Baltic J. Modern Computing, Vol. 14 (2026), No. 3, 538-555
https://doi.org/10.22364/bjmc.2026.14.3.03

Introducing a Standard Process Framework to Translate Big Data into Actionable Population Insight Products for Healthcare

Wikje E. BERENDS-HOEKSTRA[1,2], Stella WEILAND[1,2], Sandra BROUWER[3], Adriëlla VAN DER VEEN[1], Sereh M.J. SIMONS[4], Bas BOLMER[5], Frederike JÖRG[6], Lilian L. PETERS[1,2]

1 University of Groningen, University Medical Center Groningen, Department of Primary and Long-term Care, Cure and Care in the Community Context, Groningen, the Netherlands
2 Amsterdam University Medical Centre location Vrije Universiteit Amsterdam, Midwifery Academy Amsterdam Groningen, Department of Midwifery Science, Amsterdam, the Netherlands
3 University of Groningen, Department of Health Sciences, Community and Occupational Health, University Medical Center Groningen, Groningen, The Netherlands
4 Datapoort, University Medical Center Groningen, Groningen, the Netherlands
5 Lifelines, Groningen, the Netherlands
6 University of Groningen, University Medical Center Groningen, University Center Psychiatry, Groningen, the Netherlands

w.e.berends-hoekstra@umcg.nl, s.weiland@umcg.nl, sandra.brouwer@umcg.nl, a.l.van.der.veen@umcg.nl, s.m.j.simons@umcg.nl, b.r.j.bolmer@lifelines.nl, f.jorg@umcg.nl, l.l.peters@umcg.nl

ORCID 0009-0006-6686-1809, ORCID 0000-0003-3708-6672, ORCID 0000-0002-3819-4360, ORCID 0009-0001-1880-3851, ORCID 0009-0004-8438-0617, ORCID 0000-0002-9965-0702, ORCID 0000-0003-1720-1231, ORCID 0000-0003-2342-0799

Abstract. Big data analytics (BDA) of healthcare data holds potential for developing and implementing Population Insight Products (PIPs) that could potentially enhance healthcare quality, efficiency, and cost-effectiveness. A stakeholder-driven framework for their adoption remains however absent. We developed CRISP-PIP, a standardised, participatory framework extending the CRoss-Industry Standard Process for Data Mining for Medicine (CRISP-MED-DM) with FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, using a four-stage mixed-methods design. A survey examined perceived relevance of PIPs; a conceptual framework was modelled; a research group discussion explored usability and adoption factors; and thematic analysis informed final framework refinement. Five themes emerged: the importance of a structured framework, basic resource requirements, stakeholder communication and cooperation, role and responsibility definition, and user-friendliness and transparency. CRISP-PIP offers a step-by-step, participatory approach to developing and implementing sustainable PIPs, that could bridge the gap between data scientists, healthcare providers, policymakers, and researchers through FAIR and open science principles.

Keywords: Big data analytics, open science, FAIR data principles, population insights, healthcare policy, healthcare research

1. Introduction

The volume of healthcare data is increasing exponentially, accounting for up to one-third of all the electronic data globally (Abedjan et al., 2019). 'Big data' is fast becoming a key instrument for improving healthcare processes and outcomes (Dash et al., 2019; Al Teneiji et al., 2024). Big data analytics (BDA) has been shown to be valuable for collecting and analysing data characterised by volume, variety, velocity, and value (Al Teneiji et al., 2024). It leverages cutting-edge technologies to transform healthcare by generating descriptive, diagnostic, prescriptive and predictive population insights (Nilashi et al., 2023; Al Teneiji et al., 2024).

These insights can support healthcare providers, policy makers, and researchers identifying disease trends and enable improvements in healthcare quality and efficiency (Abedjan et al., 2019; Subrahmanya et al., 2022; Al Teneiji et al., 2024;, Manski-Nankervis et al., 2024). Additionally, with these insights healthcare costs can be reduced by optimising resource allocation and improving the effectiveness and efficacy of healthcare delivery (Hong et al., 2018). Finally, BDA can translate complex healthcare data into actionable, data-driven tools called population insight products (PIPs), including real-time dashboards, predictive models, or automated reports. A healthcare system based on such PIPs could facilitate the shift towards informed and shared decision-making in healthcare between multidisciplinary healthcare providers and patients. This aligns with the growing emphasis on patient-centred, personalised, and participatory healthcare (Coughlin et al., 2018; Subrahmanya et al., 2022).

Despite these promising applications, the healthcare sector has not fully adopted BDA due to multiple challenges (Abedjan et al., 2019; Al Teneiji et al., 2024). Adoption can be influenced by technological (e.g., complexity, compatibility), environmental (e.g., government policies and legislation), organisational (e.g., management or financial support), and user-related domains (e.g., intention to adopt or use BDA) (Al Teneiji et al., 2024). Within these domains, multiple barriers hinder adoption, including limited access to data, fragmented healthcare infrastructures, data heterogeneity, governance issues (e.g., consent for secondary data use, data access costs and privacy protection), data quality, technical barriers to data use, capacity, and lack of BDA knowledge (Canaway et al., 2022; Furstenau et al., 2023; Manski-Nankervis et al., 2024).

To fully leverage the potential of BDA for the development and implementation of PIPs in healthcare, these challenges must be addressed (Manski-Nankervis et al., 2024). BDA frameworks have been developed to provide a structured approach to monitor the data processing phases and facilitate the adoption of BDA by systematically addressing these challenges (Chapman, 2000). The CRoss-Industry Standard Process for Data Mining (CRISP-DM) methodology is a widely used example of such frameworks (Chapman, 2000), with CRISP-MED-DM being an extension for the healthcare domain (Niaksu, 2015). This framework consists of six cyclical and iterative data processing phases, from problem understanding, data understanding and data preparation to modelling, evaluation, and deployment (Chapman, 2000; Niaksu, 2015). Yet, the perspectives and experiences of potential stakeholders regarding CRISP-MED-DM have not been sufficiently assessed. Furthermore, the existing BDA frameworks do not fully

address the aforementioned challenges to effectively adopt BDA in healthcare. Specifically, CRISP-MED-DM does not fully integrate key open science principles for data sharing and collaboration, and it lacks identification of stakeholders (e.g., healthcare providers, policy makers, researchers or data scientists) to facilitate effective collaboration. Therefore, a framework is needed that addresses these limitations and aligns with the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles (Wilkinson et al., 2016).

