FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

MARINA - A Mobile App foR medIcal anNotAtion

Tiago Antunes

Master in Informatics and Computing Engineering

Supervisor: António Monteiro

Co-supervisor: Sílvia Rêgo

January 17, 2025

MARINA - A Mobile App foR medIcal anNotAtion

Tiago Antunes

Master in Informatics and Computing Engineering

January 17, 2025

Contents

List of Figures

List of Tables

Abbreviations and Symbols

AI	Artificial Intelligence
AMT	Amazon Mechanical Turk
CSI	Crowdsourcing-Based Software Inspection
DSR	Design Science Research
FDA	Food and Drug Administration
GDPR	General Data Protection Regulation
ICT	Information and Communication Technologies
M4JAM	Money for Jam
MARINA	Mobile App foR medIcal anNotAtion
MARS	Mobile App Rating Scale
mHealth	mobile Health
ML	Machine Learning
NCBI	National Center for Biotechnology Information
NLP	Natural Language Processing
PRO	Patient-Reported Outcome
SMS	Short Message Service
SQ	Search Query
SUS	System Usability Scale
UCD	User-Centered Design
UI	User Interface
UX	User eXperience

Chapter 1

Introduction

This dissertation presents MARINA, a mobile application designed to support the content validation of educational materials in healthcare, particularly diabetes. MARINA leverages crowdsourcing to address the limitations of traditional methods, which are time-consuming and prone to high withdrawal rates. The following sections explore the context, motivation, and challenges faced in this research.

1.1 Context

Educational materials for health play a crucial role in communicating vital information to patients and the general public. The accuracy and reliability of these materials are of paramount importance to avoid misinformation, misinterpretations, and potential harm to health [1]. Content validation, the process of ensuring that content is accurate, appropriate, and understandable for the target audience, is therefore essential in the development of educational materials for health [2].

Traditionally, content validation has relied on the expertise of medical professionals who are responsible for reviewing and approving the content [1] [3]. While this method has proven effective, it faces several challenges in a rapidly evolving healthcare landscape. The growing demand for content validation, driven by the proliferation of wellness and medical care-related projects, has placed a significant strain on the availability of specialists. This traditional method is often laborious and time-consuming, leading to high dropout rates and delays in the development of educational materials [1].

The emergence of crowdsourcing, specifically crowdsourcing of microtasks, offers a promising solution to address the challenges of traditional content validation. Crowdsourcing of microtasks involves breaking down complex tasks into smaller, more manageable tasks that can be distributed to a large crowd of workers through online platforms [3]. This approach has shown potential for generating high-quality annotations in biomedical text, as demonstrated by studies using crowdsourcing platforms such as Amazon Mechanical Turk (AMT) [3].

In addition, crowdsourcing of microtasks can offer significant advantages in terms of efficiency and scalability. The distributed nature of crowdsourcing allows for the rapid completion of tasks, potentially reducing time and cost compared to traditional methods [3]. Furthermore, crowdsourcing platforms provide access to a vast pool of workers, enabling the easy scaling of content validation efforts to meet the growing demands of health projects.

Crowdsourcing of microtasks, facilitated through mobile applications, has the potential to revolutionize the process of health content validation. By leveraging the power of collective intelligence, crowdsourcing-based mobile applications can improve the accuracy, efficiency, and accessibility of content validation, ensuring that educational materials for health are reliable and effective in empowering patients and improving health outcomes.

1.2 Motivation

Content validation of health educational materials is crucial to ensure the accuracy, suitability, and comprehensibility of information for the target audience.

As observed by Rothman et al. (2015) [2], the development of patient-reported outcome (PRO) instruments, which are frequently used in health educational materials, can take at least 24 months and cost between US$1 million and US$5 million using traditional methods [2].

The inherent limitations of traditional content validation methods highlight the need for a more efficient, scalable, and cost-effective approach. This is where microtask crowdsourcing comes in. Crowdsourcing, in general, has shown promise in obtaining insights from a diverse group of workers [2]. Microtask crowdsourcing, in particular, has demonstrated its ability to generate high-quality annotations in biomedical text [3]. This approach involves breaking down complex tasks, such as content validation, into smaller, more manageable microtasks that can be distributed to a vast crowd of workers through online platforms.

The success of crowdsourcing depends on the clarity of tasks, instructions, and descriptions, impacting the quality of the work produced [4]. However, task interfaces are often poorly designed or even buggy, preventing the completion of tasks [4]. It is essential to understand how different user interface (UI) characteristics can impact worker performance on various crowdsourcing platforms. For example, in Gadiraju et al. (2017) [4] we found that certain UI elements, such as the size of input boxes, input format validation, and autocorrection, can significantly impact the worker experience and the quality of work produced.

The rise of mobile applications, combined with the potential of microtask crowdsourcing, offers a unique opportunity to revolutionize the health content validation process. Mobile applications allow for ubiquitous participation, enabling workers to contribute to content validation tasks anywhere and anytime. This accessibility, combined with the potential for engaging user elements, such as push notifications and gamification elements, makes mobile applications an ideal platform for crowdsourcing content validation.

By developing a user-centric mobile application for crowdsourcing content validation of health educational materials, this dissertation aims to address the limitations of traditional methods while leveraging the power and efficiency of microtask crowdsourcing. Focusing on diabetes-related educational materials as a use case, this thesis aims not only to improve the accuracy and reliability of diabetes information but also to establish a foundation for a broader content validation platform that can be used for other health-related projects.

1.3 Problem

The MARINA application aims to validate educational content about diabetes, which will subsequently be used in a chatbot powered by a machine learning (ML) model with natural language processing (NLP). In this context, the quality of annotations collected through crowdsourcing is crucial. Inconsistent annotations can significantly impact the accuracy of the chatbot's ML model. For instance, if annotations about diabetes symptoms are inconsistent, the chatbot may provide inaccurate or incomplete information to users [5]. The accuracy of the chatbot relies on the quality of the data used to train it. If the educational content validated by the MARINA application contains inconsistencies, the chatbot may learn incorrect patterns and provide inaccurate information to users.

To mitigate this risk, we propose using user-centered design (UCD) to create an intuitive and easy-to-use application, which will increase task completion rates and, consequently, the quality of the data collected [4]. Additionally, the integration of automatic notifications and carefully designed UI/UX elements will encourage the users participation in the platform [4].

The MARINA application, by combining crowdsourcing of microtasks with UCD, has the potential to revolutionize the process of content validation in healthcare, making it more efficient, economical, and accessible, while ensuring the accuracy of the information provided to diabetes chatbot users.

1.4 Research questions

The following research question will guide the investigation:

- Can crowdsourcing of microtasks on a mobile app achieve levels of accuracy and reliability comparable to traditional methods of expert validation in terms of health content validation?

Chapter 2

Literature Review

This literature review provides a comprehensive overview of relevant studies related to MARINA, a mobile app for medical annotation. It focuses on key areas such as medical crowdsourcing, content validation, microtasking, and mobile app development within the healthcare sector. The review employs a narrative approach, allowing for the synthesis of diverse research findings and theories that inform the UCD of the app.

2.1 Strategy

For this thesis, I chose a narrative literature review. This approach fits well in a research project with a strong practical component, like this one. A narrative review allows a broad and flexible analysis, which is useful for assessing and synthesizing diverse topics related to MARINA, such as microtasking, crowdsourcing in healthcare, and UCD in mobile apps.

In a narrative review, I'll organize relevant studies by themes rather than systematically covering each work in the field. This choice enables me to interpret findings and trends in a way that supports the practical aims of the project. Instead of a comprehensive, exhaustive list, I'll focus on key works that illustrate important concepts and methods for designing and implementing MARINA. By doing so, the review will directly inform the app's design and development without overloading with extraneous detail.

2.2 Data sources

I selected four key databases to gather literature for this review:

- **ACM Digital Library** focuses on computing and technology. It provided research on mobile app design, UCD, and microtask crowdsourcing, which are essential for MARINA's development.

- **Google Scholar** offers broad access to multidisciplinary sources, helping me locate diverse studies related to crowdsourcing, healthcare apps, and educational content validation. Its wide scope complements more specialized databases.
- **IEEE Xplore** specializes in engineering and technology. It contributed research on mobile applications, medical technology, and usability, which are important areas for MARINA.
- **PubMed** covers medical and life sciences. It was essential for finding studies on health education, content validation, and diabetes education, grounding the project's healthcare focus in solid medical research.

Using these databases, I ensured a balanced review of technical, interdisciplinary, and healthcare literature to inform MARINA's design and implementation.

2.3 Search strings

To gather relevant literature, I used a search query (SQ) designed to capture studies directly related to MARINA's objectives. The query focuses on terms that combine medical/health topics, crowdsourcing, content validation, mobile apps, and microtasks, which are the core elements of this project.

The query is structured to filter out unrelated or irrelevant topics, such as gaming or worker-centric studies. These exclusions are essential because MARINA's purpose is content validation for educational material, not gamified tasks or workforce studies. Excluding terms like "game," "gamification," and "worker" keeps the search results aligned with the project's goals, focusing on studies relevant to healthcare, education, and crowdsourcing in non-gaming contexts.

