Towards Development of an Automated Health Coach

Rommy Márquez-Hernández[1], **Leighanne Hsu**[1], **Kathy McCoy**[1], **Keith Decker**[1], **Ajith Vemuri**[1], **Greg Dominick**[2], **Megan Heintzelman**[2]
[1]Department of Computer and Information Sciences
[2]Department of Behavioral Health and Nutrition
University of Delaware
{marquez,lhsu,mccoy,decker,kumar,gdominic,mheintz}@udel.edu

Abstract

Human health coaching has been established as an effective intervention for improving clients' health, but it is restricted in scale due to the availability of coaches and finances of the clients. We aim to build a scalable, automated system for physical activity coaching that is similarly grounded in behavior change theories. In this paper, we present our initial steps toward building a flexible system that is capable of carrying out a natural dialogue for goal setting as a health coach would while also offering additional support through just-in-time adaptive interventions. We outline our modular system design and approach to gathering and analyzing data to incrementally implement such a system.

1 Introduction

It is well-known that eating a balanced diet and engaging in regular moderate-to-vigorous physical activity (MVPA), among other healthy behaviors, promotes better health and reduces the risk of cardiovascular disease and other chronic illnesses (Tsao et al., 2022). However, people may struggle to develop and integrate healthier behaviors on their own (Kivelä et al., 2014; Willard-Grace et al., 2015). Health coaching is a behavioral health intervention that is demonstrably effective in improving motivation and confidence and is strongly associated with behavior change (Dennis et al., 2013; Eakin et al., 2007; Mahon et al., 2018; Oddone et al., 2018). Health coaches utilize behavioral theories and evidence-based strategies in a client-centered approach to help clients set goals that are challenging yet achievable, supported by action plans and coping plans that include strategies to overcome barriers like lack of time or poor weather (Kivelä et al., 2014; Oddone et al., 2018). As such, goals and dialogue are highly specific and tailored to the client.

However, human health coaching is limited by coach availability, cost to potential clients, and the retrospective nature of the feedback (Hill et al., 2015). Most attempts to automate this process thus far have been mostly limited to theoretical studies or systems with pre-scripted, non-tailored dialogues, if there is any interactivity at all (op den Akker et al., 2014, 2015; Bickmore et al., 2011, 2013; Svetkey et al., 2015; Kramer et al., 2020). While some successfully demonstrate the acceptability of automated systems, even with scripted interaction, the feedback also identifies a user desire for increased tailoring with regards to timing and response to collected user data or context during coaching sessions (Bickmore et al., 2013; Mitchell et al., 2021).

Another type of intervention, Just-in-Time Adaptive Intervention (JITAI), leverages technology to monitor a user's state and deliver support at a time when it is most needed and the user is most receptive to act upon it in the moment (Nahum-Shani et al., 2018; Schembre et al., 2018; Spruijt-Metz et al., 2015). For instance, this may include nudges like "walk around the block an extra time" if the user is out for a walk.

To our knowledge, no system has yet combined automated, interactive coaching with real-time knowledge of user progress and JITAI. To this end, we aim to build an automated system capable of helping clients set achievable goals through interactive discussion and support them in achieving those goals. We focus on physical activity (PA) coaching, but this infrastructure is modular and extendable to other health coaching areas in which goals can be clearly defined or real-time user context leveraged, including stress management or adapting to a prescribed diet.

In this paper, we present our preliminary work in building a dialogue and messaging system for an application capable of coaching a user to set and achieve goals and provide useful just-in-time messaging. We will contextualize the messaging components within the greater system architecture

we are building upon in section 3 and then detail the approach we've taken to build our proposed dialogue and messaging component in sections 4 and 5. Finally, we will describe our plans for future experiments and evaluations.

2 Related Work

The rise in popularity and availability of wearable technology and biometric sensors offers the opportunity to create similarly theoretically-driven, evidence-based behavioral interventions (DiClemente et al., 2001; Fjeldsoe et al., 2009; O'Reilly and Spruijt-Metz, 2013; Bort-Roig et al., 2014; Danaher et al., 2015; Farmer and Tarassenko, 2015; Wang et al., 2015a; King et al., 2016; Lobelo et al., 2016). However, the resulting apps generally do not adhere to the American College of Sports Medicine (ACSM) recommendation of 150 minutes per week of moderate-intensity aerobic physical activity or 75 minutes per week of vigorous-intensity aerobic physical activity (Middelweerd et al., 2014; Guo et al., 2017; Modave et al., 2015). They lack guidance establishing realistic and appropriate behavioral goals, do not assist users in modifying goals over time, display messages that are not personalized, and do not account for contextual or situational barriers, such as weather and emotional states, that can significantly influence physical activity intentions and behavior (Düking et al., 2020; Rupp et al., 2018; op den Akker et al., 2014, 2015; Muntaner et al., 2016).

Active work on JITAI systems emphasizes their basis in behavioral theory, user relevance, and actionable feedback (Wang et al., 2015b; Hardeman et al., 2019). However, most do not truly account for context or barriers and instead use simple, canned messages delivered at preset moments (Klasnja et al., 2018; Lentferink et al., 2017; Saponaro et al., 2017; Mair et al., 2022). More recently, Saponaro et al. (2021) and Ismail et al. (2022) demonstrated that individualized, contextualized JITAI nudges are significantly better received than non-JITAI nudges.

Some automated coaching systems exist (op den Akker et al., 2014), but most are limited in interactivity, and efficacy varies. Many dialogue-driven health coaching systems are largely theoretical (Bickmore et al., 2011; op den Akker et al., 2015), although a few extend to more practical implementations with varying degrees of tailoring and interactivity (Svetkey et al., 2015; Bickmore et al., 2013). Several other embodied conversational agents were included in a recent survey (Kramer et al., 2020). Of these, most rely on scripted dialogue selection, and the others provide limited text interaction at best, lacking the flexibility to adequately tailor to users' unique goals and values. A detailed comparison between text-based coaching and human health coaching was performed in Mitchell et al. (2021), demonstrating the feasibility of an automated system with a wizard-of-oz setup. While participants appreciated the automated coaching system, they lamented the lack of tailoring and context sensitivity. Some analysis work has been done on counseling dialogues (Pérez-Rosas et al., 2017, 2018; Althoff et al., 2016) and, more recently, on coaching dialogues (Gupta et al., 2020a,b, 2021). This latter work is ongoing, but focuses primarily on post-dialogue SMART goal summarization and health coach assistance rather than interactivity. Thus, to our knowledge, no one has yet developed a system to conduct a dynamic, flexible, and interactive coaching dialogue.