To address this gap, we aim to develop a participatory and integrative CRISP-PIP (Cross-Industry Standard Process for Population Insight Products) framework tailored to the development and implementation of PIPs as an extension of CRISP-MED-DM, enhanced by FAIR data principles. By providing a standardised, step-by-step framework developed with potential stakeholders, we seek to overcome the challenges of BDA in healthcare to support the structured development and implementation of actionable PIPs. In doing so, we aim to support the effective sharing, integration, and (re)use of healthcare data while fostering collaboration among diverse stakeholders in healthcare.

2. Related works

2.1. CRISP-DM extensions

Since its introduction, CRISP-DM has been the most widely adopted BDA framework across various domains (Schröer et al., 2021). It has been extended for the medical domain with CRISP-MED-DM (Niaksu, 2015). More recently it has been adapted for machine learning (ML) applications with the CRISP-ML(Q) framework (Studer et al,. 2021), which extends CRISP-DM with ML-specific quality assurance tasks (Studer et al., 2021).

2.2. FAIR principles in healthcare

Besides these BDA frameworks, recent literature has focused on operationalising FAIR principles in practice to enable the sharing and reuse of (research) data (Wilkinson et al., 2016; Sinaci et al., 2020; Inau et al., 2023). For the healthcare domain specifically, Sinaci et al. (2020) developed a FAIRification workflow addressing the specific interoperability (e.g., using standardised terminologies) and governance requirements (e.g., data deidentification or pseudonymisation) of healthcare and research data.

3. Methods

3.1. Design

This mixed-methods study consisted of four stages (Figure 1). In the first stage, an online survey was conducted to gain insight into experiences and perspectives of potential stakeholders on BDA for the development, implementation, and future utilisation of PIPs. Potential stakeholders were defined as all relevant actors in healthcare who affect or are affected by the development and implementation of PIPs.

Through this survey, we aimed to assess the perceived relevance and willingness for the development and implementation of PIPs.

In the second stage, we developed a conceptual, step-by-step framework for the development and implementation of PIPs, in which we integrated FAIR data principles in CRISP-MED-DM (Chapman, 2000; Niaksu, 2015; Wilkinson et al., 2016). Specific strategies for applying FAIR principles were identified for each phase of the framework by a subset of the author team (WB, AV, LP).

In the third stage, we evaluated the conceptual framework and explored perspectives on its usability by conducting a semi-structured research group discussion among potential stakeholders. Through this discussion, we furthermore assessed adoption factors influencing the development and implementation of PIPs by BDA in healthcare.

In the fourth stage, the insights from Stage 3 were used to refine the conceptual framework. This resulted in a final, participatory, and integrative framework tailored to the development and implementation of PIPs in healthcare.

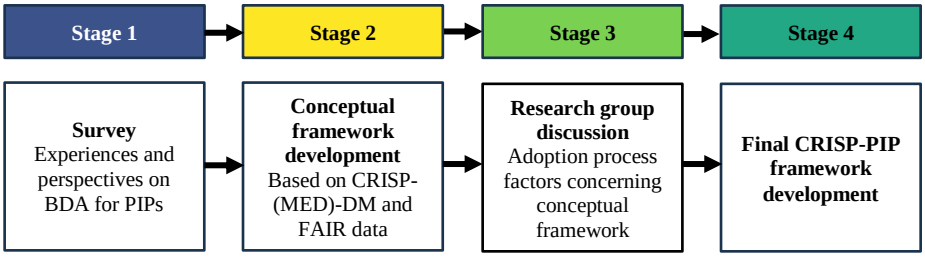

Figure. 1. Overview of the four-stage study design illustrating the steps leading to the development of the final CRISP-PIP framework for the development and implementation of PIPs in healthcare. The mixed-methods design integrates input from potential stakeholders through a survey in stage 1, the development of a conceptual framework in stage 2, a research group discussion to evaluate the conceptual framework in stage 3, and subsequent the refinement into the final participatory, integrative CRISP-PIP framework (phase 4).
BDA, Big Data Analytics; CRISP-MED-DM, CRoss-Industry Standard Process for Data Mining; FAIR, Findable, Accessible, Interoperable, and Reusable; PIP, Population Insight Product.

The study followed the CROSS (Checklist for Reporting Of Survey Studies) for reporting surveys (Sharma et al., 2021), and the COREQ (COnsolidated criteria for REporting Qualitative research) guidelines for reporting focus groups (Tong et al., 2007), where applicable.

3.2. Study population and setting

Given the exploratory, proof-of-concept nature of this study, recruitment of participants was restricted, through convenience sampling, to research- and healthcare-affiliated stakeholders. Eligible participants were invited to Stage 1 and included employees from I) the Department of Primary and Long-term Care (University Medical Center Groningen, UMCG), II) the Department of Health Sciences (UMCG), III) the Rob Giel Research Center (UMCG), and IV) Lifelines (a multigenerational prospective population-based cohort and biobank). Some of these employees combine research

activities with clinical practice. Invitations to participate in the survey were sent by email to 265 eligible participants. Prospective participants were provided with details about the survey's purpose and content (Supplementary Material II).

For Stage 3, all employees of the research section of the Department of Primary and Long-term Care (UMCG) were invited to participate during a research meeting. This department has researchers from the fields of general practice, elderly care, and midwifery science, some of whom combine their research activities with clinical practice. Participants were approached and informed about Stage 3 participation by email, after which informed consent was collected from each participant. Participants were informed of the purpose, the voluntary nature of their participation, their right to withdraw, and the recording. All participants agreed to participate.

