AMP ACTA MÉDICA PORTUGUESA
A Revista Científica da Ordem dos Médicos

Managing Data in Screening Programs: Challenges and Solutions

Gerir Dados em Programas de Rastreio: Desafios e Soluções

Hugo MONTEIRO ✉[1], Mariana OLIVEIRA[1], Ricardo MARTINHO[2], Carlos MARTINS[3]

Acta Med Port (In Press) ▪ https://doi.org/10.20344/amp.23363

ABSTRACT

Population-based screening programs are vital public health initiatives that enable the early detection of diseases, significantly reducing both morbidity and healthcare costs. As these programs expand, the management of the extensive data they generate becomes increasingly complex, highlighting the need for structured digital solutions. This narrative review article presents a pragmatic framework aimed at clarifying big data analytics tailored to the needs and practices of healthcare professionals and administrators, focusing on effective integration into routine screening workflows. To achieve effective data utilization, the process begins with systematic archiving, which involves cloud-based storage solutions capable of securely maintaining various data formats in compliance with regulatory standards, thus ensuring long-term accessibility and continuity. Subsequent real-time processing of screening data facilitates rapid decision-making and patient management by providing immediate validation and analysis, essential for maintaining the responsiveness of screening services. Transformation processes play a critical role in converting diverse data inputs into standardized, consistent formats, enabling seamless communication and exchange among multiple healthcare systems. Integration further builds upon this standardization, merging data from different healthcare providers and diagnostic centers into centralized analytical platforms. This unified approach enables comprehensive patient monitoring and supports predictive modeling for early identification of at-risk individuals. Advanced analytics, particularly process mining and predictive techniques, reveal inefficiencies within screening workflows, highlighting areas needing improvement. These methods help healthcare managers to streamline operations, optimize resources, and enhance overall program performance. Real-time visualization tools provide administrators with continuous, practical insights into operational dynamics, despite existing challenges related to data governance and system interoperability. This article illustrates these concepts through concrete examples from the colorectal cancer screening program in Northern Portugal and the response to the COVID-19 pandemic. The colorectal cancer screening scenario demonstrates how structured data management significantly boosts operational efficiency and healthcare accessibility. Meanwhile, the COVID-19 experience highlights the importance of having flexible digital infrastructures capable of quickly adapting to unexpected crises. Finally, ongoing investments in digital infrastructure, professional training, and comprehensive data governance are crucial for sustaining these improvements. This review provides clear, actionable knowledge to support healthcare professionals in adopting big data analytics effectively within preventive healthcare programs.

Keywords: Big Data; Data Management; Diagnostic Screening Programs; Public Health

RESUMO

Os programas de rastreio populacional permitem a deteção precoce de doenças, contribuindo para a redução da morbilidade e custos. Contudo, à medida que ganham escala, o enorme volume e heterogeneidade dos dados exigem soluções digitais robustas. Este artigo de revisão narrativa oferece um enquadramento pragmático para a aplicação de *big data* em rastreios, adaptado às práticas clínicas e de gestão com foco na integração eficaz nos fluxos de trabalho. A utilização eficaz dos dados inicia-se com o seu arquivo sistemático em plataformas de armazenamento na nuvem, seguras e em conformidade com as normas regulamentares, capazes de preservar múltiplos formatos de informação e garantir a acessibilidade a longo prazo. O processamento em tempo real possibilita decisões céleres sobre convocatórias, confirmação de diagnósticos e triagem subsequente, mantendo a capacidade de resposta dos serviços. Os processos de extração-transformação-carregamento normalizam dados provenientes de sistemas heterogéneos, assegurando consistência semântica e interoperabilidade. Esta normalização suporta a integração de registos clínicos, laboratoriais e administrativos em repositórios analíticos digitais que se encontram centralizados, permitindo uma visão longitudinal do percurso do utente. Com esta infraestrutura, é possível agilizar técnicas de mineração de dados e de modelação preditiva que ajudam a identificar indivíduos de maior risco, antecipar picos de afluência e identificar barreiras operacionais. Estes dados apoiam a alocação de recursos e o redesenho de processos, contribuindo para ganhos de eficiência e equidade. O recurso a ferramentas de visualização interativas traduz informação complexa em relatórios dinâmicos intuitivos, facilitando a monitorização contínua de indicadores e a tomada de decisões baseadas na evidência. Persistem desafios de governação de dados, financiamento e capacitação, exigindo políticas claras, formação contínua e mecanismos de auditoria regular a estes sistemas digitais. Este artigo ilustra estes conceitos através de exemplos concretos do programa de rastreio do cancro colorretal na região Norte de Portugal. Os investimentos sustentados em infraestrutura digital, formação profissional e governação de dados são essenciais para assegurar a sustentabilidade e o impacto operacional destes programas. Esta revisão oferece orientações práticas para apoiar profissionais de saúde na adoção eficaz de análises de *big data* em iniciativas de prevenção.

Palavras-chave: Big Data; Gestão de Dados; Programas de Rastreio; Saúde Pública

INTRODUCTION

Framing the problem

Population-based screening programs are essential public health initiatives aimed at early disease detection, particularly in asymptomatic populations. By enabling timely interventions, they significantly enhance health outcomes, reduce morbidity, and decrease long-term healthcare costs. Diseases such as cancer, diabetes, and cardiovascular conditions benefit greatly from structured screening efforts, as early detection and interventions are associated with reduced mortality

[1] Faculdade de Medicina. Universidade do Porto. Porto. Portugal.
[2] Centro de Investigação em Tecnologias e Serviços de Saúde (CINTESIS). Escola Superior de Tecnologia e Gestão. Instituto Politécnico de Leiria. Leiria. Portugal.
[3] Centro de Investigação em Tecnologias e Serviços de Saúde (CINTESIS@RISE). Universidade do Porto. Porto. Portugal.
✉ **Autor correspondente:** Hugo Monteiro. hugo.filipe.b.monteiro@gmail.com

Revisto por/*Reviewed by*: Thiago Gonçalves dos Santos Martins
Recebido/*Received*: 11/05/2025 - **Aceite/*Accepted***: 06/08/2025 - **Publicado *Online*/*Published Online***: 16/12/2025

and a lower burden on healthcare services.[1,2] However, as programs expand, they generate massive and diverse datasets that are increasingly difficult to manage using traditional systems. Without scalable digital infrastructures, the success of screening efforts risks becoming a logistical burden, delaying interventions, and straining public health services.