The SQ in the table 2.1 allows for flexibility in finding studies on crowdsourced content validation and microtasking for healthcare without unnecessary distractions. This targeted approach supports a well-focused literature review, guiding the selection of studies that directly inform MARINA's design and functionality.

Table 2.1: Last stage of search expression

SQ	Search expression
1	(((medical OR health) AND crowdsourcing AND "content validation") OR ((mobile OR app) AND microtask) OR (microtask AND crowdsourcing)) NOT game NOT gamification NOT "worker"

2.4 Inclusion and exclusion criteria

I applied specific inclusion and exclusion criteria to refine the literature search and ensure relevance to MARINA's goals.

- **Inclusion criteria**
  - Search engine results after executing the SQ: I only included studies that appeared directly in response to the tailored SQ. This ensures each study aligns with MARINA's focus on healthcare, crowdsourcing, and content validation.
  - Date of publication since 2014: I restricted the review to publications from 2014 onward to capture recent research, given the fast-paced evolution of technology in mobile apps, crowdsourcing, and healthcare.
  - Full-text access: Only studies with full-text access were included. This ensures I can fully assess each study's methods and findings, maintaining a high standard for analysis and relevance.
  - Published in English or Portuguese: I included publications in English and Portuguese, maximizing comprehension and relevance for this thesis.
- **Exclusion criteria**
  - Not directly related to the subject of this thesis: Studies that didn't directly address MARINA's themes—medical crowdsourcing, content validation, microtasks, or mobile app design—were excluded to keep the focus tight.
  - Not duplicated: Duplicates were removed to avoid redundancy, keeping the review efficient and streamlined.
  - Not work-in-progress publications: I excluded work-in-progress publications to ensure the review only includes completed research, providing robust and validated findings relevant to MARINA's objectives.

2.5 Screening process & results

The screening process started with 545 unique records retrieved from the four databases (ACM Digital Library, Google Scholar, IEEE Xplore, and PubMed). After applying each filter step-by-step, I progressively narrowed down the results to ensure relevance and quality, as summarized in the table.

1. Date Filter: I first applied a filter for publications from 2014 onward, excluding 58 records and reducing the pool to 487. This ensured the review focused on recent research.

2. Language Filter: Next, I applied a language filter, excluding 12 non-English/Portuguese records. This step left 475 records, all accessible for analysis.

3. Title Screening: I screened titles to remove studies unrelated to MARINA's scope (e.g., those focusing on unrelated fields or applications). This step excluded 391 records, leaving 84.

4. Abstract Screening: I then reviewed abstracts for relevance to core themes like crowdsourcing, content validation, and mobile app design in healthcare. This led to the exclusion of 30 more records, reducing the count to 54.

5. Conclusion Screening: Finally, I reviewed the conclusions of each remaining study to confirm relevance and eliminate any that didn't fully align with the thesis goals. This excluded 32 records, resulting in 22 studies for the final review.

The table 2.2 summarizes the database sources, the initial query results, and the final number of studies included, illustrating the focused approach used to identify relevant literature.

Table 2.2: Total number of papers retrieved and included

	Initial SQ	**Included**
ACM	103	7
Google Scholar	325	5
IEEE	44	8
PubMed	15	2
Total	487	22

2.6 Eligibility

The eligibility criteria guided the selection of studies most relevant to the objectives of this narrative review. Studies were considered eligible if they focused on areas central to MARINA's design and purpose: medical crowdsourcing, content validation, microtasking, or mobile app development for healthcare applications. Inclusion was limited to recent, completed studies published from 2014 onward, with full-text access in English or Portuguese to maintain a high standard for depth and accessibility.

In addition to the studies identified through database searches, six references from the initial sources cited in the thesis proposal were included. These references provided foundational information that aligned with the project's goals.

This approach ensured that the review was both current and comprehensive, covering essential themes in UCD for microtask crowdsourcing in healthcare. The full list of selected references can be consulted in the bibliography.

2.7 Selection

The selection process focused on identifying studies that directly inform MARINA's design as a crowdsourcing app for content validation in healthcare. Starting from an initial pool of 545 unique records, I applied structured filters (detailed in the section 2.5) to refine the selection based on criteria such as date, language, and relevance.

Each study was evaluated for its practical relevance to MARINA's core themes: medical crowdsourcing, content validation, microtasking, and mobile app design in healthcare. Studies that offered insights into UCD, educational content validation, and healthcare technology were prioritized. This selective approach ensured that each included study contributes directly to the goals of this project, supporting a focused and actionable literature review.

2.8 Content assessment and analysis

This section examines the selected references to evaluate their relevance and contribution to the development of MARINA. The analysis focuses on the yearly distribution of sources, the primary application domains they address, and a detailed examination of key articles. These insights provide a foundation for understanding existing approaches, identifying gaps, and aligning the design of MARINA with current trends and challenges in crowdsourcing and medical content validation.

2.8.1 Yearly Distribution of Sources

The selected articles were evaluated and categorized by year of publication and source type. Source types include journal quartiles (Q1–Q4), conferences, and book sections.

Conferences make up the majority of publications, reflecting the fast-paced and iterative nature of research in technology and medical crowdsourcing.

High-impact journal articles, particularly those in Q1 and Q2, play a complementary role. They provide rigorously validated peer-reviewed information, ensuring that research is grounded in well-established frameworks and methodologies.

Over time, the trend shows an increase in Q1 and Q2 journal articles, especially after 2019. This suggests a growing maturity and acceptance of crowdsourcing and mobile health (mHealth) topics in academic discussions. The single book section from 2015 adds breadth, offering a broader perspective on related themes.

The accompanying bar chart 2.1 illustrates the distribution of publication types by year. It highlights the balance between emerging research and established work, with a clear shift toward high-impact journals in recent years.

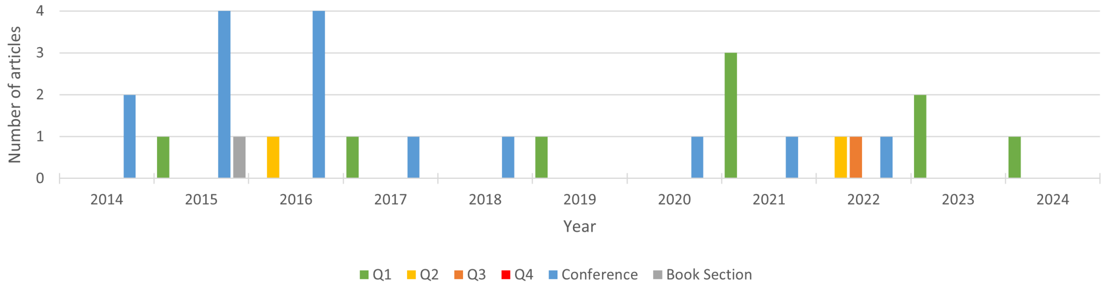


Figure 2.1: Distribution of publication types by year

2.8.2 Main Application Domains

The reviewed literature was also analyzed by its primary application domains. This categorization highlights how the research intersects with the objectives of developing a microtask crowdsourcing app for medical content validation.

- Crowdsourcing and Microtasking dominate the reviewed literature. These fields provide the theoretical and practical foundation for designing a platform that engages users to complete small, focused tasks efficiently.
- Content Validation aligns directly with the thesis's goal of ensuring accurate and reliable educational materials, particularly in the health domain.
- Health offers essential context for addressing the challenges of medical content validation.
- Mobile reflects the technical aspect of the thesis but shows limited direct research. This highlights a gap and emphasizes the contribution of the thesis to this area.

The pie chart 2.2 illustrates the proportional focus on each application domain. It underscores the multidisciplinary nature of this thesis, linking technology, healthcare, and crowdsourcing to achieve its objectives.

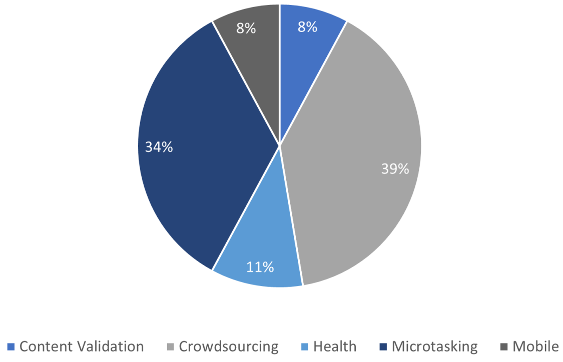


Figure 2.2: Main application domains

2.8.3 Article Analysis

For a comprehensive understanding of the reviewed literature, each article was systematically analyzed using a set of six key questions. These questions focus on the core aspects of the research, providing insight into its motivation, solution, methodology, conclusions, limitations, and potential future directions. By answering these questions for each article, we can better assess how each study contributes to the overall field of microtask crowdsourcing for medical content validation.

The six questions addressed for each article are:

1. Why was the work done?
2. What was done?
3. How was it validated?
4. What was concluded?
5. Limitations?
6. Trends for future research?

The answers to these questions provide a detailed picture of the state of the research and allow for a structured comparison of how each article contributes to the development of a microtask crowdsourcing app for content validation in healthcare. Some articles, however, lacked sufficient information to fully answer all six questions.

The complete table with answers to these questions for each article can be found in the table A.1. The table organizes the articles by reference (R), question number (Q) and corresponding answers (A).