Human coaching dialogue adheres to a specific procedural structure and language. Generation for dialogue in this context has added constraints that increase its complexity compared to other domains. While the intents and data content will be provided by the respective messaging policies, it is also important to incorporate personalization (Cawsey et al., 1997, 2000; Marco et al., 2006; Colineau and Paris, 2011), empathy (Prendinger and Ishizuka, 2005), and additional constraints for a health domain, as well as constrained generation (He and Li, 2021; Miao et al., 2019; Mou et al., 2015; Li and Sun, 2018) and style transfer (Jin et al., 2020; Toshevska and Gievska, 2021).

3 Automated Coaching System

The automated coach brings together ideas and solutions in human behavioral theory, physical activity monitoring, cooperative multi-agent architectures, and natural language processing to build an integrated approach to reducing sedentary behavior while increasing users' overall physical activity.

The automated coach is designed to ultimately take over certain basic tasks from a traditional human coach: it will be able to meet with users on a weekly basis to negotiate goals and talk about the user's progress. A number of schemas exist for creating a well-defined goal; we discuss these in section 4. The system may also leverage user

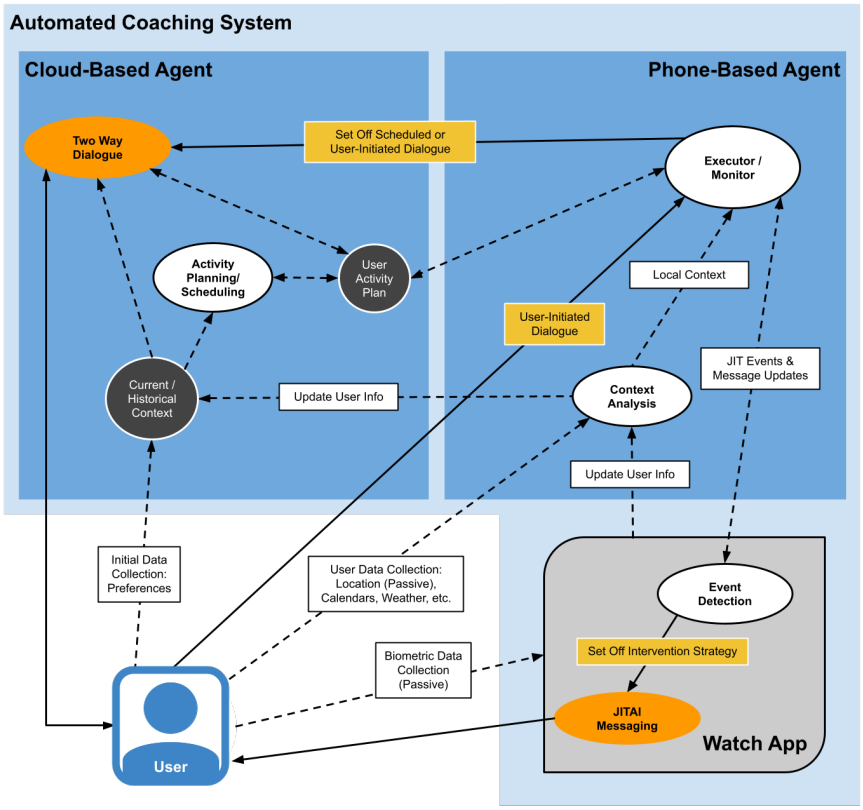

Figure 1: Interaction between the User and the Automated Coaching System: white/orange ovals represent major modules, gray circles represent stored information, rectangles represent actions and data transfer between modules.

qualities or other initially-provided information to quickly adapt to users' preferred intervention strategies. Unlike a traditional human health coach, it will be available for user-initiated coaching sessions at any time and will be able monitor the user's real-time goal progress. This will allow the automated coach to better support the user, as well as allow it to send personalized just-in-time messaging when needed.

Unlike typical one-size-fits-all solutions for just-in-time interventions such as "remind the user to walk at 10 min before the hour if they have not yet reached 250 steps," which do not work particularly well (Saponaro, 2020), the action plan and user preferences are compiled into a context- (state-) sensitive strategy enumerating possible conditions under which the coach should nudge with particular message content (remind, congratulate, suggest, or otherwise interact with) their user. Positive (or negative) reactions to these nudges are used to adaptively learn a better nudging strategy by taking into account both timing and message content.

The automated coaching system is built on a multi-agent architecture (Graham et al., 2003) for privacy and scalability, and further extended by our contributions toward coaching domains with an eye toward individualization, data integration, domain flexibility and (agent) behavior transferability (Vemuri et al., 2021). As can be seen in Figure 1, each user is allocated two personalized agents: one autonomous cloud-based agent running continuously on a server that is responsible for data collection, learning, dialogue understanding, and generation; and an app-based agent on the user's smartphone that handles local data collection from the user's smartwatch, user dialogue interface, and summary graphics. Although we have previously built-out versions of this using the FitBit Charge platform, true JITAI is not possible with that due to lack of real-time sensor access (Vemuri et al., 2021). Our current version uses the Apple Watch platform, which allows for an on-watch app that can be configured by the phone agent to detect specific, context-sensitive intervention events (prolonged inactivity, walking, stress) for JITAI. A centralized Dashboard agent (not pictured) is also used for running trials to give at-a-glance access to current participant status.

Agents are able to process data and communicate with each other concurrently. Processing can

be triggered by communications, state/sensed context/goal changes, and also pre-scheduled agent behaviors. Just-in-time notifications are triggered directly on the watch or phone (depending on the type), and do not require dialogue responses. Dialogues can be initiated by the phone agent at scheduled times, or by the user.