3.3. Data collection

The survey was conducted via Microsoft Forms. The survey consisted of three sections: demographics, perspectives on the relevance, future use, and interpretation of PIPs, and experiences regarding the development of PIPs. Questions were partly based on the System Usability Scale (Brooke, 2013). Consensus regarding the survey questions was reached between WB, LP and AV (Supplementary Material I). The survey was conducted between 23 July 2024 and 2 December 2024, with two reminder emails sent at regular intervals to encourage participation.

The research group discussion took place on 19 September 2024. Four members of the study group were present; one moderator (WB) and three assistant moderators (LP, AV and SS). The moderator introduced each step of the conceptual framework and the FAIR data principles (Supplementary Material II), asked the questions and led the discussion. One of the assistant moderators (AV) was responsible for taking notes, audio recording, and ensuring that all relevant questions were posed. The aim of Stage 3 was to evaluate the conceptual framework. The main question asked was: "What is your opinion on the conceptual step-by-step framework, and what would you like to add or remove to the framework?". Sub-questions were posed to provide deeper insights (Supplementary Material III). At the conclusion of the discussion, participants were asked whether certain topics should be discussed further or were missing. The moment the questions no longer provided any new information, the discussion was ended. Subsequently, the audio recordings were transcribed and pseudonymised.

3.4. Data analysis and reporting

Responses to the survey were analysed and described in absolute numbers and percentages. The survey design disallowed participants to leave questions unanswered. Duplicate responses from the same participants were removed. The quantitative analyses of the online survey data were performed using R version 4.4.1.

Transcription data from Stage 3 were managed and analysed using Excel. An inductive thematic analysis was conducted to find the main themes, following a six-step process (Naeem et al., 2023). WB and AV transcribed the data, familiarised with the data, and selected quotations, independently selected keywords, coded the data, and developed themes, which were discussed with a third researcher (LP) to reach consensus. Subsequently, WB, AV, and LP conceptualised by interpretating the keywords, codes, and themes. Lastly, a conceptual model was developed by WB, AV, and LP.

4. Results

4.1. Stage 1: Survey

The survey was filled in by 39 women and 19 men between 23 July 2024 and 2 December 2024, with a response rate of 21.9%. Table 1 shows the characteristics of the participants.

Table 1. Characteristics of survey participants (n = 58)

	n (%)
Sex	
Female	39 (67.2)
Male	19 (32.8)
Age (years)	
18 –25	4 (6.9)
26 – 35	20 (34.5)
36 – 45	14 (24.1)
46 – 55	11 (19.0)
> 55	9 (15.5)
Position	
Postdoctoral[a]	14 (24.1)
PhD candidate[b]	12 (20.7)
Associate professor	6 (10.3)
Full professor	6 (10.3)
Research associate	6 (10.3)
Assistant professor	4 (6.9)
Research coordinator	3 (5.2)
Other[c]	7 (12.1)

[a] Types of postdoctoral functions that were mentioned were: medical epidemiologist-microbiologist, general practitioner, research coordinator, senior methodological advisor, and (programme) advisor
[b] PhD, Doctor of Philosophy; PhD candidates also included doctors in training as a PhD
[c] Other functions were: data manager, project leader, programme manager, senior service desk assistant, team leader data

Most participants (52%) agreed that PIPs are relevant for conducting their work, Figure 2. A strong majority (67%) wanted to use or keep using PIPs in the future. Of the studied participants, 45% had experience in interpreting PIPs.

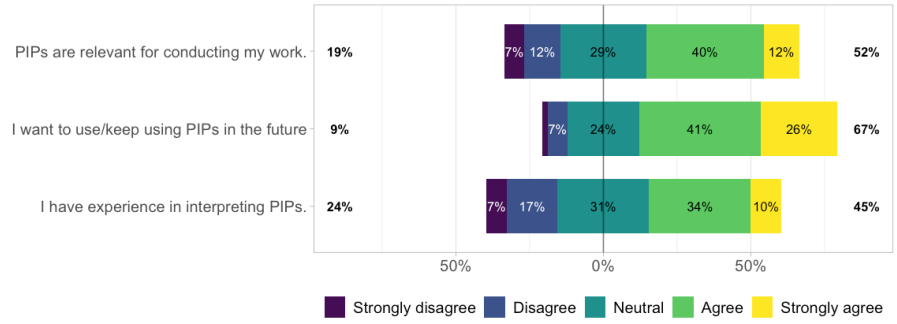

Figure. 2. Participants perspectives on the relevance, future use, and interpretation of PIPs (n = 58). Responses are shown on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The percentages reflect levels of agreement with each statement.

Of all participants, 27.6% previously contributed to the development of some kind of PIP. Results for each example type of PIP are shown in Figure 3.

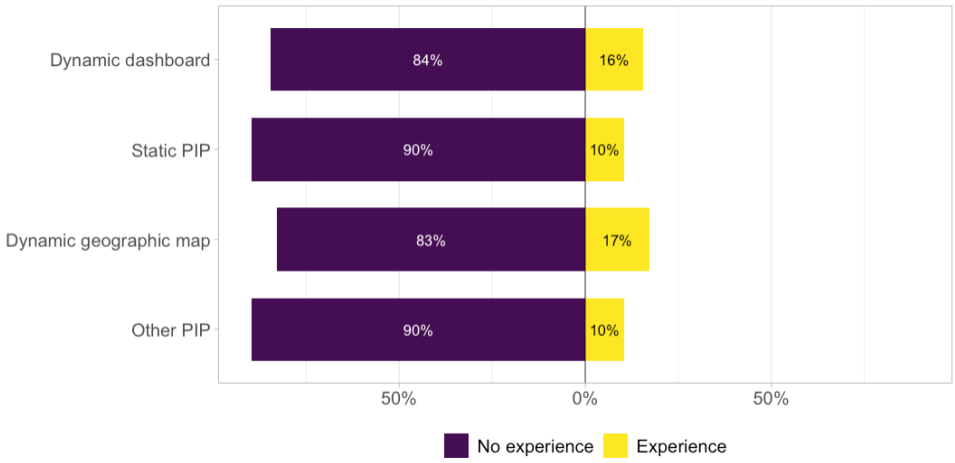

Figure. 3. Experiences regarding the development of different types of PIPs (n = 58). The percentage of participants reporting experience versus no experience for each example type of PIP is shown.

