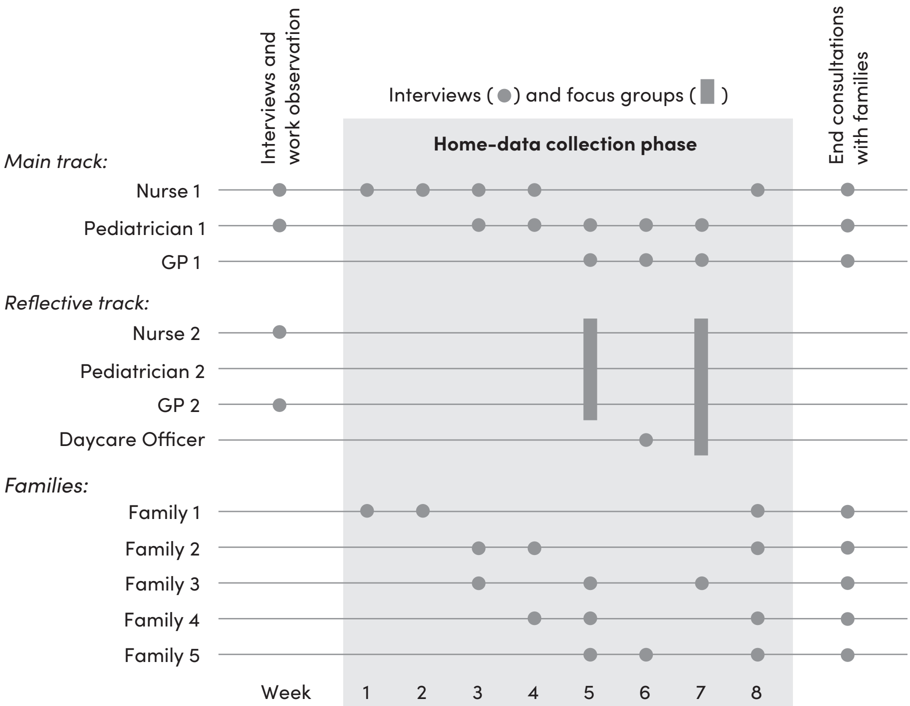

Figure 7 Study design; timeline of observations, interviews and focus groups with participants.

their data with their healthcare professional, while we repeatedly interviewed the healthcare professionals and families on their experiences. Prior to this phase, we held interviews and work observations with healthcare professionals. At the end of the study, all families met their healthcare professional in an end consultation, where they could share and discuss their experiences with each other. All visits and interviews were executed by one or two researchers, and are described in more detail in the Data Collection section. We deliberately phased in the different professionals at different moments in the study, (Figure 7; e.g., GP 1 started 5 weeks later than Nurse 1), to have more time to fix possible technical issues, and to extent the period to iteratively generate and test insights throughout the study.

This project was sponsored by and situated at Philips Design. The study was approved by the Philips internal ethics committee board and by the Dutch Medical Ethics Committee at Maxima Medical Center.

DATA COLLECTION

During the study, the parents used a custom-build toolkit to collect and share data of their baby with their healthcare professional. Next to these 'baby-data', we use qualitative data from observations, audio-taped interviews and focus groups with the healthcare professionals and the parents. In this section, we will describe the toolkit, the family visits, and the sessions with the healthcare professionals during the study.

The Toolkit

We created a custom-built toolkit, which enabled the families to collect data (Figure 8). The toolkit consisted of a base station that could be placed in a central place in the house, and several portable data trackers. We decided to bring physical objects in the homes instead of a digital tracking environment (e.g., a smartphone app). This allowed parents to think outside their frame of reference that is often strongly influenced by smartphone apps, and allows for in-situ use (e.g., place a sleep button in the bedroom of the baby). Furthermore, it makes it easier for other caregivers (e.g., grandparents, babysitters) to engage in the data collection.

Figure 8 Open-ended toolkit, including a base station and portable data trackers.

The toolkit was open-ended and flexible in use, aiming to empower patients to track what they considered relevant to share with their healthcare professional, and at the same time, also enabling healthcare professionals to request specific data collection. We adopted several trackers in the toolkit covering a broad range of functions (Table 7); families determined the meaning of the trackers by providing a label for the particular tracker that they wanted to use, through the screen on the base station of the toolkit. By doing so, they created their own tracking tool, tailored to their own needs and grounded in their own experiences. For example, the rotary button had been used to capture "crying intensity" from "whining" (minimum value) to "screaming" (maximum value), and the push button had been used to capture a "poo diaper", a "feed", or "belly time". The particular use of the trackers could be changed at any point in time during the study to meet evolving needs. A timeline-visualization of the data was presented on the screen in the toolkit, where parents could adjust or delete incorrect data points, and make annotations on data points if desired. In addition, there was a tab on the screen where parents could send and receive messages with the healthcare professionals and the research team. More details on the development, design and potential usage of this toolkit can be found in our earlier paper on this work (Kollenburg et al., 2018).

Data tracker	Saves...	Tracking type
Push button	When it is pressed	Manual
Rotary button	When and at what position it is pressed (0-100)	Manual
Toggle switch	When it is switched on and off	Manual
Text module	Text messages	Manual
Video module	Short videos	Manual
Audio module	Short samples of environmental sound intensity and pitch	Automatic (only activated when parents provide a label for this tracker)

Table 7 Types of data trackers in the toolkit.

We expected the open-ended character of the toolkit to facilitate discussion on the meaning of data and negotiation of roles among the parents and healthcare professionals. The way the toolkit was designed and used may even be considered as disruptive, first because it strongly empowers parents, second because the communication between parents and healthcare professionals was all at distance and a-synchronous (apart from an end consultation visit at the end of the study). This enlarged the role of data within the relation and communication between parents and healthcare professional, allowing for a thorough understanding of the value and practices of sharing home-collected data. It has to be noted though, that some healthcare professionals and families did organize visits during the study on their own initiative, to address medical issues that the toolkit could not account for. Also, some healthcare professionals and parents made phone calls during the study for extra clarification. These acts were in themselves informative, showing a need for synchronous and in-person communication about the data.

Family Visits

The researchers visited the families three times: at the start, midterm, and at the end of the study (Figure 7). In the first visit, the parents were asked what they wanted to share with their professional as a start, and the researchers helped to install the toolkit accordingly. We instructed the parent how to adapt the labels of the data trackers in case their needs evolved over time, and showed how to send messages to the healthcare professional and researchers. We emphasized that in case of medical or urgent questions, they should follow the regular procedure (e.g., call the doctor's office).

Midterm and at the end of the study, we visited the families again for a semi-structured interview. The interview scripts are included in Appendix B. In this interview, we asked the parents to generally reflect on their experience on the data tracking and sharing, and we asked them to elaborate on specific data points or messages, to obtain a deeper understanding of their process. Sometimes, new needs emerged during the midterm interviews, which was often followed up by an adjustment in the labeling and use of the data trackers.

Sessions with Healthcare Professionals

Work Observations

At the start of the study, we visited four of the participating healthcare professionals (Figure 7) and observed their interaction with (other, not included) patients during clinical visits. During these observations, which lasted approximately three hours, we focused on the information exchange between the healthcare professional and the patient, and the questions the professionals asked, reflecting their information needs. Subsequently, there was a short interview on their work routines (see Appendix B for the script). They also filled in a questionnaire on their attitude towards technology (Rosen, Whaling, Carrier, Cheever, & Rokkum, 2013). We ended up not using these scores in the analysis, as they did not add much information to the qualitative data.

Interviews during the Data Collection Phase

During the families' data collection phase, we visited the professionals weekly for an interview on their experiences (Figure 7). These interviews included several standard questions, such as "which insights do you take from the collected data so far?" (for more details, see Appendix B), and several additional specific questions. These specific questions were defined in weekly researcher team meetings, based on current data-collection practices (for example, when a family changed a label for a particular data tracker) and insights from other previous interviews.

We aimed at five interviews per professional, to have sufficient time to get used to the interface and the possibilities of the data tracking, and to allow for several iterations in the data collection and visualization. Eventually, we interviewed the GP only three times, but aided by the input from the other professionals, our insights obtained from the GP were saturated (three researchers agreed on saturation).

Focus Groups (Reflective Track)

To validate our insights in a broader group of healthcare professionals, we held focus groups with four other professionals (Figure 7). We included all professions that were represented in the main track, and additionally a children's daycare officer.

Both focus group sessions lasted for two hours. In the first session, we specifically discussed the data of one family, to encourage a discussion on the possibilities and implications of the having access to these home-collected data. The discussion was guided by a pre-defined list of questions, including "What home-collected data would be useful?" and "How would you like the data to be presented?" (for more details, see Appendix B). Only near the end of the session, we presented the insights we obtained from the healthcare professionals of the main track, to avoid self-fulfilling prophecy. In the second session, we continued the discussion, based on the most recent insights from the study. As our insights were further developed, we discussed more concrete topics, including which data the healthcare professionals would need given specific health issues, and a role-play with and without the availability of data.

End Consultations with Families

At the end of the study, we held consultation meetings with families and their healthcare professional together to reflect on the study, which was joined by one or two researchers. Most of the families did not see their healthcare professional during the study; they had only communicated via the toolkit. Consequently, this consultation meeting was a natural moment to share experiences on the data collection and sharing. During these sessions, the researchers only observed, not intervened.

DATA ANALYSIS

All interviews, focus groups and end consultation meetings (in total 36 hours) were audio recorded and transcribed verbatim. We analyzed these data during the course of the study, and captured main insights on notes on a wall in the project room, serving as a dynamic and shared mind map. These insights, combined with the home-collected data by the families, the message dialogs between the parents and the healthcare professionals, and the technical maintenance of the toolkit, was used as input in weekly multi-disciplinary team meetings. Here, we combined our technical-, design- and HCI-oriented perspectives to decide upon new design interventions, new data trackers, new data visualizations, or specific questions and topics for the upcoming interviews or focus groups. The design iterations are described in our earlier work (Kollenburg et al., 2018).

After the study, the transcripts were analyzed again, completely and more thoroughly, using thematic analysis (Boyatzis, 1998). This process was led by one researcher, and guided by iterative discussions with the research team. The analysis consisted of two phases, see Appendix C. First, we analyzed the data as a whole, not taking into account the temporal dimension in the data. That is, we initially omitted the fact that some quotes originated from the same participant at different points in time. In this phase, we focused on capturing data opportunities or challenges, for either healthcare professionals, parents, or their interaction. Thus, as shown in Appendix C, all codes were falling under one of these categories. From this analysis phase, Themes 2 and 3 emerged. The software package Nvivo (version 12) was used as coding tool. Twenty percent of the data was coded by an independent researcher, resulting in an inter-rater reliability of Kappa=0.75. All disagreements were resolved by discussion.

In the second data analysis phase, we restructured the data per family and healthcare professional over time, and analyzed their *journeys* over time. In this analysis, we focused on how tracking practices evolved over time, and how health issues and the understanding of these issues evolved over time. We also included the message dialogs between parents and the healthcare professionals in the analysis. We again applied thematic analysis, this time using an inductive approach. This resulted in Theme 1. See Appendix C for the mapping of initial codes and family journeys to final themes and subthemes.

RESULTS

To provide an overview of the families, including their situation, health issues and tracking behaviors, we describe family vignettes in Textbox 1.

Family 1
The baby of this family had no health issues, yet the family was clearly struggling with caring for their baby. In the study, they initially focused on feeding; they tracked the moment and the amount of milk that the baby was drinking. After a few days, this was supplemented with tracking their baby's sleep. Throughout the data collection, the mother was increasingly focusing on how difficult it was for the baby to fall asleep. For example, she made extensive notes about the location that the baby ultimately fell asleep (e.g., bed, living room, on her lap, etc.), the things she had tried to calm her baby down, and how her baby responded to that. To the nurse, the shared data often provided more questions

than answers. Therefore, she made several phone calls with the mother during the study, in which some personal issues arose, such as the parents not always being aligned on how strictly to apply the bed routines. As the study was progressing, the focus shifted from the baby's behavior to the parents' experiences.

Family 2

The baby of family 2 was born underweight, which caused the parents to have a strong focus on feeding. Before the study had started, the parents already extensively tracked their baby's behavior from birth, using a mobile app. The parents showed to approach their problems systematically. They were well able to reflect on their experiences, and targeted the data collection to specific issues. In the beginning of the study, they tracked the length of the feeding moments and the type of milk (breastfeeding, pumped milk in a bottle, or artificial milk). After about a week, this evolved into tracking the length of the pumping, the amount of the expressed milk, and the satisfaction of the baby after the feed. In the end, the parents were reassured that their baby received sufficient nutrition, and they developed an increasing understanding of their baby's signals. As the parents showed substantial ownership over the data collection and their baby's wellbeing, the nurse did not need to take an active role in this process.

Family 3

The baby of family 3 had multiple past hospitalizations, but at the start of the study all medical issues were resolved. Still, the parents had some non-medical questions. The family started off with tracking 'belly time' (that is, the baby practicing to lift her head op while laying on her belly), feeds and sleep. This reflected their interest in finding the right moment for belly time – not right after the feed such that she would spit up, but also not too late such that she would be too tired. Furthermore, they were struggling with whether or not to feed their baby at night when she would wake up. The pediatrician showed a strong interest in temporal relations between crying, feeding, defecation, and spitting up, to understand cause and effect. Furthermore, she was interested in the relation between the baby's behavior and the parents' experiences. Even though there was no clear need with this particular family, out of curiosity, she requested the family to track additional aspects, such as defecation (i.e., the moment of changing a poo diaper) and notes on the parents' experiences.

Family 4

The parents of family 4 were struggling with excessive crying of their baby, and the data-collection revolved around unraveling the cause of this crying. The GP and the parents suspected the baby was suffering from acid reflux, for which the GP had prescribed antacid medicine a few days before the study had started. At the start of the study, the mother showed difficulty to decide what to track, as in her perspective there was so much going on, and she still could not explain the crying herself. She mainly wanted to show the GP how often and how intense her baby was crying, and give her additional information that could potentially explain the crying. The family started off with tracking crying intensity and feeds, after a while supplemented with moments of spitting up. The GP above all reflected on the information overload that she experienced while assessing the large amounts of data, painfully contrasting with the family's need for recognition. Near the end of the study, also the pediatrician met this family and assessed their data. She additionally asked the family to keep track of a paper-based crying diary, aiming to get more insight, as well as providing the parents an additional tool to regain control over their situation. Even though no satisfying explanation was found on the baby's crying, towards the end of the study, the crying had decreased, and the families' struggles slowly resolved.

Family 5

Also for the baby of family 5, acid reflux was suspected by the GP and the parents, and they had started medicine prior to the study. The parents were motivated to collect and share home-data, as they felt that during prior clinical visits, the seriousness of their baby's problems did not fully come through. The family decided first to track feeds and 'reflux-behavior', which was their means of expressing the intensity of their baby's behavior while she was struggling with acid reflux. In the parents' perspective, their baby was suffering from silent reflux, such that the reflux-behavior did not only show by spitting up, but also less obvious signals as stretching and being uncomfortable after the feed. After a few days, they additionally tracked their baby's sleep, as they felt that the acid reflux was keeping her from her sleep. For a certain period they also wrote notes on how and where the baby had fallen asleep. As the study progressed, their focus gradually shifted from acid reflux to sleep. For the GP, it was hard to capture the parents' evolving question, as well as understanding the parents' actual worries, from solely examining the data. We observed a tension between the parents' need for recognition, and the actual understanding of the GP through assessing their data. As time progressed, the acid reflux symptoms decreased, and the baby's sleep improved.

Textbox 1 Family vignettes.

Table 6 provides an overview of the final themes and subthemes that emerged from our data, which we will discuss in more detail below. We will denote quotes from healthcare professionals of the main track with "1", and the reflective track with "2".

Themes	Subthemes
Theme 1 The role of the family's questions and problems when interacting with data.	1.1. Sharing data is not equal to sharing problems.
	1.2. The importance of focus when collecting data.
	1.3. Data for exploratory versus confirmatory purposes. • Explicit or implicit questions and the role of data. • Professional- versus parent-initiated tracking.
Theme 2 The impact of data-sharing on parents and their relation and communication with healthcare professionals.	2.1. Impact of data on parents can be both positive and negative.
	2.2. Mismatching expectations between parents and health-care professionals on the use of data.
	2.3. Data are valuable input in a conversation.
	2.4. Advantages and disadvantages of a-synchronous com-munication.
Theme 3 Some information can, and some cannot, be captured by data.	3.1. Data can provide objective information.
	3.2. Data can provide temporal (and thus cause-effect) insights.
	3.3. Data may be hard to interpret.
	3.4. Data often lack, although sometimes provide, context and background information.
	3.5. Data usually do not capture the (often essential) experi-ence of parents.

Table 8 Overview of final themes and subthemes

THEME 1: THE ROLE OF THE FAMILY'S QUESTIONS AND PROBLEMS WHEN INTERACTING WITH DATA.

For all healthcare professionals, the parents' question, reflecting their perception of their baby's problem and the situation at home, played a central role when interacting with the families and their home-collected data. Across the families and over time, we observed various and evolving questions. This theme describes how these questions played different roles in the data-discovery, or in other words, how data played different roles in the question-discovery.

Subtheme 1.1.: Sharing data is not equal to sharing problems.

All healthcare professionals agreed that *"the question of the patient cannot be replaced by data"* (Pediatrician 2). This was indeed observed during the study, as in none of the families, the problem naturally followed from the data. For example, Family 5

But what is their problem??

General Practitioner

used the trackers to capture feeds and sleep of their baby, and the intensity of their baby's behavior while struggling with acid reflux. The interviews with the family revealed that they believed that the baby's acid reflux caused her extensive crying and limited sleep, and they were worried about the potential impact that would have on the baby's development. When the GP was assessing their data, she was overwhelmed and repeatedly questioned: *"but what is their problem?"* Although the parents believed they had communicated their problem well, only after the GP explicitly requested the family to phrase their question and worries in a message, the problem came across and provided a lens to effectively assess the data.

Subtheme 1.2.: The importance of focus when collecting data.

The open-ended toolkit facilitated the collection of a broad range of data. Yet, the healthcare professionals clearly expressed a need for focusing the data collection around the parents' question. For example, the nurse (1) explained: *"Everything I do is driven by the question of the parent. So I can ask many things out of my own curiosity, but I actually want to respond to their questions or worries."* One of the main reasons for focus was to protect the parents. The GP (1) stated: *"We should not create new problems. They visit us with a specific question, and if you track too broad... We should avoid somatization; if you give them too much, they will see problems everywhere."* Furthermore, it would help avoiding information overload for themselves, illustrated by the GP (1) reflecting: *"I only have 10 minutes per patient. (...) Parents want to share all sorts of details, well, I don't care, I want to get to the core of the problem. (...) I want to know as little as possible, as focused as possible, as fast as possible."*

Subtheme 1.3.: Data for exploratory versus confirmatory purposes.

As parents' questions are key to guide and focus the data collection and review practices, we explore how different questions have guided this in practice in our study. From our results, two cases emerged: either data examinations guided the forming of the question (i.e., the data was used exploratory), or reversed, the parents' question guided the data examinations (i.e., the data was used confirmatory). We found that the purpose for which the data were used, that is, exploratory versus confirmatory, was related to the availability of an explicit question, and to the preference for professional- versus parent-initiated tracking, which we will discuss in more detail below.

Explicit or implicit questions and the role of data.

All families visited their healthcare professional some time before the study with a feeding-related problem, worry or question. However, most families were not able to clearly pinpoint their own problem at the start of the study. This was illustrated by most families struggling with initially setting up the toolkit in such a way that the data trackers fairly captured their situation. This was particularly hard for those families who were facing a mixture of issues, often regarding crying, feeding, and sleeping. They typically started off with tracking the baby's behavior very generically.

For some families, the act of tracking itself – including deliberately determining the labels of the trackers, as well as using the trackers to capture the behaviors accordingly – facilitated them in identifying their own problem. For example, Family 2 had started with tracking the duration of the bottle feeds, evolving into the satisfaction of the baby after the feed. By doing so, this reassured them that the baby was drinking sufficient amounts of milk, and they learned to interpret their baby's signals and distinguish between hunger and need to suck. The nurse was not actively involved in this process; she was just generally commenting on the data and reassuring them that they were on a good track. Basically, she supported the parents just by being the recipient of their data.

For Family 1 it was also hard to define their problem, but in this case, only after active participation of the nurse, through collaboratively reflecting on the data, their problem became clear. Interestingly, the mother reflected in the beginning of the study on the fact that she was very broadly tracking and sharing data: *"Now I possibly burden the nurse with irrelevant data, unless I have a specific question about it.* (...) *On the other hand, by checking my routines, she might indicate anomalies or suboptimal routines, and recommend me to do things differently that I would not think of."* Later, when the nurse was scrolling through the data, she reflected: *"I can respond on many things, but I actually want to respond to the question of the parent."* At several moments during the study, the nurse decided to make a phone call, to understand what the parents themselves perceived as a problem. She explained: *"The mother usually says everything is fine, but she does have questions. When I call her, we are talking for 45 minutes, and there are a lot of topics we discuss."* Here, the shared data helped the nurse to ask relevant questions, which turned out to be very helpful, particularly because the parents were not able to make their own problem explicit. The mother reported: *"Multiple times in the phone call, I thought we were finished, but then the nurse asked about specific data, and that reminded me of other questions that I had."* During the end consultation in the study, the mother explained the nurse that *"by asking all these questions about the data, you gave me a new perspective on my situation and my baby. I was not aware, I just did not see it."* So for this family, the active involvement of the nurse through sharing and collaboratively discussing the data unraveled their problems, and helped them to solve the problems accordingly. In this case, the data had a clear exploratory purpose.

This was contrasted by cases where there was a much clearer question at hand. For example, towards the end of the study, a very specific question emerged with Family 2. The mother's breastmilk declined, and she wanted to know if she could increase this. Through sharing when she was breast pumping, combined with the amount of milk that she expressed, the nurse could give her personalized and effective advice.

Because there was a clear question available, the data was used confirmatory rather than exploratory, in a sense that the question was leading for the data examination.

The questions of all families showed to be more explicit and specific over time, resulting in increasingly specific tracking practices. This reflected a process where tracking, sharing and discussing their baby's data, facilitated them to discover their actual problem, and by that understanding, problems had often (partly) resolved.

Professional- versus parent-initiated tracking.

Another aspect that was highly related to the exploratory versus confirmatory use of data, was the preference for professional- versus parent-initiated tracking. In general, the main drivers for professional-initiated tracking were high time pressure and a need to focus. The main drivers for parent-initiated tracking were to gain insights in the experiences of the parents and facilitating them to discover and determine their problem themselves.

Time pressure provoked a need for the professionals to initiate and specify the data collection themselves, opposed to letting the parents decide what to track and share, to avoid data overload. This often led to confirmatory use of data. For example, the GP (1) explained: *"I want to get to the core of the problem. I want to know how often she cries, whether the cry is related to acid reflux, and how much she spits up. That's it. I don't care about the rest."* Healthcare professionals were very keen on the idea of a fixed data tracking plan for a specific health issue. Contrary, when discussing this with the families, they were not merely positive. For example, the mother of Family 5 refuted: *"If I would have been limited to track only crying, the problem would not come across. My baby's problem is expressed in other behavior, for example in limited sleep."*

For the exploratory use of data, with the purpose of unraveling a problem, parent-initiated tracking was recognized as a very helpful tool. Particularly for the nurses working in preventative care, often there is no clear question at hand, so a fixed or a professional-initiated tracking plan would be infeasible and pointless. Parent-initiated tracking shows the healthcare professional what the parents come up with themselves, revealing essential information on the parents' perception of their baby and problem. For example, the fact that some families decided to track their baby's sleep, and soon after started to supplement this sleep data with extensive notes on how and where their baby had fallen asleep, revealed their struggle with falling asleep rather than sleep itself. When we provided an additional feature to the dashboard that facilitated parents to formulate their question and illustrate this with specific data points (see our previous paper for more details (Kollenburg et al., 2018)) this was much appreciated by both the healthcare professionals and parents.

The fact that we observed different preferences for who initiates the data collection, should also be understood in terms of the different professions and clinical practices we included in our study. For example, the GPs face high time pressure as

Parents are really focusing on the things they are tracking.

Nurse

Yeah, and then we need to explain what is normal. Well, no one is normal.

Pediatrician

their role is to quickly identify problems. Contrary, the nurses experience less time pressure, and due to their preventative care role, they are often facing less severe problems in earlier stages.

THEME 2: THE IMPACT OF DATA-SHARING ON PARENTS AND THEIR RELATION AND COMMUNICATION WITH HEALTHCARE PROFESSIONALS.

Tracking, sharing and discussing data obviously impacted parents. The subthemes below describe this impact in more detail, including how the data affected their relation and communication with their healthcare professionals.

Subtheme 2.1.: Impact of data on parents can be both positive and negative.

Healthcare professional reported both potential positive and negative impact of tracking data on parents. Throughout the study, the healthcare professionals clearly highlighted that tracking data may give parents control over their situation, insight and reassurance. For these reasons, asking patients to keep a (paper) diary was already a common practice for some healthcare professionals. At the same time, they emphasized the potential negative effects that data tracking might have. The nurse (2) explained: *"Parents are really focusing on the things they are tracking"*, and the pediatrician (2) added: *"Yeah, and then we need to explain what is normal. Well, no one is normal."* Being obsessive about the tracking, and the behavior that is tracked, was mentioned as a key risk of data collection by parents of newborns.

Subtheme 2.2.: Mismatching expectations between parents and healthcare professionals on the use of data.

While the parents in our study in general acknowledged and accepted that their healthcare professional would not have unlimited time to review their data, they still often showed to be disappointed when some information did not come across well. For example, the mother of family 3 commented slightly frustrated on their pediatrician's message: *"We already do all the things she recommends, and that should be clearly visible in the data. Did she even look at our data?"* This disappointment was most frequent when the information that was obvious to the parents themselves, still was not picked up by the healthcare professionals.

Furthermore, some families reported to be disappointed when the study had finished and they had to stop collecting and sharing data. For example, when the GP (1) suggested to stop the data collection for family 4, because she felt the need to track was over, the mother reported *"That is a pity. For us it does not feel resolved."* Apparently, their needs and goals according to the data tracking and sharing were different.

Subtheme 2.3.: Data are valuable input in a conversation.

Both healthcare professionals and parents reported the value of using the data in a dialog with each other. Often, the problematic behavior (e.g., crying, spitting up) is not visible during consultation meetings. This is frustrating to parents, and inconvenient to healthcare professionals. Indeed, parents explained to be very willing to share home-collected data to prove the seriousness of what they experience at home. The mother of family 5 explained: *"Earlier, the GP had sent me home because 'it was all within the normal range', but I wish I would have had the home-collected data back then to prove: check this, this is not normal, is it?"* Also, healthcare professionals themselves highlighted the value of using data as supporting evidence when talking to parents, for example, the GP (1) explained: *"From this overview I can immediately see that this baby is not suffering from acid reflux. The data provide helpful evidence to explain this to the parents."*

Subtheme 2.4.: Advantages and disadvantages of a-synchronous communication.

During the study, the parents could track and share data, and send messages at any time. The healthcare professionals mostly checked the dashboard and responded once a week, during or right after an interview with the research team, where the data was actively explored and discussed. For the parents, this a-synchronous communication was helpful and convenient, illustrated by the mother of Family 3 explaining: *"It is nice that at a specific moment, when something happens, you can immediately capture it and share it"*. It gave them a feeling of control, and they were all fine with waiting on a response for a while. On the other hand, the healthcare professionals typically reflected on the disadvantages of a-synchronous communication. For example, the nurse (1) explained, while being overwhelmed by the data: *"This feels cumbersome; a home visit would be much more effective and efficient in this situation. Scrolling through the data I get more and more questions, whereas in a home visit I would immediately get answers."* Also the GP (1) repeatedly emphasized the need to have direct communication with the parents, supplemented with the observation and physical examination of the baby, to clarify the problems. She strongly stated that just home-collected data were not sufficient to understand the situation, let alone to provide appropriate care. A dialog with the patient – possibly based on past data and messages – would fill in the gaps instantly.

THEME 3: SOME INFORMATION CAN, AND SOME CANNOT, BE CAPTURED BY DATA.

This theme describes the potential added value and limitations of data for healthcare professionals.

Subtheme 3.1.: Data can provide objective information.

All healthcare professionals highlight the value of having access to objective information through data. For example, the pediatrician (1) explains: *"It is notoriously hard to get reliable information from anamnesis, for example on how often a baby spits up or cries. Home-collected data really adds value by providing objective information."*

The healthcare professionals explain that this objective information is not only used to have a better understanding of the situation, it also helps to support claims towards the parents. This information is likely to reassure them, as with parents of newborns, situations are often perceived more serious than the objective situation shows.

Subtheme 3.2.: Data can provide temporal (and thus cause-effect) insights.

The healthcare professionals showed a vast interest in temporal aspects of the data. *"Serial information, for example how long the baby cries, and how it relates to defecation and feeds, remains hidden in the anamnesis. Home-collected data could be really helpful to understand cause and effect"* (Pediatrician 2). Having an overview over time also helps to find potential solutions. For example, the nurse (1) reported while exploring the data: *"Look, this day the baby did not easily fall asleep, and the next day she did. Now this is really more information than I would usually have. So now I wonder, did the parents do something different that day?"*

Subtheme 3.3: Data may be hard to interpret.

Seemingly paradoxical to data potentially providing objective information (subtheme 3.1), the healthcare professionals often reflect on issues with face validity of the home-collected data. The GP (2) explained: *"The instruction for data collection needs to be clear and to the point. Otherwise, you risk that you think for example they track crying, but actually they are measuring something else."* This highlights the need for clear agreement on what to track, and what the threshold is for "*pushing the button*". The difficulty of interpreting the data was clearly illustrated by the case of family 4, whose data were shared with both the GP (1) and the pediatrician (1). The pediatrician concluded from the data that the baby was clearly struggling with some form of acid reflux, whereas the GP strongly doubted the baby having acid reflux. The ambiguity of the data was experienced by other families too, for example the mother of family 1 reflected: *"Maybe I should define the data-trackers together with the nurse. I might have different expectations of my baby. (...) She could explain me what is normal, and I could track accordingly."* In an attempt to solve the ambiguity, Family 4 shared short videos of their baby crying, to illustrate what they meant by minimum and maximum crying intensity.

Subtheme 3.4.: Data often lack, although sometimes provide, context and background information.

The healthcare professionals showed time and again the need for background information, in order to accurately interpret the data. For example, when the GP (1) was reading the data: *"Spitting up, up to 2 hours after the feed. Yeah, but then I also need to know how much milk she has had, if it was breastmilk or artificial milk, and if it was thickened. I really need this information in order to accurately interpret the spitting data."* And later: *"Here the baby cried less, but she might have been at her*

How long the baby cries before she falls asleep is not meaningful in itself, it depends on how the parents perceive the length and intensity of the crying. Some parents feel 10 minutes crying is long, other parents are fine with even an hour crying

Nurse

grandparents or the daycare. That information is really required here. (...) We need to have the bigger picture." At the same time, particularly due to the open-ended nature of our data trackers, data showed to *provide* background information in some cases. For example, family 3 provided very precise labels, i.e., the meals were tracked on a range from "0 grams" to a maximum value of "62.5 grams of vegetables" – corresponding to half a jar. From these labels, the pediatrician inferred that the family was very consistent and precise, and this background information helped her to understand that also the other collected data would likely be reliable.

The need for background information varied across the different professions. This was clearly illustrated by the different responses on the same movie showing the baby of family 4 is crying. The GP (1), who was already familiar with this family, stated: *"This is too much information, I really don't need this"*, and contrasting the pediatrician (1), who was meeting this family for the first time, stated: *"This is definitely helpful, it illustrates what happens, how the parents respond to the situation, the setting at home."*

Subtheme 3.5.: Data usually do not capture the (often essential) experience of parents.

In line with theme 1, where we reflected on the value of understanding the parents' problem in order to guide data explorations, we found that for healthcare professionals behavioral data are often not sufficient. Behavioral data only acquires meaning through the experience of the parents. For example, the nurse explained: *"How long the baby cries before she falls asleep is not meaningful in itself, it depends on how the parents perceive the length and intensity of the crying. Some parents feel 10 minutes crying is long, other parents are fine with even an hour crying."* During the study, the healthcare professionals often requested the parents to express their experiences supplement to the data, for example the GP (2) explained: *"If the parents would indicate how easy or tough they experienced that day, additional to the length of crying, that would be really helpful to understand the situation."* The pediatrician (2) added: *"It also guides us where to focus on in the care process: the health issues of the baby, or the perception and well-being of the parents. If a mother cannot cope with half an hour crying, she might suffer from a post-traumatic depression."*

DISCUSSION

This study was designed to understand the evolving practices and value when sharing customized home-collected data with healthcare professionals. More specifically, we provided parents of newborns with an open-ended toolkit that gave them control over what data to collect about their baby. They could share these data with their own healthcare professional, including nurses in preventative care, general practitioners and pediatricians. By observing data-collection, -sharing and -review practices of five families over several weeks, and by regularly interviewing these parents and their healthcare professionals, we learned how experiences, needs and expectations regarding home-collected data evolved over time. It also enabled us to explore how these data-practices affected, and were affected by, communication between parents and healthcare professionals.

In short, our results show the value of sharing home-collected data for both parents and healthcare professionals, for example by providing objective information, or to use in a conversation. At the same time, we find that this value is only capitalized when parents' and healthcare professionals' expectations are aligned. Particularly our iterative approach reveals the critical and evolving role of the parents' question, in order to effectively share and review data. Our results also show the value of customizing data-collection, allowing parents to actively engage and express themselves, which shows to be an informative process in itself for healthcare professionals to observe. In this section, we will discuss these results in more detail.