4 Approach

Bickmore et al. (2011) and Bickmore et al. (2013) described human health coaching as the gold standard for automated coaching systems to aim for. Such a system relies on a rich library of information representing user data, preferences, and coaching knowledge and principles. The dialogue and messaging system architecture is outlined in section 5. This system will interface with the agent architecture described in section 3 for timing and information for messages as can be seen in Figure 1. It will be capable of handling one-way JITAI messages (see section 5.3) and two-way, interactive health coaching dialogues.

We frame the coaching dialogue as a task-oriented dialogue. However, most task-oriented dialogues consist of rigidly defined simple tasks (e.g., booking a flight or negotiating a price) defined by a few parameters that the system needs to elicit from the user to complete the task. Dialogue policy, which determines each system intent (e.g., request information, offer a suggested value) and directs the dialogue, is similarly simplistic and limited: a task is complete when the parameters have been filled and an operation or query happens successfully (e.g., a flight is successfully booked). Parameters can be modified or updated until the task is completed. Additionally, there is no need for information to carry over from session to session; once a price is agreed upon, for instance, the task is complete, and there are no further exchanges on the subject.

Health coaching instead centers around a reflective discussion to achieve a more loosely-defined objective: setting a well-defined goal with realistic strategies for completing it. The dialogue is completed when the goal and strategies are not only fully defined by their parameters, but sufficiently motivated and supported to improve the user's success. The latter is accomplished not through filled parameters, but a series of reflective questions to ensure the user has thought their goal through thoroughly. This goal is then revisited at the subsequent coaching session, where a new goal may be set or a new coping plan may be created to assist the client in overcoming unforeseen barriers. Additionally, understanding barriers or support systems requires some representation of a health coach's world knowledge.

To ensure that the top-down approach aligns with practice and data, we also examined coaching dialogues. Due largely to patient privacy and protection, few publicly available datasets exist within the health coaching domain. There is one recently released dataset containing health coaching dialogues conducted via SMS text message (Gupta et al., 2020a). This dataset is tagged with a two-level labeling structure. One level covers stages and phases, breaking down the overall weekly dialogue into goal setting and goal implementation stages, which further break down into phases such as refining, anticipating barriers, negotiation, and follow up. Additionally, they identify SMART goal components, which break down a goal into Specificity, Measurability, Attainability, Realism, and Time-bound components.

This dataset released after development on our system had begun, and the coaching paradigm is different from our face-to-face data. Due to their curtailed nature, text messages often lack certain nuances, context, and cues compared to verbal interaction (Mitchell et al., 2021). Messages tend to be more curt and elaborate less, which affects the style of questions that need to be used to elicit the same information. Discussion of action plan, barriers, and coping strategies is thus unsurprisingly significantly more limited in this dataset, which focuses more on the goal parameterization. However, since our target coaching format is also text message-like, it will still be crucial for designing a text message coaching session, as well as for training the natural language understanding components described briefly below in section 5.1.

To ensure that our automated system is rooted in core health coaching concepts and behavior change theory, we examined coach training materials and guidelines provided by our health coaching team or publicly available online. These included outlines as well as coaching roleplay videos. Based on these materials, we developed a dialogue model described below in section 5.2. This model was further refined by examining data collected through a tangential, developmental study. We will refer to this data as dataset 1.

4.1 Data and Annotation

Dataset 1, currently being collected through BeSMART feasibility trial (Heintzelman et al., 2022), closely mimics the face-to-face coaching sessions that the coaching team regularly conducts with clients. Clients meet with their coach one-on-one initially for approximately an hour, and then subsequently for twenty or thirty minutes, generally with at least a week between meetings. For the short feasibility trial, our coaches did not receive information about goal progress between meetings, but we intend to correct this in subsequent studies (see section 7), as the proposed automated coaching system will have access to users' goal progress and other context.

Data collection for dataset 1 is ongoing, but the existing data is being broadly analyzed to further refine the dialogue model. The data was collected over video call, and the audio was automatically transcribed. The video was discarded for patient privacy, and the audio was kept only for quality control; the transcripts were manually cleaned for major mistranscriptions only. Verbal fillers, restarts, and exchanges consisting only of repeated acknowledgements will be automatically removed in a preprocessing step later, prior to data annotation.

Our coaches use a slightly different strategy to that of Gupta et al. (2020a). In addition to developing SMART goals, our health coaches utilize FITT (Frequency, Intensity, Time/duration, and Type of activity) and the W5 (What activity, Where, When, Who is supporting or accompanying, and Will any preparation be needed) to better assist clients in visualizing how their goal and action plan (details and strategy of how to achieve the goal) will fit into their daily schedule.

5 Dialogue and Messaging System

In this section, we detail the dialogue and messaging model and how the natural language interfaces will be built upon it. We will focus primarily on the dialogue model, as the one-way JITAI messaging is largely driven by the multi-agent architecture described previously and requires no interaction and much less tailoring of wording than the eventual dialogue system.

The overall dialogue system architecture is shown in Figure 2. We chose a traditional, modularly built dialogue system over an end-to-end neural network because the latter is unable to handle the level of complexity, control, and constraint that a health coaching system requires. This system is a modified dialogue state architecture. The dialogue schema and models are hierarchical. The modularity of this system allows for an evolving implementation. The current policy and generation modules are fully rule-based, which allows us to ground the overall structure of the dialogue in theory and coaching protocols. However, these will be incrementally swapped for dynamic, data-driven, learned implementations as the rest of the dialogue system develops to support them.

5.1 Natural Language Understanding & Dialogue State

During a dialogue exchange, for a given user input, the Natural Language Understanding (NLU) module identifies a number of different levels of slot and message labels, conditioned upon the system's prior request, if any. These labels update the dialogue state tracker, which keeps track of the information that has been provided by the user. It effectively captures the history and current knowledge state of the system based solely on the user's messages. This knowledge state representation is multilayered. At the top, the user directly or indirectly conveys an intent. In system-initiated, scheduled dialogues, the system is expected to direct the flow of conversation, determining when to move onto the next subdialogue. On the other hand, our system will eventually also accommodate user-initiated dialogues, which would start with a new intent without a prior message. A given input will also have one or more dialogue acts (e.g., whether the user is requesting information, setting parameters for their goal, or suggesting a possible coping plan to overcome a barrier). At a more fine-grained level, we will need a classifier to identify the task-specific labels (e.g., goal or action plan components, barriers, or support figures). Sentiment and uncertainty analyses will be added later to direct the policy and generation to produce clearer, more appropriate, or empathetic messaging.