Chapter 3

State of the Art

This chapter reviews the key concepts and methods relevant to this dissertation. It covers the importance of content validation in healthcare, limitations of traditional methods, and the potential of crowdsourcing and microtasking. It also examines relevant platforms, technologies, and design principles, ending with a critical discussion to guide MARINA's development.

3.1 Context and Importance of Content Validation in Healthcare

Content validation, an essential process in healthcare, ensures that educational materials for patients are accurate, understandable, and unbiased. Reliable medical and wellness information is critical for patient safety and to prevent misinformation, which can lead to harmful health decisions [6] [7]. This process guarantees that the information shared is factually correct, tailored to the target audience, and free from misleading content [1] [6].

The increasing digitalization of medical information makes this process even more critical, requiring new approaches to ensure the quality and reliability of health and educational materials [6] [8]. The internet has flooded patients with medical information that is not always trustworthy, making content validation essential to filter this information and guide patients toward credible and accurate sources [8].

The importance of content validation in healthcare is highlighted by several factors:

- **Patient Safety**: Inaccurate medical information can lead to poor decisions with severe consequences [1] [6]. Validation reduces this risk [6].

- **Legal and Ethical Responsibility**: Healthcare professionals and institutions have a legal and ethical duty to provide accurate and complete information to their patients. Content validation helps fulfill this responsibility [9].

- **Trust and Credibility**: Educational materials validated by experts strengthen patient trust in the information provided and enhance the credibility of healthcare institutions [6].

- **Improving Health Literacy**: Validated educational materials contribute to improving the population's health literacy, empowering patients to make informed decisions about their healthcare [1].

Content validation is essential in several areas of healthcare, including:

- **Patient Education Materials**: Brochures, websites, videos, and other educational materials must be validated to ensure that information about diseases, treatments, and prevention is accurate and understandable [1] [7].
- **Healthcare Professional Training**: Content validation is essential to ensure the quality of courses and training materials for healthcare professionals [9].
- **Clinical Research**: Content validation is important to ensure the accuracy and reliability of data collection instruments in clinical trials [3].
- **Health Mobile Applications**: Apps that provide medical information or health tracking to patients must have their content validated to ensure safety and effectiveness [10].

3.2 Traditional Methods of Content Validation

Traditional methods of health content validation typically involve expert reviews [2] [11]. This process usually includes a small group of subject matter experts who review the content for accuracy, completeness, relevance, and appropriateness [2] [5]. However, this method can be time-consuming and costly, often leading to high dropout rates due to the workload placed on the experts [2] [11].

Some traditional content validation methods include:

- **Peer review**: In this method, the content is reviewed by one or more subject matter experts. Reviewers provide feedback on the accuracy, completeness, relevance, and clarity of the content [2].
- **Focus groups**: Focus groups involve gathering a small group of people from the target population to discuss the content. Focus group participants can provide feedback on their understanding of the content, as well as any concerns or suggestions they may have [2].
- **Interviews**: Interviews can be conducted with subject matter experts or members of the target population to gather feedback on the content [2].
- **Usability testing**: Usability testing involves observing users as they interact with the content. This can help identify any issues with the usability or clarity of the content [12].
- **Surveys**: Surveys can be used to gather feedback from a larger sample of the target population [12].

It is important to note that inconsistencies in annotations are common, even when highly experienced clinical experts annotate the same phenomenon. This may be due to inherent biases of the experts, judgment errors, lapses, and other factors [5].

While traditional methods of content validation can be effective, they can also be challenging to implement and scale [2] [5] [11]. As the volume of health content continues to grow, there is an increasing need for more efficient and scalable solutions for content validation [5] [11].

3.3 Crowdsourcing and Microtasking as Emerging Solutions

Crowdsourcing and microtasking are emerging as promising solutions to address the challenges of traditional content validation in healthcare. Crowdsourcing platforms, such as AMT, have been widely used by researchers in various fields, including healthcare, for tasks like data processing, data extraction, transcription, and sentiment analysis. Recent studies have shown the reliability and potential of AMT as a low-cost, time-efficient tool for analyzing health-related data [13].

Microtasking, a subset of crowdsourcing, involves breaking down large tasks into smaller, more manageable components [10] [14] [15]. This allows a large number of workers, usually online, to contribute to the completion of the task as a whole [6]. In the healthcare domain, crowdsourcing and microtasking have been used for a variety of tasks, including image annotation, data transcription, data curation, and content validation [13] [14] [16].

3.3.1 Advantages for content validation in healthcare

- **Efficiency and speed**: Crowdsourcing can significantly reduce the time required to validate content by distributing small tasks to a large number of workers [3] [11] [13].

- **Cost reduction**: Crowdsourcing platforms can offer a more economical solution than traditional content validation methods [3] [13].

- **Diversity of perspectives**: The participation of a diverse crowd of workers can provide broader perspectives and identify potential issues that may be overlooked by a smaller group of experts [3] [9] [13].

- **Scalability**: Crowdsourcing platforms can easily handle large volumes of data, making them suitable for large-scale content validation projects [3] [13].

- **Flexibility**: Crowdsourcing platforms allow researchers to define and adjust tasks according to their specific needs [3].

- **Innovation**: Crowdsourcing platforms can foster innovation by leveraging the collective knowledge of the crowd [8].

3.3.2 Applications in content validation in healthcare

- **Creating training datasets for machine learning algorithms**: Crowd workers can annotate unstructured data, such as social media posts, to create training datasets for machine learning models used in pharmacovigilance [11] .

- **Identifying disease mentions in PubMed abstracts**: Crowd workers can identify and annotate disease mentions in biomedical texts, creating valuable resources for research [3].

- **Validating predicted gene-mutation relationships in PubMed abstracts**: Crowdsourcing can be used to validate predictions made by natural language processing systems, improving the accuracy of results [3].

- **Extracting medical relationships**: Crowd workers can extract relevant relationships from biomedical texts, such as gene-disease interactions or drug side effects [3].

- **Evaluating health mobile apps**: Crowdsourcing can be used to assess the usability, content, and overall quality of health mobile applications [12].

3.3.3 Challenges and considerations

- **Data quality**: It is essential to ensure the quality of annotations provided by crowd workers [3] [6] . This can be done through quality control mechanisms such as qualification tests, majority voting, expert review, and gold standard data [3] [6] [13].

- **Ethical issues**: Ethical aspects of crowdsourcing, such as fair compensation for workers, privacy protection, and preventing exploitation, must be considered [11].

- **Task design**: Tasks should be well-defined, easy to understand, and suitable for non-expert workers [3].

- **Crowd management**: Effective crowd management is crucial for the success of a crowdsourcing project, including recruitment, communication, and feedback [3].

3.4 Crowdsourcing Platforms and Relevant Technologies

Crowdsourcing has become increasingly popular in various fields, including healthcare. Several platforms and technologies facilitate this process, each with its specific features and strengths. In this section, we explore some of the most relevant crowdsourcing platforms and technologies, focusing on their capabilities and applicability to health content validation.

3.4.1 Microtasking Platforms

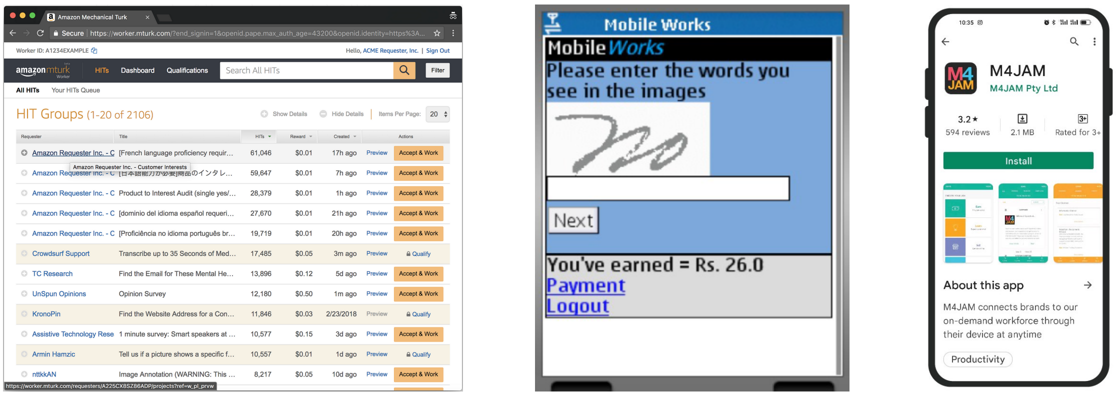
Microtasking platforms are a subcategory of crowdsourcing platforms that focus on breaking down large tasks into small units, called microtasks. These microtasks can be completed quickly and independently by a large number of workers, known as crowdworkers. Microtasking platforms are particularly well-suited for tasks involving data collection and annotation, making them relevant for the validation of medical content. Here are some notable microtasking platforms:

- **AMT**: One of the most popular microtasking platforms, MTurk offers a large user base and an easy-to-use interface for creating and managing microtasks. It has been widely used in various domains, including healthcare, for data collection, annotation, and sentiment analysis. Its scalability and cost-efficiency make it a potential tool for large-scale content validation [13].