4.2. Stage 2: Conceptual framework development

The conceptual framework was designed based on the CRISP-MED-DM and the FAIR data principles (Chapman, 2000; Niaksu, 2015) and was extended to focus on the

development and implementation of PIPs. This conceptual framework is visualised in Supplementary Material II and was presented to the participants during Stage 3.

The conceptual framework was split into six phases. Phase 1 of the cyclical, iterative framework introduces the establishment of objectives and context of the PIP, including the identification of stakeholders, ethical, legal, and financial conditions for the development and implementation of the PIP. Phase 2 focusses on data understanding, emphasising the identification, collection, and exploration of relevant data. In Phase 3, data is prepared and organised for analysis. Phase 4 covers the development of the data science method, including selecting, designing, and testing the method. In Phase 5, the PIP is evaluated, assessing its effectiveness, efficiency, and stakeholder satisfaction. Phase 6 outlines the planning of the implementation, monitoring and reporting of the PIP within existing systems.

We incorporated the FAIR data principles, such as open access documentation in Phase 1, Digital Object Identifiers (DOIs) and registration of datasets with standardised metadata in Phases 2 and 3. Transparent reporting of model development was added in Phase 4, and evaluation in Phase 5. In Phase 6 detailed implementation documentation was incorporated to support reproducibility and reuse.

4.3. Stage 3: Research group discussion

The group included nine men and 12 women (57%). The majority were aged between 26 and 45 years (67%). Participants held the function of full professor (n=2), associate professor (n=2), assistant professor (n=3), postdoctoral (n=3), PhD candidate (n=6), research associate (n=1), or data specialist (n=4). The median work experience was 11 years (IQR: 3.0-18.0). The duration of the total session was 60 minutes.

Based on the thematic analysis, the five main themes that we identified were 1) the importance of the framework, 2) the basic requirements for developing a PIP, 3) the importance of communication and cooperation between stakeholders, 4) the importance of defining the roles and responsibilities of stakeholders, and 5) the importance of user-friendliness and transparency to stakeholders.

1. The importance of the framework

Participants indicated that that the framework is recognisable in the current practice and makes clear which phases are essential for data processing. They stated that this is important for awareness of the data processing life cycle among insight-seekers without in-depth knowledge of data analytics.

"This roadmap clarifies which phases are important, which is very nice. So even a policy maker who already has data and thinks we can make a graph out of it, he probably doesn't understand data and how this is going to come to a graph. So, who can see all these steps to know what is needed." (ID21)

Additionally, it was stated that the framework is not only relevant for analysing multiple linked datasets, but also for analysing a single dataset.

"Especially, if you want to link multiple data sources these steps might become even more important, as they introduce an additional step for those links. But you have to do the same steps if you only have a single database." (ID21)

2. The basic requirements for developing a PIP

Participants indicated that the basic conditions when developing a PIP – e.g., availability and accessibility of data, human and financial resources, ethical and legal aspects, and time – must be met. It is important to assess not only what data is needed, but also whether it is available.

"Of course, you are not going to ask a scientific question without already knowing that you can find that data somewhere. You are not going to write grant applications without knowing that the data is there." (ID 3)

"And that has consequences for staffing and finances. Well, finances are always the biggest problem in my opinion." (ID8)

3. The importance of communication and cooperation between stakeholders

Participants commented that the model is recognisable for data analysis, with a clear iterative aspect that is in line with practice, in which data managers work closely with researchers. Participants expressed that although streamlining recurring processes is often complex, this model can be used effectively through collaboration and continuous optimisation. Participants emphasised that communication between data scientists and other stakeholders should be an integral part of the framework to ensure alignment in data preparation and interpretation.

"You see the iterative aspect of the model reflected in practice, because data managers and researchers have a lot to do with this, going back and forth. How do you do this efficiently? Streamlining this is sometimes quite complicated. There is a lot of functions and collaboration behind it." (ID6)

"What I miss in the framework is the communication between the data scientist and the stakeholder with the question. Often there is an application, and the data scientist must figure out what the exact question is. I think it would be useful to create a demo before you prepare the data. Then we could have a chat with the stakeholders to see if that's what they mean. Then you avoid going back and forth about changing formats." (ID4)

4. The importance of defining the roles and responsibilities of stakeholders

Participants emphasised that, in addition to stakeholder identification, it is also important to clearly define the stakeholder's roles and responsibilities within the framework. In practice, participants have often experienced gaps in professional knowledge – such as those between specialists in prediction models and healthcare providers. This can complicate collaboration, requiring both parties to invest time and effort to understand each other's processes and perspectives.

"So here you do have a gap to bridge. This can be complex because you must understand from each other what is happening. In other words, the general practitioner must understand how it is on the data side and the data scientist must understand how it is in the general practitioner practice. That's tricky and that takes time." (ID22)

Participants indicated that, ideally, the involved data specialists completely unburdened other stakeholders (e.g., healthcare providers and researchers) of certain data analytic responsibilities (e.g., data cleaning). In practice, however, this is often different.

"In our department, it's almost the other way around, that the data scientist is very good at advising on data as experts, but a lot of work still lies with the executive researcher [example stakeholder]. And that has consequences for staffing and finances." (ID8)

5. The importance of user-friendliness and transparency to stakeholders

Participants indicated that the framework should explicitly incorporate the provision of a manual for the stakeholders alongside the PIP, to guide stakeholders in understanding and interpreting the end-product.

"Midwives [example stakeholders] sometimes have a question, but the registry data [i.e., EHR data] is so complicated. It would be good to have a manual." (ID2)

Furthermore, participants stated that it is essential to be transparent about the source and description of the data, in the context of disinformation. Participants also indicated that data quality should be consistently monitored throughout each phase of the framework, from data collection to implementation and evaluation.