While these understanding components will be based on the some of the same techniques used in summarization (Gupta et al., 2020b, 2021) or dialogue state tracking (Young et al., 2010), the policy and generation components allows an additional advantage of requesting confirmation to reduce mistakes in the summarization and dialogue state tracker. These labels will feed into the dialogue policy and eventually add to the knowledge

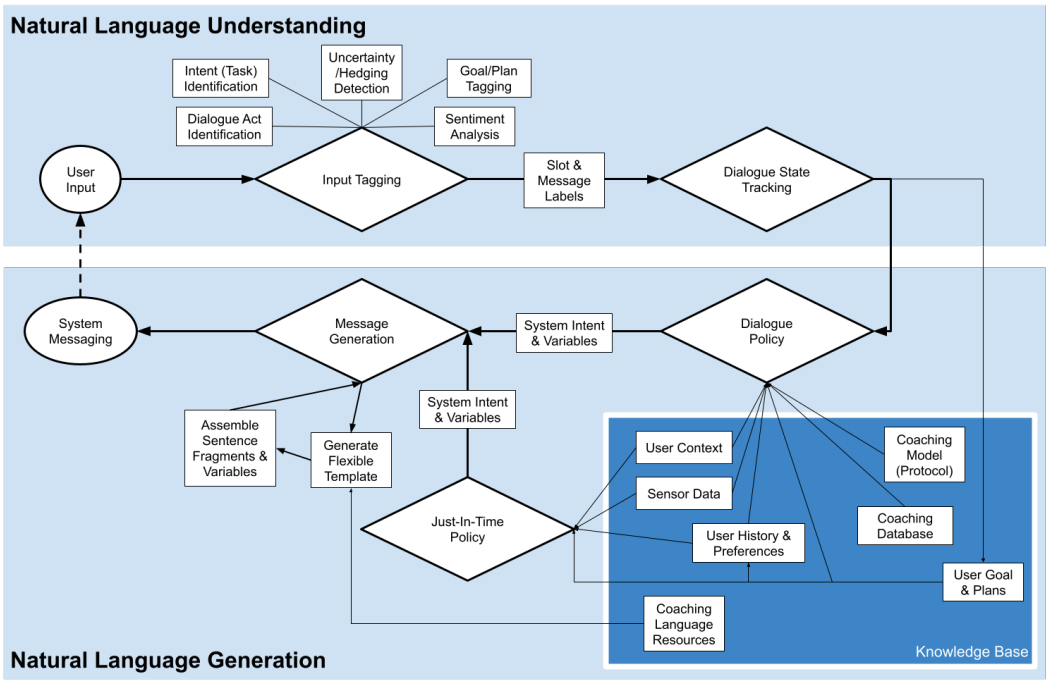

Figure 2: Dialogue State and Messaging Architecture: diamonds represent major modules, rectangles represent information transfer and functionality for those modules. The knowledge base is a logical representation of the context, history, and other information stored across the agents and apps in the automated coaching system.

base by updating the cloud-based agent. The NLU components can be built separately and combined to produce a multi-layered representation of the dialogue state. We will test a variety of features, including word embeddings and system intents of the previous turn, across a variety of machine learning classifiers. Gupta et al. (2020b) found that phase/stage classification (roughly equivalent to our intent/task classification) was more accurate when the SMART components were included as features, so we will test certain labels as potential features for other labels as well.

A particular challenge lies in understanding barriers. These are not necessarily unique to each user; barriers such as lack of time/space are common. We plan to build an expandable database, seeded initially with categorized examples from our own dataset, to represent barriers and potential solutions.

5.2 Dialogue Policy

Health coaching dialogues follow a particular pattern of subdialogues, which we refer to as the "backbone". Coaches establish rapport and get to know their clients before discussing anything goal-related, building up knowledge about their client that will help the coach guide the goal-setting subdialogue that follows. Coach and client establish a specific and realistic goal and an action plan to achieve it, discuss anticipated barriers and brainstorm resolutions and coping strategies, and discuss how the client's support network may help them in achieving this goal, either by reminding them or joining in the physical activity or holding them accountable for it. In follow-up sessions, emphasis is placed on exploring patient success and developing coping strategies for previously unanticipated barriers. Coaches thus guide clients in establishing a structure and pathway for success. This strategy guides our policy development.

The hierarchical dialogue policy is a key component in allowing us to direct the conversation in a sensible manner by supplying intents to the generation system. These intents dictate the kind of information the system wants to request from the user, such as slot values, clarifications, or confirmations. It draws from the knowledge base as needed, which contains user history and data and coaching knowledge. The high-level backbone, rooted in coaching theories and protocols, remains the same in all iterations. It will comprise multiple subdialogues for the goal-setting process (e.g., past goal progress and reflection, (re)negotiation of new goals, barrier resolution, etc.) and direct the flow of conversation sequentially through each of

these as the previous subdialogue completes. Each subdialogue will also have a policy, which will be rule-based initially. However, in time, we would like to expand the subdialogue policies to be more flexible. In freeform dialogue, users may provide more information in a response than was initially asked for. Humans naturally adjust their intents accordingly to avoid asking for the same information or to focus instead on discussing the extra information. We will conduct experiments to learn efficient, flexible strategies from the coaching transcripts to complete the subdialogue task. In later iterations, we will refine on collected dialogues and further incorporate the user's context, sensor data, and history to dynamically plan the ongoing dialogue for a more fluid and natural conversation.

5.3 Just in Time Adaptive Intervention (JITAI) Messages

JITAI messages are one-way messages that do not require a response and have their own policy. This policy is mostly driven by the agent architecture and will be based on the user's context, sensor data, and history, allowing us to implement logic for the timing of different types of JITAI messages. The logic will identify moments such as "the user has achieved a weekly goal" and "the user planned to exercise in the morning but it is supposed to rain" with their associated JITAI messages as well as identify whether it would be appropriate to send at that time. These topics have been explored before (Hardeman et al., 2019; Nahum-Shani et al., 2018; Mair et al., 2022; Ismail et al., 2022; Mutsuddi and Connelly, 2012; Muller et al., 2017).