Effective health programs are guided by rigorous frameworks, which uphold evidence-based and regularly updated protocols that reflect the advances in research. In cancer screening, for instance, the application of new detection technologies such as genomics and liquid biopsies has improved awareness and accessibility to screening options. However, these do not come without challenges, such as managing the risks of overdiagnosis and false positives. Balancing the benefits against potential harms requires a nuanced approach, guided by robust data systems and transparent methodologies.[3-5] Adhering to guidelines and aligning with population health goals ensures maximum benefit with minimal harm.[6]

Integrated data infrastructure enhances program effectiveness by tracking participation, compliance, and outcomes, while also addressing disparities driven by socioeconomic and demographic barriers. However, achieving a high level of integration demands sophisticated tools capable of handling the multifaceted challenges of what is often referred to as 'big data'.[7,8]

Big data in healthcare

As screening programs expand, the resulting data increases not only in volume but also in complexity, posing significant challenges for storage, integration, and analysis.[3,9] These datasets, sourced from various clinical, administrative, and diagnostic systems, present challenges rooted in the core characteristics of big data, that healthcare professionals must increasingly navigate as part of routine practice. Integrating big data techniques into screening programs enhances their efficiency and scalability, but only if such techniques are well understood and embedded into everyday workflows.[10,11] Digital platforms facilitate data collection, integration, and processing, enabling real-time analytics that drive proactive decision-making (Fig. 1). This capacity transforms data from a passive record into an active tool for clinical and managerial insight, guiding early interventions and optimizing resource use. One way to approach this is by viewing data flows as a continuous pipeline, from capture to insight. In this model, each phase, from data generation to final analysis, must be intentionally designed to avoid fragmentation and delays that could compromise patient care. The challenges can be analyzed through the five key characteristics of big data: volume, velocity, variety, veracity, and value. Each of these traits presents specific requirements for healthcare systems.

- A high **volume** of data is generated while harnessing big data in healthcare, which entails leveraging diverse datasets, including electronic health records (EHRs), imaging, and laboratory results.
- **Velocity** in data processing is essential for real-time decision-making and timely follow-up with patients.
- Data **variety** requires interoperable systems to integrate the diversity of data sources, from electronic health records to mobile health devices.
- Ensuring data **veracity** is critical, as inaccuracies and inconsistencies can lead to erroneous decisions and adverse patient outcomes.
- All the while, extracting **valuable** insights – which are important to provide meaningful perceptions that support preventive healthcare, refine screening strategies, and guide policy decisions - requires advanced analytics.

For healthcare professionals, understanding these concepts is no longer optional. They underpin the technologies that now influence everything from triage to follow-up scheduling. Tools like predictive analytics can flag patients at higher risk and prompt timely actions. Meanwhile, process mining techniques optimize workflow efficiency, reduce delays in patient follow-ups and resource allocations.[12] Advanced computational statistical models, including machine learning (ML) and artificial intelligence (AI), further enhance the capacity to uncover actionable patterns, refine protocols and improve program outcomes.[13,14]

However, the transition to big data-driven infrastructure does not come without challenges and requires more than access to software. It requires an alignment between data systems and clinical practice, ensuring that tools support – not replace – professional judgement. Therefore, integrating diverse data sources requires adherence to standardized data models.[15] Privacy regulations such as the General Data Protection Regulation (GDPR) introduce stringent requirements to safeguard sensitive health information. Furthermore, while advanced analytic methods hold promises, they must also address potential biases and maintain transparency to ensure patient trust and equity in healthcare delivery.[16]

To address these challenges, screening programs must implement structured, data-driven strategies that prioritize security, interoperability, and analytical rigor. Digital platforms play a key role in supporting the various needs of screening programs, including:

- Clinical needs: Facilitating patient care through EHRs and decision support systems.[17]

- Administrative needs: Optimizing scheduling, resource allocation, and reporting.[18]
- Process monitoring needs: Tracking workflow efficiency and program performance.[19]
- Data management: Ensuring data quality, security, and compliance with regulatory standards.[20]
- Financial oversight: Managing costs, billing, and financial analytics.[21]

By leveraging these principles in everyday operational realities, screening programs can improve operational efficiency, enhance equity in healthcare access, and maximize the impact of preventive interventions.[14] Drawing on lessons from Portugal's colorectal cancer screening program, this narrative review aimed to synthesize recent evidence and practice-based insights into a five-phase data management framework that can strengthen preventive medicine services. Colorectal cancer remains one of the leading causes of cancer death in Europe, presenting the need for robust screening strategies supported by effective digital systems. We present a practical overview of big data integration in real-world screening programs, highlighting both the persistent barriers and the solutions that have contributed to sustainable, outcome-oriented models.[22]

Handling data

To overcome the data management challenges of modern screening programs, we propose a structured, five-phase framework (archiving, processing, transformation, integration, analytics – Fig. 2) capable of addressing the unique challenges posed by the nature of this data.[3] In the following sections, we detail how each phase contributes to building resilient, interoperable, and insight-driven screening programs – drawing from real-world implementations in Northern Portugal.