VALUE AND EFFECTIVE USE OF HOME-COLLECTED DATA

Our results reveal the potential value of sharing home-collected data, as well as the requirements that need to be met in order to effectively capitalize this value. First, we find that home-collected data can provide healthcare professionals with rich insight into parents' experiences as well as into actual health problems through objective and continuous information on the babies' behavior collected in situ. This information adds value to current means of information exchange in clinical visits, e.g. parental reports or physical examination. However, in order to accurately interpret these home-collected data, it is important that there is agreement on how the data are tracked (i.e., which thresholds are applied for tracking particular behavior, e.g., when does crying count as crying) and that sufficient background information is available (e.g., where the baby was at a particular moment, or which type of milk was used). In prior work, also ambiguity of data has been recognized as a main barrier for healthcare professionals (West et al., 2018), and contextual information has been shown to be improve sensemaking of data (Raj et al., 2019). Furthermore, like in prior work (see *Chapter 2* and *Chapter 3*, and Mentis et al., 2017), our participants reflect on data potentially enhancing effective communication. Data facilitate both parents and healthcare professionals to justify and illustrate their claims. For satisfactory use of data in a collaborative sense, it is very important though, that expectations on engagement with the data and goals of data-sharing are aligned. Resembling prior literature showing that patients and healthcare professionals' expectations are not always aligned (Kientz et al., 2009; Lordon et al., 2020; Zhu et al., 2016), our results show that parents typically overestimate the potential and self-explanatory capability of data, and often expect higher levels of engagement and understanding of the healthcare professionals. Healthcare professionals are much more reluctant and express a need to avoid excessive data tracking, to lower the burden to process it, but also to prevent obsessive tracking, potentially resulting in rumination or hypochondria with parents (Gaunt et al., 2014; Lupton, 2013). Lastly, while the a-synchronous communication (i.e., sharing data and messages that are not immediately processed by the recipient) is often considered as convenient by our participants, particularly healthcare professionals emphasize the value of synchronous communication on these data (i.e., in phone calls or visits), facilitating them to immediately get answers to their questions, thus being more efficient.

Beyond these benefits and requirements, which are largely in line with prior literature, our contribution lies in the iterative approach in a realistic setting with an open-ended toolkit. This allowed us to observe how data gradually changed traditional roles and practices, and what collaborative practices revolving around these data can look like. As sharing patient-generated data is a relatively new trend, currently, consensus is lacking on who should initiate tracking, how one should decide what exactly to track, who reviews the data, and how responsibility and ownership are distributed (Chung et al., 2016). We facilitated our participants to collaboratively converge to desired practices. Parents iteratively developed their own tracking tools according to their own situation and needs. Observing this process in itself showed to be highly informative for healthcare professionals, as it revealed much about the family's situation, struggles, and perceptions. In the remainder of this section, we will discuss evolving experiences, needs and expectations more extensively, including alignment between parents and their healthcare professionals, and the role of the parents' questions for effective use of data.

ALIGNMENT BETWEEN PARENTS AND HEALTHCARE PROFESSIONALS

Parents' and healthcare professionals' expectations regarding home-collected data were not always aligned, and this alignment may occur on three distinct aspects (see Figure 9). First, parents and healthcare professionals may or may not be aligned on their understanding of the family's situation and the health issues of the baby. The parents' experiences did not easily come across through sharing data, even though these data were customized by the parents themselves, and often supplemented with annotations and messages. Second, parents often had higher expectations of the potential of sharing data in order to solve their issues. And third, parents typically expected higher investment of the healthcare professionals than actually observed.

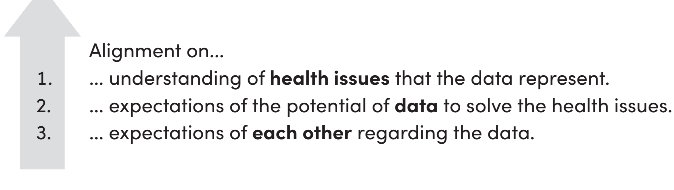

Figure 9 Different aspects of alignment between parents and healthcare professionals, on (1) health issues, (2) data and on (3) each other.

By asking all these questions about the data, you gave me a new perspective on my situation and my baby. I was not aware, I just did not see it.

Mother to nurse

Over the course of the study, we observed that some family-healthcare professional pairs were converging towards an increasing alignment, while others were diverging. Successful alignment of the families and healthcare professionals in our study typically started with aligned expectations of each other regarding the data, followed by alignment on expectations regarding the data itself, and finally, a shared understanding on the meaning of the data and the family's situation could be achieved. Factors that contributed to this convergence included an investment of the healthcare professional in terms of time and commitment, in between synchronous communication about the data and their expectations, and families' high self-reflective capacity. For example, when families took explicit ownership over their data and their problems, this made mutual expectations relatively clear from the beginning, often resulting in a satisfying process for both the family and the healthcare professional. Other families were more awaiting of their healthcare professional to gain value from the data, but when there was substantial communication and investment of the healthcare professional, this led to a shared understanding of data and health problems. In some other cases, families and healthcare professionals were diverging throughout the study, for example due to limited or ineffective communication; their mutual expectations were not aligned, and they did not reach a shared understanding of the meaning of the data. This process was often frustrating, and typically leaded to disappointed parents and information overloaded healthcare professionals.

Not only alignment showed to enable effective collaborative use of data, it also worked the other way around; tracking, sharing and discussing data in some cases enabled alignment. Particularly, being actively engaged in tracking forced parents to thoughtfully capture and express what they experienced, which is an initial step towards effective communication with healthcare professionals. Tracking potentially makes patients more aware of their actual situation and problems (Lavallee et al., 2020; Orji et al., 2018), it gives the healthcare professionals cues for an effective dialog (see *Chapter 2* and *Chapter 3*, and Lordon et al., 2020), and it facilitates talking about patients' values (Chung et al., 2015). Thus, tracking tools can play an important role in enabling alignment. The design of tracking tools should actively engage patients and allow them to express themselves, and they should trigger explicit discussion on goals and expectations on the use of data between patients and healthcare professionals.

DISAMBIGUATING DATA AND THE FAMILY'S PROBLEMS

In order to effectively use home-collected data, our results highlight the importance of understanding the families' perceived problem and their questions to the healthcare professional. Focusing on the question was considered essential to avoid encouraging worries with parents and information overload with healthcare professionals. Somewhat unexpectedly, this focus did not naturally follow from the fact that the parents themselves customized their data collection. Data turned out to be rarely self-explanatory, and disambiguating data was needed in order to achieve a shared understanding of the meaning of data and the underlying problem.

In some cases, data were used in an exploratory way, where disambiguating data helped unraveling the problem. In other cases, when healthcare professional gained sufficient understanding of the problem perceived by the parents, data could be employed to solve these problems, i.e., used in a confirmatory way. It might be tempting to conclude that confirmatory use of data requires standardized tools, which captures pre-determined data given a particular health issue. We proposed this concept in the study to the participants, and while this idea was indeed appreciated by the healthcare professionals, the parents were much more skeptical and doubted whether standardized tools would be sufficient to express their problems. In other cases, where data are used exploratory, flexible or customizable tools are a tempting solution. However, in the exploratory cases in our study, when there was not yet a clear question available, also some guidance and focus had shown to be effective. For example, when parents were asked to formulate their experiences explicitly, and to illustrate this with data, this process has shown to be a catalyst in unraveling problems and solving them accordingly.

So, tools that facilitate sharing home-collected data would ideally have a hybrid form, with a possibility to adopt standardized and pre-determined tracking plans, at the same time, flexible enough to allow for customized input. Adding to prior work where healthcare professionals have dismissed patient-initiated tracking as often clinically irrelevant or unreliable (Chung et al., 2016; Zhu et al., 2016), our results show that patient-initiated tracking can be highly informative for healthcare professionals, particularly when problems are yet less clear. When patients initiate tracking, or even stronger, when they define their own tracking tools (e.g., by formulating their own labels or by annotating data-points), this may comprise rich information about their problems and daily lives. Therefore, we argue that tools should allow for both patient- and professional-initiated tracking, and give particularly patients some room for customization. It has to be noted, though, that deciding upon what to track to effectively meet tracking goals may be a difficult process. Prior literature points to typical pitfalls of quantified-self tracking practices, such as tracking too many things, or focusing on outcomes rather than triggers (Choe et al., 2014). It may be beneficial to make the process of defining tracking tools a collaborative act, where healthcare professionals discuss the implications of different tracking-practices with patients.

LIMITATIONS AND FUTURE WORK

Based on the time and resources available for this study, and to limit the burden on participants, we limited the number of participants and the length of the study. Still, we found that the period of about five weeks for sharing home-collected data was suitable and sufficient, as for all families near the end of the study the insights were saturated, from the researchers' perspective as well as for the families and healthcare professionals. Most of the data collection terminated naturally before the end of the study. In terms of the number of participants, we could only include three healthcare professionals in the main track due to our highly labor-intensive approach. An additional four healthcare professionals were included in the reflective track to validate our results with a larger sample. This limited number of participants has both disadvantages and advantages. We observed substantial differences in attitudes towards the value of these data and desired practices across the healthcare professionals in the main track. This means that our obtained results were not saturated and may not be fully representative for a larger group of healthcare professionals, especially in areas outside newborn care. However, our learnings largely resonated with the healthcare professionals in the reflective track and with findings in prior literature, indicating that our results are valid. Including a small number of participants also provided benefits, as it allowed for closely and attentively following the data collection, sharing and reviewing practices, experiences, and resulting (information) needs and communication between parents and healthcare professionals. We were able to quickly respond to evolving needs with new concepts and data visualizations. Thereby, we deepened our understanding of the implications of sharing home-collected data.

The toolkit and dashboard were specifically developed for this study, so the participants were by design presented with tools and information unfamiliar to them. Furthermore, the prototypes we worked with were inevitably immature. This might have reinforced the perceived burden of dealing with data collection, sharing and reviewing, and should be validated in future research with further developed tools.

To reflect on generalizability of our results, we should consider the context of this study: parents and their newborns. Caring for newborns comprises high uncertainty and responsibility (Lupton, 2011), as newborns are vulnerable and can only limitedly express themselves. Therefore, parents may experience a relatively high need for control and reassurance, compared to other patient groups such as chronically ill people. It may be expected that other patient groups are potentially less dedicated and might have lower expectations, possibly changing the value and practices of sharing data.

Future research considering patient-generated data may benefit from following participants over time. Our results show that by doing so, we obtained a deeper and more nuanced understanding of participants' needs when working with data. It shows that peoples' needs and perceptions can indeed change, and it reveals how healthcare professionals and patients may converge towards shared expectations and understanding.

CONCLUSION

This chapter reflects on a study where we attentively empowered parents of newborns to share customized home-collected data with their healthcare professionals, and followed their practices and experiences over time. We contribute to established research in this field with a highly ecologically valid approach, allowing for a nuanced understanding of the – continuously evolving – value and practices when sharing home-collected data. Indeed, many things had changed over the course of the study. We observed how parents, by tracking their baby's behavior and re-defining their tracking tools, improved their understanding of their problem. Healthcare professionals were observing this process and discussing the evolving data with parents, and this enabled alignment of expectations and a shared understanding of the meaning of data and the family's situation. Because the value of home-collected data emerged and evolved within this process of defining data-trackers and discussing these data, we argue that tracking tools should provide room for customization of data-collection, and facilitate collaborative practices for deciding upon what to track and what the data mean. When patients and healthcare professionals are aligned on what to expect from each other and from the data, this enhances a shared understanding of health issues and solving them accordingly.

I do not change my
the tool is much
experiences, and
better explained.
past shows that
in particular
large influence. My
to try to be more
over the week, and
session with the
about other factors.

Coach

advice. But now
clearer, the factors,
correlations are
Data from the
physical activity,
exercise, has a
advice is still
physically active
to have an intake
client to learn more

Tailoring Transparency to Expertise: Health Professionals' Need for Transparency in Representing Self-Tracking Data of Clients

CHAPTER 5

Tailoring Transparency to Expertise: Health Professionals' Need for Transparency in Representing Self-Tracking Data of Clients

In the previous chapters, we have gained rich insight in the potential value of health data for coaches and the coaching process. Our results show that coaches are not merely positive about the impact of data. For example, they are concerned that data will overemphasize the behavioral aspects of the client, distracting from the client's experiences (*Chapter 2*). Furthermore, we have observed that data occupy coaches and client with another task: making sure the other sees the data in the right perspective (*Chapter 3*). Thus, data do not only bring value, they also bring extra work, and alignment of expectations is critical in this regard (*Chapter 4*). The coaches in our studies suggested that data not necessarily help, but may also harm the coaching process, and they showed to be motivated to avoid potential negative effects of data as much as possible (*Chapters 2, 3* and *4*).

HEALTH COACHING IN THE FRAMEWORK OF AUTOMATION; AUTOMATE THE *ANALYSIS* OF DATA?

For a closer investigation of the coaches' perspective on the benefits and risks of data, we draw from the automation framework of Parasuraman and his colleagues (2000). They distinguish between four different types of automation, which we will apply to health coaching (see Figure 10). The first type is *information acquisition*; in terms of health coaching this refers to collecting health data. Coaches were generally appreciative of *high* levels of automation of this task; they valued the use of wearable devices for collecting information that they would otherwise have limited access to. Coaches also expressed a clear preference for *low* automation on the tasks *decision selection* and *action implementation*, which in health coaching may refer to selecting the coaching advice or motivational statement and actually give this to the client. Coaches expressed aversion to e-health technologies autonomously sending out recommendations or motivational messages, as it would lack a holistic view on the client and it may trigger negative emotions, value judgements or obsessive behavior.

There is one more type of automation that Parasuraman and his colleagues (2000) present, that is *information analysis*. For health coaching, this refers to analyzing or modelling health data, for example in order to detect trends and correlations, possibly resulting in actionable insights (e.g., drinking less coffee will likely improve your sleep quality). Coaches did not have a clear preference for the level of automation of this particular task. They were not necessarily against automating this data analysis, but they were skeptical on the feasibility. They argued that every

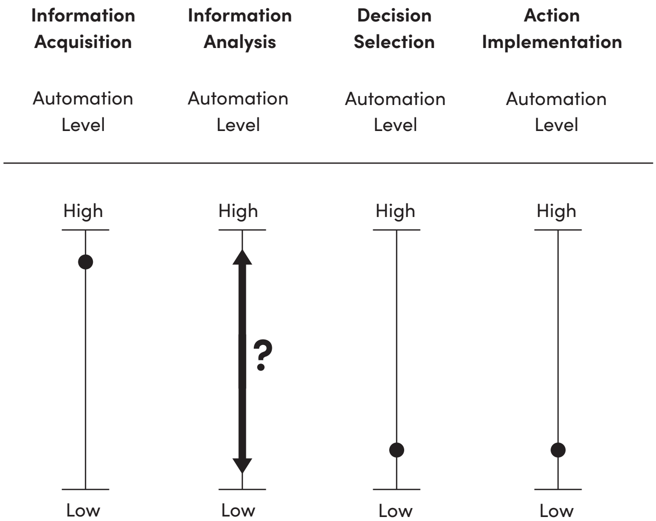

Figure 10 Based the conceptual framework of Parasuraman et al. (2000). This represents the different types of automation in the health coaching process, and coaches' preferred levels of automation on each of these types.

client is unique, and therefore they did not expect standardized models to work. For one, because models rely on measurable behavior as proxy for general health and wellbeing, and they expect this would not capture the complete picture required to serve as effective input for health coaching. The discussion whether analysis of data should be automated is also raised by Chung and her colleagues (2019). They observe that so far, research mainly revolves around the question whether or not to automate *data tracking*, and they suggest extending our focus towards the automation of *data analysis*. Particularly, they point to the tension between efficiency on the one hand, that requires high levels of automation, and complete information on the other hand, as automation might obscure essential information for the coaching process.

We should note, however, that the framework of automation as presented by Parasuraman and his colleagues (2000) gives an overly simplified view on interaction with technology, as also discussed in *Chapter 1*. Tasks cannot be simply be pulled apart in subtasks and distributed across humans and technology, and automating certain tasks not necessarily alleviates the user's workload, rather it changes it (Dekker & Woods, 2002). In the context of health coaching, it is likely that a complex interplay between coaches and data will arise. For example, high automation of data analysis may imply that coaches seek to understand and check the analysis methods, outcomes may inspire coaches and guide their own thinking, it may shift coaches' focus to experimenting with certain behaviors to find cause and effect (c.f., Karkar et al., 2016), among other things. In turn, coaches' input into these analyses may make it more complete, appropriate, and accurate.

So far, in the previous chapters, we did not apply models that extensively analyzed clients' health data. This limits our understanding of the dynamics between coaches and data, especially when data take a more prominently role. Therefore, in the next chapters, we will increasingly focus on data by analyzing them through data-driven models. This will broaden our understanding of coaches' perceptions, attitudes and needs from lower levels to higher levels of automation of information analysis. We argue it is vital to allow coaches to interrogate, judge and control how data are used, even, or especially, when data are more extensively analyzed. First, we will elucidate different types of coaching situations, where data and coaches play substantially different roles.

COACHES AND DATA TAKING DIFFERENT ROLES IN DIFFERENT COACHING SITUATIONS

Our findings in the previous chapters clearly show that oftentimes, data serve as input for conversation and exploration; data were not meaningful in themselves, yet they gained meaning when talking to the client about her goals, motivations and experiences. In these cases, health issues and goals typically had a subjective nature, that is, they 'existed' in terms of the experience of the client, and thus problems and progress were only accessible through conversation with the client. For example, wellbeing and fitness levels can mean very different things to different people. Or, health goals can revolve around increasing self-esteem, even though these problems are often initially presented as weight-loss goals. In contrast, in other cases health issues and goals had a more objective nature. For example, we encountered clients who were recovering from injuries, who wished to increase their breast milk production, or who had specific sports goals such as running a marathon. While, of course, there are subjective experiences underlying these goals, it is relatively easy to interpret data and measure progress without extensive discussion with the client. In such cases, data analysis can add considerable value to the coaching process.

This is not a dichotomous distinction; all health goals have objective and subjective aspects, to a more or lesser extent. The level of objectivity does have important implications for the focus in the coaching process, and thus for the use of data. When goals have a largely subjective nature, data are most likely used in exploratory ways, facilitating the interpersonal exchange between coach and client where goals and experiences are unraveled. When goals have a more objective nature, then data can have a more straightforward purpose and use. When goals and progress are, to a large extent, measurable, it is worthwhile to utilize data in well specified ways. For example, coaches may seek to optimize training programs, or find causes for complaints.

EXPLORING EFFECTIVE COLLABORATION BETWEEN COACHES AND DATA-DRIVEN MODELS

Thus, the unique contributions of coaches and data may substantially shift across these different coaching situations. In cases where health goals and progress are to some extent measurable, the health coaching process may largely benefit from data analysis. After all, models perform generally better on quantitative prediction tasks than humans (Grove et al., 2000). Still, we want coaches to actively engage in this process, such that they can add their unique perspective, based on their domain knowledge and their view on the client. In the remainder of this dissertation, we seek to understand coaches' experiences and needs when faced with data-driven models, to be able to understand what constitutes effective and satisfying means of interaction between them.

Based on our findings in the previous chapters, we make several assumptions of what is needed to facilitate effective collaboration between coaches and data when data are more extensively analyzed. First, we expect that it is important for coaches to be recognized in their expertise, and to account for their investment in the client and the coaching process. This means that, when presenting them data-driven models, we should provide them sufficient and appropriate information about how these models work. This allows them to compare this with their own understanding of the client and their knowledge on health and wellbeing in general, and as such, fairly judge the competence and accuracy of the model, not resulting in under reliance nor over reliance of such systems. We will take a first step towards this goal in the current chapter. Second, and we will save this for *Chapter 6*, we expect it is important to facilitate coaches to interact with the model, as they may add their unique perspectives on the client.

Specifically, in this chapter we investigate the effect of model transparency on coaches' levels of trust and acceptance of data-driven health recommendations. Transparency potentially allows coaches to compare data-driven models with their own understanding of the client and their knowledge on health and wellbeing in general, and as such, fairly judge the competence and accuracy of the model, not resulting in under reliance nor over reliance of such systems. Furthermore, we explore whether knowledgeable coaches have different needs compared to laypeople regarding transparency. We conducted an online study using a data-dashboard with different levels of transparency, and used participants with various expertise levels in both health coaching and modeling methods. Our results indicate that highly experienced coaches indeed have different needs than laypeople and less experienced coaches when it comes to transparency. This study serves as a first step in illustrating the importance of designing interfaces recognizing the knowledge level of the user.

This chapter is derived from:

Rutjes, H., Willemsen, M. C., & IJsselsteijn, W. A. (2020). Tailoring Transparency to Expertise: Health Professionals' Need for Transparency in Representing Self-Tracking Data of Clients. *HUMANIZE Workshop at the ACM Conference on Intelligent User Interfaces (IUI'20)*. Cagliari, Italy.

INTRODUCTION

Wearable self-trackers are readily available and increasingly used. They are not only beneficial to their users, for example by enhancing self-awareness (Kersten – van Dijk et al., 2017) and positively influencing physical activity levels (Brickwood, Watson, O'Brien, & Williams, 2019), but also offer potential benefits for health professionals. For one, self-tracking data potentially provide longitudinal and in-situ insights from a client's daily life. Current self-tracking devices assist users mainly by collecting data and presenting simple data summaries, for example, active minutes per hour or per day, calorie intake per meal, or sleep duration. There is ample room for more advanced modeling approaches beyond summarizing and visualizing self-tracking data (Ohlin & Olsson, 2015). Health professionals have expressed a need for more extensive data processing, which allows them to quickly identify trends, correlations and critical events. Using such inferences, in turn, will facilitate professionals to propose more effective lifestyle recommendations or medicine adaptions (Chung et al., 2015; Raj et al., 2017; West et al., 2018). Prediction models, relating a client's behaviors to her well-being measures, would facilitate this process.

At the same time, health professionals are critical towards the competence of systems to provide meaningful support, and are anxious about losing control over the coaching process (see *Chapter 2*). They express a need for transparency in order to verify and trust such systems' decisions (Bussone, Stumpf, & O'Sullivan, 2015; Gagnon et al., 2016). However, prior literature shows that, paradoxically, too much detail in transparency can erode trust (Kizilcec, 2016; Springer & Whittaker, 2019), particularly when the provided details are conflicting with a user's mental model. Users with high domain knowledge are known to be more critical towards supporting technology (Fogg & Tseng, 1999) and less attentive to explanations (Lim, Dey, & Avrahami, 2009). Possibly, their richer mental models increase the risk that transparency reveals a mismatch with that mental model. Technical literacy, i.e., knowledge on the inner workings of a system, on the other hand, potentially makes users more lenient towards transparency provided by such systems.

We set out to test whether both domain and modelling expertise moderates the influence of transparency level on trust in those systems. We developed a health-coaching dashboard presenting self-tracking data of a (fictitious) client, including a recommendation considering the optimal coaching advice. Our recommendation was based on a regression model linking the client's behaviors to well-being measures, which was revealed in the transparent dashboard. We measured the participants' trust in the dashboard and the congruency between their own coaching advice and the dashboard's recommendation.

In the remainder of this section, we will discuss related work, and present our research questions.

HEALTH PROFESSIONALS' NEEDS AND ATTITUDES TOWARDS SUPPORT SYSTEMS

Health professionals (e.g., dieticians, sports coaches, and clinicians caring for chronically ill people) potentially benefit from the self-tracking health data of their clients. For example, self-tracking data provide professionals with in-situ information from their daily lives (Yoo & Choudhury, 2019), it helps to understand the effects of training loads (Cardinale & Varley, 2017), and it provides them with more reliable and objective information on lifestyle behaviors than clients' self-report (see *Chapter 2, 3* and *4*). Furthermore, self-tracking data facilitate understanding behavioral trends over time (Raj et al., 2017) and causal relations between behaviors and symptoms, which, in turn, supports diagnosis and personalized treatment plans (Chung et al., 2015).

Despite the efforts of designers of self-tracking devices to summarize and visualize the data in meaningful and understandable ways, interpretation of self-tracking data is not a straightforward task (Choe et al., 2014). For clients, it is hard to relate presentations of their health data to their health-related questions and gain actionable insights (Choe, Lee, et al., 2017; Choe et al., 2014; Raj et al., 2017). Health professionals use and evaluate self-tracking data differently than their clients (Raj et al., 2017). For example, they asses the data not from a first-person perspective, such that they miss the lived experiences related to the data. Furthermore, their level of clinical knowledge is higher. Still, interpreting self-tracking data is not easy for professionals either. The data summaries and visualizations that are currently often shown in the interfaces of self-tracking devices cannot seemingly be used in coaching practices (Raj et al., 2017; West et al., 2018). To understand and effectively coach a client, they are often looking for trends, critical events, and correlations in the data (Chung et al., 2015; Mentis et al., 2017; Raj et al., 2017; West et al., 2018), and more extensive data processing is needed to support them in this process, e.g., prediction models that link a client's behaviors to her well-being. When not presented effectively, self-tracking data potentially add even to the professional's workload, when they need to make an effort in processing the data transforming it into meaningful information. Indeed, this is considered as one of the ironies of automation (Bainbridge, 1983).

Even though health professionals generally express a need toward more extensive processing of self-tracking data, they are also hesitant to rely on automated analysis and technological support for this purpose. First, health professionals are skeptical about whether support systems are capable of making meaningful interpretations of self-tracking data, for example, because they believe that such data are unreliable and ambiguous (West et al., 2018), or because the measurements put too much emphasis on a client's behavior rather than lived experiences (see *Chapter 2*). Using inaccurate or incomplete data as input, professionals believe that it is not feasible to automatically interpret these data meaningfully. Moreover, health professionals indicate that support systems such as Electronic Health Records often inhibit informal and unstructured communication, which is a highly valuable process in healthcare and coaching (see *Chapter 2, 3* and *4*, and Marcu et al., 2014). Furthermore, when the data involve more measures than those relevant for the health problem at hand, they fear the emergence of new health problems, as a result of rumination and obsessively focusing on irrelevant data (Gabriels & Moerenhout, 2018; Kollenburg et al., 2018). Thus, health professionals frequently question the competence of supporting technology to make a meaningful contribution to the health coaching process. They become cautious when such systems are too obtrusive, and want to stay in control over the coaching process and the role that self-tracking data take inthat process.

When people are asked what they need in order to trust an intelligent agent, transparency is mentioned as an important aspect (Glass, McGuinness, & Wolverton, 2008). User studies in clinical contexts show that when clinicians are collaborating with decision support systems, clinicians explicitly request transparency in order to verify the system's decision (Bussone et al., 2015; D. Wang, Yang, Abdul, & Lim, 2019). Also when using machine learning in healthcare, transparency has been addressed as an important prerequisite (Gui & Chan, 2017; Wiens & Shenoy, 2018).

TRANSPARENCY AND TRUST

Transparency is associated with many benefits, for example, it can improve users' mental models of a system, which enhances system adoption (Cramer et al., 2008; Muramatsu & Pratt, 2001) and improves task performance (Lim et al., 2009). It can also help to lower algorithmic anxiety and improve perceived fairness and level of control (Jhaver & Antin, 2018). Users particularly value explanations that are well-interpretable and in line with their own mental models (Eslami, Krishna Kumaran, Sandvig, & Karahalios, 2018), supplementing current knowledge (Coppers et al., 2018; Desai et al., 2019).

An important benefit of transparency is that it is a means to foster trust. Trust is one of the main drivers of system reliance (J. D. Lee & See, 2004), or more specifically, of adoption of a system's recommendation (Xiao & Benbasat, 2007). Trusting beliefs can be distinguished into three categories: *competence* (the ability to do what the user needs), *benevolence* (acting in the user's interest), and *integrity* (being honest and adhere to moral values) (Mayer, Davis, & Schoorman, 1995; McKnight, Choudhury, & Kacmar, 2002). W. Wang and Benbasat (2007) have shown that various types of explanations (i.e., how-, why- and trade-off-explanations) may be used to positively influence specific trusting beliefs (i.e., respectively, *competence*, *benevolence,* and *integrity*). As health professionals mainly doubt the competence of support systems, we will focus on *how*-explanations in our transparent conditions. Furthermore, previous studies have used various dependent measures to test the effect of transparency, e.g., trust, perceived credibility, system acceptance or recommendation adoption. In the current study, we decided to measure trust directly, as well as the congruence of the participant's coaching advice with the system's recommendation.

Regardless of the benefits, transparency is not solely for the better, and will not foster trust in and of itself (Bannister & Connolly, 2011). User studies on transparent systems show mixed results on how transparency affects users' perceptions and beliefs. Kizilcec (2016) studied students' trust in a grading system, where part of the students was presented with information about the procedure of the grading, while others received additional detailed information about specific grades. Whereas procedural information fostered students' trust, the detailed information eroded trust, particularly when the outcome violated the students' expectations (Kizilcec, 2016). Springer and Whittaker (2019) studied the effect of transparency on trust, through users' evaluations of an emotional text analysis algorithm. They similarly showed that high transparency, where the algorithm provided feedback on word-level rather than on document-level, resulted in underestimation of the performance of the system (Springer & Whittaker, 2019). Transparency reveals a certain complexity, and when that is not in line with the user's mental model on the inner workings of the system, this potentially results in dismissing the system. In line with this, Eslami and her colleagues (2018) found that when people are presented with transparent reasons why they are receiving certain advertisements, too much detail in the explanation can be experienced as creepy. Furthermore, it can cause users to be overly critical on details of the explanation that do not match their self-image (Eslami et al., 2018).

In these previous studies, different levels of transparency have been determined and used, for example, global versus local explanations (Klein, Hoffman, & Mueller, 2019), and procedural transparency versus additional data (Kizilcec, 2016). These classifications roughly make the same distinction; there is a medium transparency level on the one hand, where only general procedures or rationales are given, and a high transparency level on the other hand, where additional specific, often numerical, information is provided. The level of detail in transparency seems to be a critical factor influencing trust. Providing more detail increases the risks of revealing a mismatch between a user's mental model of how a system works, and a system's actual inner workings. To conclude, while health professionals request high transparency as an important prerequisite for trust and adoption of support systems, prior literature indicates that high transparency potentially erodes trust, especially when the transparency conflicts with the user's expectations.

THE ROLE OF EXPERTISE

Domain experts have richer mental models than laypeople and novices on the domain the system is working on. Prior literature shows that users who are familiar with the topic, are more critical on support systems, and perceive them as less credible (Fogg & Tseng, 1999). It may even happen that expert users reject those systems, while agreeing with the outcome, because they perceive the explanation as too simplistic (Cramer et al., 2008). Users who have low self-reported domain knowledge are more likely to adhere to systems as a result of transparency than users with high self-reported domain knowledge (Schaffer, O'Donovan, Michaelis, Raglin, & Höllerer, 2019). Interestingly, this effect disappears when using actual knowledge as a predictor. This highlights the importance of how knowledgeable users perceive themselves, rather than their actual knowledge, for understanding their attitude towards support systems.

Knowledge about the domain (e.g., health) is essentially different from knowledge about modelling the domain (e.g., regression models, decision trees, etc.). Lim, Dey and Avrahami (Lim et al., 2009) showed that high levels of domain knowledge makes users less attentive to explanations, even while these explanations actually covered the modelling method rather than the domain. In this sense, high domain knowledge actually risks making users less diligent to learn from explanations, possibly resulting in inaccurate understanding of the inner workings of a system. High modelling knowledge, on the other hand, potentially makes users more lenient towards support systems, and this has indeed been shown to positively influence trust in algorithmic decisions (Cheng et al., 2019).

One may argue that mental models of experts represent a deeper awareness of the true complexity of the domain. Contrary, for prediction tasks, it has been shown that experts can be too rigid in their judgement, trying to fit each and every single case, potentially resulting in overfitting (Tazelaar & Snijders, 2004). For quantitative prediction tasks, human decisions are found to be often inferior to algorithmic decisions (Grove et al., 2000). Tazelaar and Snijders (2004) have found that, even though algorithms outperform experts, systems where algorithms are supplemented with an average across experts' decisions outperform single use of algorithms. This illustrates the value for experts to collaborate with algorithms when making quantitative decisions. Health coaching is a highly social and contextualized activity (see *Chapter 2, 3* and *4*, and Jones & Wallace, 2005), still, decisions regarding what to advise to the client potentially benefit from insights based on the client's self-tracking data, which is, in essence, a quantitative prediction task.

RESEARCH QUESTIONS

Taken together, health professionals – considered as experts in the health-coaching domain – may considerably benefit from systems supporting them with assessing their clients' self-tracking data. Prediction models, connecting the client's behavioral measures to her health or well-being, supports professionals in understanding the implications of certain coaching advice. Professionals express a need for this more extensive modelling of self-tracking data in order to use such data effectively, and they indicate transparency is an important prerequisite for them to trust and use such systems. While this desire is clear, it is less clear how health professionals actually deal with high transparency levels in practice. Transparency brings the risks of revealing redundant details resulting in information overload, or details unaligned with the user's mental model. This may be particularly critical for domain experts who have well-developed mental models, and potentially erodes trust. This results in the following research questions:

1. How do different levels of transparency (non, medium and high) affect trust in health coaching support systems based on clients' self-tracking data, and to what extent will the user's advice be aligned with the system's recommendation?
2.a. How does the user's level of domain expertise moderate the effect (in RQ1) of transparency on trust and advice congruency?

Orthogonal to domain expertise in health coaching, also modelling expertise might play a role. In our study, we show self-tracking data in a dashboard, and use a regression model (revealed in the transparent dashboard) for predicting a coaching advice. So, experience with viewing self-tracking data, and familiarity with regression modeling, potentially influences how users perceive the dashboard, and how they respond to the level of transparency. For example, users familiar with regression models would probably value seeing a detailed regression equation, whereas others might be deterred by such a level of detail. This results in the following research question:

2.b. How does the level of modelling expertise (i.e., expertise with regression modelling, and experience in viewing self-tracking data) moderate the effect (in RQ1) of transparency on trust and advice congruency?

METHODS

PARTICIPANTS AND PROCEDURE

We recruited 111 participants in total. We targeted different groups of participants, to make sure our data would show sufficient variance on both domain expertise in health coaching, as well as expertise on regression models. First, we recruited 21 health coaches (working with clients on physical exercise, diet, or lifestyle) through the alumni network of a sports academy, and via e-mails and phone calls to coaching practices. We expected them to have high expertise in health coaching. Second, we recruited 26 students (3^{rd}-year) from a sports academy, expected to be novices in health coaching. Lastly, we recruited 64 Bachelor students (mostly 2^{nd}-year) through mathematics and statistics courses in university, either via e-mail or during class. This group was expected to have very limited experience with health coaching, yet possess moderate to high levels of knowledge of regression models.

The 111 participants (50 females, 2 preferred not to say their gender) were on average 23 years old (*SD*=5.38). All participants independently filled in an online survey; some students did this during class. The survey started with an informed consent form, after which we presented the non-transparent dashboard, to measure trust and congruency as a baseline. Thereafter, we presented the (medium or high) transparent dashboard, and again measures trust and congruency. Finally,

Figure 11 Screenshot of the dashboard. The bar charts represent various behavioral and well-being measures of the (fictitious) client over a week, including sleep duration and minutes awake, coffee and alcohol consumption, light physical movement, exercise, sedentary time, steps and subjective sleep score. In the top left corner a recommendation was provided (see Table 9 for details).

we measured demographics and expertise. Completing the questionnaire took on average 20 minutes. We compensated health coaches with a €10 voucher for their participation. Among every three students, we raffled one €10 voucher. This study was approved by the local ethics committee at Eindhoven University of Technology, Human-Technology Interaction group.