- **MobileWorks**: A managed microtasking platform, MobileWorks focuses on ensuring high-quality work through the curation of its crowdworkers and implementation of quality control mechanisms. Its emphasis on quality can be especially beneficial for sensitive tasks like medical annotation, where accuracy is crucial [6] [9].

- **Money for Jam** (M4JAM): A microtasking platform focused on emerging markets, M4JAM provides employment opportunities to workers in developing countries [6] [17]. Its accessibility and diverse user base can be advantageous for gathering perspectives from a wider range of individuals.

- **mClerk**: A microtasking platform designed for data collection and validation [6] [9] [17]. Its specialization in data collection makes it well-suited for creating annotated datasets for content validation purposes.



Worker ID: A1234EXAMPLE

Hello, ACME Requester, Inc. | Sign Out

HIT Groups (1-20 of 2106)

Requester	Title	HITs	Reward	Created	Actions	
Amazon Requester Inc. - C	[French language proficiency requir…	61,046	$0.01	17h ago	Preview	Accept & Work
Amazon Requester Inc. - C	[日本語能力が必要]商品のインタレ…	59,647	$0.01	7h ago	Preview	Accept & Work
Amazon Requester Inc. - C	Product to Interest Audit (single yes/…	28,379	$0.01	1h ago	Preview	Accept & Work
Amazon Requester Inc. - C	[dominio del idioma español requeri…	27,670	$0.01	21h ago	Preview	Accept & Work
Amazon Requester Inc. - C	[Proficiência no idioma português br…	19,719	$0.01	20h ago	Preview	Accept & Work
Crowdsurf Support	Transcribe up to 35 Seconds of Med…	17,485	$0.05	3m ago	Preview	Qualify
TC Research	Find the Email for These Mental He…	13,896	$0.12	5d ago	Preview	Accept & Work
UnSpun Opinions	Opinion Survey	12,180	$0.50	1m ago	Preview	Accept & Work
KronoPin	Find the Website Address for a Con…	11,846	$0.03	2/23/2018	Preview	Qualify
Assistive Technology Rese	1 minute survey: Smart speakers at …	10,577	$0.15	3d ago	Preview	Accept & Work
Armin Hamzic	Tell us if a picture shows a specific f…	10,557	$0.01	1d ago	Preview	Qualify
nttkkAN	Image Annotation (WARNING: This …	8,217	$0.05	10d ago	Preview	Accept & Work

Amazon Requester Inc. - Customer Interests

https://worker.mturk.com/requesters/A225CX8SZ86ADP/projects?ref=w_pl_prvw

(a) AMT

Please enter the words you see in the images

You've earned = Rs. 26.0

(b) MobileWorks

M4JAM connects brands to our on-demand workforce through their device at anytime

(c) M4JAM

Figure 3.1: Example of microstasking platforms

3.4.2 Relevant Technologies

Beyond crowdsourcing platforms, several technologies support the microtasking process and content validation. These technologies facilitate the creation, management, and analysis of crowdsourcing data:

- **NLP**: A branch of artificial intelligence focused on enabling computers to understand, interpret, and manipulate human language [15]. In the context of crowdsourcing, NLP can be used to automate tasks such as entity extraction, sentiment analysis, and text summarization. This automation can improve the efficiency and accuracy of content validation by identifying errors or inconsistencies in the text [15].

- **Deep Learning**: A type of machine learning that uses artificial neural networks to learn complex representations from data [15] [18]. In the context of crowdsourcing, deep learning can be used to develop predictive models that automate tasks such as task assignment and quality control [18]. This automation can improve the efficiency and scalability of crowdsourcing platforms.

- **User-Centered Interfaces**: Well-designed user interfaces are crucial to the success of microtasking platforms. Easy-to-use, intuitive, and mobile-accessible interfaces can increase user participation and improve the quality of collected data [4] [10] [17]. Specific considerations for medical content validation platforms may include clear presentation of instructions, feedback mechanisms, and the ability to incorporate multimedia features [3].

- **Gamification**: Gamification involves applying game design elements to non-game contexts to increase user engagement and motivation [15]. On microtasking platforms, gamification can be used to make tasks more enjoyable and rewarding, which can lead to higher completion rates and better data quality.

- **Data Validation and Reliability**: Ensuring the quality of crowdsourced data is essential, especially in the sensitive context of healthcare. Various methods can be used to validate and improve the reliability of data, including majority voting, expert gold standards, and machine learning algorithms [3] [5]. The selection of the most appropriate method depends on the specific tasks and data quality requirements.

3.5 User-Centered Design Considerations for Microtasking Applications

UCD is crucial for the success of any application, especially for microtasking platforms where user participation and performance are key [6] [15]. In the context of a mobile app for medical annotation, such as MARINA, UCD considerations become even more important due to the sensitive nature of the information and the need to ensure data reliability.

3.5.1 The Importance of UCD in Microtasking

Microtasking platforms rely heavily on voluntary participation of a large number of users. If the app is complex, confusing, or frustrating, users can quickly abandon it, compromising the quality and speed of data collection [6]. A UCD ensures the app is:

- **Intuitive and easy to use**: The interface should be simple, with clear and concise instructions, so that users can easily understand the tasks and how to complete them [3] [19].
- **Engaging and rewarding**: The app should provide a positive experience for users, motivating them to participate and contribute high-quality data [6] [17]. This can be achieved through gamification elements, positive feedback, and an effective reward system [15].
- **Accessible to a wide range of users**: The design should reflect the needs of users with varying levels of digital literacy, skills, and devices [9].

3.5.2 Specific Considerations for MARINA

Given the nature of MARINA as an application for medical annotation, the following UCD considerations are particularly important:

- **Clarity and accuracy of the instructions**: The instructions for the annotation tasks must be extremely clear, precise and unambiguous to minimize the risk of errors or misinterpretations [3] [19]. Detailed examples and a glossary of medical terms may be helpful.
- **Effective context management**: Microtasks should be presented with the appropriate context to allow for accurate annotations [10]. In the case of medical text annotations, this may involve presenting entire sentences or paragraphs, rather than just isolated words or phrases.
- **Feedback and quality control**: Feedback mechanisms and quality control are essential to ensure the reliability of the data [5] [6] [9]. This may include peer review, comparison with a reference standard, or the use of machine learning algorithms to identify and correct errors [11] [15].
- **Data privacy and security**: MARINA will handle sensitive medical information, so data privacy and security are of utmost importance [6] [11]. The application should comply with relevant data protection regulations, such as the General Data Protection Regulation (GDPR), and implement robust security measures to protect user information.

3.6 Validation and Reliability of Crowdsourcing Data

The validation and reliability of data generated through crowdsourcing are crucial to ensure the quality and trustworthiness of the results, especially in sensitive areas like healthcare. Several factors can influence data quality, including the variability in workers' expertise, inherent platform biases, and task complexity.

Data validation in crowdsourcing can be approached through multiple strategies aimed at mitigating potential errors and biases:

- **Redundancy and Voting**: Involving multiple workers in the same task and aggregating their responses through voting mechanisms (e.g., majority, consensus) is a common technique to improve accuracy [3]. However, the effectiveness of voting depends on the individual quality of the workers.

- **Specialized Workers**: Platforms like AMT allow for the selection of workers based on their qualifications, prior experience, and ratings [3] [13]. Recruiting workers with specific expertise in the medical field can enhance data reliability.
- **Qualification Tests**: Implementing qualification tests before assigning tasks helps assess workers' understanding of instructions and their ability to perform the task accurately [3].
- **Gold Standard and Benchmarking**: Comparing crowdsourced data with a reference dataset ("gold standard") previously annotated by experts allows for evaluating the accuracy and performance of the system [3].
- **Evaluation of Annotation Learnability**: Analyzing the consistency and "learnability" of annotations provided by workers can help identify and remove noisy data, improving the quality of the final dataset [5].
- **Consensus Methods**: In scenarios with multiple experts, simple majority voting to determine consensus may lead to suboptimal models [5]. Investigating more sophisticated methods for aggregating expert opinions is essential to obtain high-quality models.

3.7 Critical discussion

Validating health-related content is crucial, but traditional methods face challenges such as high costs and slow processes. Crowdsourcing and microtasking emerge as promising solutions to address these issues. However, applying these approaches to medical content validation is still an emerging field with several areas requiring further investigation.

- **Validation and Reliability of Crowdsourced Data**: A critical issue is ensuring the quality of crowdsourced data. It is essential to establish robust quality control mechanisms to minimize errors and biases inherent to non-expert participation [6] [13]. Future research should deepen the understanding of factors influencing data quality, exploring task design strategies, worker selection, and result aggregation to optimize accuracy and reliability. Strategies like qualification tests [3], voting systems [3] [13], and agreement thresholds among workers [13] are examples of approaches that could be implemented in the MARINA platform to ensure data quality.
- **User Personalization and Engagement**: Usability and user-centered design are essential for the success of microtasking platforms [10] [18]. Developing the MARINA application should prioritize creating an intuitive and user-friendly interface, considering users' needs and preferences. Incorporating personalization elements, such as task adaptation based on

cognitive profiles, can improve user performance and satisfaction. Studies have demonstrated the feasibility of personalizing microtasks through cognitive testing and user interaction analysis [18], suggesting that these techniques could be integrated into MARINA to optimize user experience and data quality.