Archiving: ensuring data longevity and compliance

Effective long-term data storage is critical for screening programs, as data must be retained for patient follow-ups, epidemiological studies, and quality control processes. Given the scale of these programs, traditional storage methods are insufficient, requiring cloud-based solutions that provide scalability, redundancy, and compliance with data protection regulations such as GDPR. Screening data often spans a wide range of formats, including structured EHRs, unstructured imaging files, and laboratory tests, all of which require tailored storage solutions. Modern storage solutions, such as cloud-based infrastructures, can enhance data accessibility and scalability. These systems should incorporate role-based access permissions and embedded security protocols to ensure compliance with confidentiality standards and data protection regulations.[23]

One of the persistent challenges in data archiving is managing legacy systems. Older databases were often designed without significant future expansions in mind, leading to difficulties when integrating with modern digital platforms. Additionally, storage costs increase over time, especially as screening programs expand their reach and generate more data, including high-resolution medical imaging.[24] Screening data in our experience grew from just a few megabytes per month to over dozens of GB within five years – making cloud scalability a non-negotiable requirement.

Addressing these challenges requires initiative-taking budget planning for storage infrastructure, periodic system upgrades, and interoperability frameworks that allow seamless data retrieval from both old and new systems.

Real-time processing: unlocking insights and enhancing efficiency

Timely data processing is a critical requirement in screening workflows, where delays in analyzing test results can impact early intervention efforts. Screening data flows continuously, from appointments to test results, which requires immediate validation and feedback loops.[25]

The complexity of screening workflows means that multiple healthcare professionals interact with patient data daily, requiring constant access to updated information. For example, radiologists reviewing mammograms, lab technicians processing blood tests, and administrative teams managing patient scheduling all rely on rapid and error-free data flow. To support this, automated data validation mechanisms should be implemented to ensure that input from various sources is accurate and actionable, reducing the risk of administrative errors and clinical misinterpretations. For example, real-time lab results are obtained by connecting lab machines to the same digital platform for screening management, errors in analysis are automatically provided to human technical experts and reviewed. System monitoring is critical to identify process bottlenecks and enable timely corrective action

Transformation: standardizing data for consistency and usability

One of the most significant obstacles in screening data management is the variety of data formats and sources. Each screening modality – whether mammography, colonoscopy, or diabetic retinopathy detection – generates data in different structures, requiring transformation processes that standardize information for further analysis. Without transformation, screening data remains fragmented – hindering cross-institutional insights and automated reporting.

To facilitate this, standardized health data exchange frameworks such as HL7[26] and FHIR[27] must be employed, allowing seamless communication between different healthcare information systems.[28,29] Additionally, data preprocessing techniques should be used to clean, normalize, and structure raw data, ensuring that irrelevant or erroneous information is filtered out before integration. A key part of this process is ensuring semantic consistency, meaning that medical terminologies and classifications remain uniform across different systems. The alternative is to extract and manipulate data later, decreasing efficiency and increasing errors.

A major challenge in transformation is the presence of inconsistencies in primary data input formats, such as variations in date formats or different character sets in administrative records. Delays often stem from the lack of uniform transformation rules, especially when merging legacy data with evolving standards.

Integration and connectors: unifying data sources for holistic insights

After transformation, integration brings together data from diverse systems into a unified analytical environment, enabling patient monitoring, performance evaluation, and predictive insights. Population-wide programs rely on integration to combine data from hospitals, labs, registries, and administrative platforms. Integration efforts ensure that screening data is not siloed and simply stored but rather accessible for holistic patient monitoring, predictive modeling, and program evaluation. Data from more than 30 institutions was consolidated and connected daily, changing reporting lags from weeks to minutes, making it possible to monitor thousands of different processes close to real time.

To achieve this, data lakes and relational databases should be used to store and unify data, allowing healthcare providers to retrieve both individual patient records and aggregated population-level insights. Additionally, data connectors are deployed to enable automated merging of screening data across multiple platforms, minimizing the need for manual reconciliation. These connectors facilitate cross-institutional interoperability.

However, integration presents challenges, particularly when connecting digital systems that rely on different technical standards. Numerous systems, including database management systems (DBMS), use proprietary protocols and data structures, thus making data exchange intricate and time-consuming. For example, differences between structured databases (such as those using SQL) and more common file formats like spreadsheets often require additional steps to align and prepare the data. Technical mismatches caused delays, errors, and duplications, slowing down analytics pipelines.

Analytics: unlocking actionable insights

The goal of screening data management is to extract valuable insights that improve program efficiency and patient outcomes. Process mining techniques are particularly useful in screening programs, as they allow managers to identify inefficiencies in operational workflows, such as delays in test processing, scheduling gaps, or bottlenecks in follow-up procedures.[30] Tools like Power BI,[31] MicroStrategy,[32] and R[33]/Python[34]-based dashboards enable healthcare administrators to monitor screening program performance dynamically.

Several challenges persist despite these advancements. The absence of clear application programming interface (API) integrations, standardized data governance policies, and well-defined analytical workflows often delay the implementation of advanced solutions. Additionally, the risk of biases in machine learning models must be carefully managed to prevent disparities in screening access and outcomes. Lastly, the effectiveness of data-driven insights is dependent on high-quality, clean, and well-structured datasets, which remain a significant challenge in real-world healthcare environments.

By following the principles described above, we will provide evidence of the experience of the implementation of such techniques in screening programs and solutions found in alignment with best practices.[35]

Lessons learned: data-driven health interventions

In the northern region of Portugal, screening programs were implemented using the discussed methods. These programs covered most of the eligible population over more than ten years and included both oncological and non-oncological areas.

Data archiving and storage challenges

Archiving has played a critical role since the inception and evaluation of screening programs. However, legacy systems have struggled to scale effectively, leading to bottlenecks as screening volumes increased. Initially, available server storage seemed sufficient. However, as the screening program expanded, growing storage needs and renewed processing demands underscored the need for strategic planning in acquiring more robust digital systems.