THE DASHBOARD

We developed a dashboard for health coaches showing (fictional) self-tracking data of a client named Sam (see Figure 11). We kept the visualization and type of self-tracking data similar to current commercially available tools, to make the case as realistic as possible. On top of that, we recommended a coaching advice for Sam, suggesting that based on the data, a coaching advice regarding exercise would potentially be most beneficial (see Table 9). The fully functional dashboard was built in Tableau and embedded in the survey. The participants were asked to imagine that Sam would visit them tomorrow asking for coaching advice, and to use the dashboard to prepare for his visit. Sam's problem was described as follows: he is often awake at night and feels tired during the day. He tracked various data for a week, partly automatically through his health watch (i.e., total time in bed and actual sleep duration, light physical activity, exercise, sedentary time, and steps), partly manually (i.e., subjective sleep score, and coffee and alcohol consumption). The data was shown in interactive bar charts; hovering over the bars provided numerical details and highlighted the other measures of that particular day (see Figure 11).

Recommendation in the dashboard	Transparency level		
	Non	Medium	High
The dashboard generates the following recommendation for Sam: Exercise has the most positive influence on Sam's sleep duration. Thus, advice regarding exercise has the largest potential to improve his sleep duration.	X	X	X
Explanation of the recommendation Based on data from a large group of people, the influence of life-style factors (such as less use of coffee, more exercise, or more time in bed) on sleep duration was determined.		X	X
This is shown in the following formula: Sleep [hours] = Time in bed [hours] – 1 – 0.06 x Coffee [number of cups] – 0.09 x Alcohol [number of glasses] + Exercise [hours] + 0.2 x Light physical activity [hours]			X
The dashboard has calculated this for Sam. It is found that when Sam starts exercising more, this is expected to have the largest positive effect on his sleep duration.		X	X

Table 9 Different levels of transparency in explaining the recommended coaching advice.

LEVELS OF TRANSPARENCY

All versions of the dashboard provided the same self-tracking data, along with a recommended coaching advice for Sam in the top left corner of the dashboard (see Figure 11). The recommendation stated that a coaching advice regarding exercise would have the most positive influence on Sam's sleep duration. A regression model was underlying this recommendation, which is a common model for this type of task. The non-transparent dashboard provided the recommendation without explanation. The medium-transparent dashboard explained the recommendation by providing an abstract rationale for regression (see Table 9). In the high-transparent dashboard, the rationale for regression was supplemented with more details on how the influence of lifestyle factors on sleep are quantified, by showing the regression equation (see Table 9). As the literature discussed in the introduction shows, gradually increasing transparency levels can have many appearances. We decided to provide the regression equation as maximum transparency, as it both reveals a quantification of the effects in this particular case (through the coefficients) as well as something on the general procedure, (c.f., Klein et al., 2019). In the spirit of (W. Wang & Benbasat, 2007), our explanation comprises a *how*-explanation (opposed to for example a *why*-explanation), targeting competence-related trust issues of health coaches. The regression equation itself was informed by general guidelines and studies on sleep.

All participants first examined the non-transparent dashboard, after which their initial trust and initial advice was measured. Subsequently, half of the participants was randomly assigned to the medium-transparent dashboard, the other half to the high-transparent dashboard, again followed by a trust and congruence measurement.

		Item-scale correlation	
		T1	**T2**
1.	I feel comfortable using this dashboard	0.73	0.82
2.	I believe that these types of dashboards are effective	0.83	0.85
3.	I think using the dashboard's recommendation leads to a positive outcome	0.79	0.78
4.	I do not trust this recommendation	−0.48	−0.61
5.	I believe this recommendation is useful	0.63	0.73
6.	If I were a health coach, I would not intend to use similar dashboards in the future	−0.70	−0.71
	Chronbach's alpha	*0.78*	*0.84*

Table 10 Questions measuring Trust used on a summative scale.

This enabled measuring increase or decrease in trust (*DeltaTrust*) as well as the congruence between the coaches' advice and the recommended coaching advice, as a result of the increased transparency level. The repeated measures allowed us to measure more precisely, which was worthwhile considering the relatively small sample. Note that the difference between medium- and high-transparency, though, is measured between-subject.

MEASURES

Congruence with Recommended Coaching Advice

After the examination of the non-transparent dashboard, the participants picked a topic from a list where their own coaching advice would mainly focus on (i.e., sleep, coffee, alcohol, light physical activity, exercise, sedentary time, or steps). This allowed us to measure whether their advice was congruent with the recommendation of the dashboard (suggesting to focus on exercise). Additionally, the participants were asked to write down their coaching advice in an open text field, to give them room to elaborate on their advice.

After the participants were presented with the transparent dashboard (i.e., the second time they assessed the dashboard), the participants were asked to reconsider their coaching advice as a second measure of congruence with the transparent dashboard.

Trust

Trust was measured twice, as a baseline after the non-transparent dashboard (*T1*) and after the transparent dashboard (*T2*), by six questions on a 7-point Likert-scale (1 = totally disagree, 7 = totally agree; see Table 10). As previous literature indicates health professionals mainly doubt the competence of support systems, we focused the questions around competence-aspects. We reversely coded questions 4 and 6.

Basically, I stick to the same advice as before. I do like it that I now know how the tool has generated its advice. This allows me to explain it to Sam if he asks for that.

Coach

1. Domain expertise: Working experience as coach *“How much experience do you have working with individual clients?”*	
No experience	62 (56%)
<1 year	13 (12%)
1-3 years	19 (17%)
3-5 years	9 (8%)
5-10 years	4 (4%)
>10 years	4 (4%)
2.† Expertise self-tracking data *“How much experience do you have with viewing data from self-trackers/wearables? For example, because you or someone you know is wearing one.”*	
A lot	19 (18%)
A little	54 (52%)
No experience	31 (30%)
3.† Expertise regression models *a. “The dashboard that you just viewed [...]. To what extent are you familiar with regression models?”*	
Very familiar	33 (32%)
Somewhat familiar	54 (52%)
Not familiar	16 (16%)
b. “How many courses did you follow in which regression was discussed?”	
0	15 (15%)
1	25 (24%)
2	37 (36%)
3	16 (16%)
4 or more	10 (9%)
c. “In the regression model below, can you indicate which predictor (A or B) has the most influence on the outcome of Y? $Y = 5 + 3A + 2B + e$”	
I don’t know	31 (30%)
A	47 (46%)
B	0 (0%)
It depends on the unit and variance of A and B	25 (24%)
† As we included question 2 and 3 in the questionnaire only after the first group of students (n=7) already participated, these questions are only filled in by the remaining 104 participants.	

Table 11 Frequencies (and percentages) of the participants’ distribution over the expertise variables.

The items were combined on a summative scale, resulting in a Chronbach’s alpha of 0.78 and 0.84, respectively at (T_1) and (T_2).

Expertise

Domain expertise was measured by the number of years they have been working as a coach (see Question 1, Table 11). For the analysis, the categories 3-5 years, 5-10 years and >10 years of working experience as a coach are combined into one category, because of the limited number of observations in each of these cells. For the sake of clarity, when discussing the results, we refer to the different levels of domain expertise as *laypeople*, *novices*, *intermediates* and *experts*, respectively for having no, less than a year, one to three years, and more than three years of working experience as a coach.

Regarding modelling expertise, we asked the participants how much experience they had with viewing self-tracking data (see Question 2, Table 11). Furthermore, we measured the participants' expertise in regression models. For this, we combined both self-reported expertise with more objective measures of expertise (Question 3a-c, Table 11), because prior research has found effects of self-reported expertise rather than actual expertise on how people respond to transparent systems (Schaffer et al., 2019). Yet, in the analysis, we did not find differences when using solely question 3a (self-reported expertise), or a combination of the questions 3a-c. Therefore Questions 3a-c were combined on a summative scale (Cronbach's alpha = 0.64).

Domain expertise (i.e., working experience as a coach) and expertise on viewing self-tracking data were negatively correlated ($r=-0.42$, $p<0.001$). This was moderated by age; age was negatively related to expertise on viewing self-tracking data, and positively related to domain expertise. This correlation did not largely impact multicollinearity of the predictors in our analysis, as the maximum VIF = 2.72 (mean VIF = 1.68).

Regression modelling expertise was quite evenly distributed over the levels of domain expertise, see the scatterplot in Figure 12, facilitating an estimation of both types of expertise separately.

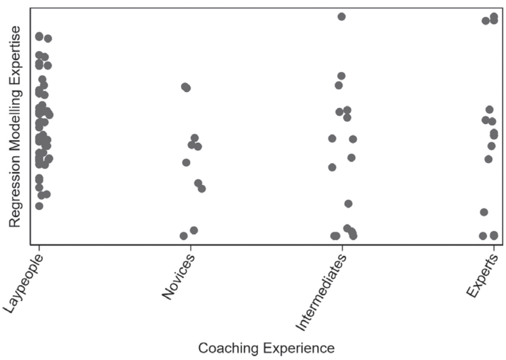

Figure 12 Jittered scatterplot, showing the regression modelling expertise per level of domain expertise (in years of working experience)

OPEN QUESTIONS AND DEMOGRAPHICS

Besides writing down their coaching advice, as explained in Section 'Congruence with Recommended Coaching Advice', the participants were asked to write down in an open text field whether they missed any information in the dashboard that was vital for them to be able to give appropriate coaching advice. At the end of the questionnaire, there was an open text field for general questions and remarks. Furthermore, we measured age and gender, as these variables potentially influence trust in automation (Hoff & Bashir, 2015).

RESULTS

After outlier analysis, we noticed that the measures of one participant highly influenced our results. The answers in the open text fields revealed that this participant did not correctly understand the assignment, so we omitted this data point, and continued the analysis with 110 participants.

Trust and coaching advice congruency are, as expected, related. Trust is significantly higher among those participants that gave coaching advice congruent to the dashboard's recommendation, compared to those who did not, both the after the non-transparent dashboard ($\text{difference}_{\text{Trust}}=0.29$, $p=0.04$) and after the transparent dashboard ($\text{difference}_{\text{Trust}}=0.58$, $p<0.01$).

INITIAL TRUST AND COACHING ADVICE CONGRUENCY

As a baseline, we measured initial trust and congruence with the dashboard's recommend coaching advice, when participants first encountered the (non-transparent) dashboard. We predicted initial trust using a linear regression model (Model 1, Table 12), where age, gender and expertise levels were used as predictors. Note that we did not include transparency level (and its interaction with expertise level) as predictor, because initial trust was measured after the non-transparent dashboard, making transparency level not relevant at this point. We predicted initial congruency using a logistic regression model (Model 2, Table 12), such that the probability on a congruent advice (binary variable; 1=congruent, 0=incongruent) was explained in terms of the predictors age, gender, expertise level. Such modelling of initial trust and congruency provides an understanding of whether there are any systematic differences across participants when assessing the dashboard, regardless of the transparency level.

After the initial examination of the dashboard, 45% of the participants' coaching advices was congruent with the dashboard's recommendation, to focus on exercise (see Figure 13). Initial congruency did not significantly differ across gender, age, and levels of expertise, illustrated by the lack of significant predictors of initial congruency in Model 1, Table 12. Furthermore, we found that participants with more expertise regarding regression models, showed lower initial trust in the non-transparent dashboard (*$\beta=-0.25$, $p=0.009$*, see Model 2, Table 12), possibly indicating that these

Using these data supports making a decent plan, in consultation with the client, and focus on what is important.

Coach

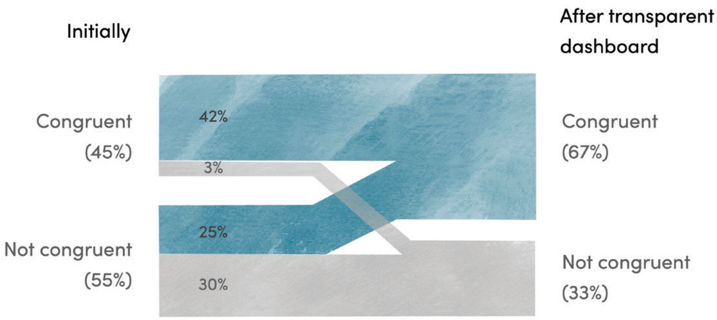

Figure 13 Percentages of participants giving (non-) congruent coaching advice with the system's recommendation. Left represents initial congruence (after seeing the non-transparent dashboard); right represents congruence after seeing the medium/high transparent dashboard.

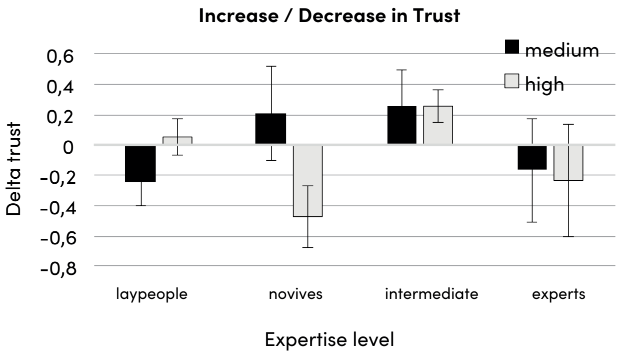

Figure 14 Average increase / decrease in Trust after the (medium or high) transparent dashboard, relative to the non-transparent baseline.

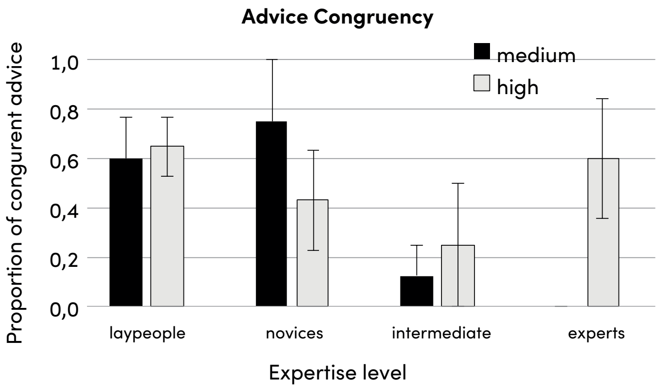

Figure 15 Fraction of participants that switched from incongruent to congruent advice, after assessing the (medium or high) transparent dashboard.

	1. Initial Trust†	**2. Initial Congruency†**	**3. Delta Trust††**	**4. Switched to Congruency After Transparent Dashboard†††**
	(linear regression)	*(logistic regression)*	*(linear regression)*	*(logistic regression)*
	Adj. R2 = 0.039	*Pseudo R2 = 0.069*	*Adj. R2 = 0.026*	*Pseudo R2 = 0.161*
	Coefficient (S.E.)	Coefficient (S.E.)	Coefficient (S.E.)	Coefficient (S.E.)
Intercept	2.48 (0.48)	0.89 (1.24)	-0.27 (0.28)	0.51 (0.36)
Transparency Level	n.a.	n.a.	0.14 (0.14)	-0.45 (0.73)
Working Experience as Coach (0 years)				
(<1 year)	-0.20 (0.31)	-1.43 (0.88)	-0.09 (0.19)	-0.70* (0.30) (continuous)
(1-3 year)	-0.36 (0.28)	-0.37 (0.68)	-0.10 (0.17)	
(>3 years)	0.09 (0.37)	-0.31 (0.91)	-0.41 (0.22)	
Working Experience as Coach (0 years) * Transparency Level				
(<1 year)	n.a.	n.a.	-1.22** (0.40)	1.19* (0.59) (continuous)
(1-3 year)	n.a.	n.a.	-0.16 (0.31)	
(>3 years)	n.a.	n.a.	0.10 (0.34)	
Expertise Self-Tracking Data	0.08 (0.10)	-0.33 (0.25)	0.04 (0.06)	
Expertise Self-Tracking Data * Transparency Level	n.a.	n.a.	-0.02 (0.12)	
Expertise Regression	-0.25** (0.09)	-0.10 (0.24)	-0.02 (0.06)	
Expertise Regression * Transparency Level	n.a.	n.a.	0.08 (0.11)	
Age	-0.00 (0.02)	-0.03 (0.06)	0.01 (0.01)	
Gender (Male)				
(Female)	-0.25 (0.19)	-0.05 (0.46)	0.03 (0.11)	
(Prefer not to say)	0.76 (0.63)	0.00 (n.a.)	-0.06 (0.37)	

p-value significance: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.
† The initial measures (Model 1 & 2) are based on the non-transparent dashboard, so the Transparency Level variable is not applicable.
†† Delta Trust is calculated by: Trust in the transparent dashboard, minus the (baseline) Trust in the non-transparent dashboard.
††† Model 4 is performed on a subset of the data: we aim to predict congruent advice for those participants with initial incongruent advice (n=62). Overfitting issues occurred when estimating coefficients for all predictors. Therefore, we simplified the model to the variables of our main interest, and treated Working Experience as Coach as a continuous variable.

Table 12 Results of (linear / logistic) regression on Initial Trust (Model 1), Initial Congruency with the (non-transparent) dashboard's recommended coaching advice (Model 2), Delta Trust (Model 3) and Congruency with the (medium / high) transparent dashboard's recommendation, provided initial incongruency (Model 4).

participants are more critical towards such dashboard. However, the low adjusted R2 (=*0.039*, Model 1, Table 12) signals that this model explains very little variance. The other variables did not significantly influence initial trust.

TRUST AND CONGRUENCY AFTER THE TRANSPARENT DASHBOARD

After the participants assessed the transparent dashboard, we again measured trust and congruence with the recommended coaching advice, and we analyzed the difference with the baseline (i.e., non-transparent dashboard).

Trust after the Transparent Dashboard

We model *Delta Trust* (Model 3, Table 12), representing the increase (or decrease) in trust compared to the initial trust measurement, as a result of the transparency in the dashboard. For this, we used a linear regression model, where transparency level, age, gender and expertise levels were used as predictors. We added the interaction terms combining transparency level with expertise level (domain, modelling, as well as self-tracking expertise), to understand whether the effect of transparency level on *Delta Trust* varied over the different expertise levels.

On average, trust did not increase (nor decrease) significantly after viewing the transparent dashboard, reflected in the non-significant intercept in Model 3 (Table 12). Also, the non-significant main effect of transparency level shows there was on average no significant difference between the medium- and high-transparent dashboard. Yet, we did find a significant interaction between domain expertise and transparency level. For novices, i.e., participants with little (<1 year) working experience as a coach, medium-transparency fostered trust, and high-transparency eroded trust. The low adjusted R2 (=*0.026*, Model 3, Table 12), though, signals that the explained variance is small[11]. Figure 14 shows this detrimental effect of high-transparency on trust for this group. This detrimental effect only occurred for novices, and not for the other expertise levels, and this difference across expertise levels was significant (e.g., for comparison novices vs. laypeople, β=-1.22, p=*0.003*).

Furthermore, trust did not significantly vary over different levels of expertise (averaged over transparency level), reflected by the lack of significant main effects of all three expertise measurements (domain, modelling and self-tracking data) in Model 3, Table 12. Furthermore, modelling expertise did not moderate how transparency level affected *Delta Trust*. Also, age and gender had no significant influence on *Delta Trust*.

This result indicates that the level of domain expertise indeed matters when faced with transparency, but the effect is not linear with expertise. Actually, we see that

11 When dropping the non-significant factors from the model, that is, only keeping Working Experience and Transparency Level as predictors (including their interaction), increases the adjusted R2 to 0.081. Still, we report on the full model to give more insight. The qualitative conclusion remains the same across the simple and the full model.

in the novice-stage of expertise, high transparency erodes trust. Given the limited explained variance of *Delta Trust*, the evidence for this effect is not strong. Yet, when moving onto analyzing the congruency of advice, we observe a similar pattern.

Congruency of Coaching Advice after the Transparent Dashboard

We aimed to understand which participants changed their coaching advice after assessing the transparent dashboard. First, as can be seen in Figure 13, a very small proportion of the participants (3%) switched from a congruent to an incongruent advice with the system. This group is too small for further analysis, but does indicate that the increased transparency of the dashboard did not have substantial adverse effects. A much larger proportion (25%; see Figure 13), switched from an incongruent to a congruent advice, after assessing the transparent dashboard. To better understand the effect of transparency on congruence, we analyzed in Model 4 only those participants that were initially incongruent (55%, n=62), and predicted whether they switched their advice or not, as a function of their domain expertise and transparency level, using logistic regression. Because we fitted this model on a smaller dataset, overfitting issues occurred, restricting us to a simplified model including only the main variables of interest, and coding the variable *Working Experience as Coach* as continuous.

We found that congruency of the advice with the transparent dashboard (provided initial *in*congruency) was less likely for participants with higher levels of domain expertise. For example, Figure 15 shows that the congruency of the participants with 1-3 or >3 years of working experience is considerably lower than the participants with less domain expertise. This implies that participants with higher expertise are, regardless of the level of transparency, less likely to switch their coaching advice. This effect was significant (β=-0.70, *p=0.019*, Model 4, Table 12). Furthermore, for experts, switching to a congruent coaching advice was more likely after the assessing the high-transparent dashboard, and contrary, for novices, switching to a congruent coaching advice was more likely after the medium-transparent dashboard. This interaction effect was significant (β=1.19, *p=0.044*, see Model 4, Table 12).

Due to the restrictions of the simplified model, we could only test for domain expertise as moderator and not for modelling expertise (i.e., on regression modelling and self-tracking data) regarding congruency of the advice.

CONCLUSION

In conclusion, on average, trust did not significantly increase (nor decrease) after viewing the transparent dashboards. Also, the particular transparency levels (i.e., medium and high) did on average not influence trust nor congruency. So, in answer to RQ1, we did not find main effects of transparency level on trust nor congruency. However, we found differences when discriminating between different domain expertise levels. In answer to RQ2a, our results show that domain expertise is indeed an important moderator of how transparency influences trust and congruence. In short, novices show higher trust in medium transparency compared to high transparency, whereas experts are more likely to give congruent advice when presented with high transparency. Laypeople and intermediates seem to be indifferent to the transparency level. This indicates that experts might benefit from a high level of detail in transparency, whereas for novices this potentially erodes trust.

Furthermore, in answer to RQ2b, we did not find evidence for the moderating effect of modelling expertise, i.e., expertise in regression models or viewing self-tracking data. Apparently, this did not influence how participants perceived transparency. However, high expertise in regression models did lower the initial trust in the non-transparent dashboard. Furthermore, we did not observe any effect of age and gender on trust and congruence.

DISCUSSION

This chapter reports on a study designed to investigate the importance and implications of transparency of model-based health coaching recommendations. Specifically, we investigated whether transparency would affect users' trust in the system's recommendations, and whether required levels of transparency would vary with the user's expertise level. Below we discuss our results in the light of prior research, and use participants' responses to the open questions to further explain our quantitative findings. Furthermore, we will discuss the limitations of our study, as well as the implications for the design and future research of transparent decision support systems.

DOMAIN KNOWLEDGE MATTERS

Our results imply that transparency – and particularly the level of detail in transparency – matters for users with domain knowledge, and that novices and experts show opposite needs for transparency (which we will discuss in more detail under 'Novices' and Experts' Needs for Transparency are Different').

Laypeople were more likely to follow the system's recommendation than participants with more domain knowledge, and they were neutral to the level of detail in the transparency; their trust and congruence levels did not significantly change across the medium and high-transparent dashboards. One explanation could be that they have less developed intuitions and skills on health coaching and fewer expectations of the role self-tracking data might play in that context. This potentially makes them less engaged with the dashboard and the presented client, and thus, also less attentive to the explanation of the recommendation. Similarly, Springer amd Whittaker (2019) argue using the Elaboration Likelihood Model that people only engage in central processing when a system's outcome is counterintuitive or unexpected, otherwise, peripheral processing routes are used, resulting in less deliberate and critical reflection. Our results seem to suggest that a similar process may be at play

Next to the data from the tool, I would like to talk to Sam and learn more about him; his age, medical history, medication, physical and psychosocial well-being, daily life and work environment, personal goals. How does Sam score on perceived work pressure / satisfaction, psycho-social functioning, etc?

Coach

in our sample of laypeople; a lack of intuition and skills makes users less engaged and critical to a system's explanation provided through transparency.

In addition, our qualitative data, where participants elaborated on their coaching advice, show differences across different domain expertise levels. Participants with higher levels of domain knowledge reflected more extensively on the presented data, the applicable guidelines for a healthy lifestyle, and their approach when coaching the client. In fact, the participants with the highest levels of working experience as a coach (>3 years), used significantly more words in their coaching advice than the less experienced participants, showing richer and more detailed domain knowledge and a willingness to share that information. Prior research (Lim et al., 2009) has already suggested that domain knowledge potentially decreases the attention experts pay to an explanation, and our qualitative data indeed indicate that for domain experts, the information presented in the dashboard triggers an extensive elaboration based on their own domain knowledge, instead of the dashboard's explanation. Contrary, the participants with little or no working experience as health coach more often shortly referred to the dashboard's explanation when arguing their coaching advice, e.g., *"regarding the formula above, exercise does have the most impact on his sleep duration"*. In line with previous literature (Schaffer et al., 2019), this suggests a need for a strong domain-oriented explanation for users who are knowledgeable on the domain.

Regarding modelling expertise, orthogonal to domain expertise, we find that users with more expertise on regression modelling – a proxy for the technical literacy required in determining complex relationships between multiple variables – show lower initial trust towards our (non-transparent) dashboard, indicating that they are overall more critical towards such systems. Yet, when assessing the transparent dashboard, modelling expertise did not significantly influence the increase or decrease of trust after the transparent dashboard. This contradicts prior research (Cheng et al., 2019), where a positive relation has been found between technical literacy and trust in algorithmic decisions. However, similar to (Cheng et al., 2019), we find that modelling expertise does not moderate the effect of transparency level on trust. We conclude that, at least for the level of modelling expertise of the participants in our study, and the complexity of our model, modelling expertise does not seem to influence trust in transparent support systems.

NOVICES' AND EXPERTS' NEEDS FOR TRANSPARENCY ARE DIFFERENT

In prior work, health professionals expressed a clear need for high transparency, in order to validate and trust systems that support them with modelling self-tracking data. While this need is clear, it was less clear whether and to what extent health professionals would be able to understand and effectively utilize high transparency in practice.

Our results confirm previous findings that too much detail can be detrimental for trust (Kizilcec, 2016; Springer & Whittaker, 2019), and add to this that this is most likely the case for novices, who have some, but not extensive domain knowledge. Interestingly, our data show clear opposite effects for novices and experts. For novices' (i.e., participants with some, but less than 1 year of working experience as health coach), a high level of detail eroded trust, and decreased the likelihood of giving advice congruent with the dashboard's recommendation. They were more likely to give advice congruent with the medium-transparent dashboard than with the high-transparent dashboard, and their trust levels increased after assessing the medium-transparent dashboard, and dropped after assessing the high-transparent dashboard. This may be explained by information overload; for novices, having limited working experience, it typically takes more effort to come up with a coaching advice than for experts, making them more susceptible for information overload.

For experts (i.e., participants with at least 3 years of working experience, some even more than 10 years), this detrimental effect of a high level of detail in the transparency on trust did not show. Also, none of the experts switched from an advice that was incongruent to congruent as result of the medium-transparent dashboard, yet, some of them did so as result of the high-transparent dashboard. This may implicitly point to a higher level of self-efficacy for the expert group, trusting their own domain expertise, and fitting the system recommendations to their own insights, rather than the other way around. Even though the group of expert participants was rather small, the interaction effect of transparency level and domain expertise was significant for both trust and congruence, and the trend over an increasing level of domain knowledge was clear. In sum, novices seem to be better off with less detail, for experts this is less critical.

LIMITATIONS

We decided to increase transparency of the dashboard's recommendation by providing either a general explanation of regression, or a detailed regression formula. While this does provide more information and potentially fosters insight in the system's inner workings, it may not have been an intelligible explanation to all participants. In future research, we may rely more on recent advances in visualizations of quantitative models that are more intuitive and interpretable, (c.f., Lundberg & Lee, 2017; Ribeiro, Singh, & Guestrin, 2016). In addition, correlation-based analyzes such as regression has been criticized by others (Karkar et al., 2016), as this does not reflect good self-experimentation practice. It would have been more useful, and thus more valuable for coaches, when the client had deliberately changed in his behavior over a certain period of time, allowing for a systematic and reliable comparison of the health effects.

By increasing the level of transparency, inevitably also the level of detail, and likely the level of complexity, increases. This confound hinders understanding what essentially caused the change in trust in the system and congruency of the advice. In future work, we plan to check our manipulation explicitly by measuring the users' understanding of the system and the perceived transparency, to validate whether users actually perceive the intended transparency as transparent and intelligible. Yet, even though from the current result it may be hard to conclude what caused the differences in trust, we at least found clear differences across expertise levels. This fact points to the importance of assessing these groups separately, when studying the implications of and users' need for transparency and level of detail in this transparency.

CRITICAL ASSESSMENT OF RECOMMENDATIONS RATHER THAN TRUST

While aiming for a realistic client case and regression model predicting an optimal coaching advice, we faced the inevitable ambiguity in health coaching; there is no such thing as a singular optimal coaching advice. The recommendation in the dashboard argued, based on the regression equation, that a coaching advice focusing on exercise would potentially be most beneficial to the client. However, some participants indicated that *"the client's exercise levels are already according to the guidelines"* and that therefore, it would be more feasible to coach the client on improving other lifestyle aspects, such as drinking less coffee. Some other participants elaborated in their coaching advice to distribute the exercise moments more evenly over the week, perhaps even dividing it into shorter workouts, to gain more benefit on sleep during the week. This shows that multiple coaching advices could potentially make sense for this client, thus allowing users with domain expertise to argue against the system recommendations (i.e., give an incongruent advice), given the available data. Domain experts will likely have a better understanding of the real-world implications of the recommended coaching advice, and in that sense, being critical as a result of transparency is desirable, and not something that we should try to avoid. This is not to say that experts are always effective in making decisions; there is clear evidence that systems often outperform human experts in quantitative tasks (Grove et al., 2000; Tazelaar & Snijders, 2004). What constitutes an effective collaboration between humans and computers in such tasks is still an open research question.

Rather than designing transparency that aims at maximizing trust per se, transparency may play a vital role in calibrating trust. It may help to avoid over and under reliance, by improving a user's understanding of the trends and correlations in the data, facilitating a deliberate consideration and a fair and well-argued use (i.e., adoption or rejection) of a system's recommendation. This may be particularly fruitful (and critical) when users have high domain knowledge, because it facilitates effective use of both the user's knowledge and the system's potential added value. Nevertheless, prior literature indicates that trust calibration through transparency is not self-evident, particularly when systems are uncertain (Lim & Dey, 2011). In those cases, transparency may easily lead to under-trust, because it emphasizes the limitations of the system.

Aim for at least 8000 steps a day. I give this advice because it is not feasible for everyone to exercise daily, but increasing steps is.

Coach

LEVELS OF EXPERTISE, LEVELS OF TRANSPARENCY

It is notoriously hard to recruit experts with multiple years of professional experience, as they are scarce and often busy. Yet, our results signal the importance of including this group, as they do respond differently to transparency than their less experienced peers. This argues for making the extra effort to recruit and study this group. In future work, we would like to use more explicit measures of assessing the expertise of our participants, and the different ways in which they attend to and process the information made available to them. Such measures could include explicit tests of information retention and interpretation (e.g., chunking), but may also include eye- or mouse-tracking as proxy measures.

Additionally, our results emphasize the value of considering transparency as something that can be manipulated gradually, and not just binary (i.e., transparent versus non-transparent). Systems may offer too much or too little transparency, and this may substantially influence a users' perception. A parameterized approach to transparency is necessary in order to improve understanding of the factors underlying transparency and their impact on the user's experience.

Explainability, intelligibility and transparency of AI systems have received increasing attention over recent years, mainly aiming to support users overcome their reasoning biases and misconceptions, (c.f., D. Wang et al., 2019). Our findings suggest that a more transparent system is not necessarily a better system for all users. In fact, high levels of transparency may erode trust for less experienced users. We conclude that transparency becomes particularly critical, as well as worthwhile, when users' domain knowledge is increasing. Disclosing the right kind of information, and the right type of explanation, depends on the audience's needs and level of expertise. Effective use of decision support systems requires an adaptive, personalized approach to transparency.

I think I'm able
previous races are
and which not.
runners so well, I'm
was easy or difficult

Coach

to indicate which
representative
Because I know the
able to say if a race
and why.

Running Coaches’ Interactions with a Marathon
Prediction Tool

CHAPTER 6

Running Coaches' Interactions with a Marathon Prediction Tool

In *Chapter 5*, we took a first step towards engaging health coaches in the analysis of health data, by giving them insight in how data-driven recommendations are generated and observe how this affected their levels of trust and acceptance. Yet, we relied on a one-directional interface in which the model was explained, whereas we also want to engage coaches more actively to deploy their complementary knowledge into the model. Therefore, in this chapter, we build an interactive tool, in which we invite coaches to adapt the model. In addition, in contrast to the previous chapter, in this chapter we use a real model and let coaches work with real data of their own clients. This allows coaches to use knowledge of the specific client at hand, and it increases the ecological validity of the results.

In this chapter we specifically focus on a prediction task; that is, setting a challenging yet realistic finish time for a runner's next marathon. While we acknowledge that health coaching covers a much broader domain, we decided to focus on such a specific task as a first step, to be able to use a data-driven model with good performance. Models are shown to perform generally well on this task, which potentially supports running coaches who still often rely on their own experience and intuition for this. Allowing coaches to interact with the model can benefit the model and the coaches: Running coaches can harness additional, deep knowledge about a runner (current form, type of motivation, personal context etc.) that is typically absent from prediction models, at the same time, models can support coaches to make unbiased predictions. Thus, this marathon prediction task provides a well-defined coaching context where both models and coaches have complementary contributions, enabling us to study such a mutually beneficial interaction.

Through pilot interviews, think aloud sessions and an online study, we gradually developed a meaningful means of interaction, driven by user needs and the nature of the task at hand. Our results show that coaches deployed rich knowledge when working with the model, and those who were able to interact with the model displayed higher levels of trust and acceptance in the resulting predictions. Furthermore, model accuracy generally improved when coaches adapted the model. We also found that coaches were more critical on the model's prediction when working with their own runners, compared to runners unfamiliar to them. This study demonstrates the value of targeted user studies for evaluating interactive machine learning support systems, and shows the benefits of model interactivity for both coaches and model performance.

This chapter is derived from:
Rutjes, H., Willemsen, M. C., Smyth, B. & IJsselsteijn, W. A. (Under review). Running Coaches' Interactions with a Marathon Prediction Tool.