While there is a consensus that the focus of JITAI messaging is to provide the user with the support they need at the time they need it, so that they can accomplish the goal of (in our case) increasing their PA, there is not much focus placed on categorizing the messages themselves other than to say that they are personalized/tailored messages that are motivational or encouraging (Nahum-Shani et al., 2018; Mair et al., 2022; Ismail et al., 2022). In order to preserve clarity, we have separated our JITAI messages into two main categories: anticipatory (which aim to to reduce barriers such as the weather, time of day or year, time management, planned meetings or events) and opportunistic (which provide encouragement and motivation at moments when there is perceived dwindling enthusiasm or when the user could take advantage of times they already unknowingly partake in activity). Anticipatory messaging can be planned in advance and delivered to the user at appropriate times that can be determined without complex sensing data (e.g., in the morning before leaving for work). The timing of opportunistic messaging is much more delicate as they must be delivered "in the moment" to be effective.

An example scenario for an anticipatory message would be to send the participant a message, while they are getting ready to leave their house to go to work, that lets them know that they are about to encounter one of their barriers and reminds them of the strategy they had already planned.

> *Remember to pack your umbrella! You planned to walk during your lunch break and there is a 50% chance of rain this afternoon.*

The purpose of this message is to anticipate a barrier that could cause the participant to fail at their goal for that day if not corrected in time.

Similarly, an example scenario for an opportunistic message would be to send the participant a message if their heart rate is going down and they only have 5 minutes left to finish their goal for the day.

> *Don't give up now! You only have 5 minutes left to go!*

The purpose of this message is to encourage the user at an opportunistic time to finish the goal they had set for themselves for that day.

5.4 Message Generation

Once either the dialogue or JITAI policy has determined the overall intent and data content of a given message, the next step is message generation. Due to the fact that we are implementing a complex task-oriented dialogue system, the fact that we have limited health coaching data and the importance of preserving the coaching language (e.g., the coaches must remain positive and encouraging through interactions, and there are guidelines for things that coaches should or should not say), the message generation will at first remain template-based. While Neural Natural Language Generation (NNLG) models have been improving greatly, they have many pitfalls when it comes to task-oriented dialogue systems. These include introducing hallucinated content (Reiter, 2018; Erdem et al., 2022), poor sentence planning and discourse operations (Reed et al., 2018), and not approximating human

generated text on complex problems (Wiseman et al., 2017; Erdem et al., 2022), especially in situations with a limited dataset.

While using a template-based method will help us avoid these pitfalls, they can be too structured and repetitive, which can hurt the user experience. Therefore, in an effort to introduce variety to our wordings over time, we plan to use what we call "flexible message templates." We first begin by identifying sentence fragment sections that we can put together to form our flexible templates. Each flexible template is made up of sentence fragment sections and any needed variables (e.g., proposed goal, dates, proposed strategies). Then, each sentence fragment section within the flexible template is replaced by one of multiple sentence fragment options that will together create a relatively unique message. We call them flexible templates both because each sentence fragment section could be used for multiple different templates and because in generating our templates in this way, we can create multiple different ways of saying the same message despite the overall generation being templated.

As we gather data during the collection of dataset 1 (as mentioned in section 4.1), we are looking to augment the number of flexible templates that cover the same purpose and content. However, since dataset 1 is speech-based while our system is text-based, we will need to handle the inherent differences between text and speech interactions and what that will mean for how our automated coach will need to differ from the human coach. As was encountered in Mitchell et al. (2021), during text-based interactions the health coaches felt like they could not have conversations that were as in-depth and nuanced because they were not just missing auditory input but also visual (e.g, body language, facial expressions, etc.). Additionally, they found that the health coaches found it hard to transition to a text platform because they had difficulties connecting to the user when they received short and ambiguous responses. As a result, we will make two assumptions: (1) the messages in text-based conversation need to be more direct and (2) the user is less likely to elaborate on little input.

Once we have augmented the flexible templates, instead of randomly selecting which flexible template to use in any given instance, we will explore ways to select the best template based on the conversation history and the users past reactions. We look to consider features such as message structure variability (e.g., if the last message had a prepositional phrase at the beginning, the next message should not), missing information (e.g., if we need to know three pieces of information, how many have already been given and what is remaining), and vocabulary variability (e.g., back-to-back messages should not use similar wording). We are taking inspiration from Razavi (2021), whose dialogue manager LISSA uses the user's last response to choose the best next response from multiple possible options.

In order to add more variety to the automated coach's speech, we aim to incorporate text style transfer techniques in order to affect the tone of the output by making adjustments in the emotion portrayed and politeness without needing to affect the content (Jin et al., 2020; Toshevska and Gievska, 2021). This requires user sentiment components for the NLU and dialogue policy and allows for the creation of a more empathetic, likable coach (Prendinger and Ishizuka, 2005).

Once we have more data following additional trials, we would also like to use information retrieval and constrained generation techniques to automate the generation of our flexible templates and sentence fragments. Recently, constrained generation research has put a focus on lexical constraints (He and Li, 2021; Miao et al., 2019; Mou et al., 2015; Li and Sun, 2018), which suits our needs in preserving the coaching language where we need to put soft and hard constraints on keywords or sentence formats that must be in the output and those that cannot appear in the output.

6 Formative Evaluation

The first iteration of our Dialogue and Messaging System will be both derived (as described above) and evaluated on dataset 1. We are aware that evaluating a system on the data that was used to derive it will bias it. However, since this will only be the rudimentary Dialogue and Messaging System, we do not believe the risk to be too great, since we will be further refining and evaluating the system with further trials. The evaluation needs to be separated into two parts: evaluating the NLU component and evaluating the message generation.