However, legacy databases posed difficulties in retrieval and integration due to outdated formats and limited compatibility with modern digital platforms. Solutions included targeted infrastructure upgrades and interoperability frameworks to bridge gaps with legacy databases.

Data integration and standardization

Digital platforms have facilitated the expansion of local implementations to broader populations by standardizing data processes and enhancing transparency. Real-time dashboards provide regional health program managers with immediate insights into underperforming areas, allowing targeted interventions to boost screening uptake.

A major challenge (Table 1) in data integration stemmed from disparate DBMS, inconsistent protocols, and variations in language formats and file structures. Lack of clear connectors to data forced manual data reviews, slowing issue detection and requiring high technical expertise. Overcoming these obstacles required continuous standardization efforts and signaled the need to accelerate the adoption of interoperability frameworks such as HL7 and FHIR.

Nevertheless, these platforms also supported the automation of routine processes, freeing up resources for strategic tasks and enabling faster deployment of updated workflows, providing more time to audit processes, replying to queries from local program coordinators or other healthcare professionals and even providing the opportunity to innovate. To provide a concrete figure, automating data collection of eligible patients can represent a task taking weeks to less than a few hours or even minutes. This enabled more accurate cost analysis and adjustments, ensuring resource allocation closely matched demand, thereby avoiding overspending or underspending. These operational improvements, combined with standardized data pipelines, automation of routine processes, and structured training for key users improve capacity in scaling the framework across diverse institutional settings.

Advanced analytics and data utilization

Advanced analytics enabled descriptive, comparative, and scenario-based analysis, helping transform raw data into strategic decisions. Tools like Power BI allowed managers to explore large datasets and deliver tailored insights (example of dynamic dashboards presented in Fig. 3). For instance, visualizations of patient journeys helped identify drop-outs in screening adherence, allowing targeted outreach to improve follow-up rates.

Furthermore, process mining techniques revealed workflow bottlenecks and reduced operational delays through timely alerts and resource reallocation. This capacity to identify bottlenecks helped reduce the number of operational delays by facilitating timely adjustments in response to workflow dynamics. For example, whenever unexpected disruptions to services were detected, different resources responsible for 'upstream' or 'downstream' activities could be warned of the delays and adjust accordingly.

The robust data ecosystem also enabled more equitable access to healthcare services. Vulnerable populations, previously underrepresented, were provided free-of-cost screening tests, diagnostics, and treatment plans through the National Health System, supplemented by private care procurement agreements. By leveraging these tools and strategies, program managers ensured that the entire eligible population was included within a span of less than three years for each screening program. This comprehensive inclusion significantly enhanced the accessibility and impact of screening services, particularly in underserved regions.

Response capacity during COVID-19 and inclusive program maintenance

The adaptability of digital platforms proved essential in maintaining screening continuity during the COVID-19 pandemic. The ability to rapidly analyze large datasets and modify workflows ensured service resilience. The crisis exposed gaps in integration and coordination, reinforcing the need for interoperable, crisis-resilient systems. At the onset of the pandemic, disruptions to routine appointments and diagnostic follow-ups threatened to stall screening activities. Yet, digital infrastructure – equipped with centralized dashboards and automated data connectors – helped reschedule and reorganize screening appointments with minimal delays. Real-time dashboards helped coordinators detect service gaps and adjust schedules, avoiding prolonged care delays.

The pandemic experience reinforced that flexible data architectures, continuous workflow auditing, and stakeholder collaboration are key components in sustaining screening effectiveness under extreme circumstances. Going forward, building crisis-resilient infrastructure and workflows, with these tools and application of the concepts presented in this paper, will be instrumental for protecting public health and minimizing disruptions to essential services, even beyond COVID-19.

Training and capacity building for digital platforms

Even with strong digital platforms, success hinged on training healthcare professionals to use them effectively. Training programs were developed to enhance user proficiency in handling screening management tools and business intelligence (BI) dashboards.[36] Business intelligence tools had to be adapted to different user needs, from clinical staff to regional managers. Providing structured training sessions helped bridge digital skill gaps, ensuring seamless adoption and use of these technologies. Structured training, aligned to user roles, helped balance depth with practicality. Based on user profile and providing key users with the capacity to train others was essential to scale up faster digital platform usage.

Data driven screening programs

Big data strategies significantly improved efficiency, responsiveness, and equity in screening programs. However, the transition from traditional data handling to advanced digital ecosystems is not without challenges (Fig. 2), particularly in the context of critical healthcare scenarios where rapid adaptability is essential. Often, barriers such as limited funding and digital skills, not the technology itself, hinder implementation. This highlights the need for targeted investments in capacity-building and digital literacy.[37]

By integrating real-time dashboards, program managers could identify underperforming areas and bottlenecks, implementing targeted interventions with minimal delay. For instance, during periods of increased demand, such as flu seasons or post-pandemic recovery, quick access to updated data prevented the system from becoming overwhelmed, allowing for adequate distribution of resources. However, reliance on digital tools introduces vulnerabilities, including potential system downtimes, data input errors, and risks related to data integrity. Implementing strong data governance frameworks and redundancy measures is critical to mitigating these risks. Furthermore, privacy concerns and compliance with regulations like GDPR add layers of complexity, particularly in the context of cross-platform data sharing. Balancing transparency with data security remains a delicate and ongoing challenge.

