

Quality Assurance and Testing Strategies in a Digital Platform for Screening and Early Diagnosis of Oral Cancer

Leandro Pedrosa¹, Luiza de Oliveira Costa², Gabriel Freitas²,
Tallya Jesus Sousa Barbosa², Giovanna Vargas Barbosa³, Renata Dutra Braga²,
Rejane F. Ribeiro-Rotta⁴

¹Programa de Pós-Graduação em Ciências da Saúde
Faculdade de Medicina - UFG - Goiânia - GO - Brasil

²Instituto de Informática – UFG – Goiânia – GO – Brasil

³Instituto Acadêmico de Ciências Agrárias e Sustentabilidade
- UEG - Sao Luis de Montes Belos - GO - Brasil

⁴Faculdade de Odontologia – UFG - Goiânia - GO - Brasil

{leandropedrosa, luizaluiza, freitas.gabriel, tallyabarbosa}
@discente.ufg.br, {renatadbraga, rejanefff}@ufg.br,
Giovannavargas2003@outlook.com

Abstract. Research Context: Software testing in healthcare systems is critical, as failures can compromise diagnoses, professional trust and patient safety. In severe conditions, such as oral cancer, the absence of robust validation strategies may lead to delays or clinical errors with serious consequences. **Scientific and/or Practical Problem:** Despite advances in AI-based healthcare platforms, gaps remain in the adoption of integrated testing strategies that ensure quality, usability, and security in critical systems. These shortcomings can hinder the reliability of AI-driven tools in clinical environments. **Proposed Solution and/or Analysis:** This study reports the testing practices applied in the Sobre-Vidas Platform-“Oral Cancer”, designed to support early diagnosis through AI. A multidimensional testing strategy was implemented, encompassing back-end (unit, integration, performance, and security) and front-end testing (usability, responsiveness, accessibility and functional), supported by tools including Python, JUnit, SonarQube, Locust, and Cypress. **Related IS Theory:** The work is grounded in software quality models, particularly ISO/IEC 25010, and in socio-technical system theory, which emphasizes the interplay between technical reliability and user acceptance within digital health environments. **Research Method:** A qualitative case study approach was adopted, combining practices such as test driven development (TDD) and automated API testing with validation in real world healthcare contexts. Trials were conducted with physicians, dentists, and health secretaries in two municipalities in the state of Goiás, with feedback collected through observation, interviews, and usability sessions. **Summary of Results:** The need for adjustments in workflows, terminology, and interface design was revealed, highlighting that continuous and integrated testing processes are essential for ensuring clinical effectiveness, safety, and system adoption. **Contributions and Impact to IS Area:** This work demonstrates how integrated testing strategies, aligned with established quality frameworks, can enhance trust and adoption of AI-driven healthcare platforms. It broadens

the discussion on software testing as a socio-technical enabler of safety and innovation in digital health.

1. Introdução

A transformação digital em saúde tem redefinido a forma como os sistemas de informação são concebidos, implantados e utilizados em diferentes níveis de atenção. A incorporação de tecnologias digitais (registros eletrônicos em saúde, telemedicina e plataformas inteligentes de monitoramento) amplia o acesso, favorece a integração entre serviços e fortalece a gestão clínica e administrativa. No entanto, esses avanços também acarretam riscos associados à complexidade tecnológica, tornando essencial a implementação de estratégias sistemáticas de qualidade e segurança do *software* [Din 2025].

Nesse contexto, a interoperabilidade desponta como requisito indispensável para garantir o fluxo contínuo e confiável de informações entre sistemas heterogêneos, condição fundamental para o sucesso de iniciativas em saúde digital [Bates and Samal 2018]. O padrão HL7® FHIR® tem se consolidado como referência internacional ao permitir a troca padronizada de dados clínicos, promovendo a consistência semântica e a integridade informacional [Nan and Xu 2023]. Sem estratégias de validação e testes, a interoperabilidade pode se transformar em um vetor de risco, com potenciais falhas de integração capazes de comprometer a qualidade do cuidado.

A inteligência artificial representa outra fronteira de inovação, com destaque para algoritmos de aprendizado de máquina e modelos de linguagem que vêm sendo aplicados em triagem, apoio à decisão clínica e monitoramento de populações de risco [Mehraeen et al. 2025]. Apesar do potencial, falhas não detectadas nesses sistemas podem gerar vieses, diagnósticos incorretos e impactos adversos na saúde do paciente. Por isso, a adoção de práticas de *Software Quality Assurance* (SQA) é determinante para assegurar não apenas a acurácia algorítmica, mas também a confiabilidade, a ética e a aceitação clínica de soluções digitais baseadas em Inteligência Artificial (IA) [Bernardo et al. 2024].