We mentioned above that basing our text-based system on the speech-based dataset 1 will affect the message generation by forcing us to make two assumptions: (1) the messages in text-based conversation need to be more direct and (2) the user is

less likely to elaborate on little input. These two assumptions will also affect the evaluation of our generated messages since we cannot evaluate on whether the two messages (one from dataset 1 and one generated by our system) are equivalent. Instead, we will need to evaluate on whether both messages ask the user for the same information given the same prompt.

Evaluating the NLU component could also be complicated due to the same assumptions. In this case, since our system is expecting more direct messages, the NLU component would expect that the user's response would be more straightforward. However, we can see *what* the NLU component can correctly recognize and this could be a worse-case situation. In addition, we can evaluate it on whether it reacts correctly to a message. Therefore, we will be evaluating the system on whether it correctly identifies the parameters it is expecting and on whether the policy correctly prompts the message generation.

7 Next Steps

Once we have a working Dialogue and Messaging System, we plan to lead two trials in order to evaluate and improve the system: Trial Alpha and Trial Beta.

Trial Alpha. In this trial, we will be generating a dataset we plan to release and evaluating our two-way dialogue. As with the collection of dataset 1, we will once again be collecting data from real user-human health coach interaction. This time, however, all interactions will be text-based and the human coach will have the same information as our automated coach will.

We expect the data labelling will function similarly as it did for dataset 1 (as described in the section 4.1). However, we hope that the data will be much cleaner and much more catered to the text domain. As previously mentioned, there was another dataset released in 2022 by Gupta et al. (2020a), but it does not cover barrier resolution and strategy negotiation. Therefore, we believe this dataset of labelled data will be very helpful in improving future health coaching research.

As far as evaluating our system goes, we will not need to evaluate our system around base assumptions like we will have to do for the Formative Evaluation. Therefore, the evaluation will be focused on four factors: (1) given two messages (one from dataset 2 and one generated by our system), is the content of both equivalent?, (2) is the language from the generated messages appropriate for a health coach? (3) are the parameters the NLU component is expecting reasonable?, and (4) does the NLU component correctly identify the parameters, and does the policy correctly prompt the message generation? To answer all these questions we will be using both standard metrics, such as BLEU-4 (Papineni et al., 2002), and human health coach manual evaluation.

Trial Beta. This trial is the first time that users will be using our system. It will serve to evaluate both our two-way dialogue and our JITAI messaging. By this point we hope to assess (1) how users respond JITAI messages and timing, (2) how users respond to our automated coach as opposed to the human coach in two way dialogue, (3) how successful the NLU component is at properly understanding the user, and (4) how successful the automated coach is when compared to the human coach. The goals and focus of this trial are subject to change based on the results of Trial Alpha.

8 Conclusion

Increasing engagement in MVPA and reducing sedentary behavior is a national priority for improving cardiovascular health. While wearable PA monitors show promise in initiating PA change, they do not assist the user in updating their PA goal, nor do they provide personalized messaging to assist the user in overcoming barriers to PA. Human coaching, following sound theoretical models of behavior change, has been demonstrated to be effective, but is hard to scale and misses the potential of "just-in-time" behavior suggestions and encouragement, as the coach is not always readily available.

Our Automated Coaching System is an integrated system to provide personalized, evidence-based, just-in-time feedback as well as interactive coaching including goal (re)negotiation, targeted at increasing PA and reducing risk for cardiovascular disease. Our system focuses on PA, but this infrastructure is modular and extendable to other health behaviors, including stress management and sleep hygiene.

Acknowledgements

This research was partially informed by data collected through National Institutes of Health award R21-AG056765-01.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Timothy W. Bickmore, Daniel Schulman, and Candace Sidner. 2013. Automated interventions for multiple health behaviors using conversational agents. *Patient Educ Couns.*, 92(2):142–148.

Timothy W Bickmore, Daniel Schulman, and Candace L Sidner. 2011. A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. *Journal of biomedical informatics*, 44(2):183–197.

Judit Bort-Roig, Nicholas D. Gilson, Anna Puig-Ribera, Ruth S. Contreras, and Stewart G. Trost. 2014. Measuring and influencing physical activity with smartphone technology: A systematic review. *Sports Medicine*, 44(5):671–686.

Alison Cawsey, Ray B. Jones, and Janne Pearson. 2000. The evaluation of a personalised health information system for patients with cancer. *User Modeling and User-Adapted Interaction*, 10:47–72.

Alison Cawsey, Bonnie Webber, and Ray Jones. 1997. Natural language generation in health care. *Journal of the American Medical Informatics Association*, 4(6):473–482.

Nathalie Colineau and Cecile Paris. 2011. Motivating reflection about health within the family: the use of goal setting and tailored feedback. *User Modeling and User-Adapted Interaction*, 21:341–376.

Brian G. Danaher, Håvar Brendryen, John R. Seeley, Milagra S. Tyler, and Tim Woolley. 2015. From black box to toolbox: Outlining device functionality, engagement activities, and the pervasive information architecture of mHealth interventions. *Internet Interventions*, 2(1):91–101.

Sarah M Dennis, Mark Harris, Jane Lloyd, Gawaine Powell Davies, Nighat Faruqi, and Nicholas Zwar. 2013. Do people with existing chronic conditions benefit from telephone coaching? a rapid review. *Australian Health Review*, 37(3):381–388.

Carlo C. DiClemente, Angela S. Marinilli, Manu Singh, and Lori E. Bellino. 2001. The role of feedback in the process of health behavior change. *American Journal of Health Behavior*, 25(3):217–227.

Peter Düking, Marie Tafler, Birgit Wallmann-Sperlich, Billy Sperlich, and Sonja Kleih. 2020. Behavior change techniques in wrist-worn wearables to promote physical activity: Content analysis. *JMIR mHealth and uHealth*, 8(11):e20820.

Elizabeth G Eakin, Sheleigh P Lawler, Corneel Vandelanotte, and Neville Owen. 2007. Telephone interventions for physical activity and dietary behavior change: a systematic review. *American journal of preventive medicine*, 32(5):419–434.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Andrew Farmer and Lionel Tarassenko. 2015. Use of wearable monitoring devices to change health behavior. *JAMA*, 313(18):1864.