Another key area of discussion is equitable access. Results show that big data tools enabled significant improvements in reaching underserved populations. This raises the question of how such systems can maintain inclusivity as they scale up further. Critical situations, such as natural disasters or sudden economic downturns, could exacerbate existing disparities. Proactive measures, such as incorporating geographic information systems to identify and address spatial inequities, could further strengthen the inclusiveness of these programs.[38]

Predictive analytics and process mining have greatly optimized screening workflows, but biases inherent to historical data can inadvertently perpetuate healthcare disparities.[39] Transparency in algorithm development and continuous model auditing are necessary to maintain fairness in care delivery. Additionally, maintaining historical data consistency is important, as structural changes in data catalogs over time can introduce analytical gaps, affecting long-term decision-making. Similarly, while process mining has proven effective in identifying inefficiencies, its reliance on comprehensive and accurate workflow data means that gaps in data coverage can lead to incomplete or misleading insights. Considering the time that these programs have been present, changes in the data catalogue or table structure might have introduced historical gaps in knowledge.

The COVID-19 pandemic highlighted both the strengths and weaknesses of digital screening infrastructures. While rapid adaptation and workflow flexibility ensured continuity, the crisis underscored the need for enhanced intersectoral collaboration, interoperability, and contingency planning to ensure seamless service delivery during future disruptions. Further underscoring the significance and potential for broader application, these results have been previously presented at European Public Health congress.[40]

Further research is needed to validate this framework in different national and regional contexts, particularly regarding its operational impact and scalability. Exploring how it performs across varied digital infrastructures and screening program designs will help determine its broader applicability. In parallel, assessing the role of advanced analytics and real-time monitoring in improving clinical outcomes remains a critical area for continued investigation.

CONCLUSION

Successful screening programs depend on integrating digital infrastructure with skilled professionals, ensuring that technology enhances – rather than replaces – clinical decision-making. Ongoing investment in training and capacity-building will ensure that expertise keeps pace with technological advancements.

Leveraging big data management techniques can further enhance program effectiveness. Patient engagement improves service quality and fosters trust, while quick, valuable insights and their integration into the decision-making process enable targeted interventions, ensuring equitable access and efficient resource allocation.

The discussion published here demonstrates the amount of investment required, not only in technological tools but also in highly skilled professionals. There is a further need to research how to innovate while compelling workers to keep up with innovative technology (like new generative AI tools), as well as how to procure technologies that are as future-proof as possible, so that there is significant value for money in a relevant period.

Finally, integrating process mining techniques into digital platforms supports continuous monitoring and optimization of workflows.[40] This initiative-taking approach enhances responsiveness and sets the stage for AI-driven automation, enhancing scalability, equity, efficiency, and impact of screening programs in tackling public health challenges. Health authorities should prioritize structured data management as a strategic pillar in scaling equitable and resilient screening services.

ACKNOWLEDGEMENTS

The authors acknowledge the Regional Coordination Team for Oncological Screening of the Portuguese National Health Service for their operational support and availability during the conceptualization of this review. We further acknowledge the dedication of all experts and healthcare professionals involved in the implementation, improvement, and daily execution of screening programs. Special thanks are also due to the teams responsible for developing and maintaining the digital platforms that enable these services.

The authors declare that AI-assisted technologies were used solely to generate code for workflow figures; no AI tools were used in writing, editing, or reviewing the manuscript.

AUTHOR CONTRIBUTIONS

HM: Data interpretation, writing and critical review of the manuscript.

MO, RM, CM: Writing and critical review of the manuscript.

All authors approved the final version to be published.

PROTECTION OF HUMANS AND ANIMALS

The authors declare that the procedures were followed according to the regulations established by the Clinical Research and Ethics Committee and to the Helsinki Declaration of the World Medical Association updated in October 2024.

ETHICS

The study was conducted in accordance with relevant ethical guidelines. No direct patient involvement occurred, and all data used were anonymized and obtained from authorized institutional datasets. All procedures complied with legal regulations and were approved by the Health Ethics Committee of the Regional Health Administration of Northern Portugal (CE/2023/96).

DATA CONFIDENTIALITY

The authors declare having followed the protocols in use at their working center regarding patients' data publication.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

FUNDING SOURCES

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

1. World Health Organization. Cancer prevention and control in the context of an integrated approach. Geneva: WHO; 2017.
2. World Health Organization Regional Office for Europe. Screening programmes: a short guide. Increase effectiveness, maximize benefits and minimize harm. Copenhagen: WHO/Europe; 2020.
3. Sweeney SM, Hamadeh HK, Abrams N, Adam SJ, Brenner S, Connors DE, et al. Case studies for overcoming challenges in using big data in cancer. Cancer Res. 2023;83:1183-90.
4. Mazzucco W, Stracci F, Gatta G, D'Argenzio A, Bidoli E, Carone S, et al. Cancer registries and data protection in the age of health digital interoperability in Europe: the perspective of the Italian Network of Cancer Registries (AIRTUM). Front Oncol. 2022;12:1052057.
5. American Association for Cancer Research. Screening for early detection. 2024. [cited 2025 Apr 24]. Available from: https://cancerprogressreport.aacr.org/progress/cpr24-contents/cpr24-screening-for-early-detection/.
6. World Health Organization. World Health Organization report on health and wellness. 2024. [cited 2025 Apr 24]. Available from: https://www.who.int/data/gho/data/major-themes/health-and-well-being.
7. Li L, Novillo-Ortiz D, Azzopardi-Muscat N, Kostkova P. Digital data sources and their impact on people's health: a systematic review of systematic reviews. Front Public Health. 2021;9:645260.