"It is also good nowadays to describe your data. In the context of fake news. Include the stakeholder in what kind of data it is. Where did it come from? So, what is the source?" (ID2)

"I would like to be a bit more explicit about the quality of the data. One piece of data is of better quality than another piece of data. Not only is quality important in the beginning, but also when you are going to implement it or evaluate it, you must make it clear what you are looking at. How valid is it? You must take quality through the whole cycle." (ID17)

4.4. Stage 4: Final CRISP-PIP framework development

Based on the results of Stage 3, we adjusted the data processing phases of the conceptual framework and refined the FAIR data principles per phase. The resulting final CRISP-PIP framework is visualised in Figure 4.

In Phase 1 of the CRISP-PIP framework, we added the stakeholder engagement and the definition of roles and responsibilities to enhance collaboration among stakeholders. To ensure transparency, we added data quality checks and reporting throughout all phases. Moreover, in Phase 6, we added the development of a manual of the PIP for stakeholders. A global description of the steps within each phase of the CRISP-PIP framework is shown in Table 2. For a detailed description we refer to Supplementary Material IV.

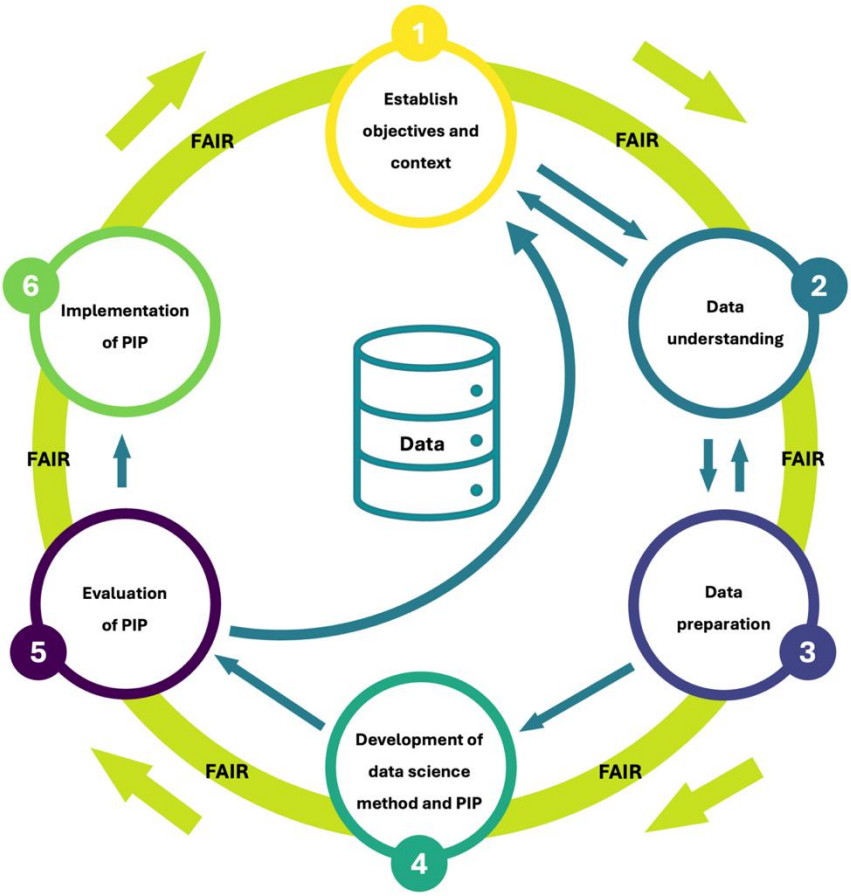

Figure. 4. Overview of the six iterative phases of the CRISP-PIP framework for the development and implementation of PIPs, which were based on the CRISP-MED-DM and the FAIR principles (Chapman 2000, Niaksu 2015, Wilkinson, Dumontier et al. 2016). Phase 1 introduces the establishment of the objectives and context of the PIP, including stakeholder identification and engagement, and ethical, legal, and financial conditions. Phase 2 focusses on data understanding, emphasising the identification, collection, and exploration of relevant data, In Phase 3, data is prepared and organised for analysis. Phase 4 covers the development of the data science method and the PIP, including its selection, design, and testing. In Phase 5, the PIP is evaluated, assessing its effectiveness, efficiency, and stakeholder satisfaction. In Phase 6 the implementation, monitoring and reporting of the PIP within existing systems are planned.
CRISP-MED-DM, Cross-Industry Standard Process for Data Mining extension for the medical domain; CRISP-PIP, CRoss-Industry Standard Process for Population Insight Products; FAIR, Findable, Accessible, Interoperable, and Reusable; PIP, Population Insight Product.

Table 2. Global description of each phase of the CRISP-PIP framework, the application of FAIR data principles, and the stakeholders involved in each phase of the framework. Suggestions for practical tools, drawn from previous literature, are included for some steps where relevant

Steps		FAIR data principles	Stakeholders
	PHASE 1. ESTABLISH OBJECTIVES AND CONTEXT		
1.1	Identify and engage stakeholders of the PIP	- Ensure that the project plan objectives and documentation are findable - Publish documentation on open-access platforms (e.g. GitHub [2]) - Use standardised coding and classifications, such as the International Classification for Primary Care (ICPC) or the International Classification of Diseases (ICD) in the documentation	In this phase, all relevant stakeholders for the project are identified and actively involved in the process.
1.2	Define roles and responsibilities of stakeholders (e.g., using a Responsible, Accountable, Consulted, Informed – RACI – matrix [1])		
1.3	Establish the objectives of the PIP		
1.4	Establish the context (situation) in which the PIP will be implemented		
1.5	Determine data mining objectives (including quality metrics and validation criteria)		
1.6	Determine the type and shape of the PIP		
1.7	Plan activities (including data quality assessments)		
	PHASE 2. DATA UNDERSTANDING		
2.1	Determine what data is needed	- Register datasets with clear metadata, using a standardised method - Use a unique Digital Object Identifier (DOI) - Make the raw data accessible via repositories with clearly defined access rights	The stakeholders involved in this phase are mainly the relevant data scientists, data analysts, but also the insight-seekers and other members of the project team who have knowledge about the data (e.g., epidemiologists).
2.2	Determine minimal dataset (technical)		
2.3	Prepare data collection		
2.4	Collect initial data		
2.5	Explore data and assess potential biases or missing values		
2.6	Determine data quality through predefined metrics and documentation of limitations		