Brianna S. Fjeldsoe, Alison L. Marshall, and Yvette D. Miller. 2009. Behavior change interventions delivered by mobile telephone short-message service. *American Journal of Preventive Medicine*, 36(2):165–173.

J. Graham, K. Decker, and M. Mersic. 2003. Decaf: A flexible multi-agent system architecture. *Autonomous Agents and Multi-Agent Systems*, 7(1–2):7–27.

Yi Guo, Jiang Bian, Trevor Leavitt, Heather K Vincent, Lindsey Vander Zalm, Tyler L Teurlings, Megan D Smith, and François Modave. 2017. Assessing the quality of mobile exercise apps based on the american college of sports medicine guidelines: A reliable and valid scoring instrument. *Journal of Medical Internet Research*, 19(3):e67.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020a. Human-human health coaching via text messages: Corpus, annotation, and analysis. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 246–256.

Itika Gupta, Barbara Di Eugenio, Brian D Ziebart, Bing Liu, Ben S Gerber, and Lisa K Sharp. 2020b. Goal summarization for human-human health coaching dialogues. In *FLAIRS Conference*, pages 317–322.

Itika Gupta, Barbara Di Eugenio, Brian D Ziebart, Bing Liu, Ben S Gerber, and Lisa K Sharp. 2021. Summarizing behavioral change goals from sms exchanges to support health coaches. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–289.

Wendy Hardeman, Julie Houghton, Kathleen Lane, Andy Jones, and Felix Naughton. 2019. A systematic review of just-in-time adaptive interventions (jitais) to promote physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, 16(1):31.

Xingwei He and Victor OK Li. 2021. Show me how to revise: Improving lexically constrained sentence generation with xlnet. In *Proceedings of AAAI*, pages 12989–12997.

Megan P Heintzelman, Gregory M Dominick, Ajith Vemuri, and Keith Decker. 2022. Development of the be smart feasibility trial to increase physical activity in midlife adults. In *ANNALS OF BEHAVIORAL MEDICINE*, volume 56, pages S253–S253. OXFORD UNIV PRESS INC JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA.

Briony Hill, Ben Richardson, and Helen Skouteris. 2015. Do we know how to design effective health coaching interventions: a systematic review of the state of the literature. *American Journal of Health Promotion*, 29(5):e158–e168.

Tasnim Ismail, Dena Al Thani, et al. 2022. Design and evaluation of a just-in-time adaptive intervention (jitai) to reduce sedentary behavior at work: Experimental study. *JMIR Formative Research*, 6(1):e34309.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020. Deep learning for text style transfer: A survey. *arXiv preprint arXiv:2011.00416*.

AC King, EB Hekler, and LA Grieco. 2016. Effects of three motivationally targeted mobile device applications on initial physical activity and sedentary behavior change in midlife and older adults: A randomized trial. *PLOS ONE*, 11(6):e0156370.

Kirsi Kivelä, Satu Elo, Helvi Kyngäs, and Maria Kääriäinen. 2014. The effects of health coaching on adult patients with chronic diseases: a systematic review. *Patient education and counseling*, 97(2):147–157.

Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B Hekler, and Susan A Murphy. 2018. Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of heartsteps. *Ann behave med*.

Lean L Kramer, Silke Ter Stal, Bob C Mulder, Emely de Vet, and Lex van Velsen. 2020. Developing embodied conversational agents for coaching people in a healthy lifestyle: Scoping review. *Journal of medical Internet research*, 22(2):e14058.

Aniek J Lentferink, Hilbrand KE Oldenhuis, Martijn de Groot, Louis Polstra, Hugo Velthuijsen, and Julia EWC van Gemert-Pijnen. 2017. Key components in ehealth interventions combining self-tracking and persuasive ecoaching to promote a healthier lifestyle: A scoping review. *J Med Internet Res.*, 19(8):e277.

Jingyuan Li and Xiao Sun. 2018. A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation. *arXiv preprint arXiv:1806.07000*.

Felipe Lobelo, Heval M. Kelli, Sheri Chernetsky Tejedor, Michael Pratt, Michael V. McConnell, Seth S. Martin, and Gregory J. Welk. 2016. The wild wild west: A framework to integrate mHealth software applications and wearables to support physical activity assessment, counseling and interventions for cardiovascular disease risk reduction. *Progress in Cardiovascular Diseases*, 58(6):584–594.

Susan Mahon, Rita Krishnamurthi, Alain Vandal, Emma Witt, Suzanne Barker-Collo, Priya Parmar, Alice Theadom, Alan Barber, Bruce Arroll, and Elaine Rush. 2018. Primary prevention of stroke and cardiovascular disease in the community (prevents): Methodology of a health wellness coaching intervention to reduce stroke and cardiovascular disease risk, a randomized clinical trial.

Jacqueline Louise Mair, Lawrence D Hayes, Amy K Campbell, Duncan S Buchan, Chris Easton, and Nicholas Sculthorpe. 2022. A personalized smartphone-delivered just-in-time adaptive intervention (jitabug) to increase physical activity in older adults: Mixed methods feasibility study. *JMIR formative research*, 6(4):e34662.

C. Di Marco, P. Bray, H.D. Covvey, D.D. Cowan, V. Di Ciccio, E. Hovy, Joan Lipa, and C. Yang. 2006. Authoring and generation of individualized patient education materials. *AMIA Annu Symp Proc*, pages 195–199.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6834–6842.

Anouk Middelweerd, Julia S Mollee, C Natalie van der Wal, Johannes Brug, and Saskia J te Velde. 2014. Apps to promote physical activity among adults: a review and content analysis. *International Journal of Behavioral Nutrition and Physical Activity*, 11(1).

Elliot G. Mitchell, Rosa Maimone, Andrea Cassells, Jonathan N. Tobin, Patricia Davidson, Arlene M. Smaldone, and Lena Mamykina. 2021. Automated vs. human health coaching: Exploring participant and practitioner experiences. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1).

François Modave, Jiang Bian, Trevor Leavitt, Jennifer Bromwell, Charles Harris III, and Heather Vincent. 2015. Low quality of free coaching apps with respect to the american college of sports medicine guidelines: A review of current mobile apps. *JMIR mHealth and uHealth*, 3(3):e77.