8. Fragala MS, Shiffman D, Birse CE. Population health screenings for the prevention of chronic disease progression. Am J Manag Care. 2019;25:548-53.
9. Awrahman BJ, Aziz Fatah C, Hamaamin MY. A review of the role and challenges of big data in healthcare informatics and analytics. Comput Intell Neurosci. 2022;2022:5317760.
10. Akyüz K, Cano Abadía M, Goisauf M, Mayrhofer MT. Unlocking the potential of big data and AI in medicine: insights from biobanking. Front Med. 2024;11:1336588.
11. Batko K, Ślęzak A. The use of big data analytics in healthcare. J Big Data. 2022;9:3.
12. Aversano L, Iammarino M, Madau A, Pirlo G, Semeraro G. Process mining applications in healthcare: a systematic literature review. PeerJ Comput Sci. 2025;11:e2613.
13. Olawade DB, Wada OJ, David-Olawade AC, Kunonga E, Abaire O, Ling J. Using artificial intelligence to improve public health: a narrative review. Front Public Health. 2023;11:1196397.
14. Pastorino R, De Vito C, Migliara G, Glocker K, Binenbaum I, Ricciardi W, et al. Benefits and challenges of big data in healthcare: an overview of the European initiatives. Eur J Public Health. 2019;29:S23-7.
15. Mensah E, Goderre JL. Data sources and data tools: preparing for the open data ecosystem. Public Health Inf Inform Syst. 2020:105-27.
16. Borges do Nascimento IJ, Marcolino MS, Abdulazeem HM, Weerasekara I, Azzopardi-Muscat N, Gonçalves MA, et al. Impact of big data analytics on people's health: overview of systematic reviews and recommendations for future studies. J Med Internet Res. 2021;23:e27275.
17. Alexiuk M, Elgubtan H, Tangri N. Clinical decision support tools in the electronic medical record. Kidney Int Rep. 2024;9:29-38.
18. Abdalkareem ZA, Amir A, Al-Betar MA, Ekhan P, Hammouri AI. Healthcare scheduling in optimization context: a review. Health Technol. 2021;11:445-69.
19. Pingili R. How workflow optimization improves patient care. Int J Res Comput Appl Inf Technol. 2024;7:1192-206.
20. Bernardi FA, Alves D, Crepaldi N, Yamada DB, Lima VC, Rijo R. Data quality in health research: integrative literature review. J Med Internet Res. 2023;25:e41446.
21. Cleverley WO, Cleverley JO, Parks AV. Essentials of health care finance. Massachusetts: Jones & Bartlett Learning; 2023.
22. Mendes D, Figueiredo D, Alves C, Penedones A, Costa B, Batel-Marques F. Impact of the COVID-19 pandemic on cancer screenings in Portugal. Cancer Epidemiol. 2024;88:102496.
23. Mehrtak M, SeyedAlinaghi S, MohsseniPour M, Noori T, Karimi A, Shamsabadi A, et al. Security challenges and solutions using healthcare cloud computing. J Med Life. 2021;14:448.
24. England PH. Retention, storage and disposal of mammograms and screening records. 2018. [cited 2025 Apr 24]. Available from: https://www.gov.uk/government/publications/breast-screening-manage-mammograms-and-records/retention-storage-and-disposal-of-mammograms-and-screening-records.
25. Schulz WL, Durant TJ, Torre Jr CJ, Hsiao AL, Krumholz HM. Agile health care analytics: enabling real-time disease surveillance with a computational health platform. J Med Internet Res. 2020;22:e18707.
26. Health Level Seven International. Introduction to HL7 standards. 2025. [cited 2025 Apr 24]. Available from: https://www.hl7.org/implement/standards.
27. Fast Healthcare Interoperability Resources. Welcome to FHIR®. 2025. [cited 2025 Apr 24]. Available from: https://build.fhir.org/.
28. Centers for Disease Control and Prevention. Implementing public health interoperability. 2025. [cited 2025 Apr 24]. Available from: https://www.cdc.gov/data-interoperability/php/public-health.
29. Williams E, Kienast M, Medawar E, Reinelt J, Merola A, Klopfenstein SA, et al. A standardized clinical data harmonization pipeline for scalable AI application deployment (FHIR-DHP): validation and usability study. JMIR Med Inform. 2023;11:e43847.
30. Van Der Aalst W. Data science in action. New Mexico: Springer; 2016.
31. Microsoft. Microsoft power bi. 2025. [cited 2025 Apr 24]. Available from: https://powerbi.microsoft.com/en-us.
32. Strategy. MicroStrategy. 2025. [cited 2025 Apr 24]. Available from: https://www.strategysoftware.com/.
33. Dalgaard P. R Development Core Team (2010). R: a language and environment for statistical computing. 2010. [cited 2025 Apr 24]. Available from: https://research.cbs.dk/en/publications/r-development-core-team-2010-r-a-language-and-environment-for-stats.
34. Rossum V. Python 3 reference manual. Scotts Valley: CreateSpace; 2009.
35. Union CotE. Council recommendation on strengthening prevention through early detection: a new EU approach on cancer screening replacing Council Recommendation 2003/878/EC. Off J Eur Union. 2022;100:1-10.
36. Santos MY, Ramos I. Business intelligence-da informação ao conhecimento. Lisboa: FCA–-Livros de Informática; 2017.
37. Brossard PY, Minvielle E, Sicotte C. The path from big data analytics capabilities to value in hospitals: a scoping review. BMC Health Serv Res. 2022;22:134.
38. Attah RU, Gil-Ozoudeh I, Garba B, Iwuanyanwu O. Leveraging geographic information systems and data analytics for enhanced public sector decision-making and urban planning. Magna Sci Adv Res Rev. 2024;12:152-63.
39. Munoz-Gama J, Martin N, Fernandez-Llatas C, Johnson OA, Sepúlveda M, Helm E, et al. Process mining for healthcare: characteristics and challenges. J Biomed Inform. 2022;127:103994.
40. Monteiro H, Oliveira M, Reis J, Tavares F. Optimizing colorectal screening in Portugal with process mining. Eur J Public Health. 2024;34:ckae144.2110.