	PHASE 3. DATA PREPARATION		
3.1 3.2 3.3 3.4 3.5 3.6 3.7	Prepare data Extract structured and unstructured data from datasets Select data Clean data while documenting modifications Construct data (ensuring consistency) (e.g., using harmonisation steps [3]) Integrate data (including quality assessments) Format data with standardised procedures	- Same as in phase 2 - Make the script publicly available on, for example Github	The stakeholders involved in this phase are mainly the data analysts and data scientist, but also the insight seekers and other project members. In general, this phase requires the most time in the entire framework.
	PHASE 4. DEVELOPMENT OF DATA SCIENCE METHOD AND PIP		
4.1. 4.2. 4.3. 4.4. 4.5.	Determine data science method, considering data reliability and representativeness Generate test design (including data quality and validation steps) Develop the PIP based on the data science method Test the model (including quality assurance protocols) Prepare the PIP for interoperability	- Document and publish the methods and techniques used, including model architecture, parameters and algorithms used - Document the model training and model performance, so that others can reuse the model in similar contexts	The stakeholders involved in this phase are the data analysts, data scientists and members of the project team who have knowledge about the data (e.g., epidemiologists). However, even in this phase, it is important to involve insight-seekers and other project members in the progress.
	PHASE 5. EVALUATION OF PIP		
5.1 5.2 5.3	Evaluate results (including data quality, accuracy, and validity) Review the process (including review of quality) Determine next steps	- Publish evaluation results in a report/article, if possible, with a DOI and publicly available - Use standardised and validated evaluation methods - Publish evaluation methods scripts	The stakeholders involved in this phase are basically all stakeholders; from insight-seekers, developers to policy makers, identified in the first phase.

PHASE 6. IMPLEMENTATION OF PIP			
6.1	Develop a manual for the PIP	- Make documents and implemented PIPs findable (e.g., by using a DOI) - Describe how others can access it - Document the implementation, such as the implementation methods used, environment requirements and licence conditions, so that the PIP can be applied in other environments	The stakeholders involved in this phase are the insight-seekers, data scientists, data analysts and data managers, and others involved in the project.
6.2	Plan implementation of PIP (e.g., using the Reach, Effectiveness, Adoption, Implementation and Maintenance - RE-AIM - framework [4])		
6.3	Plan monitoring and maintenance (e.g., using RE-AIM framework [4])		
6.4	Develop end report (including quality findings and limitations)		
6.5	Review the project: lessons learned		

DOI, Digital Object Identifier; FAIR, Findability, Accessibility, Interoperability, Reusability; ICD, International Classification of Diseases; ICPC, International Classification of Primary Care; PIP, Population Insight Product; RACI, Responsible, Accountable, Consulted, Informed; RE-AIM, Reach, Efficacy, Adoption, Implementation, and Maintenance framework

5. Discussion

In this study, we propose a participatory and integrative framework intended to support the adoption of big data analytics (BDA) in the development and implementation of population insight products (PIPs) in healthcare. The standardised, step-by step framework may offer a structured approach for developing PIPs that could potentially enhance healthcare quality, efficiency, and cost-effectiveness. By integrating the experiences and insights of potential stakeholders, an existing framework (CRISP-MED-DM) and the FAIR data principles, our framework proposes a holistic approach to mitigating the barriers to adopting BDA for the development and implementation of PIPs and advancing open science.

Our survey findings among healthcare researchers showed that only 24% has experience in developing PIPs, while most participants indicated that they would like to use PIPs in the future (67%) and found PIPs relevant to their work (52%). According to the Technology Acceptance Model, behavioural intention to adopt a technology, such as those in BDA, depends on an individual's attitude towards using it (Davis and Davis, 1989; Rahimi et al., 2018). This attitude is influenced by the perceived ease of use and perceived usefulness, which are therefore key drivers of technology adoption (Davis and Davis, 1989; Rahimi et al., 2018). The strong interest in adopting PIPs suggests that researchers recognise their value and perceive them as useful, urging broader implementation. However, a clear gap remains between those interested in using PIPs and those experienced in developing them. To bridge this gap, we integrated insights from our research group discussion to develop the CRISP-PIP framework into a user-centred, participatory framework for developing and implementing PIPs.

The research group discussion revealed five themes reflecting the perspectives of potential stakeholders on PIP development and implementation using BDA. First, our

findings align with previous research showing that standardised frameworks support BDA adoption by enhancing stakeholders' understanding and awareness of data processing phases (Al Teneiji et al., 2024).

Second, participants highlighted basic requirements for PIP development, such as human and financial resources, aligning with prior research on regulatory barriers, and the extensive time and effort required for successful BDA (Hong et al., 2018 ; Al Teneiji et al., 2024). While governance issues remain an ongoing debate, the European Health Data Space introduces binding regulation for the (re-)use and exchange of electronic health data across the European Union (European Union, 2024).

The third and fourth themes reflect prior research showing that clear roles, mutual understanding and interdisciplinary collaboration between stakeholders (e.g., data specialists, policy makers, healthcare providers, and researchers) are necessary to address knowledge gaps and reduce fragmentation (Hong et al., 2018, Al Teneiji et al., 2024). To facilitate this, we recommend identifying stakeholders deserving or requiring attention based on the stakeholder theory (Mitchell et al., 1997). To define tasks and responsibilities, we recommend using a RACI (Responsible, Accountable, Consulted, Informed) matrix (Brower et al., 2020). These stakeholders-related themes also reflected the iterative nature of collaboration. This aligns with research indicating that healthcare process frameworks – though often conceptualised as stepwise and linear – are, in practice, iterative in the translation of research into practice (Nilsen, 2015).