INTRODUCTION

Marathon running is increasingly popular and as weeks of training reach the inevitable conclusion of race-day, runners and their coaches must confront the challenging but important task of establishing an appropriate goal-time and a suitable pacing strategy to achieve it. This is as true for recreational runners as it is for more experienced athletes and professionals, because the marathon's punishing 42.2km distance leaves no room for complacency. An overly conservative approach can lead a runner to under-perform just as being too ambitious can translate into serious pacing problems later in the race, including *'hitting the wall'* (Smyth, 2018). By establishing a challenging but achievable goal-time and a suitable pacing strategy, coaches and runners can plan to optimize performance on race-day.

Tracking running-related data has become increasingly commonplace, with athletes tracking their races and training sessions using a variety of wearable devices and sensors. These data are easily and routinely shared with coaches and other runners, for example though popular platforms like Strava[12] and RunKeeper[13]. In addition, many races provide official race-time data for 5-kilometer intervals throughout the race. These so called *5k split-times*, including finish times, are usually publicly available through race websites and collectively these data offer the potential to support runners and their coaches when training for and planning their next marathon. For example, models learned from training session information (Doherty et al., 2020; Keogh et al., 2019; Tanda, 2011) or marathon performances of similar runners (Smyth & Cunningham, 2017), have been shown to accurately predict marathon finish-times.

Although coaches usually have access to these data, they typically do not analyze them, nor do they actively seek the support of models to do so. Instead, they rely on their experience and intuition, fine-tuned from years of coaching (see *Chapter 2*, and D. Collins et al., 2016; Lyle, 2010). Indeed, running coaches have access to a broad range of domain knowledge and knowledge about the runner that is typically not incorporated in models, such as the runners psychological traits, their motivations to run, (c.f., Ogles & Masters, 2003), injury history, and personal context. Yet, when setting an appropriate target finish time for a next marathon, coaches may benefit from data-driven models because it has been shown that predictive models typically outperform humans (Grove et al., 2000) on such tasks. Human judgement is susceptible to a variety of biases (Kahneman et al., 1982), particularly people tend to

12 www.strava.com
13 www.runkeeper.com

over-rely on too many, and sometimes irrelevant, factors when making predictions (Salzinger, 2005), leading to sub-optimal predictions. Running coaches may be subject to this bias, especially when they know their runners very well and when other factors potentially influencing marathon performance are largely available. On the other hand, models too can benefit from leveraging the personal knowledge, experience, and intuitions of coaches, which may improve the interpretation of previous running performances and can be a relevant input for predicting future performances. Thus, we argue that there is an opportunity to promote and facilitate effective collaboration between data-driven models and coaches, by utilizing their complementary forms of knowledge, and to the benefit of coaches and runners.

The current work presents a user study of an interactive machine learning (IML) system designed to support running coaches with planning upcoming races for their athletes. Through our iterative approach, including pilot interviews (n=2), think-aloud sessions (n=7), and an online study (n=71), we gradually developed a meaningful and effective interaction. The results show that allowing coaches to interact with the system benefits the system and the coaches; it shows improved levels of acceptance and trust from coaches, and it reveals the potential of coaches' feedback to improve model accuracy. Moreover, the combination of our quantitative and qualitative approach provides rich insight in what constitutes effective interaction; it shows what coaches do, need and value when interaction with a support tool.

In the remainder of this chapter, we will first discuss related work and describe our research questions. We will describe the three iterations of our study with running coaches, and we conclude the chapter by discussing the results and drawing implications for future studies and the design of interactive systems.

RELATED WORK AND RESEARCH QUESTIONS

This work sits at the intersection between two important research areas. On the one hand it is concerned with leveraging sensor data that can be collected from real-world activities (in this case sporting activities and specifically marathon running) in order to identify meaningful patterns and make actionable recommendations. On the other hand, the work draws on ideas from *Interactive Machine Learning* (IML) and the exciting opportunity that exists to engineer human-in-the-loop ML systems that are enhanced by various forms of interactivity and user feedback.

Running and Data

Running is increasingly popular as a recreational sport (Scheerder, Breedveld, & Borgers, 2015), and participation in endurance events such as the marathon has been steadily growing over the last few decades (Vitti, Nikolaidis, Villiger, Onywera, & Knechtle, 2020). Recent studies highlight how most runners are using wearable monitoring to track and optimize their running performances; estimates vary from 75% (Pobiruchin, Suleder, Zowalla, & Wiesner, 2017) to 86% (Janssen, Scheerder, Thibaut, Brombacher, & Vos, 2017) of the runners. Tracking activities in this way can provide runners with useful feedback, including instructional, motivating or challenge-oriented feedback (Dallinga et al., 2018). Sharing these data with others is also increasingly commonplace and even actively encouraged by popular platforms such as Strava[12] and RunKeeper[13].

Running-related data may be particularly useful when preparing for a race. It allows runners, and their coaches, to learn from their previous performances, by identifying successful training patterns while avoiding those that appear to be problematic. In an attempt to support runners and coaches, several data-driven models have been developed to generate performance predictions for future events. Perhaps the most widely known and used endeavor is Riegel's early work on finish-time prediction in which the finish-times of shorter races are used to predict longer events (Riegel, 1977), using linear regression techniques. More recent efforts have explored more advanced machine learning approaches (Claudino et al., 2019) and predicting performance using wide sets of data, including training determinants (Doherty et al., 2020; Keogh et al., 2019) and past race performances of similar runners (Smyth & Cunningham, 2017).

When preparing for a long distance race such as a marathon, it is particularly important to plan a suitable pacing strategy, or pace profile (Foster, Schrager, Snyder, & Thompson, 1994), for during the race. There are three basic strategies, including an *even-split* where the pace is kept constant during the race, a *positive-split* where the runner starts faster and ends slower, and a *negative-split* where the runners starts slower and ends faster. Inexperienced runners often show a positive-split, meaning that they start too fast, often resulting in exhaustion towards the end the race, sometimes referred to as *hitting the wall* (Stevinson & Biddle, 1998). An even-split or slightly negative-split is associated with optimal finish times. A good pacing strategy is of key importance, as given the same training preparation and fitness of the runner at the start, a suitable pacing strategy can still greatly influence a runner's performance. The actual race-plan is directly derived from the target finish time, combined with the pacing profile, which results in the actual pace that a runner needs to achieve at every moment of the race.

Smyth and Cunningham (2017) developed a Case-Based Reasoning (CBR) model to predict suitable target finish times and pacing strategies for marathons. Briefly, CBR is a machine learning technique that solves new problems by reusing the solutions for similar problems that have been solved previously and stored in *cases* in a *case base* (Aamodt & Plaza, 1994). Smyth and Cunningham (2017) applied this idea to marathon running in an effort to predict the finish-time of a runner based on the races of similar runners. They did this by building a case base of past races, pairing a non-personal-best race (nPB) for a runner with a subsequent personal-best (PB) race, and predicting a finish-time for a new target runner by reusing the PB races from runners with similar nPB races to the target runner. Moreover, in addition to predicting a finish-time for the target runner, the system also recommended a pacing plan based on the pacing profiles of the PB races for these similar runners. The advantages of this approach are two-fold. Cross-validation studies demonstrate that it is capable of generating reasonably accurate finish-time predictions (Smyth & Cunningham, 2017), but in addition, the idea of reusing past races from similar runners is intuitively appealing (Aamodt & Plaza, 1994) making it more straightforward to explain to coaches and runners alike. For these reasons we adopt this approach as the starting point for the work described in this chapter.

On the Acceptance of Interactive Machine Learning

While there is mounting evidence that it is possible to build accurate and effective performance prediction models, it is less well studied how users interact, or wish to interact, with such models, and whether these models will be accepted and useful to coaches. In fact, studies in healthcare settings reveal challenges for the adoption of clinical decision support systems. Healthcare professionals have expressed several barriers for use, including a lack of fit in their current workflows making it inefficient to use, and skepticism about the competence and reliability of such system (Devaraj et al., 2014; Khairat, Marc, Crosby, & Al Sanousi, 2018). Health coaches have expressed similar concerns, and add that it potentially distracts from the personal experiences of a client, which they consider to be an essential part of effective coaching (see *Chapter 2*).

One of the problems with existing machine learning approaches is that, all too often, their *black-box* nature makes them rather impervious to interrogation, and this lack of transparency only exacerbates any inherent trust issues. Interactive Machine Learning provides one potential solution strategy in this regard, by creating a more collaborative setting in which a system and coach can interact more openly and, crucially, where coaches can provide feedback to guide, and potentially improve, the operation of the underlying model. In IML systems, human feedback is incorporated in the model training process (Dudley & Kristensson, 2018; Fails & Olsen, 2003). There is a broad range of possible interactions, for example feature selection, model selection or model steering (Dudley & Kristensson, 2018). Enabling users to interact with machine learning models has been shown to be a beneficial process, as it allows for incorporating users' knowledge, potentially improving the accuracy of the models as well as users' trust (Amershi, Cakmak, Knox, & Kulesza, 2014; Cai, Reif, et al., 2019; Stumpf et al., 2007, 2008).

In an attempt to understand the nature of users' interactions with systems, and their needs when interacting, Stumpf and her colleagues (2007) performed a study where participants were asked to provide unrestricted feedback to a machine learning system. Their results show that participants were most willing to provide feedback, and that the type of feedback was often richer and more complex than currently facilitated. Examples include suggestions to employ feature combinations, or adding rules that would fundamentally change the algorithm (Stumpf et al., 2007). In a follow-up study by Stumpf and her colleagues (2008), they show that incorporating this rich user feedback is challenging, yet resulting in improved accuracy of the model. Also Amershi and her colleagues (2014) show, by reviewing user studies in the field of IML, that users wish to express rich domain knowledge in order to improve models, beyond just generating labels to data. They emphasize the value of explicitly studying users' interactions with models, in order to improve the design of interactive systems (Amershi et al., 2014). Thus, with this work we aim to understand running coaches' interactions with support systems, and the degree to which model interactivity influences their acceptance and trust in those systems.

1. What do coaches wish to contribute to the model; what constitutes meaningful and effective interaction?
2. Does model interactivity improve coaches' acceptance of the model's recommendation and their trust in the model?

Very often, studies in IML use fictional tasks or unrepresentative participants as proxy for realistic situations. It is notoriously difficult to recruit a large number of participants who have sufficient knowledge and experience to perform a domain-specific task. Thus, studies with intended end-users on realistic tasks typically remain small in sample size, (c.f., Boukhelifa, Bezerianos, & Lutton, 2018). In studies involving larger samples, participants are typically drawn from tools as Amazon Mechanical Turk, (c.f., Boukhelifa et al., 2018), which not only forces the task to be high-level to make it understandable to all participants, it also implies that the dynamics of the interactions between users and system not necessarily resemble a real-life situation where domain-experts are invested in the system and the outcome. It is likely that this affects the results. Furthermore, the extent to which users trust and accept a support system has shown to be subject to personal investment in the task (Beck, McKinney, Dzindolet, & Pierce, 2009; Hoff & Bashir, 2015). Specifically, when participants are *not* personally responsible for a certain task, they tend to rely more on decision support systems (Beck et al., 2009). Therefore, it may be expected that coaches' interactions with a support tool – and their tendency to rely on it – depends on whether they work with data of their own pupils, or unfamiliar runners. When working with their pupils' data, not only this enables them to employ their knowledge about this particular runner, it also makes them more invested in the task. After all, it reflects a process where they actually engage in in real-life, as they prepare their runner for their next marathon. This potentially results in stronger – and perhaps more biased – opinions regarding the recommendation, and a stronger need to interrogate and control the model. Accordingly, we consider the following questions:

3. Do coaches show different acceptance levels when considering familiar or unfamiliar runners in the model?

4. Is the effect of model interactivity on acceptance (RQ2) different when coaches are interacting with data of familiar runners, compared to unfamiliar runners?

We extend existing work, by using a substantial sample of knowledgeable and invested participants. Adopting coaches' own pupils in the systems allows for high ecological validity, at the same time, comparing coaches' interactions across familiar and unfamiliar runners provides insight in how user's interactions and evaluations differ when working on a realistic task compared to a more fictitious task.

Accuracy of Predictions by Models and Humans

Besides coaches' trust and acceptance, there is an additional perspective on whether or not model interactivity is to be considered successful. That is, we examine the model's accuracy, and the extent to which the predictions improve when coaches interact with it. In IML in general, incorporating user feedback is recognized as beneficial to improve system performance (Dudley & Kristensson, 2018; Fails & Olsen, 2003). Indeed, as coaching is an inherently interpersonal (see *Chapter 2* and *Chapter 3*) and complex (Bowes & Jones, 2006) process, coaches potentially contribute unique and important knowledge that can add considerable value to a ML system trained using instances that may be limited to a narrow range of observations and/or sensor data. They know their runners, including their current form, character, personal context and injury history.

At the same time, model predictions have been shown to generally offer equal or better performance then human predictions (Meehl, 1954). This holds for quantitative prediction tasks within a variety of contexts, including diagnosing diseases, predicting student performance or the fit of a job applicant (Grove et al., 2000), and this may be true for the task of marathon finish time prediction, which is, after all, a quantitative prediction task. When working on such tasks, coaches may be subject to biases, (c.f., Kahneman et al., 1982), for example, a tendency to incorporate too many cues in their predictions (Salzinger, 2005) which may result in over-fitting. Observing coaches' interactions with the model, specifically, what they change and how they explain their contributions, provides insight in the knowledge that coaches aim to bring in. Furthermore, we can measure how these contributions actually affect the prediction accuracy. Accordingly, in this work we consider the following additional questions:

5. Is model accuracy improved by coaches interacting with it?
6. What do the coaches contribute to the model; what domain knowledge are they aiming to express?

Aiming for a broad understanding of coaches' interactions with support tools, we perform several iterations of user studies, and analyze how running coaches engage with an interactive support tool using relevant data of their runners. We evaluate this using both human-centered and model-centered evaluation metrics (c.f., Boukhelifa et al., 2018). In the remainder of this chapter, we will describe our pilot interviews, think-aloud sessions and online study, and reflect on the findings. This study was approved by the local ethics committee at Eindhoven University of Technology, Human-Technology Interaction group. We preregistered this study on AsPredicted[14].

ITERATION 1: PILOT INTERVIEWS

In order to design a meaningful and effective interaction enabling coaches to contribute what they find relevant (RQ1), we started with pilot interviews. We interviewed two running coaches, first, to validate whether setting a target finish time and pace strategy was indeed an important pre-race task ahead of an upcoming marathon. Second, we were interested in how they typically approached this task and any opportunities that might exist to provide support and assistance. More specifically, we asked them to reflect on two paper prototypes of interactive interfaces, to assess which type of interactions were likely to be helpful and intuitive.

METHODS

We recruited two running coaches from our personal network; they participated on a voluntary basis. The interviews were semi-structured and lasted for approximately an hour. We began with general questions about their coaching practices, followed by more specific questions about how they prepared their runners for an upcoming marathon, and their current use of data. Afterwards, we briefly shared our plans to support coaches with data-driven models, and we showed two paper prototypes (see Appendix D) of interfaces of interactive systems. We asked them to reflect on the usefulness and usability of these systems, and their ability to provide the type of inputs they required.

At this point we had not developed a working system, but our intent was to use and extend the CBR model of Smyth and Cunningham (2017). There are two possible ways to interact with this model, which we worked out in paper prototypes. First, a coach could indicate which of the previous marathons for the query runner (for whom we make the prediction) should be included and used as the basis for this prediction. In this prototype (Appendix D, Figure I), we showed the previous races (year, location, finish time, and 5k split-times) and asked coaches to select those they considered to be most representative for the upcoming race. Second, a coach could indicate which similar cases selected from the case-based should be used for the prediction. In this prototype (Appendix D, Figure II) we asked coaches to set filters on the features (age, gender and finish time), and to select those

14 Please find our preregistration document at: http://aspredicted.org/blind.php?x=ye6dt4.

cases (i.e., runners) from the case-base that they viewed as most relevant to the query runner.

RESULTS

The interviews established that the task of setting a suitable target finish time, and a pacing strategy to achieve it, was indeed a relevant and important priority for coaches as they prepared their runners for an upcoming marathon. Coach 2 explained her practice this way: *"I first set the finish time, then I calculate the 5k splits. I give the runner the option to start slightly faster, but I emphasize to listen to their bodies."* Both coaches explained that they mainly used their own experience when setting the finish time, and they highlighted this to be individualized with respect to a given runner. For example, Coach 1 reflected: *"Training schemes and predictions are just a guidance, there is a large individual component. You have to know people, that is often forgotten."* Coach 2 explained how the motivations of a runner were important to take into account: *"We work very individually. And it all starts with the question: why do you want to run?"*.

The coaches demonstrated a clear focus on the runner's experiences rather than their data per se, when looking back at past races, or when planning ahead for future races with a runner. Interestingly, Coach 2 explained: *"I recommend my runners to use Strava and a sports watch, such that they get real-time feedback during the training, that refrains them from pushing themselves too much. We [as coaches] don't use the data ourselves, it is too complicated with ownership and consent. They technically could show it to me when we meet, but that never happens."* In this way, it can be argued that data from a runner's own training is not playing a significant role in their current coaching practice.

The coaches were most keen on prototype 1 (Appendix D, Figure I). They both explained that they expected to be able to provide the required input, as Coach 1 reflected: *"I think I'm able to indicate which previous races are representative and which not. Because I know the runners so well, I'm able to say if a race was easy or difficult and why."* Prototype 2 (Appendix D, Figure II), built to select similar runners, showed to be less intuitive and useful. Coach 1 questioned the feasibility: *"How can I know which other runners are similar if I don't know them?"* and Coach 2 even doubted whether selecting other runners based on such sparse information would be effective: *"Every runner is different. Selecting some runners would not help to predict a suitable finish time, I think. It is not that simple, I ran three hours, you too, and then I ran 2,5 hours, so now you're able to do that too. You need a lot more information, for example from the training."*

While both coaches acknowledged having a limited understanding of the model, and thus have no clear expectations of its performance, they appreciated the possibility to interact with it, as Coach 1 stated: *"I like that the model incorporates my knowledge"*.

Training schemes and predictions are just a guidance, there is a large individual component. You have to know people, that is often forgotten.

Coach

CONCLUSION

We concluded that the task of setting a suitable finish time and pacing strategy was relevant and important. The results clearly pointed to prototype 1, which was built to select representative races, as the more useful in this regard. More specifically, in answer to RQ1, coaches wished to provide information on the different past performances of a runner, rather than on similarity between runners. They reported they knew their runners well, and they had a very individual coaching approach. Thus, showing past performances of a runner is expected to represent rich and meaningful information to the coaches, enabling them to make informed adaptations to the model, and appreciate this means of control.

AN IML SYSTEM TO SUPPORT MARATHON PLANNING

A CASE-BASED REASONING MODEL FOR MARATHON PREDICTION

We developed a system to support running coaches when determining a suitable target finish time and planning an appropriate pacing strategy for their runners' upcoming races by extending the model of Smyth and Cunningham (2017). Their original model generates recommendations for a runner based on their performance in a single previous marathon. We extended it by incorporating multiple previous marathons, by querying the original model for all previous marathons separately, and then combining the resulting recommendations using equal weights. Such an equal weighting policy typically leads to good performance, often outperforming human-adjusted models (Dawes & Corrigan, 1974). This highlights the value of the non-interactive condition, as not necessarily inferior to the interactive condition, nor unrealistic.

The dataset on which the model was built consisted of data from the three largest marathons in The Netherlands, as we were targeting Dutch running coaches: Amsterdam, Eindhoven and Rotterdam, from 2008 until 2019. This dataset included 63,000 race records from 24,000 unique runners. Each race record included, among other features, the 5k split-times, the finish time, age, and gender.

THE INTERACTION

Based on the results from the pilot interviews in Iteration 1, we decided to provide the running coaches in the experimental condition with controls to allow them to weight the model's inputs. Specifically, for all previous marathon races for a query runner, the coaches were able set a slider to indicate how representative a given race was with respect to predicting their upcoming performance (see Figure 16). By default the sliders were set at position 0.5, and could be changed with increments of 0.1, to values ranging between 0 and 1. Note that the default position of the sliders resembles equal weight policy of the non-interactive condition. The model output changed in real-time with slider movement, thereby providing coaches with immediate feedback and encouraging them to experiment with the model as they adjusted

Figure 16 Example of the main page of the survey, where the model prediction is presented and the coaches are asked to reconsider their advice. This example shows the interactive model; the non-interactive model did not show the sliders, nor the instruction to adapt the model.

various slider positions and observed their impact on predicted finish times. This is in line with other researchers' recommendation (Amershi et al., 2014) to facilitate users to experiment with a model before needing to commit to its output.

METHODS: ITERATION 2 AND 3

PROCEDURE IN THE STUDY

We embedded the model in an online survey, which was used for both the think-aloud sessions (Iteration 2)) and the online study (Iteration 3). The procedures for the iterations specifically will be discussed in the method sections of those iterations, here we describe the general part. The coaches (71 in total) signed an informed consent on the first page, followed by a page with a brief introduction of the study. Subsequently, the coaches selected their runners, by searching for their own pupils in our database. When there were more than five races available for a specific runner, we only presented the five most recent ones. We added the option to add marathon data by hand, if they could not find their runners in our database, or if some races were missing. The coaches worked consecutively with four runners during the study, and were randomly assigned to start with two of their own pupils they had just selected, or with two runners unfamiliar to them. After collecting data of 12 coaches in the online study, we observed a large drop out at the runner selection page. They could not find their own runners, and apparently filling in the data by hand was too much effort. Therefore, we decided to provide participants with an option to continue without familiar runners. Of the remaining 59 participants, 22 used this option and were presented with four unfamiliar runners. The final number of observations are given in Table 13.

	Unfamiliar Runner	Familiar Runner	Total
Interactive	102	50	152 (=38 coaches*4 runners)
Non-interactive	84	48	132 (=33 coaches*4 runners)
Total	186	98	284 (=71 coaches*4 runners)
Note. The observations of unfamiliar runners are overrepresented, as we gave coaches the option to work with unfamiliar runners only. Still most coaches (n=49) assessed 2 familiar and 2 unfamiliar runners.			

Table 13 Number of observations, split by condition.

We did not assign unfamiliar runners randomly to the coaches but aimed to reuse previous participants' selected runners, to make sure that these were comparable, relevant and timely cases. We also aimed to keep the number of marathons similar across the familiar and unfamiliar runners and within the coach, to avoid confounding. So, when a new participant started, we assigned them with the last previously selected runners with an equal number of marathons to their own selected runners. In the beginning of the experiment, when no previously selected runners were available, we drew runners from our own prepared list of suitable cases.

The coaches consecutively assessed all four runners. For each runner, the coaches followed the same procedure. On the first page the runner was briefly introduced by their gender, age and previous marathon performances (including year and location of races, finish times and 5k split-times, with graphs similar to the layout in Figure 16). On this page, the coaches were asked to recommend an initial finish time and a suitable pacing strategy to go with it. The purpose of this initial prediction was to provide a baseline against which to calculate acceptance (RQ2-4).

On the next page, the model presented its recommendation for the runner, including finish time and pace strategy (see Figure 16). On top of the page, we provided an explanation of the model, including the general idea of CBR and a concrete example, that was unfolded when participants clicked on it. The coaches in the interactive condition were able to change the model inputs on this page (see Figure 16). In the non-interactive condition, the sliders were not presented, and the model was thus using equal weighted inputs. After seeing the model's recommendation, we asked the coaches again for their recommendations on target finish time and pacing strategy. To summarize, for each runner, the following predictions were made by both the coach and the model, regarding finish time and pacing profile:

- $Coach_{pre}$: Coach provides initial prediction based on previous marathon performances (introductory page)
- $Model_{initial}$: Model provides initial prediction (model page)
- $Model_{adapted}$: Final model prediction after adapted by the coach (model page, NB: interactive condition only)
- $Coach_{post}$: Coach may revise their initial prediction, after assessing (and when interactive: adapting) the model's output (model page)

For the analysis, we decided to focus solely on the predictions regarding finish time. During the think-aloud sessions we observed that coaches' mostly focused on setting a suitable finish time, and the interaction for setting a pacing strategy was less intuitive and less important to them. Also in the online study, a large number of participants did not use the option to adapt the pacing strategy.

At the end of the study, after assessing all four runners, the coaches completed a questionnaire to evaluate their trust in the system, coaching experience, several self-efficacy measures, and other demographic information, as discussed in more detail below.

MEASUREMENTS

Acceptance and Trust

To answer RQ2-4, we measured acceptance and trust. For the non-interactive condition, we calculated acceptance as follows:

$$Acceptance = 1 - \left| \frac{Model_{initial} - Coach_{post}}{Model_{initial} - Coach_{pre}} \right|$$

In the interactive condition, it makes more sense to calculate to what extent the coaches accepted the final (adapted) model as follows:

Acceptance is 1 when coaches fully adopted the (adapted) model's outcome, and it is

$$Acceptance = 1 - \left| \frac{Model_{adapted} - Coach_{post}}{Model_{adapted} - Coach_{pre}} \right|$$

0 when they stick with their initial finish time. If they revised their initial prediction in the direction of the model's prediction, then this is expressed as a fraction between 0 and 1, relative to the distance between the model's and coach's initial prediction. The distribution of Acceptance is given in Figure 17. Note that in the raw data we also observed a few negative values, indicating that coaches were disagreeing even more with the model after seeing the model. We capped those values to 0, as they indicate no acceptance.

We observed that most coaches rounded their target finish time to a multiple of five minutes. For example, for one of the runners of Coach 46, the model predicted a finish time of 3h42, and the coach filled in 3h40 and indicated in open text field *"The model and I are totally aligned!"*, indicating a full degree of acceptance. This type of feedback occurred frequently. Therefore, the Acceptance is set to 1, if the final prediction of the coach is equal the model's prediction when rounded to the nearest 5 minutes.

The final distribution of Acceptance (Figure 17) shows that the majority of the coaches is either fully accepting (Acceptance = 1) or not accepting at all (Acceptance = 0) the models' output. In the analysis, we will consider both Acceptance as a continuous measure, as well as a binary (rounded at .5) measure.

To evaluate trust, we used a questionnaire consisting of 12 questions on a 7-point scale, translated from (McKnight et al., 2002). We used only the items regarding the subcategories Competence (e.g., *The model performs its role of predicting marathon performances well.*), Willingness to Depend (e.g., *I feel I can count on the model when I need a suitable finish time and pace strategy.*), and Subjective Probability of Depending – Follow Advice (e.g., *If I had to set a suitable finish time and pace strategy, I would want to use the model again.*). Factor analysis revealed two factors. The first factor loaded highest on items from the subcategory Competence (labelled as Perceived Competence, Cronbach's α = 0.917). The second factor loaded highest on items from the other two categories, i.e., Willingness to Depend, and Subjective Probability of Depending (labelled as Willingness to Depend, Cronbach's α = 0.910).

The model and I are totally aligned!

Coach

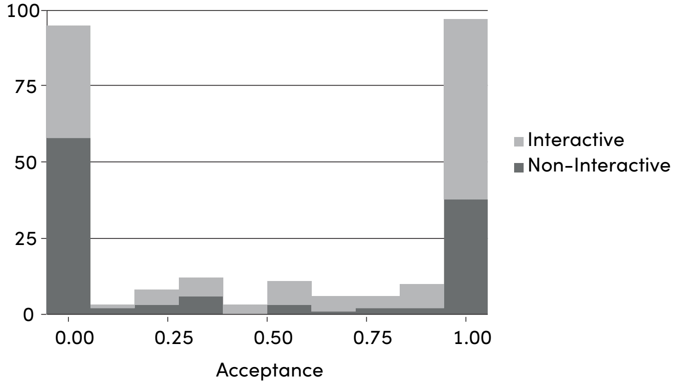

Figure 17 Distribution of Acceptance, colored by condition: interactive and non-interactive model.

Note that per coach, we measured Trust once (n=71), but Acceptance was measured four times, that is, once per assessed runner (n=71x4=284, also see Table 13).

Accuracy

In order to determine whether model accuracy was influenced by the adaptations of the coaches (RQ5), we calculated the accuracy of the predictions $Model_{initial}$ and $Model_{adapted}$. For this, we used a cross-validation approach; for all unfamiliar runners we omitted the best race in the dataset (i.e., fastest finish time) and used this as ground truth (G), after which we compared that with the model's prediction (P) based on the other races. We calculated the Percent Error as follows:

$$PercentError = 100\% * \left|\frac{G - P}{G}\right|$$

To interpret this Accuracy measure, we should consider this within the context of the task. Coaches, and thus the prediction tool, are facing the task of setting a suitable target finish time, that is challenging yet realistic. The runner's best recent race (out of a maximum of 5 most recent races) serves as a representative measure for this, after all, it shows what the runner recently has been capable of. If the model is able to determine this finish time as a realistic target based on the other recent races, we say this is an accurate prediction. Comparing the Accuracy of $Model_{adapted}$ versus $Model_{initial}$ then shows whether coaches were able to improve the Accuracy of the model's prediction.

We can also apply the Accuracy measure on the coaches' predictions, that is $Coach_{pre}$ and $Coach_{post}$. This provides insight in the quality of the coaches' predictions and whether their predictions have been improved after working with the (interactive or non-interactive) model. It also allows for explaining the coaches' prediction Accuracy in terms of covariates such as coaching experience, thus, are more experienced coaches better capable of setting suitable target finish times? We should, however, keep in mind that the task of the coaches (i.e., determining a suitable target finish time for the *next* marathon), while aligned, is not identical to the procedure of the model (i.e., predicting the best recent race based on other recent races). Therefore, we only analyze the Accuracy of coaches' predictions in an exploratory manner.

Note that we could only calculate Accuracy for the unfamiliar runners (n=186, see Table 13). It was not feasible to omit the best race from the data of the familiar runners (n=98), as the coaches would immediately have recognized that the best race of their pupils was missing. Furthermore, the calculation of the Accuracy of $Model_{adapted}$ could only be done for unfamiliar runners in the interactive condition in which they actually could adapt the model (n=102, see Table 13).

On the Contributions of Coaches to the Model

To understand what coaches aimed to express and actually contributed to the model (RQ6), we took quantitative and qualitative measures. Quantitatively, we saved the weights of the inputs set by the coaches, i.e., the slider positions that indicated how representative they consider the previous races for future predictions. Qualitatively, we asked coaches to motivate their predictions regarding finish time and pacing strategy for each runner in an open text field. Lastly, we obtained rich qualitative information from the think-aloud sessions, where coaches extensively reflected on the model and their interactions with it.

Covariates

Other variables are likely to play a role in the acceptance, trust and adaptation of models, including coaches' self-efficacy, demographic variables such as age or gender, and coaching experience. We aimed to explore, and statistically control for, the potential effects of these variables by including the following self-report measures:

- Self-efficacy measures:
  - Self-efficacy of setting target finish times by oneself (1 question, 7-point scale): *I believe I am able to set a suitable finish time and pace strategy without the help of a model.*
  - Self-efficacy of providing appropriate input in the model (interactive condition only, 2 questions, 7-point scale): *I believe I am able to determine which races are representative for a runner*, and *I believe I am able to adjust the model such that it gives a meaningful prediction.*[15]
- General computer self-efficacy (translated from (Marakas, Johnson, & Clay, 2007), factor scores based on 6 questions, 10-point scale, Cronbach's $\alpha = 0.948$).

15 In the end, we did not use this variable as predictor, as it limited the analysis to the data of the interactive condition.

- Demographics and Experience:
  - Age, gender, and education level.
  - Experience as a coach / runner, through multiple questions: years being coach, years being runner, number of runners coached, and number of marathons participated in.
  - Previous experience with using tools or models for setting target finish times (factor scores based on 3 questions, 7-point scale, Cronbach's $\alpha = 0.937$); e.g., *As running coach I often use (calculation) models*.
  - Familiarity with the runner (for familiar runners only, 1 question per runner, 5-point scale).

A relatively high correlation was found between the following five variables: age, years being coach, years running, number of runners coached, number of marathons raced. We decided to omit age, and used factor analysis on the remaining four experience measures (Cronbach's $\alpha = 0.66$) to construct a factor score as a measure of 'Coaching Experience'.

ITERATION 2: THINK-ALOUD SESSIONS

To deepen our understanding of what happens when coaches interact with the model, particularly their motives to make changes and what they need and value while interacting (RQ1), we accompanied the first batch of coaches while participating, and asked them to think aloud. We also used these sessions to check usability issues and possibly improve our model interface and survey.

METHODS

Running coaches were recruited from our personal network. After 7 sessions with coaches (of which 2 female, and mean age of 58 years), we terminated the data collection as we agreed on saturation. That is, we did not expect next sessions to add new insights substantially influencing our results. The sessions approximately lasted for 1 – 1,5 hours, and we compensated the coaches with a €10 voucher. We asked the coaches to fill in the online survey, including working with the marathon prediction tool, while thinking aloud. We presented them all with the interactive model, because we mainly sought to understand their interactions.

The think-aloud sessions were facilitated by one or two researchers. The sessions were all audio recorded and the researchers took notes of their observations during the sessions. We aimed to distract the participants as little as possible. While the participants were following the steps in the survey, we were attentive to usability issues. We more actively solicited for their thoughts and motivations for these actions at the pages where the model presented its recommendations, when they were adapting the model and providing their recommendations on finish time and pacing strategy.

RESULTS

In answer to RQ1, first and foremost, the think-aloud sessions showed that the data and the model triggered coaches to extensively reflect on the runners and their performances. The task of determining a suitable finish time and pace strategy was clearly relevant to them, and the presented data were intuitive, as they did not hesitate to start talking about the data and the runner immediately. When presented with their own pupils, it was often clear that these data were on top of their minds. Most coaches immediately filled in the target finish time which they used in their training sessions, and they often knew their runners' previous marathon performances by heart. So, our system clearly resonated with the running coaches' daily practice.

Coaches were aiming to express rich knowledge when adapting the model (RQ1 and RQ6). When interpreting data of their own pupils, they gave lengthy reflections on the runner's background, motivations (*"This runner has had a heart attack recently, if she will ever run a marathon again, it is for fun rather than performance"*), how well they were prepared for specific races, their character (e.g., *"I know this runner is stubborn"*), and their approach to races (*"We should challenge this runner, because she's too conservative herself"*). Rather surprisingly, also when assessing the data of the unfamiliar runners, their reflections and possible interpretations were about as lengthy. They showed to be eager to use all information at hand to understand the runner and her performances, such as age (*"Given the relatively old age, there is not much room for improvement"*), and how different performances were distributed over time (*"This runner improved greatly within a relatively short amount of time, that's a great achievement"*). They were trying to find possible explanations for anomalies or trends in the data (*"Here she hit the wall, maybe it was hot weather, or she started too fast"*). Thus, the results show rich domain knowledge at play, even when knowledge of the runner was limited. Overall, the coaches stated they would be more likely to accept the model's recommendation with the unfamiliar cases, compared to familiar cases, because they had relatively little information to oppose.