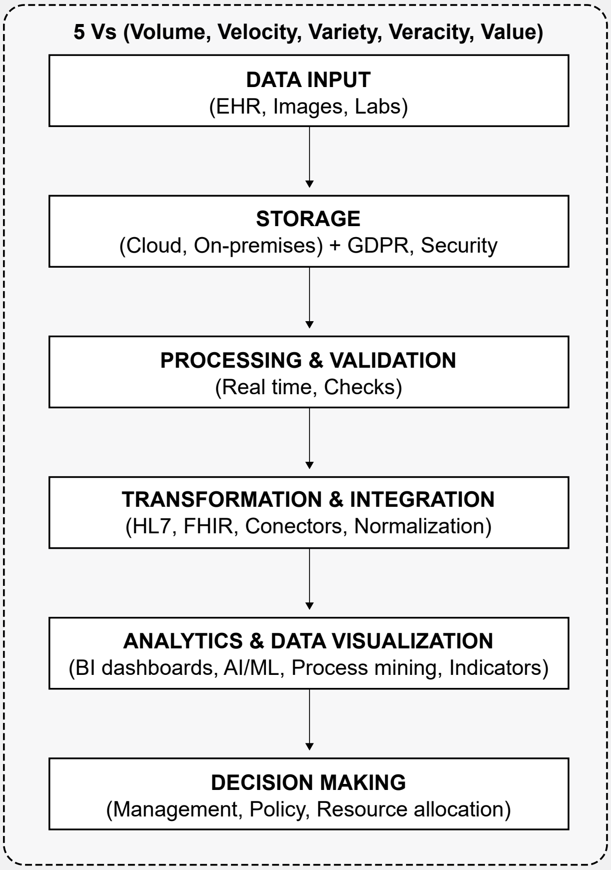
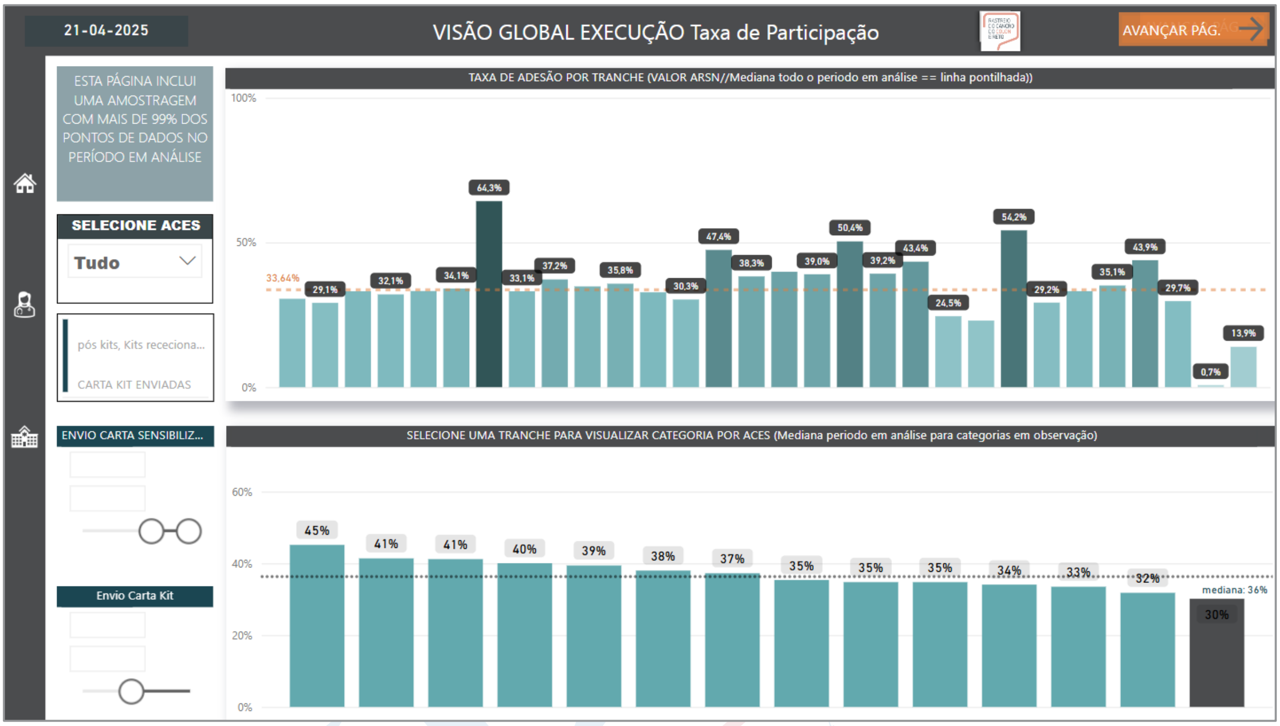

5 Vs (Volume, Velocity, Variety, Veracity, Value)
DATA INPUT
(EHR, Images, Labs)
STORAGE
(Cloud, On-premises) + GDPR, Security
PROCESSING & VALIDATION
(Real time, Checks)
TRANSFORMATION & INTEGRATION
(HL7, FHIR, Conectors, Normalization)
ANALYTICS & DATA VISUALIZATION
(BI dashboards, AI/ML, Process mining, Indicators)
DECISION MAKING
(Management, Policy, Resource allocation)

Figure 1 – The ‘data pipeline’

EHR: electronic health record; GDPR: general data protection regulation; HL7: health level seven (a set of international standards for the exchange, integration, sharing, and retrieval of electronic health information); FHIR: fast healthcare interoperability resources (a standard describing data formats and elements for exchanging electronic health records); BI: business intelligence; AI/ML: artificial intelligence/machine learning



21-04-2025

VISÃO GLOBAL EXECUÇÃO Taxa de Participação

AVANÇAR PÁG.

ESTA PÁGINA INCLUI UMA AMOSTRAGEM COM MAIS DE 99% DOS PONTOS DE DADOS NO PERÍODO EM ANÁLISE

SELECIONE ACES: Tudo

pós kits, Kits receciona...