- **Scalability and Sustainability**: MARINA should be designed to handle a high volume of content and a growing number of users. Scalability and sustainability must be carefully considered during development, including platform architecture, worker recruitment, and cost management. Insights from platforms AMT, which has shown the capacity to process large volumes of data [13], can inform the design of MARINA.

- **Ethics and Privacy**: Collecting and using health data raise ethical and privacy concerns that must be carefully addressed. Developing MARINA should ensure compliance with data privacy regulations, such as the GDPR, and implement security measures to protect user information. Practices like data anonymization and informed consent [11] should be integrated into the platform.

- **Application Domains**: While the thesis focuses on diabetes-related materials, MARINA should be flexible and adaptable to other health topics. Future research should explore generalizing the platform to different types of content, user populations, and healthcare contexts. Involving experts from various medical fields in the design process can help ensure the application's versatility.

In summary, crowdsourcing and microtasking hold significant potential for validating health content. By addressing the critical issues outlined above, the MARINA application can contribute to producing high-quality, reliable educational materials, fostering health literacy, and supporting informed decision-making in healthcare. However, it is crucial to recognize the current limitations of crowdsourcing and invest in research and development to improve its reliability, usability, and social impact.

Chapter 4

MARINA: A User-Centered Journey

4.1 Approach

The MARINA platform was designed to address the challenges of validating educational health materials, a task traditionally requiring manual, time-intensive efforts of medical experts [7]. The growing demand for content validation in projects related to well-being and healthcare, particularly in digital health [22], calls for more efficient approaches.

MARINA adopts a user-centered approach, leveraging microtasking and crowdsourcing principles [6] [11] [17] [21] . Microtasking breaks down the complex task of content validation into smaller, self-contained, and quick tasks [6] [19] [21]. This approach makes the work more accessible and suitable for mobile devices [10], enabling users to contribute during short intervals and in various contexts [17]. Crowdsourcing harnesses the collective knowledge of a large group of people [8] [17] to speed up validation and ensure content quality [3].

4.1.1 Challenges and Limitations Identified in the Literature

Existing literature highlights several shortcomings of traditional content validation methods and the need for more efficient solutions [3] [6]:

- **Time-Consuming and Labor-Intensive Processes**: Traditional methods relying on medical experts are slow and require significant effort, leading to high dropout rates [7].
- **Lack of Scalability**: Conventional approaches are not scalable to meet the growing demand for content validation [11].
- **Need for Flexibility**: Users benefit from the ability to complete microtasks at different times and locations [10] [17].
- **Role of Context**: Research on microtasks emphasizes the importance of context in completing tasks effectively [10].

4.1.2 MARINA's Improvements Over Existing Approaches

MARINA introduces several enhancements to address these issues:

- **Personalization and Adaptation**: Tasks are tailored to user profiles [6] [17] [18], which adapt content and interface design to improve task execution and information quality [18].

- **Mobile Platform**: MARINA is designed as a mobile app [4] [10] [12] [17] ensuring accessibility and flexibility for users to complete tasks anytime, anywhere, including during downtime or travel [10] [17].

- **Intuitive Interface and Notifications**: The platform offers a user-friendly interface [4] [6] [8] [10] with clear instructions and integrated UI/UX elements to increase task completion rates [4] [19] [21]. Automated smartphone notifications improve engagement and retention [22].

- **Flexible Task Management**: MARINA enables the creation and distribution of microtask lists [6] [17], ensuring tasks are matched with the right users [6] [17].

- **Focus on Content Quality**: Unlike typical crowdsourcing platforms that rely solely on worker qualifications [13], MARINA emphasizes accurate content validation by providing feedback to users [6] and employing voting methods when needed to reach consensus [3] [11]. This ensures the reliability of validated information through its profile-aware approach [6].

- **Scientific Validation**: MARINA uses a design science research methodology to assess its proposed solution through comparative analysis and crowdsourcing metrics [6] [21]. This evaluation ensures the validity, reliability, quality and utility of the platform [1] [7] [12].

4.2 Methods

This section provides an overview of the research methodology used for developing and evaluating MARINA, a mobile app designed to validate educational healthcare content, focusing on diabetes management. The detailed approach and methods are described in the table B.1, ensuring transparency, reproducibility, and validity of the results, while offering a clear understanding of the research process and outcomes.

4.3 Expected results

This dissertation focuses on presenting the MARINA mobile app, designed to support the validation of educational materials in healthcare, specifically related to diabetes. The main goal is to investigate whether using microtask-based crowdsourcing through a mobile app can achieve accuracy and reliability comparable to traditional expert validation methods. By combining crowdsourcing with user-centered design, MARINA aims to make content validation in healthcare more efficient, cost-effective, and accessible while ensuring the accuracy of information provided to a diabetes chatbot.

The expected outcomes are as follows:

- **Content validation**: The app is expected to achieve accuracy comparable to expert validation methods. This will be evaluated by analyzing the agreement between crowdsourced microtask responses and reference answers provided by experts. Comparing MARINA with traditional methods will help determine if mobile-based microtask crowdsourcing is a viable and effective alternative
- **Task completion rate**: A user-centered, intuitive app design is expected to boost task completion rates and improve the quality of collected data. Features like automatic notifications and carefully designed UI/UX elements are anticipated to encourage user participation
- **App usability**: MARINA's usability will be assessed using the System Usability Scale (SUS). The app is expected to achieve a high score, reflecting its ease of use and efficiency for users
- **Comparison with traditional methods**: The study will compare MARINA's efficiency and effectiveness with traditional content validation methods, which are often slow and have high dropout rates. MARINA is expected to prove faster, more accessible, and cost-effective
- **Contribution to health chatbots**: Content validated through MARINA will be integrated into a diabetes-focused chatbot powered by an ML model. This will improve the accuracy and reliability of diabetes-related information provided by the chatbot
- **Platform for broader health projects**: The dissertation aims to lay the groundwork for a broader content validation platform applicable to other health-related projects beyond diabetes

These expected outcomes address the need for more efficient and accessible content validation methods in healthcare. Traditional methods, like peer review, are often time-consuming and expensive. Microtask-based crowdsourcing offers a promising alternative by leveraging collective intelligence to speed up validation and lower costs. Additionally, user-centered design ensures the app is easy to use and motivates participants, leading to high-quality data.

This study aims to demonstrate the potential of crowdsourcing and user-centered design to improve healthcare content validation, with significant implications for providing accurate and reliable information. A successful implementation of MARINA will pave the way for future content validation projects in various healthcare fields.

4.4 Project plan

The project is structured around a main work package called "Dissertation," which includes all essential activities to achieve the project goals. Project management is handled using activity cards that outline each task's duration, dependencies, objectives, expected results, deliverables, and milestones.

The "Dissertation" work package is the project's core, consisting of five interconnected tasks with a defined schedule. The tasks are:

1. **Literature Review**: This task focuses on analyzing the state of the art in crowdsourcing, microtasks, healthcare content validation, and mobile applications. The goal is to identify gaps and opportunities for MARINA. The expected outcome is a comprehensive review that provides a strong theoretical basis for the project's development.

   - Duration: September 1 to December 4
   - Deliverables: "Literature Review" and "State of the Art" chapters

2. **Mobile Application Development and Testing**: The aim is to design and implement the MARINA app with an intuitive user interface and seamless user experience. Usability testing will address potential issues. The expected result is a functional and user-friendly app ready for real-world testing.

   - Duration: November 3 to March 7
   - Deliverables: MARINA app for iOS and Android, usability test report

3. **Pilot with Healthcare Professionals**: This task aims to evaluate the MARINA application in a real-world setting with healthcare professionals. The primary objective is to validate diabetes-related educational content and assess the app's effectiveness in comparison to traditional expert validation methods. By testing the app with end-users in realistic scenarios, the task seeks to establish its credibility and reliability. The expected result is a detailed evaluation of the app's accuracy and its potential to streamline the content validation process in healthcare.

   - Duration: April 8 to May 9
   - Deliverables: Pilot analysis report comparing app validation with traditional methods

4. **Pilot Analysis and App Refinements**: The purpose of this task is to analyze the data collected during the pilot phase to identify strengths and areas for improvement. This analysis will inform refinements to the MARINA application, ensuring it meets the expectations and needs of healthcare professionals. The expected outcome is an optimized and improved app that incorporates feedback from the pilot, enhancing its usability and effectiveness for content validation.

   - Duration: June 9 to July 11
   - Deliverables: Updated MARINA app, pilot analysis report

5. **Dissertation Writing**: The objective of this final task is to document the entire project process, including the literature review, methodology, results, and conclusions. This comprehensive documentation will present a critical discussion of the findings and address any limitations identified during the study. The expected result is a complete and high-quality dissertation that adheres to FEUP standards and effectively communicates the project's contributions to the field.

    - Duration: September 1 to July 11
    - Deliverables: Final dissertation

The Gantt chart 4.1 present the project plan, showing task sequences, durations, and milestones. It provides a clear timeline and aids in project management and progress tracking.