Most coaches were freely experimenting with setting the sliders, and observing their effects, before committing to their final input. Some coaches showed to be keen on adapting the model such that it would fit their own ideas, for example, one coach stated: *"I hope the model understands it now"*. Coaches could adequately explain their intentions when setting the sliders. For example, one coach stated *"By making this race very representative, I say: this is how I like the runner to approach the next race"*. Notably, coaches rarely set the sliders to zero (i.e., not representative at all), because they believed that *"even bad performances are in a way representative for their capabilities"*.

Lastly, some coaches reflected on the applicability of the model for different types of runners. For example, one coach explained: *"I think this tool is very useful for* *those runners who are highly driven by performance, but I know many other runners who mainly want to be healthy and enjoy the race."* To test the assertion that the model might be more applicable to specific types of runners' motivations, and less to others, we added an additional question in the online study, measuring the type of motivation of the familiar runners perceived by the coach (nine motivation types, translated from (Masters, Ogles, & Jolton, 1993), such as *life meaning* or *competition*, multiple answers possible).

Some small usability issues emerged, for example, the process of selecting runners in the beginning of the survey turned out to be not very intuitive. Also, the visualization of the graphs was not always clear (e.g., the labels on the axes). Based on the participants' feedback and our own observations, we improved the usability of the survey and the model interface.

As the results from the think-aloud sessions suggested that coaches are able to successfully interact with the model, we did not change the model nor the means of interaction. We will use the data from the think-aloud sessions, supplemented with the answers in the open text fields in the online study, to interpret and enrich the quantitative findings of the online study.

ITERATION 3: ONLINE STUDY

METHODS

Running coaches were recruited through social media and news messages on running platforms, by contacting running associations and marathon organizations, and via our personal network. All running coaches participated online, 68 completed the study. Of the 7 coaches that participated in the think-aloud sessions, 3 fully completed the questionnaires and could thus be included[16] in the analysis of iteration 3. This resulted in a total of 71 participating running coaches, of which 28 females, mean age was 45 years, ranging from 19 to 69 years, mean working experience as coach was 5 years, ranging from 0 to 28 years[17]. The online study lasted for approximately 30 minutes, and we compensated the online participants by raffling one €25 voucher per 5 participants, which they received by e-mail. The participants were randomly assigned to either the interactive (n=39) or the non-interactive (n=32) condition (see Table 13).

RESULTS

Coaches in the interactive condition made ample use of the interactivity option; on average they interacted with the model 25 times per runner (i.e., changing the slider positions, median = 16.5, ranging from 0-105 times). Figure 18 shows that the number of interactions is right-skewed, and that coaches' interacted most with the model

16 Excluding these 3 coaches did not change the results.

17 Four coaches indicated to have 0 years of experience as running coach, these participants probably had just started working as coach.

Here she hit the wall, maybe it was hot weather, or she started too fast.

Coach

	Acceptance		Trust - Perceived Competence		Trust - Willingness to Depend	
	Pseudo R2 = 0.183		*Adj. R2 = 0.324*		*Adj. R2 = -0.040*	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Constant	-0.058	0.204	-1.209	0.666	-0.476	0.789
Interactivity	0.266***	0.067	0.780***	0.209	0.515*	0.247
Familiar Runner	-0.121*	0.056				
Interactivity * Familiar Runner	0.193	0.111				
Runner Sequence Number	-0.013	0.022				
Computer Self-Efficacy	0.06	0.038	0.243*	0.118	0.052	0.139
Set Finish Time Self-Efficacy	0.075**	0.028	0.028	0.093	0.027	0.111
Coaching Experience	-0.098*	0.039	-0.197	0.120	0.016	0.142
Data / Model Experience	-0.100**	0.036	-0.044	0.118	-0.040	0.140
Gender (Male=0, Female=1)	0.136	0.070	0.569*	0.228	-0.028	0.270
Education (baseline: Undergraduate)						
Vocational	0.126	0.091	0.237	0.278	0.298	0.329
High school	-0.147	0.107	-1.065**	0.355	0.586	0.420
Graduate	-0.165	0.086	-0.275	0.280	0.092	0.332
Constant	-0.058	0.204	-1.209	0.666	-0.476	0.789
Interactivity	0.266***	0.067	0.780***	0.209	0.515*	0.247
Note. Similar results are obtained when fitting a multilevel logistic regression on the binary (rounded at .5) Acceptance measure. *p<0.05, **p<0.01, ***p<0.001.						

Table 14 Multilevel regression for Acceptance, random intercept per coach. Regular regression for Trust variables.

53, familiar runner, interactive condition). Even some coaches working with the non-interactive model expressed a desire to adjust the model, e.g., *"This runner coming back from a serious injury, so her old races are not representative at all. I wonder what would happen if the model would only be based on her last two races"* (Coach 4, familiar runner, non-interactive condition), even though the participants in the non-interactive condition were not aware of the existence of an interactive model. The qualitative data also revealed why some coaches deliberately decided not to accept the model. Oftentimes this was related to the model predicting a Personal Best, whereas they had a different goal for their runners, for example because they were recovering from an injury, or because of their age, as one coach explained *"it is fine to aim at a PR, but on his age, he should just treasure his fitness level and enjoy the audience cheer"* (Coach 11, unfamiliar runner, non-interactive condition).

For Willingness to Depend, the other Trust component, the effect of interactivity was significant but smaller (β = 0.515, p = 0.041, see Table 14). However, the model when assessing the first runner in the experiment. Interaction level was also high with the third runner in the experiment, which is when they switched from familiar to unfamiliar runners – or vice versa. Further analysis using multilevel regression (random intercept per coach) demonstrated that the number of interactions indeed dropped significantly for runners assessed later, but the number of interactions did not significantly vary across coach characteristics (i.e., covariates such as gender and experience), nor across familiar and unfamiliar runners.

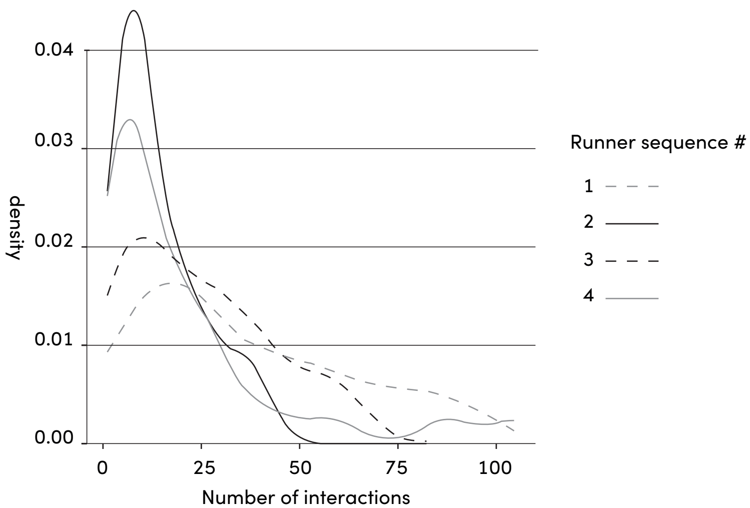

Figure 18 Distribution of coaches' Number of Interactions with the model, split by position of runner in the experiment. NB: Coaches in the interactive condition only (n=39).

Effects of Model Interactivity on Trust and Acceptance

For understanding coaches' Acceptance, we use a multilevel regression model (random intercept per coach), as the data contains repeated observations, each coach assessing four runners. For Perceived Competence and Willingness to Depend, the two Trust components, we fit regular regression models, as there is a single measurement per coach. For an overview of the regression results, see Table 14.

First, to answer to RQ2, the results show a strong positive effect of model interactivity on coaches' Acceptance and Perceived Competence of the model. Coaches in the interactive condition were more likely to accept the model's recommendation (β = 0.266, p 0.001), and their Perceived Competence of the model also increased significantly when the model was interactive (β = 0.780, p 0.001). This result is also illustrated by the qualitative data, where coaches often explicitly appreciated the ability to adapt the model, e.g., *"Without my adjustments the model did not make sense, but by eliminating the race from Eindhoven, we're getting somewhere"* (Coach itself was not significant, illustrated by the negative adjusted R2, and the effect of interactivity on Willingness to Depend was not robust[18].

Effects of Runner Familiarity on Acceptance

Our results point to a difference between familiar and unfamiliar runners regarding coaches' Acceptance. First, answering RQ3, coaches are more inclined to accept the model when working with unfamiliar runners (β = -0.121, p = 0.030, see Table 14). Indeed coaches regularly reflected in their explanations that *"having limited information about the runner makes that I now rely more on the models' recommendation"* (Coach 1, unfamiliar runner, interactive condition). When working with familiar runners, coaches showed to be much more informed and had stronger opinions, illustrated by expressions like *"I know this runner and I know what he is able to"* (Coach 37, familiar runner, non-interactive condition).

Regarding RQ4, the results suggest that model interactivity is more important when coaches are working with data of their own runners. The increase in Acceptance as result of model interactivity was stronger for familiar runners (from 0.28 to 0.51) compared to unfamiliar runners (from 0.50 to 0.61), though this effect was not significant in the regression model (β = 0.193, p = 0.082, see Table 14).

Effects of Covariates on Trust and Acceptance

Covariates play a role in both coaches' Acceptance and Trust levels (see Table 14). First regarding Acceptance, coaches with more experience were less inclined to accept the model (β = -0.098, p = 0.013). Higher self-efficacy regarding setting finish times by oneself led to increased Acceptance (β = 0.075, p = 0.008), and having more experience with data and modelling led to lower Acceptance (β = -0.100, p = 0.005). Then regarding Trust, women showed higher Perceived Competence than men (β = 0.569. p = 0.015) and lower educated coaches showed less Perceived Competence than higher educated coaches (e.g., comparing High School with Undergraduate, β = -1.065. p = 0.004). Lastly, higher scores on computer self-efficacy result in higher levels of Perceived Competence of the model (β = 0.243, p = 0.043). While these effects may not all have straight-forward explanations, all together they do suggest that coaches with different backgrounds, experience and self-efficacy levels may respond differently to support tools.

Additional Analysis on Model Adaptation

One may argue that we should be cautious in interpreting our finding that acceptance is higher in the interaction condition. Coaches in the interactive condition had the

18 Significance depended strongly on the covariates included in the model. When regressing Willingness to Depend solely on interactivity, the effect was not significant (β = 0.337, p = 0.159)

Without my adjustments the model did not make sense, but by eliminating the race from Eindhoven, we're getting somewhere.

Coach

possibility to steer the model fully towards their opinions. Accepting a model that is tuned towards one's own opinion is indeed essentially different than accepting a model that is not adapted at all, either because it was not possible, or because there was no need to. Only if coaches do take (some of) the model predictions into account when adapting and accepting the model, we show real value of model predictions for coaches. Therefore, we performed an additional analysis regarding the extent to which coaches adapted the model relative to their initial predictions, and use that to improve our understanding of Acceptance as result of model interactivity. From our data, four distinct cases emerged, illustrated by the examples provided in Table 15. While all four examples are scoring full Acceptance according to our original Acceptance measure, the actual model adaptations are quite different. In case 1, Coach 49 did not adapt the model but did change her final prediction to be similar to the model prediction, resulting in full Acceptance. In case 2, Coach 24 adapted the model somewhat in the direction of the initial prediction, and accepted that intermediate result in her final prediction. Both these cases are clear situations in which the model influences the coach's opinion. In case 3, Coach 3, adapted the model completely towards her own initial prediction, showing full Acceptance, but essentially ignoring the model entirely. In case 4, Coach 47 adapted the model towards and beyond their original prediction and fully accepted that new outcome. Case 3 is the problematic case, as it shows up as model Acceptance but in the end is the result of a coach not adapting at all to the model prediction.

Coach (Runner)	Coach Pre	Model Initial	Model Adapted	Coach Post	Acceptance	Model Adaptation
49 (3)	3:20	3:31	3:31	3:30	1	0.03
24 (3)	4:00	3:41	3:52	3:50	1	0.57
3 (4)	3:15	3:29	3:15	3:15	1	1.00
47 (4)	3:25	3:28	3:20	3:20	1	2.58

Table 15 Examples of coach's and model's predictions for target finish times (h:mm) for different cases of Model Adaptation.

We quantify the extent to which the coaches adapted the model towards their own initial prediction by:

$$ModelAdaption = \left| \frac{Model_{adapted} - Model_{initial}}{Coach_{pre} - Model_{initial}} \right|$$

This measure is given in the last column in Table 15, and illustrates that a 0 represents coaches not adapting the model at all, and 1 represents coaches adapting it to exactly their initial prediction. Based on this measure, we defined 5 cases of different types of model adaptation (see Table 15, that is, coaches not being able to adapt the model (non-interactive condition), coaches who respectively did not (c.f. case 1), to some extent (case 2), or fully adapted the model towards their own initial prediction (case 3), and lastly coaches who adapted the model towards and beyond their own initial prediction (case 4). We observe that only a minority of 23 runners in the interactive condition are like case 3 (show full adaptation of the model to their own predictions).

Analyzing the mean Acceptance across the different cases (see the last column in Table 15) using multilevel regression with these cases as predictors, we find that case 1 and 2, representing coaches who did not or only to some extent adapted the model towards their own initial prediction, show significantly higher mean Acceptance compared to the coaches in the non-interactive condition (β = 0.211, p = 0.032 and β = 0.187, p = 0.032 respectively). This shows that solely the possibility to adapt the model, regardless of using it for one's own sake, already increases coaches' Acceptance. This adds to our previous analysis of Acceptance, representing a more clean effect of interactivity on Acceptance, as it excludes those coaches who actively steered the model towards their own opinions (case 3) and beyond (case 4).

Case	ModelAdaptation range	N (runners)	Mean Acceptance
0. Non-interactive condition	n.a.	128	0.42
1. No adaptation	0.25	26	0.63
2. Some adaptation	0.25 , 0.75	37	0.61
3. Full adaptation	0.75 , 1.25	23	0.84
4. Over adaptation	1.25	7	0.96
Note. The number of runners do not add up to the total n of 71*4=284 runners due to missing values.			

Table 16 Mean Acceptance by adaptation case.

Effects of Coaches' Interactions on Accuracy

As discussed in the 'Measurements' section, we calculate the Accuracy of the model predictions by omitting the best recent time which then serves as ground truth, and that was only feasible for unfamiliar runners. Furthermore, the adapted model prediction was only available in the interactive condition (total n=102).

In answer to RQ5, the model Accuracy had indeed significantly improved by the coaches from the initial model (mean error = 3.14%) to the adapted model (mean error = 2.33%; paired t-test, p = 0.018). As coaches apparently have been able to improve the model, this raises the question what coaches actually have changed, and which domain knowledge they have been aiming to express (RQ6). Analyzing the final positions of the sliders after coaches' interactions, we found that more recent races are typically indicated as more representative (linear regression predicting slider position based on year of the race, $\beta = 0.042$, p 0.001). This resonates with the qualitative data, where coaches frequently indicated *"I decided to make older races less representative"* (Coach 46, unfamiliar runner, interactive condition). Furthermore, in line with the results of the think-aloud sessions, the qualitative data of the online study painted a rich picture of knowledge that coaches employed for their predictions and adjustments to the model. Frequently mentioned topics include a runner's mental strength, their ability to keep a constant pace, their training intensity, whether runners reached their maximal ability yet or there is something to gain, injuries, and personal circumstances unrelated to running. When coaches were assessing the data of unfamiliar runners, they had no access to other information than the runner's previous performances, age and gender. We observed, however, that coaches were actively trying to make sense of the data nevertheless, for example: *"There is clearly something going on with this lady. Maybe she stopped training, or she has a persistent injury? To make sensible prediction, I would need more information. For now, I would say, a finish time of 4 hours should be suitable"* (Coach 45, unfamiliar runner, non-interactive condition).

Exploratory Analysis on Coaches' Prediction Accuracy

Since the model Accuracy generally improved as result of coaches' interactions, this begs the question whether also coaches themselves have improved after interacting with the model. As discussed in the 'Measurements' section, we should carefully consider the context of the task when interpreting the Accuracy measure for evaluating coaches' performances. The coaches were asked to provide a "challenging yet realistic finish time" for the runner's next marathon (see Figure 16). The Accuracy measure is based on the best recent race in the dataset as ground truth. Thus, the Accuracy of coaches' predictions actually shows the extent to which the coaches' recommendations are aligned with this runner's best recent performance. We can run this analysis for both the interactive and non-interactive condition for unfamiliar runners (n = 186, see Table 13).

We find that the Accuracy measure of coaches' initial predictions (mean error = 3.77%) has improved after assessing the model (mean error = 3.41%; paired t-test, p = 0.009). We found that this improvement was similar across interactive and non-interactive conditions. This suggests that being able to interact with the model does not necessarily improve coaches' learning process. No factors or covariates (such as number of interactions, sequence number or other factors) could explain improvement in accuracy.

Additional Analysis on Runners' Motivations

Based on insights from the think-aloud sessions (Iteration 2), we hypothesized that the marathon prediction tool might be more applicable to runners driven by performance, compared to runners with other motivations such as fun or life meaning. To explore this assertion, we tested whether the Acceptance level of coaches in the online study was different across the nine possible types of runners' motivations indicated by the coach (e.g., 'life meaning', 'competition', 'recognition', 'weight concern', as described in (Masters et al., 1993)). Regression[19] reveals that Acceptance is only significantly lower for those runners whose motivation is 'life meaning' (19 out of 98 runners, β = -0.303, p = 0.013). For the other runners' motivations, we did not find an effect.

CONCLUSION

Coaches were keen on deploying their knowledge on the specific runner to improve the model, particularly by determining how representative the runner's past race performance were for future performance (RQ1). Model interactivity improved running coaches' levels of Trust and Acceptance in the model (RQ2). They showed higher Acceptance levels when working with unfamiliar runners compared to familiar runners (RQ3), and in addition, our results suggest that model interactivity is most appreciated when coaches are working with familiar runners, however, that effect is not significant (RQ4). When coaches interacted with the model, the model's prediction Accuracy improved (RQ5), and they showed to employ rich knowledge to do this, about running in general and about the specific runner (RQ6).

DISCUSSION

Long distance running, including marathons, is increasingly benefiting from technologies that can track runners' performance and support their goal setting and pacing strategies. In turn, informed by tracking data, coaches working with professional or recreational athletes can improve their coaching practices and recommendations. However, coaches' acceptance and use of data-driven or model-based predictions and recommendations is not a given, as models are typically ignorant with respect to the significant experience and expertise coaches have regarding individual runners (e.g., their motivational profiles, styles of running, injuries history) and relevant contextual variables (e.g., weather conditions, competitive pressures). In this work, through a set of user-studies we explore the ways in which coaches wish to and actually do interrogate a predictive model based on previous races of a large set of runners. We investigate how coaches interpret these data and model predictions, the extent to which they are inclined to accept and trust these predictions, and whether that is influenced by model interactivity.

We were able to provide the participating running coaches with a means of interaction that was intuitive and meaningful to them, as illustrated by their con-

19 Information on runners' motivations, perceived by the coaches, was only available for familiar runners, so this regression is based on that subset of the data (n=98, see Table 13).

siderable levels of interaction with the model, and their extensive reflections on the data, the runner and their recommendations. Similar to prior work (Amershi et al., 2014; Stumpf et al., 2007), coaches proved to be most willing to employ their own deep knowledge in order to improve the model. We believe our user-centered and iterative approach was key to shaping these interactions that enabled coaches to effectively express themselves. We built our interaction based on insights from pilot interviews, and validated and updated this by think-aloud sessions. The final interaction consisted of giving coaches control over the weights of the inputs, i.e., the extent to which they believe previous races of a runner are representative for a future performance. This may not be a typical interface of an IML system, c.f. (Dudley & Kristensson, 2018), but a first step in answer to other researchers' call for novel means of interaction, as "users are people, not oracles" (p.109, Amershi et al., 2014).

As discussed in the related work section, user studies using large samples of knowledgeable participants working on a domain-specific task are rare. It is notoriously hard to recruit participants with sufficient knowledge or expertise that are representative for intended end-users of a system. Therefore, often, less knowledgeable participants and fictitious tasks are used as a proxy for the realistic task. Our study contributes with a substantial sample of running coaches, invested in the task of marathon preparation for the runners at hand. We observed high involvement and enthusiasm of coaches during the think aloud sessions, and extensive and detailed reflections in the open text fields in the online survey. This shows that coaches cared about the task and the runner at hand, and these dynamics are likely to be different when working with less knowledgeable participants working on tasks not resembling their daily practice. More specifically, we found that coaches were significantly more inclined to accept the model's recommendation when working with unfamiliar runners, because they lacked interest and the required knowledge on the runner. Furthermore, our results suggest that the ability to interact with the model is more appreciated when working on a familiar runner, as there is more background knowledge to add, and more interest in the quality of the final prediction. These findings may have implications for generalizability of prior work based on less invested participants, and it highlights the importance of ecologically valid user tests for future evaluation studies in IML.

A classical discussion in the field of human-computer interaction, which is also underlying the mechanical versus clinical prediction debate (Grove et al., 2000), is the allocation of tasks: What should the system do, and what should the human do? Performance emerges as a key element in this discussion. Interactive Machine Learning often implicitly relies on the assumption that we can, and should, maximize the performance of a system by incorporating user feedback. User feedback is, in this light, serving the purpose of improving the model performance. Our results, however, show that users may actually enjoy steering a model, and being able to do so improves their levels of trust and acceptance. It suggests that we should shift our focus from who's prediction is more accurate, to enabling and fostering effective and satisfying collaboration between a human and a machine. We argue that having and *keeping* a human-in-the-loop is beneficial (Holzinger, 2016; Klein et al., 2017), from both a models' and a users' perspective. Aiming for fully automated systems should not be a goal in and of itself. Not only does this avoid the hassle to measure and implement typically complex and ambiguous information into the model, it actually makes users appreciative of the ability to add their specific knowledge to the model.

Interestingly, beyond the main effects of interactivity and familiarity of the runner, we found that coaches with different levels of experience and self-efficacy responded differently to our prediction tool in terms of their acceptance and trust. Perhaps the most notable effect was that coaches with more experience, while not necessarily providing more accurate predictions, typically showed lower acceptance levels, which resonates with prior findings (see *Chapter 5*, and Hoff & Bashir, 2015; Sanchez, Rogers, Fisk, & Rovira, 2014). This underlines the importance to understand the end-user and tailor to her needs when designing an interactive system, and, again, highlights it is essential to use participants representative for the intended end-users when evaluation interactive systems.

The central IML thesis is that the combination of human and algorithm offers a level of competence that is more than the sum of the parts, especially when, and because, humans and algorithms typically bring different perspectives, domain knowledge, and expertise to the task at hand. By offering a practical path to integrating these complementary forms of knowledge, we gained insight in what coaches do, need and value when interacting with support systems. We contribute with an iterative and user-centered approach, where we gradually defined an effective means of interaction that allowed coaches to meaningfully express themselves when interacting with the model. Additionally, we tested the use of the model in a realistic setting where participating coaches worked on a task and with data that was relevant to them, thus maximizing the ecological validity of our results. Beyond the results of our statistical analyses, our rich qualitative data reflected the importance of coaches' knowledge, opinions, and experiences when working with the prediction model, including their enthusiasm on the ability to adapt the model. As one coach expressed: *"Wow, cool that I can actually adapt the model!"* Thus, allowing running coaches to interact with a marathon prediction tool is a win-win for both the coaches and the model, eventually helping long-distance runners to achieve their personal best.

The needs of the
dominate the
interface, and the
interface should
design of the rest of

Donald A. Norman (1986)

users should
design of the
needs of the
dominate the
the system.

General Discussion

CHAPTER 7

General Discussion

This dissertation addresses the influence of health data on the coaching process, it identifies the complementary strengths of health coaches and health data, and it explores ways to effectively make use of those strengths in coach-data interfaces. Through a variety of studies, starting with mainly qualitative approaches, moving on to more quantitative approaches, we have shown that data influence coaching beyond merely adding information. Health data are rarely informative when inspected out of context, yet, these data can be important sources of information for a coach, particularly when contextualized within the client's narrative, and when used in interactive models that support a sufficient level of coaches' involvement and control. In this final chapter, we will first discuss our methodological contribution, and reflect on the generalizability of our results and limitations. Thereafter, we will reflect on our findings in terms of related literature, and draw main conclusions and design implications.

METHODOLOGICAL CONTRIBUTION

Our work makes some methodological contributions. We observe that in current literature, user studies on technical artefacts such as interactive systems mostly rely on tasks that are simple and general enough to allow for recruiting a large number of laypeople participants, for example through platforms such as Amazon Mechanical Turk. When studies do involve intended end-users, typically sample sizes remain small. Indeed, it is notoriously hard to recruit a large number of knowledgeable participants, as these are often busy professionals. Still, it is to be expected representative users are more knowledgeable and more invested, and thus have different needs, than participants working on a to them fictitious task.

Our results show that representativeness of users and tasks indeed affect study outcomes in important ways. Specifically, experienced coaches responded differently to the different levels of transparency of our data-dashboard compared to less experienced coaches or laypeople, i.e., people who were not working as coach (*Chapter 5*). Also coaches' level of self-efficacy on the task showed to influence the level of acceptance (*Chapter 6*). And, we found different levels of trust and acceptance when coaches were working on data familiar to them (from their own clients) compared to working with unfamiliar data (from unknown clients; *Chapter 6*). This implies that when evaluating interactive systems, results based on unrepresentative participants on proxy tasks may not generalize to realistic settings. Our results highlight that it is, however, worthwhile to create tasks and recruit participants as close as possible to the setting of interest, insofar this is feasible.

Concluding, our work contributes with a combination of in-depth qualitative methods with larger scale quantitative methods, which allowed us to gain insights in health coaches' practices and needs, before developing and testing interfaces towards health data. Moreover, we were able to involve relatively large numbers of intended end-users (i.e., health coaches) on realistic and relevant tasks (i.e., giving lifestyle advice, or preparing for an upcoming race). Almost all our technical artefacts we used were fed with realistic data from clients that were familiar to the coaches, making the data at hand timely and relevant for the coaches. We observed that when health coaches inspected data of their own clients, this naturally made them empathize with the client, and it triggered rich background information. As a result, the coaches' attitudes and needs were embedded in and influenced by the context of the client at hand. This brings substantially different dynamics compared to studies where people are assessing data not associated with such a rich and meaningful context. We will see in the remainder of this discussion that this context of the client is indeed playing an important role in our main findings.

GENERALIZABILITY AND LIMITATIONS

Regarding generalizability, first we should explain that everyone can be a 'health coach', the professional title of 'health coach' is not protected (Wolever, Jordan, Lawson, & Moore, 2016). Despite the call for more consistency in research and practices of health coaching (Wolever et al., 2016), in practice, there is large variance in quality and style, which we also have observed. We did, however, observe one clear common factor across all the health coaches we worked with, that is, they were all passionate about health and highly committed to help their clients. We believe that this has been central in our work; the need to support professionals in acknowledging their roles and expertise and supporting them to do what they find important. In that sense, our results were robust to various coaching styles and quality levels.

The client cases we encountered in our studies varied from clients with clear-cut goals (e.g., run a marathon) to clients with more ambiguous questions (e.g., losing weight that turned out to be a self-esteem issue). Similar cases, and thus similar coaching dynamics, may appear in other domains such as group sports coaching and medical domains, for example when managing chronic diseases or in mental healthcare settings. Thus, our insights are likely to generalize to those domains as well. Furthermore, our approaches and results may inform other fields beyond health, such as participatory design, and the design of explanations and interactions in artificially intelligent (AI) systems, particularly in situations where users are highly involved in the task, yet not so experienced in using data.

We recognize several limitations of our work. First, across our studies we often worked with self-made prototypes (e.g., data-dashboards, data-collection tools) to test our ideas, which were inevitably immature. This possibly enlarged the perceived burden to work with those artefacts, as not all 'hygiene factors' were met. At the same time, we also worked with off-the-shelf products, and those results were not qualitatively different. Furthermore, our findings are largely aligned with prior literature, therefore, we assume this did not influence the validity of our work.

In addition, while we aimed to influence our participants as little as possible, particularly in our qualitative work there could have been an effect of our presence as researchers. We were often bringing in self-tracking tools or data dashboards, and asked coaches and clients to use them and reflect on them. It is likely that they were inclined to respond in socially desirable ways. It would be interesting for future work to take a more ethnographical approach in studying health coaching and the implications of health data for coaches. By observing natural practices and conversations over a longer period of time, it is expected to gain more reliable insight in actual use, including appropriation (Dix, 2007), discontinued use, and non-use of tracking tools and data-driven support systems.

This brings us to another methodological reflection. When analyzing qualitative data through thematic analysis, as we did, there are two main orientations to this, relying on different philosophical assumptions (Braun & Clarke, 2019). The first, typically referred to as qualitative with a *big Q* (c.f., Kidder & Fine, 1987), is situated within a social constructionist epistemology, where researchers may activity draw from their subjective perception when interpreting the data, and as such, they take an active role in creating themes. The second, typically referred to as qualitative with a *small q* (c.f., Kidder & Fine, 1987), relies on a positivist epistemology, in which it is assumed that researchers are objective observers, and that themes emerge from the data. Very often when thematic analysis is applied the paper of Braun & Clarke (2006) is used as reference, which describes thematic analysis from a *big Q* viewpoint. Only recently, in 2019, Braun & Clarke (2019) argued that their paper from 2006 is often unjustly applied – many researchers actually take a more positivist stance, and the applied thematic analysis should be understood as qualitive with a *small q*. This is true for the research described in this dissertation as well. While we were following the procedure as described by Braun & Clarke (2006), we were not committing to their underlying philosophical assumptions. Only later we realized that we took a positivist epistemological stance, for example, by talking about "themes emerging from the data", and by calculating inter-rater reliability. Our take on thematic analysis might be much more aligned with Boyatzis' (1998) description of thematic analysis, that uses a *small q* approach (c.f., Kidder & Fine, 1987). Namely, we used structured study designs, in which we did ask open-ended questions "*to provide greater richness* [and] *to allow for the unexpected*" (p. 59, Kidder & Fine, 1987), but, in contrast to *big Q*, we were not seeking for questions but for answers. Our research questions and interview questions did not change while we were running a study. It is interesting, however, for future work to explore whether *big Q* approaches to understanding the dynamics between health coaches and health data would lead to triangulation with our results. This would validate, among other things, whether our results sufficiently resonate with the coaches' narrative, and whether the questions we asked make sense and are relevant in this context. We are hopeful that this is indeed the case, as we started with open-ended explorations on health coaching – not even mentioning health data in the very first study – and we let these results strongly inform our later studies in which we gradually narrowed down our focus on particular use cases of data.

DIFFERENT COACHING DOMAINS, DIFFERENT TYPES OF COACHES, DIFFERENT ROLES OF DATA

Across this dissertation, we encountered a range of health coaching domains, bringing in different types of health coaches and health issues. It is useful to reflect on their similarities and differences, particularly in light of the role and value that data have. At the core of our work, we have considered health coaching in terms of wellbeing and health promotion (*Chapters 2, 3* and *5*), and in addition, we have encountered health coaching in the domain of care for parents and their newborns (*Chapter 4*) and marathon running (*Chapter 6*). The dynamics of health coaching, and in particular the roles that coaches take on, are shaped according to those domains. In wellbeing and health promotion, we have seen how coaches support clients by collaboratively working towards client-centered goals, using techniques such as goal setting, education, and self-discovery (Wolever et al., 2013). While this also largely applies to the domain of caring for parents and their newborns, the clinical orientation of nurses, pediatricians and general practitioners create dynamics that are slightly less collaborative and more hierarchical. In addition, especially pediatricians and general practitioners face high time-pressure, driving the need to quickly understand the problem and move to solutions fast. The nurses in preventative care, in contrast, clearly had more time, similar to the health coaches in the other chapters (*Chapter 2, 3* and *5*), to explore possible goals and strategies, and to facilitate self-discovery. Another notable and unique aspect of the context of newborns is that the given care was at the level of the family as a whole, whereas the in the other contexts, the coaching was focused on the individual. Lastly, we encountered health coaching in the context of marathon running, where the main focus of coaches was on sports performance. Still, also other health aspects were considered, such as nutrition, sleep and stress. Marathon running coaching differs from general health coaching, in a sense that the goal is relatively clear and objectifiable, namely, run a marathon with a good finish time.

The role and value of data varies across these different domains. We will reflect on them along two dimensions: health issues with a subjective versus objective nature, and data collection initiated by clients themselves versus others.

Health issues and goals may have a more subjective or objective nature, which we already reflected on at the start of *Chapter 5*. Data may be helpful for health goals with a more subjective nature (e.g., improve self-esteem, improve wellbeing), but only when contextualized in the client's narrative. When health goals have a more objective nature (e.g., fix injuries, run a marathon), the purpose and use of data is more straightforward, and thus analyzing data through data-driven models makes more sense. It may be tempting to assume that health coaches and health data simply take more or less prominent roles, depending on the extent to which health issues are subjective or objective. Particularly, when health issues are (to a large extent, at least) objectifiable, we may think that health data says is all and that the coach's role is redundant. Still, the findings from *Chapter 6*, where we studied the relatively objectifiable coaching context of marathon running, show that our participating coaches were highly attentive to the subjective experiences behind this goal of running a marathon. They reflected extensively on who the runner was (or could be, in case of unfamiliar runners), their motivations to run a marathon, motivational strategies that were likely to be successful with these runners, and various aspects of the client's life that they believed that would influence their performance (such as particular stress factors). Coaches naturally took this perspective of the client's subjective experiences, and by doing so, they added considerable and complementary knowledge. Allowing them to add their knowledge and unique view on the client will foster higher engagement, which makes coaches more likely to consider the models' outcomes, be inspired and utilize the data's strength.

We have observed how health goals with a more subjective nature typically ask for exploratory use of data, whereas health goals with an objective nature more often require a problem-solving focus. This is in line with the findings of Chung and her colleagues (2019), who studied the collaborative use of food diary data among patients and their providers. They studied both healthy participants as well as patients with Irritable Bowel Syndrome (IBS). Their findings show that for IBS patients, the data were mostly used to identify triggers of symptoms, whereas for healthy participants, the use of data was much more open-ended, exploring alternative possibilities and discussing potential goals. In addition, based on our findings in *Chapter 4*, we recognize a pattern where in the beginning clients' problems are often relatively vague, and this becomes clearer over time, moving from exploratory to problem-solving oriented use of data. Also, time pressure plays an important role. Health coaches, especially with clinical backgrounds, often simply do not have time to extensively explore data. And indeed, shorter sessions have motivated healthcare professionals to partly transfer the data reviewing to their clients in preparation for their sessions, to make the session itself more efficient (Chung et al., 2019). To conclude, data may serve different purposes, depending on the extent to which the problem is defined and measurable, the particular phase in the coaching process, and the time available.