CARTA KIT ENVIADAS

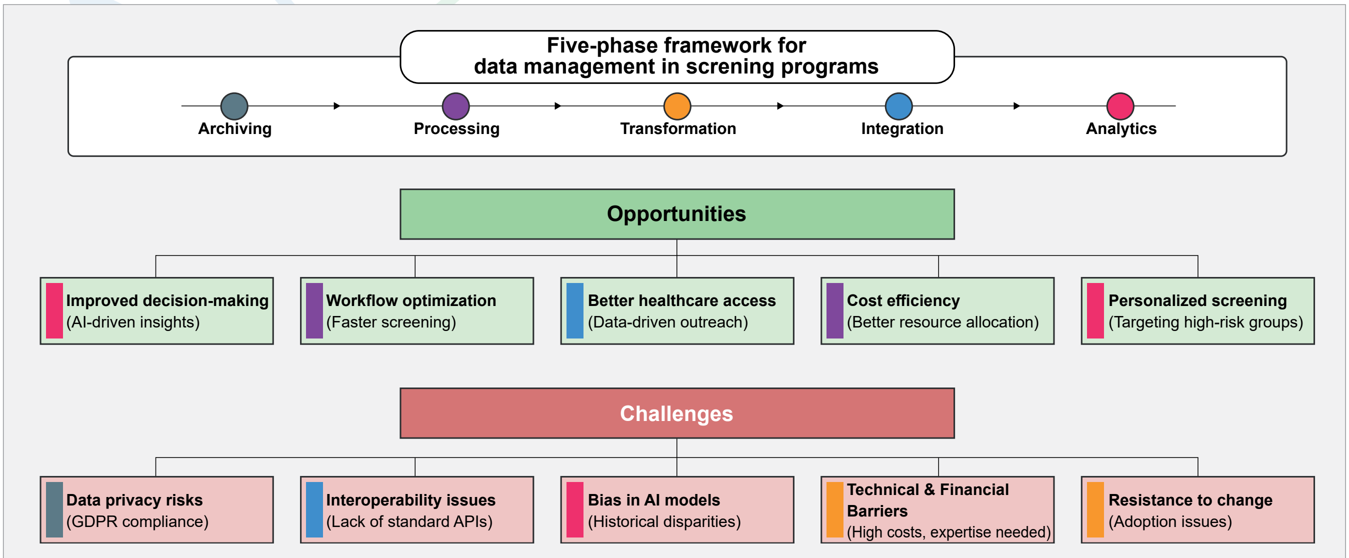
ENVIO CARTA SENSIBILIZ...

Envio Carta Kit

TAXA DE ADESÃO POR TRANCHE (VALOR ARSN//Mediana todo o periodo em análise == linha pontilhada))

SELECIONE UMA TRANCHE PARA VISUALIZAR CATEGORIA POR ACES (Mediana periodo em análise para categorias em observação)

Figure 2 – Dynamic business intelligence “BI” report with multiple calculated metrics with different management perspectives to visualize patient enrollment in the screening program. Other pages of the report can provide more granular data or wider analytical perspectives. (hidden variables due to data sensitivity issues)



Five-phase framework for
data management in screning programs
Archiving
Processing
Transformation
Integration
Analytics
Opportunities
Improved decision-making (AI-driven insights)
Workflow optimization (Faster screening)
Better healthcare access (Data-driven outreach)
Cost efficiency (Better resource allocation)
Personalized screening (Targeting high-risk groups)
Challenges
Data privacy risks (GDPR compliance)
Interoperability issues (Lack of standard APIs)
Bias in AI models (Historical disparities)
Technical & Financial Barriers (High costs, expertise needed)
Resistance to change (Adoption issues)

Figure 3 – Five-phase data-management framework (Archiving → Processing → Transformation → Integration → Analytics) and its relationship to the main opportunities and challenges of big data in screening programmes.

Each side colour bar reflects the data-management phase most closely associated with the corresponding opportunity or challenge. This mapping illustrates how each phase contributes to specific programme benefits, while simultaneously shaping the nature of the challenges encountered.

Table 1 – Summary of key challenges encountered in big data management for screening programs and the solutions implemented to address them

Concepts	Challenges in big data management	Experience
Archiving	Considering the volume of patients by the millions per year, multiple data sources are stored across servers in a widely distributed area.	It is required to define the format of data being stored; the size of the files (image files to be attached to reports can be sized by Gb or Tb); and additional budgeting might be required as technology improves not only in terms of the size of the software programs but also record size.
Processing	Servers dedicated to various aspects of data processing, as well as for image rendering, may be necessary.	If thousands of healthcare professionals work in multiple places and see patients every day, or if tests are performed daily, queries to the database to check administrative identification or validate past values are a constant.
Transformation	Primary data might be stored in wide table formats, or secondary data be the result of table merging during processing - to match records with new test results.	This requires a middle layer of specific processing to guarantee that data quality is assessed and that data values comply to regulatory standards. Indicators are also processed in this stage, where granular data is aggregated and computed to simplify signals or provide visual alerts.
Integration (connectors)	To view data from its multiple sources, where data is stored, and to transform data and provide the possibility of analyzing it, data must be integrated into a middle layer dedicated to the creation of semantic models.	Connectors from different DBMS are not mapped or established in protocols, this means that certain loaded files show differences in language formatting or type (e.g., .xls versus .xlsx) which delays the data handling process.
Analytics	Data Science is an end and a window to all the processes related to data. The abstraction of data into understandable concepts produces information and knowledge	Due to the lack of clear API and User profiles, mapped entities, or standard data protocols to ensure quality, analytics is a “delayed” product. Where it is easier to provide descriptive analysis, and more advanced methods lack timely usefulness.