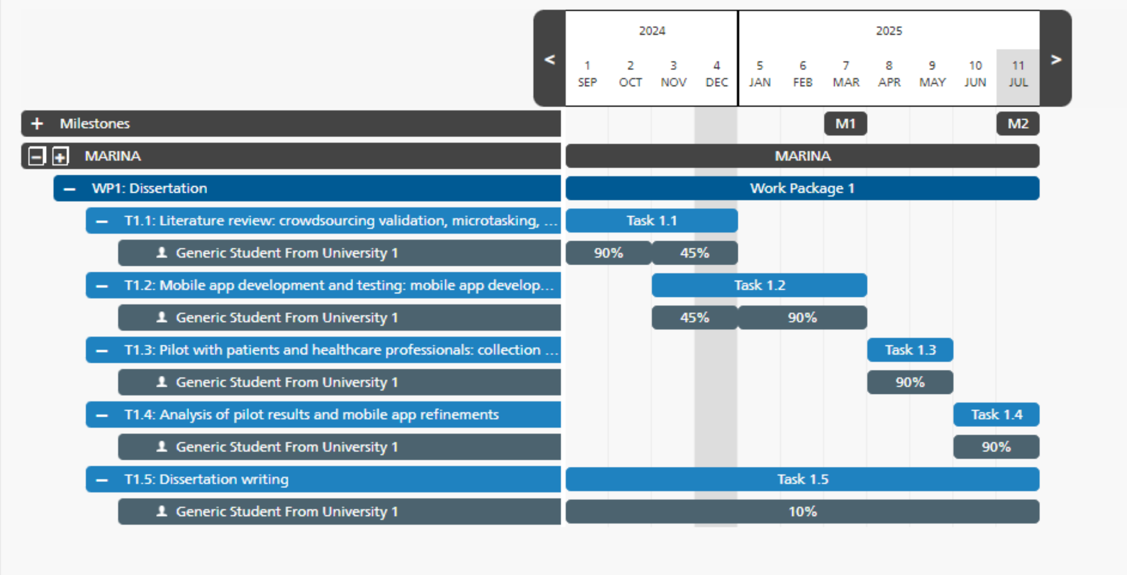


Figure 4.1: Gantt chart of the project plan

4.5 Conclusions

The MARINA project offers an efficient, cost-effective solution for validating healthcare content, focusing on diabetes. It leverages crowdsourcing via a mobile app, addressing issues of slow, costly, and high-dropout traditional validation methods. The goal is to enhance the accuracy and reliability of diabetes-related information, eventually expanding to other health domains.

4.5.1 Problem and goals

MARINA addresses the need for rapid and effective validation of healthcare educational content. Traditional methods are slow and expensive, leading to reduced accessibility and quality. The app uses crowdsourcing to validate diabetes content, which will later power a chatbot with ML and NLP features.

4.5.2 Conclusions drawn from the related work and gap analysis

Research shows growing interest in using crowdsourcing for healthcare validation but lacks focus on mobile apps for this purpose. MARINA fills this gap and explores motivational factors, task difficulty estimation, and managing workers' multidimensional skills. The project contributes to understanding these challenges within a healthcare context.

4.5.3 SMART analysis of the project goals

- **Specific**: Develop a mobile app for crowdsourced validation of diabetes content
- **Measurable**: Evaluate content accuracy using metrics like content validity index and validator agreement
- **Achievable**: The app development is feasible with available resources
- **Relevant**: The project addresses a key healthcare issue and offers benefits in cost, time, and accessibility
- **Time-bound**: The project will complete within one year, with clear phases

4.5.4 SWOT analysis of the project proposal

- **Strengths**
  - Innovative crowdsourcing approach
  - Efficient and cost-effective validation
  - Focus on diabetes for targeted development
  - Positive social impact on healthcare information quality
- **Weaknesses**
  - Relies on participant engagement
  - Needs quality assurance for crowdsourced content
  - Technical challenges in app development
- **Opportunities**
  - Expand to other health domains
  - Integrate with chatbots and digital tools
  - Build a validator community
- **Threats**
  - Low participation affecting validity

  - Risks of error or bias in crowdsourced validation
  - App development and maintenance challenges
  - Ethical issues around privacy

4.5.5 Risk assessment and contingency plan

- Low participation: A communication strategy will encourage engagement
- Content quality: Clear validation criteria and quality control mechanisms to ensure accuracy
- Technical issues: Robust planning and tools will address development challenges
- Ethical concerns: Ethical guidelines on privacy and consent

References

[1] Adeyinka Adefolarin and Gershim Asiki. Content validation of educational materials on maternal depression in nigeria. *BMC Pregnancy and Childbirth*, 22, 04 2022.

[2] Margaret Rothman, Ari Gnanaskathy, Paul Wicks, and Elektra Papadopoulos. Can we use social media to support content validity of patient-reported outcome instruments in medical product development? *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 18:1–4, 01 2015.

[3] Benjamin Good, Max Nanis, and Andrew Su. Microtask crowdsourcing for disease mention annotation in pubmed abstracts. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 20, 08 2014.

[4] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1:1–29, 09 2017.

[5] Aneeta Sylolypavan, Derek Sleeman, Honghan Wu, and Malcolm Sim. The impact of inconsistent human annotations on ai driven clinical decision making. *NPJ digital medicine*, 6:26, 02 2023.

[6] Jabu Mtsweni, Ernest Ketcha Ngassam, and Legand Burge. A profile-aware microtasking approach for improving task assignment in crowdsourcing services. In *2016 IST-Africa Week Conference*, pages 1–10, 5 2016.

[7] Geicianfran da Silva Lima Roque, Rafael Roque de Souza, José William Araújo do Nascimento, Amadeu Sá de Campos Filho, Sérgio Ricardo de Melo Queiroz, and Isabel Cristina Ramos Vieira Santos. Content validation and usability of a chatbot of guidelines for wound dressing. *International Journal of Medical Informatics*, 151:104473, 7 2021.

[8] Fernando Ressetti Pinheiro Marques Vianna, Alexandre Reis Graeml, and Jurandir Peinado. An aggregate taxonomy for crowdsourcing platforms, their characteristics, and intents. *BAR - Brazilian Administration Review*, 19:e200071, 2022.

[9] Sophia Moganedi, Njabulo Mkhonto, and Jabu Mtsweni. Evaluating the development and implementation of a profile-aware microtasking platform for crowdsourcing services. In *2016 11th International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 335–339, 12 2016.

[10] Tal August, Shamsi Iqbal, Michael Gamon, and Mark Encarnación. Characterizing the mobile microtask writing process. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. Association for Computing Machinery, 2020.

[11] Alex Gartland, Andrew Bate, Jeffery L. Painter, Tim A. Casperson, and Gregory Eugene Powell. Developing crowdsourced training data sets for pharmacovigilance intelligent automation. *Drug Safety*, 44:373–382, 3 2021.

[12] Antonio Muro-Culebras, Adrian Escriche-Escuder, Jaime Martin-Martin, Cristina Roldán-Jiménez, Irene De-Torres, Maria Ruiz-Muñoz, Manuel Gonzalez-Sanchez, Fermin Mayoral-Cleries, Attila Biró, Wen Tang, Borjanka Nikolova, Alfredo Salvatore, and Antonio Ignacio Cuesta-Vargas. Tools for evaluating the content, efficacy, and usability of mobile health apps according to the consensus-based standards for the selection of health measurement instruments: Systematic review. *JMIR mHealth and uHealth*, 9:e15433, 12 2021.

[13] Wenhua Lu, Alexandra Guttentag, Brian Elbel, Kamila Kiszko, Courtney Abrams, and Thomas R Kirchner. Crowdsourcing for food purchase receipt annotation via amazon mechanical turk: A feasibility study. *Journal of Medical Internet Research*, 21:e12047, 4 2019.

[14] Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. Break it down: A comparison of macro- and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4061–4064. Association for Computing Machinery, 2015.

[15] Allard Oelen, Markus Stocker, and Sören Auer. Creating and validating a scholarly knowledge graph using natural language processing and microtask crowdsourcing. *International Journal on Digital Libraries*, 25:273–285, 6 2024.

[16] Marcello N Amorim, Fabio R A Neto, and Celso A S Santos. Achieving complex media annotation through collective wisdom and effort from the crowd. In *2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–5, 6 2018.

[17] Shin'ichi Konomi, Wataru Ohno, Tomoyo Sasao, and Kenta Shoji. A context-aware approach to microtasking in a public transport environment. In *2014 IEEE Fifth International Conference on Communications and Electronics (ICCE)*, pages 498–503, 7 2014.

[18] Dennis Paulino, Diogo Guimarães, António Correia, José Ribeiro, João Barroso, and Hugo Paredes. A model for cognitive personalization of microtask design. *Sensors*, 23:3571, 3 2023.

[19] Michael Weiss. Designing collaborative problem-solving communities. In *Proceedings of the 10th Travelling Conference on Pattern Languages of Programs*. Association for Computing Machinery, 2016.

[20] Mengyao Zhao and Andre van der Hoek. A brief perspective on microtask crowdsourcing workflows for interface design. pages 45–46, 5 2015.

[21] Thomas D LaToza, Arturo Di Lecce, Fabio Ricci, W Ben Towne, and André van der Hoek. Ask the crowd: Scaffolding coordination and knowledge sharing in microtask programming. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 23–27, 10 2015.