Lastly, we distinguish between data collection that was initiated and customized by the client herself, or by others, for example by healthcare professionals as in *Chapter 4*, or, in *Chapter 6*, were data were tracked by marathon organizations. Data turned out to be especially informative to coaches when clients had a large say in what was tracked and how it was tracked, for example by defining the labels of the trackers (as in Chapter 4), or by deciding when to wear the tracker and whether or not to use certain manual tracking options (as in *Chapter 3*). Making tracking a deliberate act showed to reveal information about a client's motivations and perceptions on the problem. As such, even the lack of data contains information, for example, it may signal low levels of motivation, or feelings of shame or low confidence to share the data. On the other hand, automatic and unobtrusive tracking can be beneficial when aiming for complete data, for example when aiming for identification of patterns and trends, or for prediction tasks such as in Chapter 6. Thus, health data can be useful and informative in a variety of ways, depending on the coaching context at hand.

OUR MAIN FINDINGS REGARDING COACHES, CLIENTS AND DATA

To structure the discussion of our main findings, we use the triangle of coach-client-data[20] (see Figure 19). First, we highlight the right edge of the triangle, client – data, and compare the value of a client's data and self-report. We describe how both comprise unique sources of information for the coach, and how the combination of the two is most powerful. We will discuss the value of collaborative reflection, how deeper levels of reflection on the data increases the value and utility of the data. Second, we highlight the left edge of the triangle, coach – data, and compare the strengths of human coaches and data. We will discuss the extent to which data may inform coaches, and the extent to which coaches may add their knowledge into data-driven models. We will discuss ways in which coaches and models collaboratively may achieve

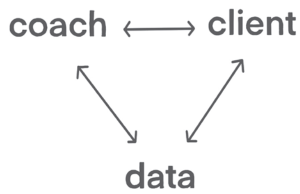

Figure 19 Triangle coach-client-data

20 The word 'data' here also refers to data-driven models. Throughout this chapter, we discuss data and models. We use data when it is relevant to talk about the information that health data comprise. We use models when there is an additional layer on top of these data, that identifies patterns or trends, or gives personalized recommendations. The distinction is not always clear, as both are highly intertwined. In general, our work addresses the value of health data, but as we learned, there is more value to gain when these are analyzed in models.

more effective coaching strategies than each on their own. Third and last, we conclude this discussion by considering the triangle as a whole, coach – client – data. We envision how each three parties can contribute from their strengths, and by doing so, bring out the best in others, ultimately leading to effective coaching strategies, serving population health.

THE SYNTHESIS BETWEEN CLIENTS' DATA AND SELF-REPORT

In all chapters, we have provided coaches with health data of their clients. And, in most chapters, the clients were directly involved to share their own stories, whether or not around these data. This enable us to contrast and compare the unique value of data, the unique value of self-report, and – perhaps most interesting – the value of the combination of data and self-report (see Figure 20).

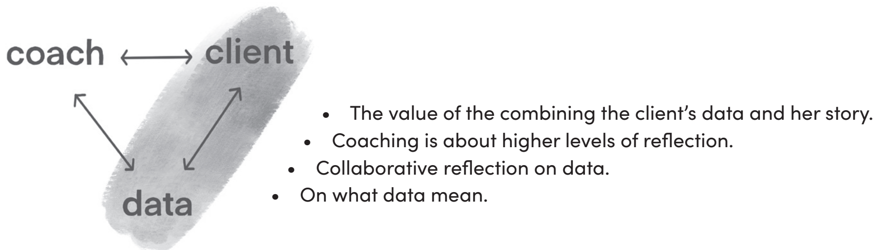

Figure 20 Topics considering the value of the client's data, her story, and their combination.

THE VALUE OF THE COMBINATION OF THE CLIENT'S DATA AND HER STORY

The main value of data emerging from our results was the objective view it provided on the client's behavior, including patterns over time that supported finding cause-and-effects. In addition, one strength of data, in particular the devices that measure them, is their ability to measure continuously and in-situ, which enabled giving timely support to the clients. This facilitated the coaches to broaden their support outside their coaching sessions. The main value of a dialog with the client, on the other hand, was the means by which this provided to access the client's experiences. A dialog was a natural way for coaches to acquire insights in their motivations and beliefs (i.e. why clients set their goals, why they were performing certain behaviors), and their daily context (i.e., practical constraints such as work schedules and access to places to exercise). Yet, our results strongly suggest that there is most value in the synthesis between data and self-report.

Often, we presented coaches with both sources of information, giving them the opportunity to talk with their clients and to assess their data. Across all our studies we observed that, while the data were valuable to coaches, they generally paid little attention to data as such. Rather, they showed a vast interest in the meaning of data through the subjective experience of the client (*Chapter 3*). In fact, they feared that the data would distract the attention from the subjective experience of the client by putting too much emphasis on behavioral aspects (*Chapter 2*). When there was limited opportunity to discuss data with clients, challenges arose with unaligned expectations, sometimes leading to frustrations (*Chapter 4*). Coaches showed, however, to be well able to weight the data and the knowledge they gained from their conversations with the client (*Chapter 6*). Zooming in on this synthesis between data and self-report, we see a process where clients and coaches collaboratively reflect on the data, which we more extensively discuss in the next sections.

COACHING IS ABOUT HIGHER LEVELS OF REFLECTION

It is known, for example from investigating practices in the Quantified Self community, that interpreting one's personal data is not straightforward (Choe et al., 2014). When clients reflect on their own data, they typically engage in lower level reflections, such as recalling certain events. Higher levels of reflection, such as using the data to make new resolutions, are rare (Choe, Lee, et al., 2017). These different levels of reflection are more explicitly described by Fleck and Fitzpatrick (2010). The authors describe the first level of reflection as simply revisiting certain data or events (R0), moving on to giving explanations or justifications for them (R1), then people may also consider alternative explanations and explore relations (R2), they may transform their initial perspective (R3), and in the highest level people critically reflect from a broader view, for example driven by social, moral or ethical considerations (R4). Tracking ones' health data mainly supports the lowest levels of reflection, by facilitating reviewing ones' health behaviors, for example in terms of step count, heart rate, sleep, or nutrition intake. While this first level is conditional to engage in higher level reflections, these higher level reflections do not come automatically. Fleck and Fitzpatrick (2010) explain that one way to support these higher levels of reflection is to engage in reflection *with someone else*. In dialog, explanations for specific data come naturally, and when someone else has a different perspective, this naturally challenges revising one's perspective.

It is exactly this process that we have observed in our studies. Coaches showed to rarely care about the behaviors as such. Instead, they sought to move as quickly as possible to higher levels of reflection and interpretation. They were keen on listening to the client's explanations and justifications of behavior, and naturally challenged them, by sharing their own perspectives. For example, when a mother shared that she found it hard to deal with her baby crying for twenty minutes before falling asleep, the nurse would reassure her by explaining that this is basically normal behavior for a baby of this age. Or, when a client would talk about cravings, the coach and client in dialog tried to unravel the underlying emotions that caused this behavior, and they would collaboratively discuss alternative ways to cope with these emotions. This suggests that the goal of coaching may be exactly this engagement in higher levels of reflection, particularly where perspectives are fundamentally changed (R3 or R4). This also connects to our finding that success is not always quantitatively measurable (*Chapter 2*), as a change in perspective may already be considered as a successful outcome of coaching. It also reveals why fully automated e-coaching applications are likely to be less effective, as it is hard to prompt these higher level reflections in digital solutions, as well as it is hard to measure whether people actually revised their perspectives in meaningful ways.

COLLABORATIVE REFLECTION ON DATA

Thus, collaborative reflection on health data is of key importance to make effective use of those data in a coaching process. Across our studies, we have seen numerous examples where coaches used the data gain insight in the client's experience (e.g., *"I see you walked here, where did you go?"*, or *"the baby cried a lot here, was that a tough day for you?"*), often followed by sharing their health-related knowledge or by simply reassuring them. The importance of collaboratively reflecting on health data is recognized in other studies too, e.g., (Mentis et al., 2017; Oygür et al., 2021; Pina et al., 2017; Raj et al., 2017; Richards, Choi, & Marcu, 2021). Results from these studies suggest some key enablers of this process. For one, there should be sufficient room to share different views on the data, and an empathetic relationship is important. This allows coaches and clients to gain a shared understanding of what the data represent and use it accordingly in the care process. In our earlier work, we have more elaborately discussed the differences in clinician's and patient's perspective on data (Rutjes et al., 2017), also highlighting some tension that these different views bring forward, for example between the comprehensiveness of data versus information overload.

ON WHAT DATA MEAN

The fact that it is so important, yet not easy, to share different views on data, illustrates that data are ambiguous and the interpretation of data is not straight-forward. This is recognized in literature in Personal Informatics too. It shows that personal data are not just numbers, and while it is tempting to see them as neutral and objective, they are inherently loaded with deeper connotations (Lupton, 2016a). For example for chronically ill people, tracking data around their disease has shown to provoke negative emotions and value judgement (Ancker et al., 2015). The very act of measuring something implicitly entails a process of optimizing, and this overemphasis on performance can have detrimental effects on health and wellbeing (Ajana, 2018; Kersten – van Dijk et al., 2017; Lupton, 2016a). Also beyond the field of Personal Informatics, the notion of objectivity of data has been criticized, e.g. (boyd & Crawford, 2012). For one, as boyd and Crawford argue, the decision of what to measure, i.e., which attributes and variables, is by definition motivated by a subjective understanding of what is important.

Indeed, in *Chapter 4* where we let clients decide themselves what to track and customize their own data, it was specifically this subjective process that was most informative to coaches, rather than the 'objective' numbers resulting from it. Thus, coaches seem to be mainly interested in the *value* rather than the *numbers*. They almost habitually sought to interpret the numbers in terms of the client's experiences and narrative. This is not to say that data were not meaningful or helpful. To the contrary, the data often sparked conversations on topics that would otherwise not have been discussed or shed new light on what the client was experiencing. Still, as it turns out, life is not an optimization problem, and so a coaching process is not an optimization problem either. Thus, merely data do not suffice, but they do challenge, reveal, inform, disrupt, trigger, and facilitate in potentially helpful ways.

HEALTH COACHES AND HEALTH DATA TEAMING UP

So far, we have seen that there is considerable value of data for coaches, while it is important that these are inspected within the context of the client's narrative. In the next sections, we will shift our focus to comparing strengths of coaches and data, or data-driven models, and discuss ways to make effective use of both strengths (see Figure 21).

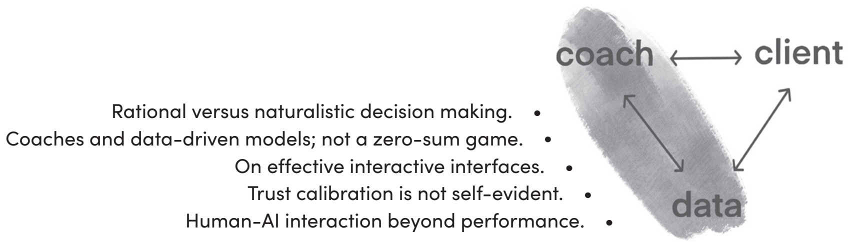

Figure 21 Topics considering how coaches and data may effectively team-up.

Considering health coaches and health data, there is a central two-fold question underlying our work. First, how and to what extent should we involve data in a coaching process? And second, how and to what extent should we involve coaches in data-driven modelling? Answers to these questions are sensitive to underlying assumptions on what effective health coaching entails. One may argue that health coaching essentially relies on naturalistic decision making (c.f., Klein, 2008), as the coaching process is too complex to capture in pre-defined structures. Others may draw from the premise that data-driven approaches generally outperform humans (c.f., Grove et al., 2000), and argue that data should play a profound role as they can substantially improve to the coaching process. We will explore both assumptions, after which we reflect on our results to contribute to this classical human-technology discussion. Thereafter, we will discuss our work in terms of interactive interfaces, trust calibration and human-AI interaction, eventually highlighting the potential of coaches and data working as a team.

RATIONAL VERSUS NATURALISTIC DECISION MAKING

Human biases and heuristics may hinder rational decision making (Kahneman et al., 1982), giving technology an opportunity to support humans and improve processes. For example, Wang, Yang, Abdul, and Lim (2019) have studied typical cognitive biases with clinicians, and have developed a medical diagnosis tool mitigating these biases. Indeed, it has been argued that doctors lack sufficient statistical literacy, for example in understanding probabilities, which may lead to suboptimal decisions (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007). The need to support professionals on their decision making is strengthened by evidence that algorithms often outperform humans in prediction tasks in various contexts (Grove et al., 2000), and this even seems to be true for professionals with large experience (Tazelaar & Snijders, 2004).

This view, however, is criticized for being too narrow. For example, Klein and his colleagues (2017) argues that evaluation metrics are too computationally driven, undermining the value of human expertise. The concept of naturalistic decision making (Klein, 2008) calls for a broader perspective, as the world is typically too complex to capture in pre-defined and measurable metrics. This is likely to be true for health metrics, as we have seen in the previous sections of this chapter. In literature on health coaching, naturalistic decision making where coaches rely on intuition is a widely accepted approach (c.f., Collins et al., 2016; Lyle, 2010), and descriptions of effective coaching strategies are dominated by interpersonal skills rather than quantitative tasks such as prediction (c.f., Wolever et al., 2013). The limited ability of computational systems to deal with tacit knowledge has been an important motivation to take on a 'human-in-the-loop', in healthcare support systems (Gui & Chan, 2017; Holzinger, 2016) and beyond (Dudley & Kristensson, 2018).

COACHES AND DATA-DRIVEN MODELS; NOT A *ZERO-SUM GAME*

While it may be tempting to pick a side in this debate, basically being enthusiastic or skeptical about data, our results provide a more nuanced picture. We have been able to identify the unique value of health coaches and health data, and we have explored ways to utilize their complementary strengths. Shifting the focus from who performs best to how they may reinforce each other reveals new perspectives on this problem. Our results show that coaches are keen on collaborating with data-driven support systems and they benefit from it, as long as, among other things, we present these systems as tools rather than something to compete with, we give coaches the opportunity to engage with those systems, and we appreciate their domain knowledge. This goes beyond filling their blind spots; coaches and data-driven models do not play a *zero-sum game*. Data-driven suggestions may inspire them by providing new perspectives, as we have seen in *Chapter 6*. It may help them understand their clients better, sometimes just by simply observing how their clients engage in data collection (*Chapter 4*) and talk about their data (*Chapter 3*). It gives coaches new material to use in their coaching, and in turn, when coaches may substantially improve data-driven models by sharing their knowledge and putting the data in new perspectives.

We echo Shneiderman's (2020) work, who suggests to leave behind our one-dimensional view on human-AI interaction, that varies from low automation where a human is in the lead, to high automation where a machine is in the lead. Instead, he proposes a two-dimensional model, envisioning that in the future both automation and human control are high. This asks for an approach where both coaches and models are supported to use their own strengths, and this requires careful design of an interface that allows for that. We argue that if we succeed in building interfaces that allow coaches to meaningfully connect their interpersonal and contextual interpretations of the client's process to specific data, this will result in a powerful tool for health coaching.

ON EFFECTIVE INTERACTIVE INTERFACES

There has been research on what users wish to 'say' and 'hear', when interacting with support systems. Regarding what users want to say, that is, what they want to express when providing feedback to a system, Stumpf and her colleagues (2007, 2008) performed studies in which they enabled users to give free-form feedback to an interactive email classification system. Users showed to be most willing to give rich feedback, which frequently went beyond simply adjusting weights of selecting features, sometimes even suggesting fundamental changes to the algorithm itself. In a similar vein, Amershi, Cakmak, Knox, and Kulesza (2014) argue that studying interaction should be more user-centered, as typical forms of interaction result in tedious tasks for its users. Amershi and her colleagues (2014) show, by reviewing a number of case studies, that users are keen on learning systems by sharing rich information beyond simply data labels, for example by demonstrating the desired behavior by examples. Regarding what users want to hear, that is, their information need from a system, Cai, Winter, Steiner, Wilcox, and Terry (2019) report on a user-study with clinicians collaborating with an AI assistant, to learn what type of information the clinicians were interested in. Cai and her colleagues (2019) found that they were interested in much higher-level information than specific model decisions, for example, its design objective and its 'world view'. The clinicians expressed a need to assess the AI's decision in similar manner as they would weigh the second opinion of a colleague, for example in terms of knowing her being typically liberal or conservative.

These insights resonate with our own findings, showing that coaches enjoy the process of interaction and appreciated the model responding to their feedback. Starting with more qualitative approaches revealing the coaches' tasks and needs therein, we have been able to develop means of interaction enabling coaches to meaningfully express themselves, which led to improved levels of trust as well as improved system performance. Potentially, this may even result in a positive feedback loop, where coaches are learning the model, resulting in improved performance of the model, in turn increasing coaches' levels of trust and engagement even more, and so on.

TRUST CALIBRATION IS NOT SELF-EVIDENT

People's trust in a system is of key importance when considering interactive systems where people and systems ought to collaboratively work on tasks. Recently, the focus has shifted from merely fostering trust to fostering *trust calibration* (c.f., Bussone, Stumpf, & O'Sullivan, 2015; Tomsett et al., 2020; Zhang, Liao, & Bellamy, 2020), that is, ideally users trust the system only when it is accurate (i.e., *justified trust*), and distrust it when it is inaccurate. System transparency and intelligibility are targeting this issue, as they potentially allow users to understand the system's limitations and facilitate users to uncover errors and potentially harmful biases in the system (Wortman Vaughan & Wallach, 2020). It is tempting to assume that by simply revealing the systems' limitations and accuracy levels would accomplish trust calibration, but user studies show that this process is not self-evident.

For example, users may overestimate their own understanding they gain from the explanations leading to over-confidence (Chromik, Krüger, & Butz, 2021), or, too detailed explanations may lead to over-reliance on the system (Bussone et al., 2015) as well as a lesser understanding due to information overload (Poursabzi-Sangdeh, Goldstein, Hofman, Vaughan, & Wallach, 2021). Furthermore, explanations regarding specific model predictions do not necessarily foster building an accurate mental model of the system as a whole (Chromik et al., 2021). In addition, transparency can backfire; research has shown that even when users agree with the outcome, if the explanation is not matching their own understanding, this may result in decreased levels of trust in the system (Cramer et al., 2008; Kizilcec, 2016; Lim & Dey, 2011; Springer & Whittaker, 2019). So, while we may think that explanations would reassure the user, it may as well trigger critique. Interestingly, even for data scientists themselves, interpretability tools are not always effective (Kaur et al., 2020). Thus, designing systems such that users have a fair and complete understanding of its inner workings and performance is far from straight-forward.

HUMAN-AI INTERACTION BEYOND PERFORMANCE

For making effective use of health data in the context of health coaching, indeed, we did not only want to convince coaches of our data-driven recommendations, we also wanted to facilitate them to disagree when they had good grounds to do so. This naturally asks for a way to objectively evaluate this, i.e.: did coaches justly adopt or reject the system's suggestion? Yet, the fact that coaching can hardly be narrowed down to an optimization problem problematizes such an objective evaluation. Our work deviates from other literature on trust calibration, that typically report on such metrics, as we had no absolute ground truth available, other than observing how satisfied coaches and clients were in the end. After all, as our results show, successful coaching is not always measurable (*Chapter 2*). Still, the value of our work lies in showing how users appreciate being involved in modelling health data, and by showing how these data can inform and facilitate them to apply effective coaching strategies. Human-AI interaction is expected to not only be applied to clear-cut quantitative tasks, but is likely to appear in more undefined and complex areas too, for example reflected in the increasing body of literature on how AI aids creative processes (Koch, Lucero, Hegemann, & Oulasvirta, 2019; Lin, Guo, Chen, Yao, & Ying, 2020; Suh, Youngblom, Terry, & Cai, 2021). Across our studies, sometimes cases emerged where clients had clear-cut questions or goals (e.g., fix injuries, run a marathon, or increase breast milk production) and there were cases where clients' issues were rather ambiguous (e.g., losing weight that turned out to have large emotional connotations related to self-esteem, or a baby's sleeping problems that turned out to be related to parents not being aligned on their approach). In the first, data were indeed often used to gain quantifiable patterns and relations, and in the latter, data were used much more exploratory to unravel a problem (c.f., Kollenburg & Bogers, 2019).

DESIGN CONSIDERATIONS

The topics we have discussed have implications for the design of health technologies that collect and present health data. First, while there is a clear call for more complete and reliable data to make it more useful for healthcare professionals, we argue that it is at least as important to present and visualize data such that it triggers higher levels of reflection, and facilitates collaborative use of such data. As explained in our earlier work (Rutjes, Kersten – van Dijk, Willemsen, & IJsselsteijn, 2018), it may be particularly incomplete or ambiguous data that helps to spark meaningful conversations and reflections. For example, when a client deliberately terminates data collection, or decides to start tracking again after a period of non-use, these may be important cues for the coach to pick up on in a coaching session. In that sense, unobtrusive and easy to use wearables may capture less information compared to wearables that make tracking a deliberate act, for example by manual tracking. This is also recognized by research prototypes where tracking is flexible and open-ended, such as in (Y.-H. Kim et al., 2017; Storni, 2011, 2014), where people are encouraged to create and shape their own tracking practices. Such technologies may facilitate higher levels of reflection to clients themselves, and for coaches such data comprise more relevant cues to understand their clients beyond their behaviors.

Furthermore, the coaches in our studies have frequently shared their concerns on the detrimental effects that data may have, which is in line with prior literature.

For example, clients may unnecessarily worry or ruminate over their data, or become obsessive about targeting 'good numbers' rather than focusing on their own bodies and intuitions. To avoid this, coaches in our studies have hinted on taking on a role of 'gatekeeper', where they have control over which data are used and how they are presented to their clients.

Lastly, we draw design implications regarding fostering effective collaboration between health coaches and health data. There is a current notion of human-in-the-loop that typically gives the human a temporary and serving role; when the system's performance is high enough, the human's input becomes redundant. We argue that we should be sensitive to situations where humans actually *like* to be in the loop, as their involvement in the modelling process also acknowledges their expertise. In line with this, ideally, interfaces between coaches and models foster a real dialog, where coaches can express their feedback to the model in ways that are natural to them. It may sound trivial to aim for users and models actually learning from each other, still, to date most interfaces are rather one-directional. One way that potentially enhances this dialog, is to more actively solicit for user's mental models of the system. An effective and simple first step may be to let the user request which information they need, and progressively disclosing this information accordingly in an adaptive interface as in (Springer & Whittaker, 2019). This allows for tailoring interfaces and explanations to a user's understanding and needs, which may even evolve over time. We have discussed this topic more elaborately in a separate paper on mental models and XAI (Rutjes, Willemsen, & IJsselsteijn, 2019). Specifically, we address the importance to understand on which level the user's concerns or information needs sit, to be able to respond to them with appropriate explanations. This may as well concern high-level information on the system's intentions (c.f., Cai et al., 2019), and in those cases, it can be more effective to communicate that the system is not intended to take over one's tasks but to support them, rather than explaining how a specific model output came to be. This relates to our final remark; when a system actively encourages users to deploy their domain knowledge, not only this knowledge itself is useful, but it may also be an effective way to communicate that the system is not designed to do it all alone (and thus, the word *e-coach* should be avoided). Rather, it highlights its limitations, and recognizes the added value of the user.

CONCLUSION: TOWARDS CONSIDERING THE SYSTEM OF COACH, CLIENT AND DATA

So far, we have focused on the right and left edge of the triangle coach-client-data. From these discussions, it clearly emerges that we cannot isolate any pair from the third. There are complex interactions at play, such as the client's interactions with data that turn out to be informative to the coach. In this final section, we consider the system of coach, client, and data as a whole (see Figure 22).

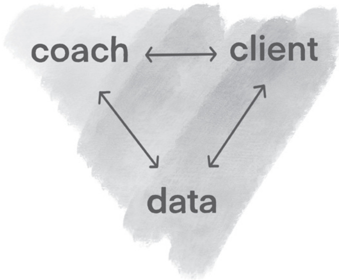

Figure 22 Considering the triangle coach – client – data as a whole.

A helpful framework for such an holistic perspective is provided by the theory of distributed cognition (Hollan et al., 2000). It considers how a combination of people and technical artefacts collaborate as a *socio-technical system* wherein knowledge and actions are shared. The theory argues the importance of taking this system as unit of analysis, rather than its parts. Drawing mainly from aviation and ship navigation contexts (c.f., Hutchins, 1995; Hutchins & Klausen, 1996), it explains how knowledge is socially distributed, how this knowledge is embedded in the spatial location of people and artifacts, and how knowledge is distributed over time, in a sense that information representations created earlier can influence later events (Hollan et al., 2000). In health coaching, indeed, we have seen that the knowledge required for effective coaching is distributed across the coach (domain knowledge on health and wellbeing), the client (knowledge about lived experiences, personal motivations and local contexts) and the data (systematic and detailed overview of measurable behaviors). Thus, this asks for a careful coordination of that knowledge, where multiple perspectives are shared. In this process, the availability and visualization of data has a critical role; it should facilitate coaches and clients to talk about these data in a natural way. It should meet both coaches' and clients' information needs. And, it should support them in sharing what they wish to share, at the same time, ensuring that the other understands the data in appropriate ways, not challenging coaches' authority nor harming the clients' self-image. It also asks for careful design of self-tracking devices, for example, allowing clients to capture their behaviors and experiences in their own customized ways, and allowing coaches to witness this process, as our results have shown how this this process had been highly informative to them. Thus, the theory of distributed cognition provides a lens that highlights how the system of coach, client and data may configure itself to acquire effective and satisfying coaching practices. For future work, it may be interesting to apply cognitive ethnography methods to understand this process even better.

Another holistic perspective derives from the field of Personal Informatics itself. There is increasing focus in literature on social tracking practices rather than *self*-tracking, for example making tracking a shared effort between patients and informal carers (Nunes & Fitzpatrick, 2015) and in family settings (Pina et al., 2017; Saksono et al., 2019). Kersten – Van Dijk and IJsselsteijn (2016) explain that self-tracking is inherently social, for example, tracking is in itself a form of self-presentation, and it facilitates social support and being hold accountable by others. Kersten – Van Dijk and IJsselsteijn (2016) also argue for a more socially constructivist point of view on data, where meaning is shaped by a socio-cultural environment. In this light, it is interesting to consider whether wearables are, and should be, presented as social actors or as tools (c.f., Fogg et al., 2009; Hancı, Ruijten, Lacroix, Kersten-van Dijk, & IJsselsteijn, 2019). For a client, both social actors and tools can be relevant and beneficial. For coaches, however, the presentation of tracking devices as social actors (for example, by calling them *e-coaches*) may easily trigger competition, as these devices are ought to work on the same tasks as coaches, and thus ask for careful coordination on how to approach this together. Our work has shown that coaches are better off with tools that support and augment them rather than social actors that compete with them. As a final remark on this issue, health data are often discussed in terms of how it transforms healthcare practices from paternalistic to participatory for clients (Pavel et al., 2013; Swan, 2009). We argue the same should apply to health coaches. For coaches to make effective use of health data, it is of key importance that e-health technologies allow for participatory rather than paternalistic practices, actively inviting coaches to engage.

In the very first paragraphs of this dissertation, we touched upon how data inherently represent information beyond the numbers, depending on the person who assesses them. Across the dissertation, we have shown how hard it is for coaches to interpret someone's personal health data without clients reflecting on them – similar to how hard it would be for you as a reader to interpret the data in this dissertation without our reflections in the discussions. In addition, we showed how coaches appreciated data-driven models that gave room for their own knowledge and reflections, for example by allowing them to implement their own value judgements on specific data. We hope that our work contributes to a better understanding of what constitutes good balance between humans and machines, where human intelligence is augmented and where human control is operationalized in meaningful ways. In addition, we hope that our work guides the design of wearable health technologies that utilize the complementary strengths of coaches, data and clients, leading to coaching strategies that are not only more effective, but also more satisfying.

Er is geen tijd te maar rustig aan.

Merel Morre

verliezen, dus doe

Biography

Heleen Rutjes was born as Heleen Muijlwijk on July 27th, 1985 in De Noordoostpolder, the Netherlands. After finishing her VWO high school degree in 2003 at the Emelwerda College in Emmeloord, she studied Psychology (BSc degree) and Science Education and Communication, track Mathematics (MSc degree) at the University of Twente, Enschede. In 2012 she graduated on Discrete Mathematics and Mathematical Programming, with a project on junction delay modelling at OmniTRANS Transport Planning Software (recently renamed as DAT Mobility), Deventer. As part of her graduation, Heleen also worked for a year as mathematics teacher at the high school Corderius College, Amersfoort. After graduation, she worked at Consultants in Quantitative Methods (CQM), Eindhoven, as a statistical consultant in the group Product and Process Innovation, where she worked on projects for various Research and Development departments of large companies. From 2015 she started a PhD project at the Eindhoven University of Technology, as part of the Data Science Flagship, a collaboration between the Data Science Center Eindhoven (DSC/e) and Philips B.V., of which the results are presented in this dissertation. Since 2021 Heleen is employed as Postdoctoral Researcher at the Human-Technology Interaction group, Eindhoven University of Technology. Heleen is married and has two children (2014 and 2017).

Acknowledgements

Het is niet zozeer van belang wat iemand doet, maar hoe iemand het beleeft. Dat is wat coaches mij vertelden. En het is waar, niet alleen als je iemand wilt coachen, maar ook als je een proefschrift wilt begrijpen. Ik heb dit onderzoek beleefd met en dankzij veel lieve en inspirerende mensen om mij heen. Dit proefschrift is het resultaat van ontelbare ervaringen die ik samen met jullie, collega's, vrienden en familie, heb meegemaakt.

De meest belangrijke drijfveer om dit onderzoek te beginnen was jij, Lieke. Jij werd geboren, jij maakte van mij een moeder, en daardoor realiseerde ik me dat ik werk wilde doen waar ik van hield en waar ik goed in was. Ik ben blij en dankbaar dat jij me hebt laten voelen wat ik moest doen: mezelf zijn. Onderzoek doen en moeder zijn was een fijne combinatie, ik ben een completer mens geworden door jou en de keuzes die ik door jou heb gemaakt. En toen, ergens halverwege Hoofdstuk 2 en 4, en 3 en 5, ben jij geboren Janne. Je hebt me laten zien hoe alles altijd goed is, en hoe tijd nooit verloren kan gaan, maar slechts goed besteed kan worden. Je forceerde je me om keuzes te maken, en om soms gewoon genoegen te nemen met wat er is. Lieve Lieke en Janne, bedankt dat jullie mij in balans hebben gehouden. Jullie hebben mij geleerd dat het niet altijd perfect hoeft te zijn. Jullie stimuleren mij om een rolmodel te zijn; niet alleen te praten over hoe het zou moeten, maar het ook echt te doen. Ik vind het heel spannend om stoer te zijn, maar ik doe het toch omdat ik het jullie ook gun, en als het lukt is het fantastisch. Ik wens dat jullie je altijd vrij zullen voelen om je hart te volgen. Ik ben ontzettend trots en dankbaar dat jullie er zijn. Door jullie is dit boek extra dik en vrolijk geworden!

Ik ben gezegend geweest met de beste begeleiders die ik me kan indenken, Martijn en Wijnand. We waren een sterk team, onze kracht zat hem in de verschillen. Gelukkig werken we bij de Human (Wijnand: de mens!) Technology (Martijn: de techniek!) Interaction (ik: de harmonie!) groep, anders hadden we het misschien niet eens zo lang met elkaar uitgehouden. Coaches hadden vaak het hoogste woord over het belang van een goede vertrouwensrelatie, en dat heb ik zelf aan den lijve ondervonden. Ik heb een rotsvast vertrouwen in jullie, niet alleen omdat ik jullie mening en visie enorm waardeer, maar ook omdat ik heb gevoeld dat jullie oprecht het beste met mij voor hebben. Het is nooit een issue geweest dat ik vrouw ben, een flapuit, zwanger, onzeker. Altijd, echt altijd, luisterden jullie naar wat ik zei en namen jullie me serieus. Dat is van onschatbare waarde geweest voor mijn zelfvertrouwen en werkplezier. Bedankt voor jullie grote vertrouwen in mij, de onuitputtelijke bron van inspiratie, de wijze raad, en het soms wel en soms niet geven van antwoorden. Bedankt voor de ruimte en stimulans om te groeien.

Many thanks to all committee members, for reading the manuscript and for giving me so much detailed and helpful feedback and thoughtful considerations. It was an honor to receive your opinions on my work, and it has substantially improved it. Een speciaal dank aan Boris de Ruyter, die mij, weliswaar wat meer op afstand, altijd gesteund heeft als derde begeleider. Bedankt dat je me de weg hebt gewezen binnen Philips, bedankt voor alle kritische vragen, en het vertrouwen dat je altijd in me hebt gehad.

A sincere and big thanks to all my HTI colleagues. I have felt at home from the start, and I have very much enjoyed your company. I have stretched my time at HTI over more than the usual number of years, so I have seen many of you coming, going, coming and going. There are too many names to mention, but you know who you are, my dear PhD peers, and also the lovely staff members. I've had countless encounters over coffee, hanging in door openings, or reaching over the monitors at our desks. All were helpful, reassuring, fun, and inspiring in one way or another. I appreciated the reflections on our work, our lives, and the world we live in. Thanks for bearing with me, even when I was ranting about broken nights or paper deadlines. It meant a lot to me. A special thanks to my paranymphs. Milou, ik ben blij dat ik je zo goed heb leren kennen de afgelopen jaren. Je hebt me enorm geholpen om rustig te blijven, trots te zijn, en te navigeren door mijn soms overvolle leven. Ik kijk er naar uit om ook de resterende tijd van jouw PhD mee te maken, en ik hoop dat we nog lang daarna onze avonturen samen kunnen delen. Anne, your critical view on things has always inspired me. I trust your opinion and I have always loved our conversations. I look forward to our collaborations in the near (or who knows far) future, I hope our paths will cross many times more! I feel blessed to have both of you backing me up at my defense. Tot slot een speciaal en groot dank aan Els, je hebt mijn PhD een fantastische kick-start gegeven. Je zorgde ervoor dat ik mijn plek vond bij HTI, je leerde me dat alle twijfels normaal zijn, zelfs nuttig, en je kritische en tegelijkertijd relativerende houding was enorm verfrissend. Ik heb je de laatste jaren enorm gemist, en ik ben blij dat we ondanks alles toch nog contact hebben kunnen houden!