Fifth, our findings correspond with the aforementioned Technology Acceptance Model, which emphasises perceived ease of use, besides the perceived usefulness, as an important adoption factor of technologies in healthcare (Rahimi et al., 2018).

In conclusion, these five themes collectively underscore the multifaceted challenges and requirements for successful BDA adoption for the development and implementation of PIPs. Previous research showed the enormous potential of BDA to enhance healthcare quality, efficiency, effectiveness and efficacy, as well as to reduce healthcare costs, and support informed, shared decision-making (Coughlin et al., 2018; Hong et al., 2018; Abedjan et al.. 2019; Subrahmanya et al., 2022; Al Teneiji et al., 2024). However, an integrated approach with interconnection of healthcare domains was lacking (Al Teneiji et al.. 2024).

Existing process frameworks address this gap only partially. CRISP-MED-DM, CRISP-ML(Q), and existing FAIRification workflows are technically oriented and were not designed for stakeholder participation (Niaksu, 2015; Studer et al., 2021; Sinaci et al., 2020). CRISP-ML(Q)'s quality assurance tasks target model risk and performance, without addressing healthcare-specific data governance or FAIR principles. FAIRification workflows do not cover problem understanding, modelling, evaluation, or deployment. Consistently, in a recent review it was found that FAIRification is often treated as an endpoint rather than a step toward a concrete, minimum viable product (Inau,et al., 2023). Additionally, a lack of structured approaches to manage stakeholder resistance or motivation was identified (Inau et al., 2023). Neither existing CRISP-DM extensions nor FAIR implementation approaches thus lead to a defined, actionable end product, nor do they address the organisational and collaborative conditions required for adoption. Our standardised CRISP-PIP framework may address this by embedding FAIR data principles, alongside the experiences and perspectives of potential stakeholders identified in this study, within a structured, participatory lifecycle that culminates in a defined, actionable product, reducing the need to redevelop existing methods and enhancing sustainable impact.

Despite this strength to now have an integrated approach, there are some limitations to our study. The framework has not yet been evaluated in real-world practice, limiting insights into its feasibility and effectiveness. We recommend empirically validating the framework by developing a PIP in future research. Additionally, the low response rate on the survey may have introduced response bias, while the convenient recruitment of participants for the survey and the research group discussion may have limited the diversity of perspectives. Additionally, as we primarily included participants from healthcare research environments, the perspectives of other potential stakeholders (e.g., policy makers and healthcare providers) were underrepresented. Therefore, we recommended empirically testing the framework on a larger scale across different healthcare contexts, incorporating a more diverse range of stakeholders to examine broader adoption process factors and further improve its valorisation (Al Teneiji et al., 2024).

In summary, we introduce the CRISP-PIP participatory framework, offering a step-by-step approach for developing and implementing PIPs using BDA in the increasingly data-driven healthcare environment. We have shown that the majority of our studied participants expressed interest to use or keep using PIPs in the future. The CRISP-PIP framework is a structured method that has the potential to bridge the gap between data scientists, healthcare providers, policy makers and researchers, fostering interdisciplinary collaboration and improving the translation of healthcare data into actionable insights.

Abbreviations

BDA	- Big data analytics;
CRISP-DM	- CRoss-Industry Standard Process for Data Mining;
CRISP-MED-DM	- CRoss-Industry Standard Process for Data Mining for the Medical domain;
DOI	- Digital Object Identifier;
FAIR	- Findable, Accessible, Interoperable, and Reusable;
ICD	- International Classification of Diseases;
ICPC	- International Classification for Primary Care;
PhD	- Doctor of Philosophy;
PIP	- population insight product;
RACI	- Responsible, Accountable, Consulted, Informed;
RE-AIM	- Reach, Efficacy, Adoption, Implementation, and Maintenance;
UMCG	- University Medical Center Groningen.

Acknowledgements

The authors are grateful for all participants of the survey and the research group discussion. Additionally, we would like to thank Datapoort, a regional partnership organisation and digital platform focused on bringing together and linking health data, for the provision of an opportunity to develop this framework, the collaboration and the contribution to the development process. Moreover, the authors are grateful for the contributions of Prof. A. de Jonge and Prof. M.H. Blanker for reviewing this paper.

Data and/or code availability

The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request. Data are located in controlled access data storage at University Medical Center Groningen.

References

Abedjan, Z., Boujemaa, N., Campbell, S., Casla, P., Chatterjea, S., Consoli, S., Costa-Soria, C., Czech, P., Despenic, M., Garattini, C., Hamelinck, D., Heinrich, A., Kraaij, W., Kustra, J., Lojo, A., Sanchez, M., Mayer, M., Melideo, M., Menasalvas, E., Roel, W. (2019). Data Science in Healthcare: Benefits, Challenges and Opportunities**:** 3-38. DOI: https://doi.org/10.1007/978-3-030-05249-2_1.

Al Teneiji, A. S., Abu Salim, T. Y., Riaz, Z. (2024). Factors impacting the adoption of big data in healthcare: A systematic literature review. *International Journal of Medical Informatics* 187: 105460. DOI: https://doi.org/10.1016/j.ijmedinf.2024.105460.

Brooke, J. (2013). SUS: a retrospective. *J. Usability Studies* **8**(2): 29–40.

Brower, H. H., Nicklas, B. J., Nader, M. A., Trost, L. M., Miller, D. P. (2020). Creating effective academic research teams: Two tools borrowed from business practice. *J Clin Transl Sci* **5**(1): e74. DOI: https://doi.org/10.1017/cts.2020.553.

Canaway, R., Boyle, D., Manski-Nankervis, J.A., Gray, K. (2022). Identifying primary care datasets and perspectives on their secondary use: a survey of Australian data users and custodians. *BMC Medical Informatics and Decision Making* 22(1): 94. DOI: https://doi.org/10.1186/s12911-022-01830-9.

Chapman, P. (2000). CRISP-DM 1.0: Step-by-step data mining guide. https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman/54bad20bbc7938991bf34f86dde0babfbd2d5a72.