[22] Muhammad Imran, Patrick Meier, Carlos Castillo, Andre Lesa, and Manuel Garcia Herranz. Enabling digital health by automatic classification of short messages. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 61–65. Association for Computing Machinery, 2016.

[23] Chul Hyun Park and Erik Johnston. Crowdsourced, voluntary collective action in disasters. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, pages 329–330. Association for Computing Machinery, 2015.

[24] Enrique Estellés-Arolas, Raúl Navarro-Giner, and Fernando González-Ladrón de Guevara. *Crowdsourcing Fundamentals: Definition and Typology*, pages 33–48. Springer International Publishing, 2015.

[25] Thomas D LaToza and André van der Hoek. Crowdsourcing in software engineering: Models, motivations, and challenges. *IEEE Software*, 33:74–80, 1 2016.

[26] Dietmar Winkler, Marta Sabou, Sanja Petrovic, Gisele Carneiro, Marcos Kalinowski, and Stefan Biffl. Improving model inspection with crowdsourcing. In *Proceedings of the 4th International Workshop on CrowdSourcing in Software Engineering*, pages 30–34. IEEE Press, 2017.

[27] Kanta Negishi, Hiroyoshi Ito, Masaki Matsubara, and Atsuyuki Morishima. A skill-based worksharing approach for microtask assignment. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3544–3547, 12 2021.

[28] Kabdo Choi, Hyungyu Shin, Meng Xia, and Juho Kim. Algosolve: Supporting subgoal learning in algorithmic problem-solving with learnersourced microtasks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2022.

Appendix A

Article Analysis

Table A.1: Research purpose description for the studies selected

R	Q	A
[1]	1	Create appropriate educational materials for maternal depression.
	2	Validate the content of educational materials in English and Yoruba on maternal depression.
	3	Experts who assessed the materials' appropriateness, relevance, clarity, and comprehensibility.
	4	Demonstration of the validation process for the English and Yoruba versions of the educational materials.
	5	Only conducted in two local government areas.
	6	Evaluate the impact of the educational materials on knowledge, attitudes, and practices related to maternal depression.
[2]	1	Traditional methods of collecting qualitative data for PRO validity are time-consuming and costly
	2	Explore social media as a tool for collecting data to support PRO validity.
	3	A panel with experts from a pharmaceutical sponsor, Food and Drug Administration (FDA) reviewer, and online data provider.
	4	Social media shows potential for collecting data to support PRO validity.
	5	Unanswered questions include the best social media type for data collection and participant representativeness.
	6	Identify key issues and gather evidence to address them.
[3]	1	Create annotated corpora for biomedical NLP research, a process that is time-consuming and costly.
	2	Investigate crowdsourcing microtasks via the AMT platform to capture disease mentions in PubMed abstracts.

R	Q	A
	3	An experiment using the National Center for Biotechnology Information (NCBI) Disease corpus to compare crowdsourced annotations with expert annotations. Multiple AMT workers annotated the same abstracts, and results were merged via a voting method.
	4	Crowdsourcing via AMT can be valuable for generating annotated corpora in biomedical NLP. The protocol replicated annotations from the NCBI Disease set with an F-measure of 0.872. Quality improved with more workers per task, but gains were minimal beyond 8 workers.
	5	The study focused on disease mentions and may not apply to other biomedical NLP tasks.
	6	Explore crowdsourcing for other biomedical NLP tasks, improve annotation quality, and integrate crowdsourcing with machine learning techniques.
[4]	1	Investigate how design choices of UI elements affect crowdsourcing worker performance.
	2	Study how microtask crowdsourcing work environments influence work quality.
	3	The article discusses a study (Study-I) on UI design choices but lacks details on methodology or results.
	4	The research aims to explore how design choices and work environments impact worker performance and output quality.
	5	The absence of methodology and results for Study-I limits interpretation of findings.
	6	Conduct further studies to examine the impact of specific UI elements and work environments on worker performance and quality.
[5]	1	Understand how inconsistent human annotations affect clinical decision support systems based on artificial intelligence (AI).
	2	The article reviews existing literature and discusses the implications of inconsistent annotations, without proposing a solution.
	4	Inconsistent human annotations present a major challenge for developing and evaluating AI systems in clinical decision support.
	5	The article does not quantify the impact of inconsistent annotations on AI system performance or suggest specific strategies to address it.
	6	Develop methods to manage inconsistent annotations, explore machine learning techniques to improve annotator agreement, and investigate the role of experts in annotation.
[6]	1	Explore the design and evaluation of a profile-aware microtask approach to improve task assignment and quality.
	2	Design and evaluate a profile-aware microtask approach using comparative analysis and crowdsourcing metrics.

R	Q	A
	3	Assess the proposed approach for relevance through comparative analysis and crowdsourcing metrics, focusing on task assignment.
	4	Worker profiling can improve task design, assignment, evaluation, and quality, but stakeholder awareness of effective microtask approaches needs further research.
	6	Future work involves implementing the approach technically, especially on mature crowdsourcing platforms.
[7]	1	Create a chatbot (BOTCURATIVO) to assist non-specialists in wound care with treatment guidelines.
	2	Build the chatbot using Google's DIALOGFLOW, with content validated by stoma care nurses.
	3	Experts validated the chatbot script using the Content Validity Index and Kappa tests.
	4	The prototype showed good usability and satisfactory content validity.
	5	More user testing is needed to improve reliability and usability checks.
	6	Future plans include voice interaction, video guides, and automatic wound detection with a phone camera.
[8]	1	The lack of standardized terminology and classifications for crowdsourcing platforms.
	2	Conduct a systematic review to consolidate existing classifications into a unified system.
	3	Analyze 13 articles discussing crowdsourcing platform classification, reducing categories from 65 to 16.
	4	Propose an aggregated taxonomy to improve understanding of crowdsourcing platforms.
	5	Methodological weaknesses in some articles may impact the reliability of the taxonomy.
	6	Refine the taxonomy through future research to standardize crowdsourcing platform classification.
[9]	1	Improve task design and assignment in microtask platforms by matching tasks to workers' skills.
	2	Develop a platform with profile recognition to specify required skills and monitor assignments.
	3	Describes the platform's design and development but lacks formal evaluation of effectiveness.
	4	Includes features to enhance task design and assignment through skill-based matching.
	5	No formal evaluation limits generalization of the platform's effectiveness.

R	Q	A
	6	Test the platform in real-world settings to assess task quality and worker satisfaction.
[10]	1	Understand how writers use mobile microtasks during document creation and how mobile interfaces fit into their workflow.
	2	Conduct a one-week field study to analyze mobile microtask usage in document creation.
	3	Involve participants who created documents using desktop text editors while integrating mobile phones for editing tasks over the course of a week.
	4	Writers used microtasks for small edits and information addition, tasks well-suited for mobile. Those using microtasks interacted with their documents more efficiently and wrote more than those editing directly on phones.
	5	The study used a controlled prompt and limited writing time, reducing the findings' relevance to less structured writing contexts.
	6	Investigate microtasks in a more natural environment by observing writers over a longer period working on personal projects.
[11]	1	Develop automated solutions for handling the increasing pharmacovigilance workload.
	2	Assess crowdsourcing as a method for creating accurate and efficient training datasets.
	3	Pharmacovigilance experts analyzed social media posts and created a reference dataset. A sample of posts was published on Amazon Turk, where workers answered questions on medical concepts. Accuracy, cost, and efficiency were measured.
	4	Crowdsourcing proved accurate and efficient, with 90% accuracy and 5% of the time required compared to the reference dataset.
	6	Explore broader applications of crowdsourcing, identify factors affecting data quality, and develop methods to improve the process.
[12]	1	Analyze the psychometric quality of mHealth app evaluation tools using the COSMIN guideline. Many apps launch with limited controls, posing risks to users.
	2	Conduct a systematic review to identify mHealth quality tools and validation studies. PubMed and Embase searches covered February to December 2019.
	3	Assess tools against the ten psychometric properties outlined in the COSMIN guideline.
	4	Assess tools against the ten psychometric properties outlined in the COSMIN guideline.
	5	A key limitation was the wide variability in tools and studies, making criteria setting difficult.

R	Q	A
	6	Future work should prioritize creating better tools or improving validation for existing ones, especially MARS.
[13]	1	Record food purchase receipts for nutritional analysis, which is costly and time-consuming manually.
	2	Use crowdsourcing via the AMT platform to annotate receipts.
	3	A consensus task where multiple turkers work on the same assignment, and their agreement is verified.
	4	The study shows that crowdsourcing can annotate food receipts accurately.
	5	The study used a small sample size and focused on limited data extraction.
	6	Explore machine learning to automate the process and improve crowdsourcing quality.
[14]	1	Large tasks can feel overwhelming because they often have a fixed structure.
	2	Study the pros and cons of breaking down macrotasks into microtasks for arithmetic, classification, and transcription.
	3	An experiment with 110 participants compared macrotasks to microtasks, with and without interruptions.
	4	Microtasks took longer but gave better results, a better experience, and handled interruptions well.
	5	The study only covered simple tasks, so it may not apply to complex ones.
	6	Test task decomposition on complex tasks and use cognitive models to study its effects.
[15]	1	Organize and represent academic knowledge in a machine-readable format due to the growing number of publications.
	2	Develop TinyGenius, a methodology for creating and validating an academic knowledge graph with NLP and crowdsourcing..
	3	Evaluate data performance using a subset of the arXiv corpus and assess usability and label quality through user evaluation.
	4	TinyGenius shows promise for validating academic knowledge through NLP and microtasks. The triple store handles data volume well, and the system's usability is strong.
	5	Participant agreement varies for microtasks, and there may be bias in selecting popular articles for evaluation.
	6	Investigate using machine learning to enhance NLP accuracy and explore methods for handling complex instructions requiring domain knowledge.
[16]	1	Improve task allocation and quality in microtask environments, particularly in developing countries.
	2	Design and evaluate a profile-aware microtasking approach to enhance task assignment using worker profiles.