I did not strictly stick to the HTI group, every now and then I broadened my horizon with other disciplines and environments. Dear Philosophy and Ethics colleagues, in particular the great thinkers Iris, Naomi and Marjolein, it has always been inspiring and fun to talk to you. Also my colleagues at the Jheronimus Academy of Data Science, JADS, thank you for making me feel at home. I have enjoyed the many interesting coffee breaks, meetings and presentations, it has broadened my view on data and the world. I look forward to collaborating more with you in the future.

Dit onderzoek is het resultaat van vele samenwerkingen – in mijn eentje was ik nooit zo ver gekomen. Bij Philips Design ben ik enorm versterkt en uitgedaagd door een paar van de beste industrieel ontwerpers ter wereld, Janne, Sander, en de jongens van Bureau Moeilijke Dingen, Joep, Paul en Pepijn. Hoofdstuk 4 is mede te danken aan jullie creativiteit, aan jullie mentaliteit van doorpakken, maken, bouwen, kiezen, regelen en gaan. Het was inspirerend om met jullie samen te werken. Ook mensen bij het Studenten Sportcentrum Eindhoven (SSC/e), de Fontys Sporthogeschool in Eindhoven, en de vele coaches die ik onderweg heb leren kennen, heel veel dank voor jullie bereidheid om met me mee te denken, mee te helpen van alles te organiseren, en bovenal, jullie mening te delen over data! Door jullie inzet en openheid heb ik dit verhaal kunnen schijven. Ik hoop van harte dat het recht doet aan jullie ideeën en dagelijkse praktijk. Ook veel dank aan mijn studenten, in het bijzonder Juliet, Tim en Ylja. Door jullie inzet zijn respectievelijk hoofdstuk 3, 5 en 6 zo mooi geworden. Geloof me, jullie zijn niet de enigen die van jullie afstuderen hebben geleerd.

Er zijn ook veel mensen die, zonder direct betrokken te zijn geweest bij mijn onderzoek, belangrijk zijn geweest voor mijn proces. Veel lieve vrienden hebben meegeleefd, gesteund, en me de nodige relativering en ontspanning geboden. Ik kan niet iedereen bij naam noemen, maar een paar mensen moet ik er even uitlichten. Katarina, je hebt me laten zien dat het mogelijk is, een PhD combineren met een gezin, en daar heb ik me altijd aan op kunnen trekken. Jouw erkenning voor zo'n beetje alles in mijn leven heeft mij altijd een enorme boost gegeven. Rogier, toevallig gingen we allebei tegelijk in de hoogste versnelling, jij voor je bedrijf, ik voor mijn PhD. Onze focus en de beperkte tijd die we daardoor samen hadden forceerde ons om alleen te bespreken wat er echt toe deed. Het was altijd precies wat ik nodig had. Dankjewel dat je altijd eerlijk bent geweest, en dankjewel dat je nooit hebt gezegd dat het teveel was wat ik wilde. En dan, lief Quenouille, samen is het altijd feest. Hoe fijn is het dat we elkaar altijd weer weten te inspireren, te helpen, uit te dagen, en gerust te stellen. Het is ontzettend fijn om altijd op jullie terug te kunnen vallen, in welke fase van ons leven we ons ook bevinden.

Lieve papa en mama, jullie zijn altijd betrokken geweest bij wat ik deed, ook al was het niet altijd makkelijk te volgen. Jullie zijn voor mij een fantastisch voorbeeld geweest in het volgen van je eigen pad, en jullie hebben me laten zien hoe je daarbij zo nodig van de gebaande paden af mag wijken. Je hart volgen, ver vooruitkijken, vertrouwen hebben. Daardoor heb ik ongebruikelijke keuzes durven maken in mijn studies en carrière, die mij uiteindelijk precies hebben gebracht waar ik wil zijn. Ik ben jullie daar ontzettend dankbaar voor. Ook alle praktische en emotionele steun aan ons gezin de afgelopen jaren hebben ervoor gezorgd dat deze fase boven alles een fijne tijd is geweest. Er is veel meer lieve familie die aan mijn zijde heeft gestaan, die ik niet allemaal kan benoemen, maar er is nog één heldin die ik wel met naam wil en moet noemen. Lieve oma Cathrien, wat ongelooflijk fijn dat jij er bent en dat we zoveel met elkaar kunnen delen. Het betekent veel voor me om nu ook het moederschap met jou te delen, en jouw bewondering en aanmoediging, vooral voor de combinatie van kinderen en werk, is ontzettend bemoedigend. Je bent een voorbeeld

in het relativeren, en de aandacht richten op dingen die er werkelijk toe doen. Het is zo ontzettend fijn en bijzonder om van jou te kunnen leren over het leven.

Tot slot, de spreekwoordelijk laatste auteur, de rol van iemand die niet zozeer meeschrijft als wel de boel van een afstandje bekijkt, de grote lijnen ziet en in perspectief plaatst, Michiel. De man, mijn man, die op de achtergrond alles mogelijk heeft gemaakt. Michiel, jij bent alles wat ik niet ben, en samen kunnen wij dus alles. Ik voel me sterk en weet dat alles uiteindelijk goed komt, als jij er maar bent. Je hebt me ontzettend goed gecoacht door de afgelopen jaren. Je gaf me een schop onder mijn kont als ik niet meer vooruit durfde, onder geen voorwaarde stond jij het toe dat ik mijzelf tekort zou doen, ultiem lief. Tegelijkertijd trapte je op de rem als ik te hard ging. Jouw idealisme en rechtvaardigheidsgevoel zijn voor mij altijd belangrijke pijlers geweest om op te varen. Ik ben ontzettend trots op en blij met alles wat wij samen hebben gemaakt, in het bijzonder onze kinderen, en hoe we ons leven leven. Je bent de beste. Zonder jou was deze PhD nooit zo leuk geweest, en het resultaat nooit zo betekenisvol geworden.

List of Publications

Rutjes, H., Willemsen, M. C., Bogers, S., Kollenburg, J. Van, & IJsselsteijn, W. A. (*Under review*). Evolving Practices and Value of Sharing Customized Home-Collected Data with Healthcare Professionals in Newborn Care: A Field Study.

Rutjes, H., Willemsen, M. C., Feijt, M. A., & IJsselsteijn, W. A. (*Under review*). The Influence of Personal Health Data on the Health Coaching Process.

Rutjes, H., Willemsen, M. C., Smyth, B. & IJsselsteijn, W. A. (*Under review*). Running Coaches' Interactions with a Marathon Prediction Tool.

Rutjes, H., Willemsen, M. C., & IJsselsteijn, W. A. (2020). Tailoring Transparency to Expertise: Health Professionals' Need for Transparency in Representing Self-Tracking Data of Clients. *HUMANIZE Workshop at the ACM Conference on Intelligent User Interfaces (IUI'20)*. Cagliari, Italy.

Rutjes, H., Willemsen, M. C., & IJsselsteijn, W. A. (2019). Beyond Behavior: The Coach's Perspective on Technology in Health Coaching. *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI '19)*. Glasgow, Scotland UK.

Rutjes, H., Willemsen, M. C., & IJsselsteijn, W. A. (2019). Considerations on Explainable AI and Users' Mental Models. *CHI'19 Workshop: Where Is the Human? Bridging the Gap between AI and HCI.* Glasgow, Scotland UK.

Kollenburg, J. Van, Bogers, S., Rutjes, H., Deckers, E., Frens, J., & Hummels, C. (2018). Exploring the Value of Parent-Tracked Baby Data in Interactions with Healthcare Professionals: A Data-Enabled Design Exploration. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '18)*. Montréal, Canada.

Rutjes, H., Kersten – van Dijk, E. T., Willemsen, M. C., & IJsselsteijn, W. A. (2018). Tell-Tale Data: The Value of Self-Tracked Data for Healthcare Professionals. *CHI '18 Workshop: A Short Workshop on Next Steps towards Long Term Self Tracking*. Montréal, Canada.

Rutjes, H., Willemsen, M. C., Kollenburg, J. Van, Bogers, S., & IJsselsteijn, W. A. (2017). Benefits and Costs of Patient Generated Data, From the Clinician's and Patient's Perspective. *PervasiveHealth '17 Workshop: Leveraging Patient-Generated Data for Collaborative Decision Making in Healthcare.* Barcelona, Spain.

Rutjes, H., Willemsen, M. C., Kersten – van Dijk, E. T., De Ruyter, B., & IJsselsteijn, W. A. (2017). Better Together: Opportunities for Technology in Health Coaching from the Coach's Perspective. In *Conference on E-Coaching for Health and Wellbeing*. Amsterdam, the Netherlands.

References

Aamodt, A., & Plaza, E. (1994). Case-Based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, *7*(1), 39–59.

Adams, A. T., Costa, J., Jung, M. F., & Choudhury, T. (2015). Mindless Computing: Designing Technologies to Subtly Influence Behavior. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*, 719–730.

Agapie, E., Chinh, B., Pina, L. R., Oviedo, D., Welsh, M. C., Hsieh, G., & Munson, S. A. (2018). Crowdsourcing Exercise Plans Aligned with Expert Guidelines and Everyday Constraints. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'18)*, 1–13.

Ajana, B. (2018). *Metric Culture: Ontologies of self-tracking practices*. Emerald Group Publishing.

Alpay, L. L., Henkemans, O. B., Otten, W., Rövekamp, T. A. J. M., & Dumay, A. C. M. (2010). E-health applications and services for patient empowerment: Directions for best practices in the Netherlands. *Telemedicine and E-Health*, *16*(7), 787–791.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine*, *35*(4), 105–120.

Ancker, J. S., Witteman, H. O., Hafeez, B., Provencher, T., Van De Graaf, M., & Wei, E. (2015). "You get reminded you're a sick person": Personal data tracking and patients with multiple chronic conditions. *Journal of Medical Internet Research*, *17*(8).

Ash, J. S., Berg, M., & Coiera, E. (2004). Some Unintended Consequences of Information Technology in Health Care : The Nature of Patient Care Information System-related Errors. *Journal of the American Medical Informatics Association*, *11*(2), 104–112.

Ayobi, A., Sonne, T., Marshall, P., & Cox, A. L. (2018). Flexible and Mindful Self-Tracking: Design Implications from Paper Bullet Journals. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '18)*.

Bainbridge, L. (1983). Ironies of Automation. *Automatica*, *19*(6), 775–779.

Bannister, F., & Connolly, R. (2011). The Trouble with Transparency: A Critical Review of Openness in e-Government. *Policy & Internet*, *3*(1), 158–187.

Baos, V., Ester, F., Castellanos, A., Nocea, G., Caloto, M. T., & Gerth, W. C. (2005). Use of a structured migraine diary improves patient and physician communication about migraine disability and treatment outcomes. *International Journal of Clinical Practice*, *59*(3), 281–286.

Barclay, L., Everitt, L., Rogan, F., Schmied, V., & Wyllie, A. (1997). Becoming a mother – An analysis of women's experience of early motherhood. *Journal of Advanced Nursing*, *25*(4), 719–728.

Beck, H. P., McKinney, J. B., Dzindolet, M. T., & Pierce, L. G. (2009). Effects of human-machine competition on intent errors in a target detection task. *Human Factors*, *51*(4), 477–486.

Beinema, T., op den Akker, H., van Velsen, L., & Hermens, H. (2021). Tailoring coaching strategies to users' motivation in a multi-agent health coaching application. *Computers in Human Behavior*, *121*.

Bensing, J. (1991). Doctor-Patient Communication and the Quality of Care. *Social*

Science & Medicine, *32*(11), 1301–1310.
Berry, A. B. L., Lim, C., Hartzler, A. L., Hirsch, T., Ludman, E., Wagner, E. H., & Ralston, J. D. (2017). Creating Conditions for Patients' Values to Emerge in Clinical Conversations. *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*, 1165–1174.
Berry, A. B. L., Lim, C. Y., Hirsch, T., Hartzler, A. L., Kiel, L. M., Bermet, Z. A., & Ralston, J. D. (2019). Supporting Communication About Values Between People with Multiple Chronic Conditions and their Providers. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '19)*, 1–14.
Bhatta, R., Aryal, K., & Ellingsen, G. (2015). Opportunities and Challenges of a Rural-telemedicine Program in Nepal. *Journal of Nepal Health Research Council*, *13*(30), 149–153.
Bogers, S., Frens, J., Kollenburg, J. Van, Deckers, E., & Hummels, C. (2016). Connected Baby Bottle : A Design Case Study Towards A Framework for Data-Enabled Design. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 301–311. Brisbane, Australia.
Boukhelifa, N., Bezerianos, A., & Lutton, E. (2018). Evaluation of Interactive Machine Learning Systems. In *Human and Machine Learning*.
Bowes, I., & Jones, R. L. (2006). Working at the Edge of Chaos: Understanding Coaching as a Complex, Interpersonal System. *The Sport Psychologist*, *20*, 235–245.
Boyatzis, R. E. (1998). *Transforming Qualitative Information: Thematic Analysis and Code Development*. Sage.
Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, *15*(5), 662–679.
Brandt, C. J., Isidora Søgaard, G., Clemensen, J., Søndergaard, J., & Bo Nielsen, J. (2018). Determinants of successful ehealth coaching for consumer lifestyle changes: Qualitative interview study among health care professionals. *Journal of Medical Internet Research*, *20*(7), 1–12.
Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.
Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597.
Brickwood, K.-J., Watson, G., O'Brien, J., & Williams, A. D. (2019). Do consumer-based activity trackers increase physical activity participation? A systematic review and meta-analysis. *JMIR MHealth and UHealth*, *7*(4).
Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. *Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI '15)*, 160–169.
Byambasuren, O., Sanders, S., Beller, E., & Glasziou, P. (2018). Prescribable mHealth apps identified from an overview of systematic reviews. *Digital Medicine*, *1*(1), 1–12.
Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., … Terry, M. (2019). Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems – CHI '19*, 1–14. Retrieved from http://arxiv.org/abs/1902.02960
Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). "Hello AI": Uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making. *PACM on Human-Computer Interaction*, *3*(CSCW).
Cardinale, M., & Varley, M. C. (2017). Wearable Training-Monitoring Technology: Applications, Challenges, and Opportunities. *International Journal of Sports Physiology and Performance*, *12*, 55–62.
Cassidy, T., Jones, R., & Potrac, P. (2009). *Understanding Sports Coaching: The Social, Cultural and Pedagogical Foundations of*

Coaching Practice (Second Edi). Routledge.

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, 1–12.

Choe, E. K., Abdullah, S., Rabbi, M., Thomaz, E., Epstein, D. A., Cordeiro, F., … Kientz, J. A. (2017). Semi-Automated Tracking: A Balanced Approach for Self-Monitoring Applications. *IEEE Pervasive Computing*, *16*(1), 74–84.

Choe, E. K., Lee, B., Zhu, H., Riche, N. H., & Baur, D. (2017). Understanding Self-Reflection: How People Reflect on Personal Data Through Visual Data Exploration. *Proceedings of the EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '17)*.

Choe, E. K., Lee, N. B., Lee, B., Pratt, W., & Kientz, J. A. (2014). Understanding quantified-selfers' practices in collecting and exploring personal data. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*, 1143–1152.

Choi, W. S., Park, J., Choi, J. Y. B., & Yang, J. S. (2019). Stakeholders' resistance to telemedicine with focus on physicians: Utilizing the Delphi technique. *Journal of Telemedicine and Telecare*, *25*(6), 378–385.

Chromik, M., Krüger, A., & Butz, A. (2021). I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. *Proceedings of the Conference on Intelligent User Interfaces (IUI '21)*, 307–317.

Chung, C.-F., Cook, J., Bales, E., Zia, J., & Munson, S. A. (2015). More than telemonitoring: Health provider use and nonuse of life-log data in irritable bowel syndrome and weight management. *Journal of Medical Internet Research*, *17*(8), 1–16.

Chung, C.-F., Dew, K., Cole, A., Zia, J., Fogarty, J., Kientz, J. A., & Munson, S. A. (2016). Boundary Negotiating Artifacts in Personal Informatics: Patient-Provider Collaboration with Patient-Generated Data. *The 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '16)*, 770–786. San Francisco, USA.

Chung, C.-F., Wang, Q., Schroeder, J., Cole, A., Zia, J., Fogarty, J., & Munson, S. (2019). Identifying and Planning for Individualized Change: Patient-Provider Collaboration Using Lightweight Food Diaries in Healthy Eating and Irritable Bowel Syndrome. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(1), 1–27.

Claudino, J. G., de Oliveira Capanema, D., Vieira de Souza, T., Cerca Serrão, J., Machado Pereira, A. C., & Nassis, G. P. (2019). Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review. *Sports Medicine – Open*, *5*(28).

Collins, D., Collins, L., & Carson, H. J. (2016). "If it feels right, do it": Intuitive decision making in a sample of high-level sport coaches. *Frontiers in Psychology*, *7*.

Collins, S. A., Fred, M., Wilcox, L., & Vawdrey, D. K. (2012). Workarounds Used by Nurses to Overcome Design Constraints of Electronic Health Records. *Proceedings of the 11th International Congress on Nursing Informatics*. American Medical Informatics Association.

Consolvo, S., McDonald, D. W., & Landay, J. A. (2009). Theory-Driven Design Strategies for Technologies that Support Behavior Change in Everyday Life. *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems (CHI '09)*, 405–414.

Coppers, S., Van Den Bergh, J., Luyten, K., Coninx, K., Van Der Lek-Ciudin, I., Vanallemeersch, T., & Vandeghinste, V. (2018). Intellingo: An intelligible translation environment. *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI '18)*.

Cordeiro, F., Epstein, D. A., Thomaz, E., Bales, E., Jagannathan, A. K., Abowd, G. D., & Fogarty, J. (2015). Barriers and Negative

Nudges: Exploring Challenges in Food Journaling. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '15)*, 1159–1162.

Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., … Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496.

Cummings, M. (2014). Man versus Machine or Man + Machine ? *IEEE Computer Society*, *29*(5), 62–69.

Dallinga, J., Janssen, M., van der Werf, J., Walravens, R., Vos, S., & Deutekom, M. (2018). Analysis of the features important for the effectiveness of physical activity–related apps for recreational sports: Expert panel approach. *JMIR MHealth and UHealth*, *6*(6).

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*(2), 95–106.

Deci, E. L., Ryan, R. M., & Koestner, R. (1999). A Meta-Analytic Review of Eperiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin*, *125*(6), 627–668.

Dekker, S. W. a., & Woods, D. D. (2002). MABA-MABA or Abracadabra? Progress on Human-Automation Co-ordination. *Cognition, Technology & Work*, *4*(4), 240–244.

Demiris, G., Iribarren, S. J., Sward, K., Lee, S., & Yang, R. (2019). Patient generated health data use in clinical practice: A systematic review. *Nursing Outlook*, *67*(4), 311–330.

Desai, P. M., Mitchell, E. G., Hwang, M. L., Levine, M. E., Albers, D. J., & Mamykina, L. (2019). Personal Health Oracle: Explorations of Personalized Predictions in Diabetes Self-Management. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems – CHI '19*.

Devaraj, S., Sharma, S. K., Fausto, D. J., Viernes, S., & Kharrazi, H. (2014). Barriers and Facilitators to Clinical Decision Support Systems Adoption: A Systematic Review. *Journal of Business Administration Research*, *3*(2).

Dey, A. K., & Abowd, G. D. (1999). Towards a Better Understanding of Context and Context-Awareness. *Computing Systems*, *40*(3), 304–307.

Dill, D., & Gumpert, P. (2012). What Is the Heart of Health Care? Advocating for and Defining the Clinical Relationship in Patient-Centered Care. *Journal of Participatory Medicine*, *4*(1).

Dinh-Le, C., Chuang, R., Chokshi, S., & Mann, D. (2019). Wearable health technology and electronic health record integration: Scoping review and future directions. *JMIR MHealth and UHealth*, *7*(9).

Dix, A. (2007). Designing for appropriation. *British Computer Society Conference on Human-Computer Interaction*, *2*.

Doherty, C., Keogh, A., Davenport, J., Lawlor, A., Smyth, B., & Caulfield, B. (2020). An evaluation of the training determinants of marathon performance: A meta-analysis with meta-regression. *Journal of Science and Medicine in Sport*, *23*(2), 182–188.

Dudley, J. J., & Kristensson, P. O. (2018). A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(2), 1–37.

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018). Bringing transparency design into practice. *Proceedings of the Conference on Intelligent User Interfaces (IUI '18)*, 211–223.

Elsden, C., Kirk, D. S., & Durrant, A. C. (2016). A Quantified Past: Toward Design for Remembering With Personal Informatics. *Human-Computer Interaction*, *31*(6), 518–557.

Epstein, D. A., Caldeira, C., Figueiredo, M. C., Silva, L. M., Lu, X., Williams, L., … Chen, Y. (2020). Mapping and Taking Stock of the Personal Informatics Literature. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(4).

Epstein, D. A., Cordeiro, F., Bales, E., Fogarty, J., & Munson, S. (2014). Taming Data Complexity in Lifelogs: Exploring Visual Cuts of Personal Informatics Data. *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS '14)*, 667–676.

Epstein, D. A., Jacobson, B. H., Bales, E., McDonald, D. W., & Munson, S. A. (2015). From "nobody cares" to "way to go!": A Design Framework for Social Sharing in Personal Informatics. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*, 1622–1636.

Eslami, M., Krishna Kumaran, S. R., Sandvig, C., & Karahalios, K. (2018). Communicating Algorithmic Process in Online Behavioral Advertising. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*.

Etkin, J. (2016). The Hidden Cost of Personal Quantification. *Journal of Consumer Research*, *42*(6), 967–984.

Fails, J. A., & Olsen, D. R. (2003). Interactive Machine Learning. *Proceedings of the Conference on Intelligent User Interfaces (IUI '03)*, 39–45.

Figueiredo, M. C., & Chen, Y. (2020). Patient-generated health data: Dimensions, challenges, and open questions. *Foundations and Trends in Human-Computer Interaction*, *13*(3), 165–297.

Figueiredo, M. C., Su, H. I., & Chen, Y. (2020). Using Data to Approach the Unknown: Patients' and Healthcare Providers' Data Practices in Fertility Challenges. *PACM on Human-Computer Interaction*, *4*(CSCW).

Fitts, P. M. (1951). *Human engineering for an effective air-navigation and traffic-control system*. Washington, D.C.

Fleck, R., & Fitzpatrick, G. (2010). Reflecting on reflection: Framing a design landscape. *Proceedings of the Australian Conference on Computer-Human Interaction (OzCHI '10)*.

Fogg, B. J. (2002). Persuasive Technology: Using Computers to Change What We Think and Do. *Ubiquity*.

Fogg, B. J., Cuellar, G., & Danielson, D. (2007). Motivating, Influencing, and Persuading Users. In *The Human-Computer Interaction Handbook* (pp. 159–172).

Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI '99)*, 80–87.

Forouzanfar, M. H., Afshin, A., Alexander, L. T., Biryukov, S., Brauer, M., Cercy, K., ... Zhu, J. (2016). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, *388*(10053), 1659–1724.

Foster, C., Schrager, M., Snyder, A. C., & Thompson, N. N. (1994). Pacing Strategy and Athletic Performance. *Sports Medicine*, *17*(2), 77–85.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280.

Gabriels, K., & Moerenhout, T. (2018). Exploring entertainment medicine and professionalization of self-care: Interview study among doctors on the potential effects of digital self-tracking. *Journal of Medical Internet Research*, *20*(1), 1–12.

Gagnon, M. P., Ngangue, P., Payne-Gagnon, J., & Desmartis, M. (2016). M-Health adoption by healthcare professionals: A systematic review. *Journal of the American Medical Informatics Association*, *23*(1), 212–220.

Gaunt, K., Nacsa, J., & Penz, M. (2014). Baby Lucent: Pitfalls of applying quantified self to baby products. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*.

Gaver, W. W., Beaver, J., & Benford, S. (2003). Ambiguity as a Resource for Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*, *5*(1), 233–240.

Gigerenzer, G. (2008). Why Heuristics Work.

Perspectives on Psychological Science, *3*(1), 20–29.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53–96.

Glass, A., McGuinness, D. L., & Wolverton, M. (2008). Toward establishing trust in adaptive agents. *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI '08)*, 227–236.

Green, W., & De Ruyter, B. (2010). The design and evaluation of interactive systems with perceived social intelligence: five challenges. *AI & Society*, *25*(2), 203–210.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30.

Gui, C., & Chan, V. (2017). Medicine learning in medicine. *Medicine & Technology*, *86*(2), 76–78.

Ha, J. F., Anat, D. S., & Longnecker, N. (2010). Doctor-Patient Communication: A Review. *The Ochsner Journal*, *10*, 38–43.

Halson, S. L. (2014). Monitoring Training Load to Understand Fatigue in Athletes. *Sports Medicine*, *44*(2), 139–147.

Hanci, E., Ruijten, P. A. M., Lacroix, J., Kersten-van Dijk, E. T., & IJsselsteijn, W. A. (2019). Are trackers social actors? The role of self-tracking on self-evaluation. *Proceedings of the International Conference on Persuasive Technology*, 31–42.

Hayes, G. R., Cheng, K. G., Hirano, S. H., Tang, K. P., Nagel, M. S., & Baker, D. E. (2014). Estrellita: A mobile capture and access tool for the support of preterm infants and their caregivers. *ACM Transactions on Computer-Human Interaction*, *21*(3).

Hermsen, S., Frost, J., Renes, R. J., & Kerkhof, P. (2016). Using feedback through digital technology to disrupt and change habitual behavior: A critical review of current literature. *Computers in Human Behavior*, *57*, 61–74.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research. *ACM Transactions on Computer-Human Interaction*, *7*(2), 174–196.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119–131.

Hong, M. K., Lakshmi, U., Olson, T. A., & Wilcox, L. (2018). Visual ODLs: Co-Designing Patient-Generated Observations of Daily Living to Support Data-Driven Conversations in Pediatric Care. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '18)*, 1–13.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

Hutchins, E. (2000). Distributed Cognition. In *International Encyclopedia of the Social and Behavioral Sciences*. (p. 138).

Hutchins, E., & Klausen, T. (1996). Distributed Cognition in an Airline Cockpit. In *Cognition and communication at work* (pp. 15–34).

IJsselsteijn, W., De Kort, Y., Midden, C., Eggen, B., & Van Den Hoven, E. (2006). Persuasive technology for human well-being: setting the scene. In *International conference on persuasive technology* (pp. 1–5). Berlin, Heidelberg: Springer.

Institute of Medicine. (2001). *Health and Behavior: The Interplay of Biological, Behavioral and Societal Influences*.

Ives, Y. (2008). What is "Coaching"? An Exploration of Conflicting Paradigms. *International Journal of Evidence Based Coaching and Mentoring*, *6*(2), 100–113.

Jacob, C., Sanchez-vazquez, A., & Ivory, C. (2020). Social, Organizational, and Technological Factors Impacting Clinicians ' Adoption of Mobile Health Tools: Systematic Literature Review. *JMIR MHealth and UHealth*, *8*(2).

Janssen, M., Goudsmit, J., Lauwerijssen, C., Brombacher, A., Lallemand, C., & Vos, S. (2020). How do runners experience personalization of their training scheme: The inspirun e-coach? *Sensors*, *20*(16), 1–19.

Janssen, M., Scheerder, J., Thibaut, E., Brombacher, A., & Vos, S. (2017). Who uses running apps and sports watches? Determinants and consumer profiles of event runners' usage of running-related smartphone applications and sports watches. *PLoS ONE*, *12*(7), 1–17.

Jhaver, S., & Antin, J. (2018). Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. *Proceedings of the ACM Conference on Human Factors in Computing Systems Proceedings (CHI 2018).*

Jo, E., Toombs, A. L., Gray, C. M., & Hong, H. (2020). *Understanding Parenting Stress through Co-designed Self-Trackers*. 1–13.

Jones, R. L., & Wallace, M. (2005). Another bad day at the training ground: Coping with ambiguity in the coaching context. *Sport, Education and Society*, *10*(1), 119–134.

Kahneman, D. (2011). *Thinking fast and slow*. London: Penguin Books Ltd.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science*, *306*(5702), 1776–1780.

Kahneman, D., & Riis, J. (2012). Living, and thinking about it: Two perspectives on life. *The Science of Well-Being*, 285–304.

Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. In *Judgment under uncertainty: Heuristics and biases*.

Karapanos, E., Gouveia, R., Hassenzahl, M., & Forlizzi, J. (2016). Wellbeing in the Making: Peoples' Experiences with Wearable Activity Trackers. *Psychology of Well-Being*, *6*(4), 1–17.

Karkar, R., Zia, J., Vilardaga, R., Mishra, S. R., Fogarty, J., Munson, S. A., & Kientz, J. A. (2016). A framework for self-experimentation in personalized health. *Journal of the American Medical Informatics Association*, *23*(3), 440–448.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Vaughan, J. W. (2020). Interpreting Interpretability: Understanding Data Scientists ' Use of Interpretability Tools for Machine Learning. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '20).*

Kelley, C., Lee, B., & Wilcox, L. (2017). Self-tracking for Mental Wellness: Understanding Expert Perspectives and Student Experiences. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '17)*, 629–641.

Keogh, A., Smyth, B., Caulfield, B., Lawlor, A., Berndsen, J., & Doherty, C. (2019). Prediction equations for marathon performance: A systematic review. *International Journal of Sports Physiology and Performance*, *14*(9), 1159–1169.

Kersten – Van Dijk, E. T., & IJsselsteijn, W. A. (2016). Design beyond the numbers: sharing, comparing, storytelling and the need for a Quantified Us. *Interaction Design and Architecture(s) Journal (IxD&A)*, *29*, 121–135.

Kersten – Van Dijk, E. T., IJsselsteijn, W. A., & Westerink, J. H. D. M. (2016). Deceptive visualizations and user bias: a case for personalization and ambiguity in PI visualizations. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (Ubicomp '16)*, 588–593.

Kersten – van Dijk, E. T., Westerink, J. H. D. M., Beute, F., & IJsselsteijn, W. A. (2017). Personal Informatics, Self-Insight and Behavior Change: a Critical Review of Current Literature. *Human–Computer Interaction*, *32*(5–6), 268–296.

Khairat, S., Marc, D., Crosby, W., & Al Sanousi, A. (2018). Reasons for physicians not adopting clinical decision support systems: Critical analysis. *Journal of Medical Internet Research*, *20*(4).

Kidder, L. H., & Fine, M. (1987). Qualitative and quantitative methods: When stories converge. *New Directions for Program Evaluation*, *1987*(35), 57–75.

Kientz, J. A., Arriaga, R. I., & Abowd, G. D.

(2009). Baby steps: Evaluation of a system to support record- keeping for parents of young children. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1713–1722.

Kientz, J. A., Arriaga, R. I., Chetty, M., Hayes, G. R., Richardson, J., Patel, S. N., & Abowd, G. D. (2007). Grow and Know: Understanding Record-Keeping Needs for Tracking the Development of Young Children. *CHI 2007 Procedings. Kids & Family*, 1351–1360.

Kim, Y.-H., Jeon, J. H., Lee, B., Choe, E. K., & Seo, J. (2017). OmniTrack: A Flexible Self-Tracking Approach Leveraging Semi-Automated Tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 67:1-67:28.

Kim, Y., Heo, E., Lee, H., Ji, S., Choi, J., Kim, J.-W., … Yoo, S. (2017). Prescribing 10,000 Steps Like Aspirin. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '17)*, 5787–5799.

Kizilcec, R. F. (2016). How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '16)*.

Klein, G. (2008). Naturalistic decision making. *Human Factors*, *50*(3), 456–460.

Klein, G., Hoffman, R., & Mueller, S. T. (2019). Naturalistic Psychological Model of Explanatory Reasoning: How People Explain Things to Others and to Themselves. *Proceedings International Conference on Naturalistic Decision Making 2019*.

Klein, G., Shneiderman, B., Hoffman, R. R., & Ford, K. M. (2017). Why Expertise Matters: A Response to the Challenges. *IEEE Intelligent Systems*, *32*(6), 67–73.

Koch, J., Lucero, A., Hegemann, L., & Oulasvirta, A. (2019). May AI?: Design Ideation with Cooperative Contextual Bandits. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems – CHI '19*, 1–12.

Kocielnik, R., Xiao, L., Avrahami, D., & Hsieh, G. (2018). Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(2), 1–26.

Kollenburg, J. Van, & Bogers, S. (2019). *Data-Enabled Design: a situated design approach that uses data as creative material when designing for intelligent ecosystems*. Eindhoven University of Technology.

Kollenburg, J. Van, Bogers, S., Rutjes, H., Deckers, E., Frens, J., & Hummels, C. (2018). Exploring the Value of Parent-Tracked Baby Data in Interactions with Healthcare Professionals: A Data-Enabled Design Exploration. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '18)*. Montréal, Canada.

Lavallee, D. C., Lee, J. R., Austin, E., Bloch, R., Lawrence, S. O., McCall, D., … Amtmann, D. (2020). mHealth and patient generated health data: stakeholder perspectives on opportunities and barriers for transforming healthcare. *MHealth*, *6*(8).

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80.

Lee, K., & Hong, H. (2018). MindNavigator. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, 1–14.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy and Technology*, *31*(4), 611–627.

Lewis, R., Liu, Y., Groh, M., & Picard, R. (2021). Shaping Habit Formation Insights with Shapley Values: Towards an Explainable AI-system for Self-understanding and Health Behavior Change. *CHI '21 Workshop: Realizing AI in Healthcare: Challenges Appearing in the Wild Workshop*.

Li, I., Dey, A. K., & Forlizzi, J. (2011). Understanding my data, myself: Supporting self-reflection with ubicomp. *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*.

Lim, B. Y., & Dey, A. K. (2011). Investigating intelligibility for uncertain context-aware applications. *Proceedings of the International Conference on Ubiquitous Computing (UbiComp '11)*.

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI '09)*, 2119–2128.

Lin, Y., Guo, J., Chen, Y., Yao, C., & Ying, F. (2020). It Is Your Turn: Collaborative Ideation with a Co-Creative Robot through Sketch. *Proceedings of the SIGCHI Conference of Human Factors in Computing Systems (CHI '20)*, 1–14.

Lordon, R. J., Mikles, S. P., Kneale, L., Evans, H. L., Munson, S. A., Backonja, U., & Lober, W. B. (2020). How patient-generated health data and patient-reported outcomes affect patient-clinician relationships: A systematic review. *Health Informatics Journal*, 1–18.

Lu, K., & Brombacher, A. (2020). Haptic feedback in running: Is it possible for information transfer through electrical muscle signalling? *TEI 2020 – Proceedings of the 14th International Conference on Tangible, Embedded, and Embodied Interaction*, 479–485.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NIPS 2017)*, 4766–4775.

Luo, Y., Liu, P., & Choe, E. K. (2019). Co-Designing Food Trackers with Dietitians: Identifying Design Opportunities for Food Tracker Customization. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '19)*. Glasgow, Scotland UK.