Cheng, C., Messerschmidt, L., Bravo, I., Waldbauer, M., Bhavikatti, R., Schenk, C. , Grujic, V., Model, T., Kubinec R., Barceló, J. (2024). A General Primer for Data Harmonization. *Scientific Data* **11**(1): 152. DOI: https://doi.org/10.1038/s41597-024-02956-3.

Coughlin, S., Roberts, D., O'Neill, K., Brooks, P. (2018). Looking to tomorrow's healthcare today: a participatory health perspective. *Intern Med J* 48(1): 92-96. DOI: https://doi.org/10.1111/imj.13661.

Dash, S., Shakyawar, S. K., Sharma, M., Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data* **6**(1): 54. DOI: https://doi.org/10.1111/imj.13661.

Davis, F. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13: 319. DOI: https://doi.org/10.2307/249008.

European Union (2024). European Health Data Space. DOI: https://doi.org/0.2875/269514.

Furstenau, L. B., Leivas, P., Sott, M. K., Dohan, M. S., López-Robles, J. R., Cobo, M. J., Bragazzi, N. L., Choo, K.K.R. (2023). Big data in healthcare: Conceptual network structure, key challenges and opportunities. *Digital Communications and Networks* 9(4): 856-868. DOI: https://doi.org/10.1016/j.dcan.2023.03.005.

GitHub (2024). GitHub Dasboard. Retrieved 16-10-2024, from https://github.com.

Glasgow, R. E., Vogt, T. M., Boles, S. M. (1999). Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health* 89(9): 1322-1327. DOI: https://doi.org/10.2105/ajph.89.9.1322.

Hong, L., Luo, M., Wang, R., Lu, P., Lu, W., Lu, L. (2018). Big Data in Health Care: Applications and Challenges. *Data and Information Management* **2**(3): 175-197. DOI: https://doi.org/10.2478/dim-2018-0014.

Inau, E.T., Sack, J., Waltemath, D., Zeleke, A.A. (2023). Initiatives, Concepts, and Implementation Practices of the Findable, Accessible, Interoperable, and Reusable Data Principles in Health Data Stewardship: Scoping Review. *J Med Internet Res*; 25: e45013. DOI: https://doi.org/10.2196/45013.

Manski-Nankervis, J.A., Hallinan, C., Cheah, R., Canaway, R. Mendonça, L. (2024). Using primary care data for research: What are the issues and potential solutions? *Australian journal of general practice* 53: 408-411. DOI: https://doi.org/10.31128/AJGP-07-23-6887.

Mitchell, R.K., Agle, B. R., Wood, D. J. (1997). Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of Who and What Really Counts. *The Academy of Management Review* **22**(4): 853-886. DOI: https://doi.org/10.2307/259247.

Naeem, M., Ozuem, W., Howell, K., Ranfagni, S. (2023). A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research. *International Journal of Qualitative Methods* 22: 16094069231205789. DOI: https://doi.org/10.1177/16094069231205789.

Niaksu, O. (2015). CRISP Data Mining Methodology Extension for Medical Domain. *Baltic J. Modern Computing* **3**: 92-109. https://www.bjmc.lu.lv/fileadmin/user_upload/ lu_portal/ projekti/bjmc/Contents/3_2_2_Niaksu.pdf.

Nilashi, M., Baabdullah, A. M., Abumalloh, R. A., Ooi, K.-B., Tan, G., Giannakis, W.-H. M., Dwivedi, Y. K. (2023). How can big data and predictive analytics impact the performance and competitive advantage of the food waste and recycling industry? *Annals of Operations Research*, 348: 1649-1690. DOI: https://doi.org/10.1007/s10479-023-05272-y.

Nilsen, P. (2015). Making sense of implementation theories, models and frameworks. *Implementation Science*, 10(1): 53. DOI: https://doi.org/10.1186/s13012-015-0242-0.

Rahimi, B., H. Nadri, H., Afshar, H. I., Timpka, T. (2018). A Systematic Review of the Technology Acceptance Model in Health Informatics. *Appl Clin Inform*, **9**(3): 604-634. https://doi.org/10.1055/s-0038-1668091.

Schröer, C., Kruse, F,, Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model, *Procedia Computer Science*, 181: 526-34. DOI: https://doi.org/10.1016/j.procs.2021.01.199.

Sharma, A., Minh Duc, N. T., Thang, T.L.L., Nam, N. H., Ng, S. J., Abbas, K. S., Huy, N. T., Marušić, A., Paul, C. L., Kwok, J., Karbwang, J., De Waure, C., Drummond, F. J., Kizawa, Y., Taal, E., Vermeulen, J., Lee, G. H. M., Gyedu, A., To, K. G., Verra, M. L., Jacqz-Aigrain É, M., Leclercq, W. K. G., Salminen, S. T., Sherbourne, C. D., Mintzes, B., Lozano, S., Tran, U. S., Matsui, M., Karamouzian, M. (2021). A Consensus-Based Checklist for Reporting of Survey Studies (CROSS). *J Gen Intern Med*, 36(10): 3179-3187. DOI: https://doi.org/10.1007/s11606-021-06737-1.

Sinaci, A.A., Núñez-Benjumea, F.J., Gencturk, M et al. (2020). From Raw Data to FAIR Data: The FAIRification Workflow for Health Research, *Methods Inf Med*, **59**(S 01): e21-e32. DOI: https://doi.org/10.1055/s-0040-1713684.

Subrahmanya, S. V. G., Shetty, D. K., Patil, V., Hameed, B. M. Z., Paul, R., Smriti, K., Naik, N., Somani, B. K. (2022). The role of data science in healthcare advancements: applications, benefits, and future prospects. *Ir J Med Sci*, 191(4): 1473-1483. DOI: https://doi.org/10.1007/s11845-021-02730-z.

Tong, A., Sainsbury, P., Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, 19(6): 349-357. DOI: https://doi.org/10.1093/intqhc/mzm042.

Wilkinson, M., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L.O., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**. DOI: https://doi.org/10.1038/sdata.2016.18.

Received June 4, 2026, revised July 18, 2026, accepted August 8, 2026