R	Q	A
	3	Use design science research (DSR) methodology, comparative analysis, and crowdsourcing metrics to assess the solution's relevance.
	4	Exploring microworker profiles can improve task design, allocation, evaluation, and task quality.
	5	Stakeholder awareness of effective microtasking approaches needs further research.
	6	Investigate stakeholder awareness of effective microtasking approaches.
[17]	1	Collecting meaningful data requires an integrated platform aligned with human activities.
	2	Explore mobile app designs for recommending microtasks in public spaces like public transport.
	3	Conduct field observations and surveys of public transport users' activities.
	4	Age and occupation influence activity choices during short leisure times, enabling personalization.
	6	Develop a system for a specific domain and study motivation, availability detection, and context-based strategies.
[18]	1	Lack of cognitive personalization in microtasks reduces performance and work quality.
	2	Create a model using cognitive testing and task fingerprinting for personalization.
	3	A case study tested four microtask types with and without personalization.
	4	Results showed improved accuracy and better task adaptation to worker abilities.
	5	Controlled conditions may limit real-world applicability of the findings.
	6	Apply the model to more tasks, integrate machine learning, and study its impact on motivation.
[19]	1	Design effective collaborative problem-solving communities.
	2	Use design patterns like "Starting in a Niche" and "Chunking" to structure and manage them.
	3	Show examples of real communities applying these patterns in practice.
	4	Applying these patterns can make communities more productive.
	6	Test these patterns with case studies or experiments.
[20]	1	Microtask crowdsourcing workflows struggle with complex interface design tasks.
	2	Three experiments explore workflow design for interface design via microtask crowdsourcing.
	3	The experiments investigate task decomposition, flexibility versus consistency, and task reviews.
	4	They aim to identify workflows that address challenges in interface design tasks.
	5	The paper describes the experiments but does not provide results or conclusions.
	6	Results from the experiments will guide the creation of improved workflows.

R	Q	A
[21]	1	Share knowledge among transient workers in microtask programming.
	2	Integrate a Q&A system into the programming environment for code-specific questions.
	3	Observe 20 crowdsourcing workers using the system over 30 hours.
	4	The system helped coordinate work but had issues like delays and duplicate questions.
	5	The study's limited scope may not reflect large-scale microtask programming.
	6	Improve response speed, merge similar questions, and link decisions to code.
[23]	1	Disasters demand coordinated relief efforts.
	2	Information and Communication Technologies (ICT) helps mobilize volunteers, report crises, and map information online.
	3	Examples include the 2011 Japan tsunami and 2010 Haiti earthquake.
	4	ICT supports crowdsourced voluntary action in crises.
	5	Risks include inaccuracies, privacy issues, and volunteer burnout.
	6	Future research must address these risks and challenges.
[24]	1	No validated or standardized tools exist to assess mobile health app quality.
	2	Review and analyze mHealth app quality tools using COSMIN guidelines.
	3	Systematic review of PubMed and Embase studies identifying tools and their validation.
	4	Most tools lack proper psychometric validation for mobile applications.
	5	Heterogeneity of tools and studies makes searches and criteria selection difficult.
	6	Develop and validate tools focusing on transparency, privacy, and data security.
[25]	1	Crowdsourcing is growing in software engineering, requiring insights into its methods and potential.
	2	Examine crowdsourcing models in software development and highlight future opportunities.
	4	Crowdsourcing could transform traditional software development methods.
	6	Study how crowdsourcing models can evolve and influence software development further.
[22]	1	Respond effectively to health-related SMS messages in disaster scenarios.
	2	Develop a crowdsourcing platform to label and categorize SMS messages.
	3	Describes the system architecture but lacks specific evaluations or case studies.
	4	Aims to improve response speed and effectiveness through expert categorization.
	5	Missing details on implementation and evaluation limit assessing scalability.
	6	Test the platform in real-world scenarios, integrate with disaster tools, and use machine learning for automation.

R	Q	A
[26]	1	Traditional software model inspection struggles with large-scale artifacts and lacks proper tool support.
	2	Introduce a Crowdsourcing-Based Software Inspection (CSI) process with tools to detect defects early in development.
	3	A controlled study comparing defect detection effectiveness between CSI and traditional pen-and-paper inspection.
	4	CSI showed promising results in defect detection performance.
	6	Explore CSI team organization, assess its impact on model quality assurance, and address confidentiality challenges.
[27]	1	Increase job opportunities for many workers.
	2	Match task assignments to workers' skills and task difficulty.
	3	A preliminary experiment showed the approach increases the number of qualified workers.
	4	The approach works, but estimating task difficulty is tough.
	5	The experiment may not reflect real-world complexities.
	6	Further research is needed to improve task difficulty estimation and expand the framework to multidimensional skills.
[28]	1	Help students improve subgoal label quality and enhance solution-planning skills.
	2	Develop AlgoSolve, a platform for peer-curated subgoal labels using microtasks.
	3	Conduct a between-subjects study comparing AlgoSolve to a baseline interface with expert feedback.
	4	AlgoSolve's learnersourcing workflow gathered high-quality labels, improving solution technique understanding.
	5	Limited participant number and focus on one problem-solving technique.
	6	Explore applying the approach to other techniques and adding expert explanations for feedback.

R - Reference, Q - Question, A - Answer

Appendix B

Approach Methods

Table B.1: Description of methods

Phase	Description
Research Design	A mixed-methods approach will be used, combining quantitative and qualitative methods. The quantitative aspect evaluates MARINA's usability and effectiveness through usage data analysis and surveys. The qualitative component explores the experiences of participants through interviews and feedback analysis, offering a comprehensive view of the impact of the app. The study is exploratory and descriptive. Initially, it explores the content validation process through a mobile platform. Then, it describes the functionality of the platform and user responses in detail.
Research Context	The study will take place in real-world conditions with MARINA users, including patients, caregivers, and healthcare professionals, both with and without prior experience with diabetes. Participants will use the app on their personal devices, which allows data collection in diverse environments. This ensures the relevance and applicability of the findings.
Population and sampling	The study participants will include healthcare professionals (e.g., nurses and doctors). A balanced and diverse sample will be selected, the size determined based on statistical and practical considerations. Participants must be over 18 years of age, provide informed consent, own a compatible smartphone and have a stable internet connection. Those without a compatible device, stable connection, or informed consent, as well as those with exclusion factors, will not be included.

Phase	**Description**
Data Collection	The tools and instruments for data collection include the following: • App Usage Data: Automatically logs user interactions, such as task completion times and feature usage frequency • Usability Surveys: Employs the validated System Usability Scale (SUS) • Content Evaluation Survey: Assesses the clarity, scientific accuracy, and completeness of MARINA's responses • Interviews/Feedback: Collects detailed insights into user experiences and satisfaction Participants will use MARINA for a defined period, completing predefined tasks. Usage data will be collected anonymously. Surveys will be administered at specific intervals, and participants will provide feedback through interviews or comments. Data collection tools will be validated. For content evaluation, a questionnaire adapted from the Suitability Assessment of Materials (SAM) will be reviewed by experts. The SUS is already validated for usability studies.
Variables and Measurements	Key Variables: • Usability: Measured through SUS and user surveys • Effectiveness: Assessed through task completion times, feature usage frequency, and satisfaction survey results • Content Quality: Evaluated via the content survey and user feedback Measurements: • Quantitative methods (e.g., scales, counts) • Qualitative methods (e.g., content analysis, feeedback)
Analysis Methods	Quantitative analysis will involve descriptive statistics (e.g., means, standard deviations) and inferential statistics (e.g., t-tests, ANOVA) using SPSS software. Qualitative analysis will focus on identifying recurring themes and patterns through content analysis. Finally, quantitative and qualitative data will be integrated to provide a comprehensive understanding.

Phase	Description
Ethical Considerations	Participants will be informed of the study's goals, risks, and benefits and will provide voluntary, informed consent. Data will be collected and processed anonymously to protect participants' privacy, securely stored, and used only for research purposes. Participants will have the freedom to withdraw consent at any time without consequences.
Limitations and Mitigation Plan	The sample may not represent all users of mobile health apps, limiting generalizability. Data collection tools might not capture all relevant aspects of content validation. Results may also be influenced by specific sample or environmental factors. To address these limitations, a diverse sample will be used, and results will be interpreted cautiously. Complementary instruments will be employed, and results will be integrated from multiple perspectives. Additionally, a detailed context analysis will be conducted, and the implications of findings for various settings will be discussed.