Lupton, D. (2011). "The best thing for the baby": Mothers' concepts and experiences related to promoting their infants' health and development. *Health, Risk and Society*, *13*(7–8), 637–651.

Lupton, D. (2013). Quantifying the body: Monitoring and measuring health in the age of mHealth technologies. *Critical Public Health*, *23*(4), 393–403.

Lupton, D. (2016a). *The quantified self*. John Wiley & Sons.

Lupton, D. (2016b). You are Your Data: Self-tracking Practices and Concepts of Data. In S. Selke (Ed.), *Lifelogging: Theoretical Approaches and Case Studies about Self- tracking* (pp. 61–79).

Lupton, D. (2020). Caring dataveillance: Women's use of apps to monitor pregnancy and children. *The Routledge Companion to Digital Media and Children*, 393–402.

Lyle, J. (2010). Coaches' Decision Making: A Naturalistic Decision Making Analysis. In *Sports Coaching E-book: Professionalisation and Practice* (pp. 27–42). Elsevier Health Sciences.

Mahajan, A., Pottie, G., & Kaiser, W. (2020). Transformation in Healthcare by Wearable Devices for Diagnostics and Guidance of Treatment. *ACM Transactions on Computing for Healthcare*, *1*(1), 1–12.

Maitland, J., & Chalmers, M. (2011). Designing for peer involvement in weight management. *Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems (CHI '11)*, 315–324.

Marakas, G. M., Johnson, R. D., & Clay, P. F. (2007). The Evolving Nature of the Computer Self-Efficacy Construct: An Empirical Investigation of Measurement Construction, Validity, Reliability and Stability Over Time. *Journal of the Association for Information Systems*, *8*(1), 16–46.

Marcu, G., Dey, A., & Kiesler, S. (2014). Designing for Collaborative Reflection. *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '14)*.

Masters, K. S., Ogles, B. M., & Jolton, J. A. (1993). The development of an instrument to measure motivation for marathon running: The motivations of marathoners scales (moms). *Research Quarterly for Exercise and Sport*, *64*(2), 134–143.

Mathison, S. (1988). Why Triangulate? *Educational Researcher*, *17*(2), 13–17.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *Academy of Management Review*, *20*(3), 709–734.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing And Validating Trust Measure for E-Commerce: An Integrative Typology. *Information Systems Research*, *13*(3), 334–359.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*.

Mentis, H. M. (2010). *Emotion Awareness and Invisibility in an Emergency Room: a Socio-Technical Dilemma*. The Pennsylvania State University.

Mentis, H. M., Komlodi, A., Schrader, K., Phipps, M., Gruber-Baldini, A., Yarbrough, K., & Shulman, L. (2017). Crafting a View of Self-Tracking Data in the Clinical Visit. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '17)*, 5800–5812.

Michie, S., Ashford, S., Sniehotta, F. F., Dombrowski, S. U., Bishop, A., & French, D. P. (2011). A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: the CALO-RE taxonomy. *Psychology & Health*, *26*(11), 1479–1498.

Mileski, M., Kruse, C. S., Catalani, J., & Haderer, T. (2017). Adopting telemedicine for the self-management of hypertension: Systematic review. *JMIR Medical Informatics*, *5*(4).

Mitchell, E. G., Heitkemper, E. M., Burgermaster, M., Levine, M. E., Hwang, M. L., Cassells, A., … Mamykina, L. (2021). From Reflection to Action: Combining Machine Learning with Expert Knowledge for Nutrition Goal Recommendations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '21)*.

Mohan, S., Venkatakrishnan, A., & Hartzler, A. L. (2020). Designing an AI Health Coach and Studying its Utility in Promoting Regular Aerobic Exercise. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(2).

Muramatsu, J., & Pratt, W. (2001). Transparent queries: Investigating users' mental models of search engines. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Informatino Retrieval (SIGIR '01)*, 217–224.

Murphy, J. W., & Franz, B. A. (2016). Narrative medicine in a hectic schedule. *Medicine, Health Care and Philosophy*, *19*(4), 545–551.

Nahum-Shani, I., Hekler, E. B., & Spruijt-Metz, D. (2015). Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology*, *34*, 1209–1219.

Niess, J., Diefenbach, S., & Woźniak, P. W. (2020). Persuasive Feedback for Fitness Apps: Effects of Construal Level and Communication Style. *Proceedings of the International Conference on Persuasive Technology*. Springer, Cham.

Nittas, V., Lun, P., Ehrler, F., Puhan, M. A., & Mütsch, M. (2019). Electronic Patient-Generated Health Data to Facilitate Disease Prevention and Health Promotion: Scoping Review. *Journal of Medical Internet Research*, *21*(10).

Norcross, J. C. (2002). *Psychotherapy relationships that work: Therapist contributions and responsiveness to patients*. New York: Oxford University Press.

Norman, D. A. (1986). Cognitive Engineering. In D. A. Norman & S. W. Draper (Eds.), *User Centered System Design*. CRC Press.

Nunes, F., Andersen, T., & Fitzpatrick, G. (2019). The agency of patients and carers in medical care and self-care technologies for interacting with doctors. *Health Informatics Journal*, *25*(2), 330–349.

Nunes, F., & Fitzpatrick, G. (2015). Self-Care Technologies and Collaboration. *International Journal of Human-Computer Interaction*, *31*(12), 869–881.

Nunes, F., Verdezoto, N., Fitzpatrick, G., Kyng, M., Gronvall, E., & Storni, C. (2015). Self-Care Technologies in HCI: Trends, Tensions, and Opportunities. *ACM Transactions on Computer-Human Interaction*, *22*(6).

O'Broin, A., & Palmer, S. (2006). The coach-client relationship and contributions made by the coach in improving coaching outcomes. *The Coaching Psychologist*, *2*(2), 16–20.

Ogles, B., & Masters, K. (2003). A Typology of Marathon Runners Based on Cluster Analysis of Motivations. *Journal of Sport Behavior*, *26*(1), 69.

Ohlin, F., & Olsson, C. M. (2015). Intelligent Computing in Personal Informatics: Key Design Considerations. *Proceedings of the International Conference on Intelligent User Interfaces (IUI '15)*, 263–274.

Olsen, J. M. (2014). Health Coaching: A Concept Analysis. *Nursing Forum*, *49*(1), 18–29.

Olsen, J. M., & Nesbitt, B. J. (2010). Health coaching to improve healthy lifestyle behaviors: An integrative review. *American Journal of Health Promotion*, *25*(1), 1–13.

Ong, L. M. L., de Haes, J. C. J. M., Hoos, A. M., & Lammes, F. B. (1995). Doctor-patient communication: a review of the literature. *Social Science & Medicine*, *40*(7), 903–918.

Orji, R., Lomotey, R., Oyibo, K., Orji, F., Blustein, J., & Shahid, S. (2018). Tracking feels oppressive and 'punishy': Exploring the costs and benefits of self-monitoring for health and wellness. *Digital Health*, *4*, 1–10.

Oygür, I., Su, Z., Epstein, D. A., & Chen, Y. (2021). The Lived Experience of Child-Owned Wearables: Comparing Children's and Parent's Perspectives on Activity Tracking. *Proceedings of the SIGCHI Conference of Human Factors in Computing Systems (CHI '21)*.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics*, *30*(3), 286–297.

Patel, M. S., Asch, D. A., & Volpp, K. G. (2015). Wearable Devices as Facilitators, Not Drivers,of Health Behavior Change. *Journal of the American Medical Association*, *313*(5), 459–460.

Pavel, M., Jimison, H. B., Wactlar, H. D., Hayes, T. L., Barkis, W., Skapik, J., & Kaye, J. (2013). The role of technology and engineering models in transforming healthcare. *IEEE Reviews in Biomedical Engineering*, *6*, 156–177.

Pichon, A., Schiffer, K., Horan, E., Massey, B., Bakken, S., Mamykina, L., & Elhadad, N. (2020). Divided We Stand: The Collaborative Work of Patients and Providers in an Enigmatic Chronic Disease. *PACM on Human-Computer Interaction*, *4*(CSCW3).

Pina, L. R., Sien, S., Ward, T., Yip, J. C., Munson, S. A., Fogarty, J., & Kientz, J. A. (2017). From Personal Informatics to Family Informatics: Understanding Family Practices around Health Monitoring. *Proceedings of the 2017 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '17)*, 2300–2315.

Pobiruchin, M., Suleder, J., Zowalla, R., & Wiesner, M. (2017). Accuracy and Adoption of Wearable Technology Used by Active Citizens: A Marathon Event Field Study. *JMIR MHealth and UHealth*, *5*(2).

Potapov, K., Vasalou, A., Lee, V., & Marshall, P. (2021). What do teens make of personal informatics? Youth responses to self-tracking practices within a classroom setting. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '21)*.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the SIGCHI Conference of Human Factors in Computing Systems (CHI '21)*.

Purpura, S., Schwanda, V., Williams, K., Stubler, W., & Sengers, P. (2011). Fit4Life: The Design of a Persuasive Technology Promoting Healthy Behavior and Ideal

Weight. *Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems (CHI '11)*, 423–432.

Puussaar, A., Clear, A. K., & Wright, P. (2017). Enhancing Personal Informatics Through Social Sensemaking. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*, 6936–6942.

Raad voor Volksgezondheid en Samenleving. (2017). *Zonder context geen bewijs*. 1–90. Retrieved from https://www.raadrvs.nl/uploads/docs/Zonder_context_geen_bewijs.pdf

Raj, S., Lee, J. M., Garrity, A., & Newman, M. W. (2019). Clinical Data in Context: Towards Sensemaking Tools for Interpreting Personal Health Data. *Proceedings of the ACM Conference on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(1).

Raj, S., Newman, M. W., Lee, J. M., & Ackerman, M. S. (2017). Understanding Individual and Collaborative Problem-Solving with Patient-Generated Data: Challenges and Opportunities. *Proceedings ACM on Human-Computer Interaction*, *1*(2).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-Agnostic Interpretability of Machine Learning. *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. Retrieved from

Richards, O. K., Choi, A., & Marcu, G. (2021). Shared Understanding in Care Coordination for Children's Behavioral Health. *PACM on Human-Computer Interaction*, *5*(CSCW1).

Riegel, P. S. (1977). Time Predicting. *Runner's World Magazin*, *12*(8).

Robinson, J. D. (2003). An Interactional Structure of Medical Activities During Acute Visits and Its Implications for Patients' Participation. *Health Communication*, *15*(1), 27–59.

Rooksby, J., Rost, M., Morrison, A., & Chalmers, M. (2014). Personal tracking as lived informatics. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*, 1163–1172.

Rosen, L. D., Whaling, K., Carrier, L. M., Cheever, N. A., & Rokkum, J. (2013). The Media and Technology Usage and Attitudes Scale: An empirical investigation. *Computers in Human Behavior*, *29*(6), 2501–2511.

Roter, D. L., & Hall, J. A. (2006). *Doctors talking with patients / patients talking with doctors: improving communication in medical visits*. Westport: Praeger Publishers.

Rutjes, H., Kersten – van Dijk, E. T., Willemsen, M. C., & IJsselsteijn, W. A. (2018). Tell-Tale Data: The Value of Self-Tracked Data for Healthcare Professionals. *CHI '18 Workshop: A Short Workshop on Next Steps Towards Long Term Self Tracking*.

Rutjes, H., Willemsen, M. C., & IJsselsteijn, W. A. (2019). Considerations on Explainable AI and Users' Mental Models. *CHI'19 Workshop: Where Is the Human? Bridging the Gap Between AI and HCI*.

Rutjes, H., Willemsen, M. C., Kollenburg, J. Van, Bogers, S., & IJsselsteijn, W. A. (2017). Benefits and Costs of Patient Generated Data, From the Clinician'ss and Patient's Perspective. *PervasiveHealth '17 Workshop: Leveraging Patient-Generated Data for Collaborative Decision Making in Healthcare*.

Saksono, H., Castaneda-Sceppa, C., Hoffman, J., Seif El-Nasr, M., Morris, V., & Parker, A. G. (2019). Social Reflections on Fitness Tracking Data. *Proceedings of the SIGCHI Conference of Human Factors in Computing Systems (CHI '19)*.

Salzinger, K. (2005). Clinical, Statistical, and Broken-Leg Predictions. *Behavior and Philosophy*, *33*, 91–99.

Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2014). Understanding reliance on automation: Effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science*, *15*(2), 134–160.

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). I Can Do Better Than Your AI: Expertise and Explanations. *Proceedings of the 24th International Conference on Intelligent User*

Interfaces (IUI '19, 240–251.

Scheerder, J., Breedveld, K., & Borgers, J. (2015). *Running across Europe: the rise and size of one of the largest sport markets*. Springer.

Schroeder, J., Hoffswell, J., Chung, C.-F., Fogarty, J., Munson, S., & Zia, J. (2017). Supporting Patient-Provider Collaboration to Identify Individual Triggers using Food and Symptom Journals. *Proceedings of the 2017 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '17)*, 1726–1739. New York, USA.

Seneviratne, S., Hu, Y., Nguyen, T., Lan, G., Khalifa, S., Thilakarathna, K., ... Seneviratne, A. (2017). A Survey of Wearable Devices and Challenges. *IEEE Communications Surveys and Tutorials*, *19*(4), 2573–2620.

Sforzo, G. A., Kaye, M. P., Todorova, I., Harenberg, S., Costello, K., Cobus-Kuo, L., ... Moore, M. (2018). Compendium of the health and wellness coaching literature. *American Journal of Lifestyle Medicine*, *12*(6), 1–12.

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction*, *36*(6), 495–504.

Smyth, B. (2018). Fast starters and slow finishers: A large-scale data analysis of pacing at the beginning and end of the marathon for recreational runners. *Journal of Sports Analytics*, *4*(3), 229–242.

Smyth, B., & Cunningham, P. (2017). Running with cases: A CBR approach to running your best marathon. *Lecture Notes in Computer Science*, *10339*.

Spelt, H. (2021). *Towards Physiology-Aware Persuasive Technology: A study on psychophysiological reactions to persuasive messages*. Eindhoven University of Technology.

Springer, A., & Whittaker, S. (2019). Progressive disclosure: Empirically motivated approaches to designing effective transparency. *Proceeding of the 24th ACM International Conference on Intelligent User Interfaces (IUI '19)*, 107–120.

Sqalli, M. T., & Al-Thani, D. (2020). Evolution of Wearable Devices in Health Coaching: Challenges and Opportunity. *Frontiers in Digital Health*, *2*.

Stevinson, C. D., & Biddle, S. J. H. (1998). Cognitive orientations in marathon running and "hitting the wall." *British Journal of Sports Medicine*, *32*(3), 229–235.

Storni, C. (2011). Complexity in an uncertain and cosmopolitan world. Rethinking personal health technology in diabetes with the tag-it-yourself. *PsychNology Journal*, *9*(2), 165–185.

Storni, C. (2014). Design challenges for ubiquitous and personal computing in chronic disease care and patient empowerment: A case study rethinking diabetes self-monitoring. *Personal and Ubiquitous Computing*, *18*(5), 1277–1290.

Stumpf, S., Rajaram, V., Li, L., Burnett, M., Dietterich, T., Sullivan, E., ... Herlocker, J. (2007). Toward Harnessing User Feedback for Machine Learning. *Proceeding of the ACM Conference on Intelligent User Interfaces (IUI '07)*, 82–91.

Stumpf, S., Sullivan, E., Fitzhenry, E., Oberst, I., Wong, W. K., & Burnett, M. (2008). Integrating rich user feedback into intelligent user interfaces. *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI '08)*, 50–59.

Suh, M. (Mia), Youngblom, E., Terry, M., & Cai, C. J. (2021). AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. *Proceedings of the SIGCHI Conference of Human Factors in Computing Systems (CHI '21)*.

Swan, M. (2009). Emerging patient-driven health care models: An examination of health social networks, consumer personalized medicine and quantified self-tracking. *International Journal of Environmental Research and Public Health*, *6*(2), 492–525.

Tanda, G. (2011). Prediction of marathon performance time on the basis of training indices. *Journal of Human Sport and Exercise*, *6*(3), 511–520.

Tazelaar, F., & Snijders, C. (2004). The myth of purchasing professionals' expertise. More evidence on whether computers can make better procurement decisions. *Journal of Purchasing and Supply Management*, *10*, 211–222.

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., … Kaplan, L. (2020). Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, *1*(4).

Tong, H. L., Quiroz, J. C., Kocaballi, A. B., Fat, S. C. M., Dao, K. P., Gehringer, H., … Laranjo, L. (2021). Personalized mobile technologies for lifestyle behavior change: A systematic review, meta-analysis, and meta-regression. *Preventive Medicine*, *148*.

Topol, E. J. (2015). *The patient will see you now: The future of medicine is in your hands*. Basic Books.

Turakhia, M. P., Desai, M., Hedlin, H., Rajmane, A., Talati, N., Ferris, T., … Perez, M. V. (2019). Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study. *American Heart Journal*, *207*, 66–75.

Van Dijk, E. T., Westerink, J. H. D. M., Beute, F., & IJsselsteijn, W. A. (2015). Unintended effects of self-tracking. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'15)*. Seoul, South-Korea.

Vitti, A., Nikolaidis, P. T., Villiger, E., Onywera, V., & Knechtle, B. (2020). The "New York City Marathon": participation and performance trends of 1.2M runners during half-century. *Research in Sports Medicine*, *28*(1), 121–137.

Vizer, L. M., Eschler, J., Koo, B. M., Ralston, J., Pratt, W., & Munson, S. (2019). "It's Not Just Technology, It's People": Constructing a Conceptual Model of Shared Health Informatics for Tracking in Chronic Illness Management. *Journal of Medical Internet Research*, *21*(4).

Vos, S., Janssen, M., Goudsmit, J., Lauwerijssen, C., & Brombacher, A. (2016). From Problem to Solution: Developing a Personalized Smartphone Application for Recreational Runners Following a Three-Step Design Approach. *Procedia Engineering*, *147*, 799–805.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI '19)*.

Wang, J., O'Kane, A. A., Newhouse, N., Sethu-Jones, G. R., & Barbaro, K. de. (2017). Quantified Baby: Parenting and the Use of a Baby Wearable in the Wild. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–19.

Wang, W., & Benbasat, I. (2007). Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs. *Journal of Management Information Systems*, *23*(4), 217–246.

Watt, A., Swainston, K., & Wilson, G. (2019). Health professionals' attitudes to patients' use of wearable technology. *Digital Health*, *5*.

Watzlawick, P., Bavelas, J. B., & Jackson, D. D. (1967). *Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes*. WW Norton & Company.

Webb, T. L., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin*, *132*(2), 249–268.

West, P., Giordano, R., Van Kleek, M., & Shadbolt, N. (2016). The Quantified Patient in the Doctor's Office. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '16)*, 3066–3078.

West, P., Kleek, M. Van, Giordano, R., Weal, M. J., & Shadbolt, N. (2018). Common Barriers to the Use of Patient-Generated Data Across Clinical Settings. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '18)*.

West, P., Van Kleek, M., Giordano, R., Weal, M., & Shadbolt, N. (2017). Information

Quality Challenges of Patient-Generated Data in Clinical Practice. *Frontiers in Public Health*, *5*.

Wiens, J., & Shenoy, E. S. (2018). Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases*, *66*(1), 149–153.

Wolever, R. Q., Jordan, M., Lawson, K., & Moore, M. (2016). Advancing a new evidence-based professional in health care: Job task analysis for health and wellness coaches. *BMC Health Services Research*, *16*(1).

Wolever, R. Q., Simmons, L. A., Sforzo, G. A., Dill, D., Kaye, M., Bechard, E. M., … Yang, N. (2013). A Systematic Review of the Literature on Health and Wellness Coaching: Defining a Key Behavioral Intervention in Healthcare. *Global Advances in Health and Medicine*, *2*(4), 38–57.

World Health Organization. (2003). *Diet, nutrition, and the prevention of chronic diseases: report of a joint WHO/FAO expert consultation*. World Health Organization.

World Health Organization. (2020). *WHO Guidelines on Physical Activity and Sedentary Behaviour*.

Wortman Vaughan, J., & Wallach, H. (2020). A Human-Centered Agenda for Intelligible Machine Learning. In M. Pelillo & T. Scantamburlo (Eds.), *Machines We Trust: Getting Along with Artificial Intelligence*.

Wu, H. S., Gal, R., van Sleeuwen, N. C., Brombacher, A. C., IJsselsteijn, W. A., May, A. M., & Monninkhof, E. M. (2018). Breast Cancer Survivors' Experiences With an Activity Tracker Integrated Into a Supervised Exercise Program: Qualitative Study. *JMIR MHealth and UHealth*, *7*(2).

Xiao, B., & Benbasat, I. (2007). E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact. *MIS Quarterly*, *31*(1), 137–209.

Yamashita, N., Kuzuoka, H., Hirata, K., Kudo, T., Aramaki, E., & Hattori, K. (2017). Changing Moods: How Manual Tracking by Family Caregivers Improves Caring and Family Communication. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '17)*, 158–169.

Yoo, D. W., & Choudhury, M. De. (2019). Designing Dashboard for Campus Stakeholders to Support College Student Mental Health. *The 13th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '19)*, 61–70. Trento, Italy.

Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science Technology and Human Values*, *41*(1), 118–132.

Zehrung, R., Huang, L., Lee, B., & Choe, E. K. (2021). *Investigating Opportunities to Support Kids' Agency and Well-being: A Review of Kids' Wearables*. 1–20. Retrieved from http://arxiv.org/abs/2104.05979

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceeedings of the Conference on Fairness, Accountability, and Transparency (FAT* '20)*, 295–305.

Zhu, H., Colgan, J., Reddy, M., & Choe, E. K. (2016). Sharing Patient-Generated Data in Clinical Practices: An Interview Study. *AMIA Annual Symposium Proceedings*, 1303–1312.

Zimmermann, H. (1980). OSI Reference Model-The ISO Model of Architecture for Open Systems Interconnection. *IEEE Transactions on Communications*, *28*(4), 425–432.

Appendices

APPENDIX A: QUESTIONNAIRES COACHES AND CLIENTS, CHAPTER 3

COACH–QUESTIONNAIRE

Coaches filled in this questionnaire twice, first halfway through the session, then at the end. Depending on the condition (*data-first* or *conversation-first*), these questionnaires were targeting their evaluations of data or conversation. We kept the questions as consistent as possible across the different sources of information (data or conversation) and timing (halfway and at the end), to allow for fair a comparison.

[Advice]	*Halfway*: What would you advice the client, and why? *End*: Do you have any additions or changes to your advice? If yes, what would you add or change, and why?
[Confident]	How confident are you that this advice will lead to a better result for the client? (5-point scale, ranging from "not confident" to "confident")
The information resulting from the {conversation / data} is:	
[Usable]*	5-point scale, ranging from "not usable" to "usable".
[Objective]	5-point scale, ranging from "objective" to "subjective".
[Clear]	5-point scale, ranging from "unclear" to "clear".
[Relevant]**	5-point scale, ranging from "relevant" to "not relevant".
[Enough]	*Halfway*: I have enough information to give the client appropriate advice. *End*: Because of the data / the conversation with the client, I have more information than before. (5-point scale, from "disagree" to "agree")
[Pers. Exp.]	I have a complete picture of the client's personal experience. (5-point scale)
[Daily Life]	I have a complete picture of the client's behavior in his/her daily life. (5-point scale)

[Supports]	{The use of data / Having a conversation with the client} supports my effectiveness as a coach. (5-point scale)
[Value]**	What was the value of {the use of data / the conversation with the client} for you as a coach? (Open-ended question)

Do to new insights, we made small updates on the questionnaire after the workshop, i.e.:
** In the workshop, we stated "useful" rather than "usable".*
*** [Relevant] and [Value] was only measured in the field study, not in the workshop.*

CLIENT-QUESTIONNAIRE

The client-questionnaire was only given to the clients in the field study, not at the workshop.

[Daily Life]	I feel like {my coach has / the data represents} a good, complete picture of my daily life and behavior. (5-point scale, from "disagree" to "agree")
[Pers. Exp.]	I feel like {my coach has / the data represents} a good, complete picture of my personal experience. (5-point scale)
[Insight]*	I feel like my coach has good insight in me as a person. (5-point scale)
[Understood]*	I feel understood by my coach. (5-point scale)
[Value]	What was the value of {the conversation with your coach / sharing data} for you? (Open-ended question)

** Only asked after conversation.*

APPENDIX B: QUESTIONNAIRES HEALTHCARE PROFESSIONALS AND PARENTS, CHAPTER 4

INTERVIEW SCRIPTS FAMILY VISITS

Midterm interview

- *How did you experience the use of the toolkit?*
  - *Which things worked well?*
  - *Which things could be improved?*
  - *Did you have doubts how to use the toolkit?*
  - *Are you receiving sufficient feedback of the toolkit?*
- *Are you able to find applications for the trackers?*
- *Which trackers do you like to use?*
  - Give tips for applying the trackers if necessary.
  - *Does the toolkit provide insight in your situation?*
  - For example, insight in trends or correlations.
- *Do you have sufficient access to the collected data?*
- *Does the toolkit affect your actions?*
  - *If yes, is that positive or negative?*
- *Did you receive a response from your healthcare professional on the data?*
  - *How did you experience that?*
- *How do you experience the communication with your healthcare professional?*
  - *Is it different than before?*
  - *Do you believe the data help your healthcare professional to understand your situation?*
- Give a summary of the first healthcare professional visits. Leave room for a response of the family.

End interview

- *How did you experience the use of the toolkit?*
  - *Which things worked well?*
  - *Which things could be improved?*
  - *Did you have doubts how to use the toolkit?*
- *Were you able to find applications for all trackers?*
  - *Which trackers did you like to use?*
- *Did the toolkit provide insight in your situation?*
  - For example, insight in trends or correlations.
- *Did you have sufficient access to the collected data?*
- *Did the toolkit affect your actions?*
  - *If yes, is that positive or negative?*
- *Did you receive a response from your healthcare professional on the data?*

    - *How did you experience that?*
- *How do you experience the communication with your healthcare professional?*
    - *Is it different than before?*
    - *Do you believe the data helped your healthcare professional to understand your situation?*
    - *What if all communication with your healthcare professional would be digital?*
- Give a summary of the recent healthcare professional visits. Leave room for a response of the family.

INTERVIEW SCRIPTS HEALTHCARE PROFESSIONAL VISITS

Pre-visit, interview on working practices after work observation
Ask healthcare professionals to reflect on the extent to which they agree with the following statements:

- *I am curious about the background and home situation of the client.*
- *I like working with data and numbers.*
- *I like to have a conversation with the client.*
- *At home, I find it hard to let go of work.*

Additionally, ask the following questions:

- *Is more information always better? Why (not)? Where is the threshold, and why?*
- *How do you experience work pressure?*
- *What is your view on technology?*

Weekly interviews during data-collection phase
Per family:

- Give the healthcare professional time to read and answer the messages of the client.
- *Which insights do you take from the collected data so far?*
- *What is the added value of the data for you?*
- *Is there any data missing?*
- *Is there any data unnecessary?*
- *How do you experience the presentation of the data?*
    - *Is it clear?*
    - *Do you need another visualization?*
    - Show how the healthcare professional can change the visualization of the data.
- *Did you recently have contact with the client, next to the data sharing in this dashboard? Why?*
- If needed, give room for composing a message to the client suggesting new

data collection.

Only after second family:

- *Are your needs regarding the data different from one family to the other? Why?*

Particularly pay attention to:

› Which clues is the healthcare professional attentive to?
› How does the healthcare professional interpret these clues?
› Where is the focus of the healthcare professional? Are things combined, trends detected, etc.?
› Which question, hypothesis or information need is underlying this focus?
› Which question, hypothesis or information need is underlying the suggested additional data collection?
› Is there a difference in approach or information need across the different families?

Final interview

- *Which insights did you take from the collected data?*
- *What is the added value of data for you?*
- *Do you now know more about your clients than before this study?*
- *Is there any data missing? Please consider this in a broad sense; it is also relevant if current technology does not yet allow for such data collection.*
- *Is there any data unnecessary?*
- *How do you experience the presentation of the data?*
    - *Is it clear?*
    - *Do you need other visualizations? Why?*
- *Did you have contact with the client during the study, next to the data sharing in this dashboard? Why?*
- *Does the data visualization meet your expectations?*
- *Would you be able and willing to adopt similar technology in your practice? How would that look like?*
- *Is your need regarding the data different from one family to the other? Can you explain the difference?*
- *What is, according to you, the connection between the care question of the client, the type of client, and the needed data? In other words, if you were to collect data of several different clients, what would be your expectations?*

Particularly pay attention to:

› Which clues is the healthcare professional attentive to?
› How does the healthcare professional interpret these clues?
› Where is the focus of the healthcare professional? Are things combined,

trends detected, etc.?
- Which question, hypothesis or information need is underlying this focus?
- Is there a difference in approach or information need across the different families?

FOCUS GROUP MATERIALS

List of predefined questions used in first focus group

1. What home-collected data would be useful?
   - Can you give (case specific) examples?
2. How would you like the data to be presented?
   - For example, would you like to create an overview yourself, of work with standard visualizations?
   - Is there any value in visualization / presentation of the parents themselves, for example a dairy?
3. What do the data bring you, and what are disadvantages?
   - For example, does it give you insight in a problem, or in a possible solution?
4. In which specific cases is data most valuable?
   - Which cases? Which parents? Which children? Which phase of care? Etc.
5. How would you expect that the home-data collection is experienced by parents? And how are you experiencing it?
   - For example, would it fit in your current working practices?

APPENDIX C: OVERVIEW PROCESS DATA ANALYSIS, CHAPTER 4

Phase I:

Codes	DATA OPPORTUNITIES	DATA CHALLENGES	
HEALTHCARE PROFESSIONALS	1) Data provide objective information 2) Data provide temporal insights → Theme 3	1) De interpretation of data is not straight-forward 2) Subjective experience (of parents) is key for a good interpretation of data 3) Data may be tracked focused around a question, or broadly with no particular focus 4) Data trigger value judgement, but there is no golden truth 5) Data give a false sense of certainty	
INTERACTION HEALTHCARE PROFESSIONALS & PARENTS	1) Data strengthen the connection		→ Subtheme 2-b
	2) Data are valuable input for a conversation	→ Subtheme 2-c	
	3) Opportunities of asynchronous communication	1) Challenges of asynchronous communication	→ Subtheme 2-d
PARENTS	1) Data give insight, control, they normalize and reassure	1) Tracking data may become obsessive	→ Subtheme 2-a

Phase II:

Families' journeys over time:

- Evolving tracking practices over time
- Evolving health issues, and according understanding, over time
- Message dialogs between

→ Theme 1

Themes and subthemes:

1. The role of the family's questions and problems when interacting with data.
   a) Sharing data is not equal to sharing problems.
   b) The importance of focus when collecting data.
   c) Data for exploratory versus confirmatory purposes.
2. The impact of data-sharing on parents and their relation and communication with healthcare professionals.
   a) Impact of data on parents can be both positive and negative.
   b) Mismatching expectations between parents and healthcare professionals on the use of data.
   c) Data are valuable input in a conversation.
   d) Advantages and disadvantages of a-synchronous communication.
3. Some information can, and some cannot, be captured by data.
   a) Data can provide objective information.
   b) Data can provide temporal (and thus cause-effect) insights.
   c) Data may be hard to interpret.
   d) Data often lack, although sometimes provide, context and background information.
   e) Data usually do not capture the (often essential) experience of parents.

Overview of the data-analysis process in Chapter 4.

APPENDIX D: PAPER PROTOTYPES, CHAPTER 6

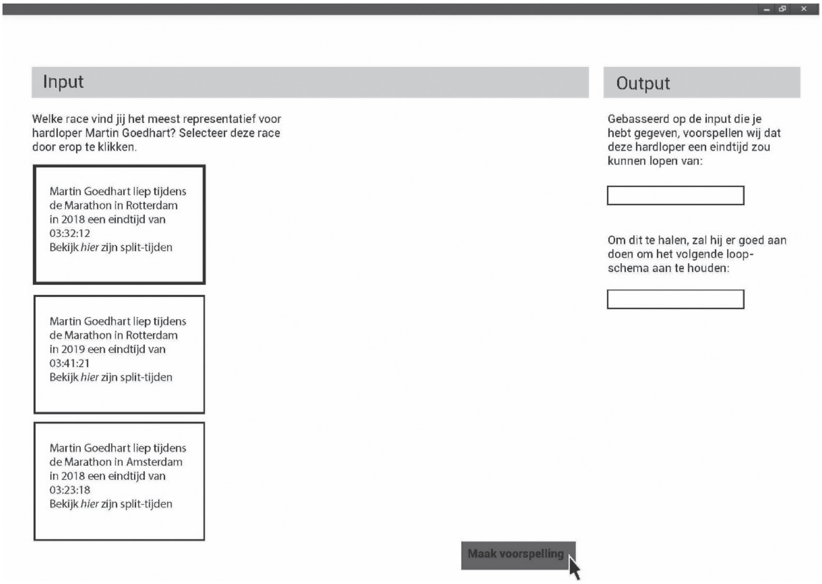

Figure I Paper prototype 1: Interface where coaches can select previous races of the query runner.

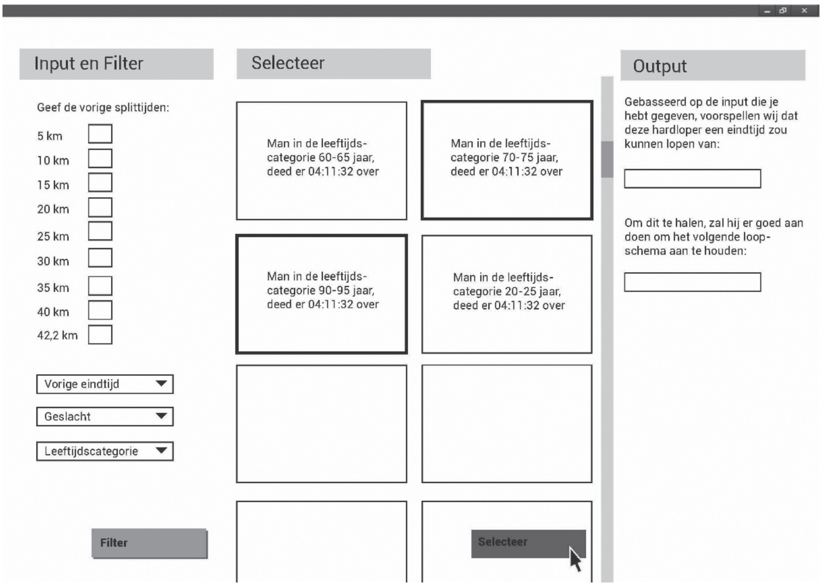

Figure II Paper prototype 2: Interface where coaches can select features and similar runners.