Lili Mou, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2015. Backward and forward language modeling for constrained sentence generation. *arXiv preprint arXiv:1512.06612*.

AM Muller, A Blandford, and L Yardley. 2017. The conceptualization of a just-in-time adaptive intervention (jitai) for the reduction of sedentary behavior in older adults. *mHealth*, 3:37.

Adrià Muntaner, Josep Vidal-Conti, and Pere Palou. 2016. Increasing physical activity through mobile device interventions: A systematic review. *Health Informatics Journal*, 22(3):451–469.

Adity U Mutsuddi and Kay Connelly. 2012. Text messages for encouraging physical activity are they effective after the novelty effect wears off? In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 33–40. IEEE.

Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462.

Eugene Z Oddone, Jennifer M Gierisch, Linda L Sanders, Angela Fagerlin, Jordan Sparks, Felicia McCant, Carrie May, Maren K Olsen, and Laura J Damschroder. 2018. A coaching by telephone intervention on engaging patients to address modifiable cardiovascular risk factors: a randomized controlled trial. *Journal of general internal medicine*, 33(9):1487–1494.

Harm op den Akker, Miriam Cabrita, Rieks op den Akker, Valerie M. Jones, and Hermie J. Hermens. 2015. Tailored motivational message generation: A model and practical framework for real-time physical activity coaching. *Journal of Biomedical Informatics*, 55:104–115.

Harm op den Akker, Valerie M. Jones, and Hermie J. Hermens. 2014. Tailoring real-time physical activity coaching systems: a literature survey and model. *User Modeling and User-Adapted Interaction*, 24(5):351–392.

Gillian A. O'Reilly and Donna Spruijt-Metz. 2013. Current mHealth technologies for physical activity assessment and promotion. *American Journal of Preventive Medicine*, 45(4):501–507.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137.

Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 2018. Analyzing the quality of counseling conversations: the tell-tale signs of high-quality counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Helmut Prendinger and Mitsuru Ishizuka. 2005. The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence*, 19:267–285.

Seyedeh Zahra Razavi. 2021. *Dialogue management and turn-taking automation in a speech-based conversational agent*. Ph.D. thesis, DAI-A 83/4(E), Dissertation Abstracts International.

Lena Reed, Shereen Oraby, and Marilyn Walker. 2018. Can neural generators for dialogue learn sentence planning and discourse structuring? *arXiv preprint arXiv:1809.03015*.

Ehud Reiter. 2018. Hallucination in neural nlg.

Michael A Rupp, Jessica R Michaelis, Daniel S McConnell, and Janan A Smither. 2018. The role of individual differences on perceptions of wearable fitness device trust, usability, and motivational impact. *Applied ergonomics*, 70:77–87.

Matthew Saponaro. 2020. *Adaptive Real-time Coaching in Free-living Conditions*. Ph.D. thesis, University of Delaware.

Matthew Saponaro, Ajith Vemuri, Greg Dominick, and Keith Decker. 2021. Contextualization and individualization for just-in-time adaptive interventions to reduce sedentary behavior. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 246–256, New York, NY, USA. Association for Computing Machinery.

Matthew Saponaro, Haoran Wei, and Keith Decker. 2017. Towards learning efficient intervention policies for wearable devices. In *Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017 IEEE/ACM International Conference on*, pages 298–299. IEEE.

Susan M Schembre, Yue Liao, Michael C Robertson, Genevieve Fridlund Dunton, Jacqueline Kerr, Meghan E Haffey, Taylor Burnett, Karen Basen-Engquist, and Rachel S Hicklen. 2018. Just-in-time feedback in diet and physical activity interventions: systematic review and practical design framework. *Journal of medical Internet research*, 20(3):e106.

D Spruijt-Metz, C Wen, and G O'Reilly. 2015. Innovations in the use of interactive technology to support weight management. *Curr Obes Rep*, 4(4):510–519.

Laura P Svetkey, Bryan C Batch, Pao-Hwa Lin, Stephen S Intille, Leonor Corsino, Crystal C Tyson, Hayden B Bosworth, Steven C Grambow, Corrine Voils, and Catherine Loria. 2015. Cell phone intervention for you (city): a randomized, controlled trial of behavioral weight loss intervention for young adults using mobile technology. *Obesity*, 23(11):2133–2141.

Martina Toshevska and Sonja Gievska. 2021. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*.

Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Alvaro Alonso, Andrea Z Beaton, Marcio S Bittencourt, Amelia K Boehme, Alfred E Buxton, April P Carson, Yvonne Commodore-Mensah, et al. 2022. Heart disease and stroke statistics—2022 update: a report from the american heart association. *Circulation*, 145(8):e153–e639.

Ajith Vemuri, Keith Decker, Matthew Saponaro, and Greg Dominick. 2021. Multi agent architecture for automated health coaching. *Journal of Medical Systems*, 45(11).

Julie B. Wang, Lisa A. Cadmus-Bertram, Loki Natarajan, Martha M. White, Hala Madanat, Jeanne F. Nichols, Guadalupe X. Ayala, and John P. Pierce. 2015a. Wearable sensor/device (fitbit one) and SMS text-messaging prompts to increase physical activity in overweight and obese adults: A randomized controlled trial. *Telemedicine and e-Health*, 21(10):782–792.

Julie B Wang, Lisa A Cadmus-Bertram, Loki Natarajan, Martha M White, Hala Madanat, Jeanne F Nichols, Guadalupe X Ayala, and John P Pierce. 2015b. Wearable sensor/device (fitbit one) and sms text-messaging prompts to increase physical activity in overweight and obese adults: a randomized controlled trial. *Telemedicine and e-Health*, 21(10):782–792.

Rachel Willard-Grace, Ellen H. Chen, Danielle Hessler, Denise DeVore, Camille Prado, Thomas Bodenheimer, and David H. Thom. 2015. Health coaching by medical assistants to improve control of diabetes, hypertension, and hyperlipidemia in low-income patients: A randomized controlled trial. *The Annals of Family Medicine*, 13(2):130–138.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Comput. Speech Lang.*, 24(2):150–174.