Dessa forma, observa-se que a transformação digital em saúde, a interoperabilidade e a inteligência artificial, quando associadas a estratégias de testes e qualidade de software, constituem um tripé essencial para a segurança do paciente, a efetividade clínica e a sustentabilidade das soluções digitais em saúde pública [Nan and Xu 2023, Stoumpos et al. 2023].

Adicionalmente, erros em sistemas de saúde podem comprometer diagnósticos, abalar a confiança de profissionais e colocar em risco a segurança do paciente. Em condições graves, como o câncer de boca, falhas tecnológicas e ausência de estratégias de validação favorecem atrasos ou diagnósticos equivocados, com consequências potencialmente irreversíveis. Nesse cenário, os testes de *software* assumem papel importante, não apenas como etapa final do desenvolvimento, mas como elemento estruturante da segurança do cuidado, sustentando a confiabilidade do sistema, a tomada de decisão clínica e a adesão dos usuários [Obigbesan et al. 2024].

O câncer de boca representa um dos dez tipos mais incidentes no Brasil, com maior prevalência em homens acima de 40 anos, especialmente tabagistas e etilistas. Apesar de ser uma neoplasia de fácil acesso anatômico para inspeção clínica, a detecção tardia

permanece um desafio, resultando em tratamentos mutiladores, custos elevados e menor taxa de sobrevivência. Nesse contexto, plataformas digitais apoiadas por IA e interoperabilidade têm potencial para ampliar o rastreamento, reduzir desigualdades no acesso e fortalecer políticas públicas de prevenção [Graili and Farhoudi 2025].

A Plataforma SobreVidas - “Câncer de Boca” foi concebida para apoiar o diagnóstico precoce de pacientes com risco para essa doença, possibilitando monitoramento e rastreamento da população na atenção primária, com encaminhamento para o nível de atenção mais adequado e maior probabilidade de cura com o auxílio da IA. No entanto, a garantia de qualidade em sistemas de saúde ultrapassa a simples execução de testes e demanda um plano integrado de qualidade de *software*, abrangendo desde a concepção de requisitos até a entrega contínua de incrementos. Diferentemente de abordagens restritivas, centradas apenas em testes funcionais, este plano adota uma visão sistêmica e preventiva, combinando definição clara de requisitos, metodologias ágeis e validação contínua em ambientes reais [International Organization for Standardization 2024, Ronchieri and Canaparo 2023].

Este artigo descreve as estratégias e práticas de qualidade aplicadas ao desenvolvimento da Plataforma SobreVidas - “Câncer de Boca”, destacando como a junção entre rigor técnico, usabilidade e validação é determinante para a segurança clínica e para a confiança dos profissionais. O estudo se fundamenta na experiência prática de concepção, prototipagem e validação em campo, incluindo pilotos realizados em municípios goianos, com *feedback* direto de dentistas, enfermeiros, médicos e gestores de saúde. Assim, a questão de pesquisa que orientou este estudo foi: de que forma estratégias integradas de testes de *software* podem contribuir para garantir a qualidade, a segurança e a aceitação de uma plataforma de saúde baseada em inteligência artificial?

Diante disso, o objetivo geral deste artigo foi relatar as práticas de testes aplicadas no desenvolvimento da Plataforma SobreVidas - “Câncer de Boca”, evidenciando como as estratégias de qualidade de software podem assegurar segurança, confiabilidade e usabilidade em sistemas de saúde. Para alcançar esse propósito, buscou-se: (i) analisar a relevância dos testes de software em sistemas de saúde, relacionando-os à segurança do paciente, à precisão diagnóstica e à confiança dos profissionais; (ii) caracterizar as metodologias de qualidade utilizadas (*Test-Driven Development* - TDD, testes automatizados de API e critérios de aceitação); (iii) demonstrar a aplicação de testes unitários, de integração, de desempenho e de segurança no *back-end*; (iv) examinar as estratégias de testes no *front-end*, incluindo usabilidade, acessibilidade, responsividade e automação funcional; (v) relatar a experiência de testes em campo, destacando as adaptações a partir dos *feedbacks* dos profissionais de saúde; e (vi) sintetizar as lições aprendidas, relacionando-as às boas práticas de garantia de qualidade em sistemas críticos e projetando próximos passos para a evolução da Plataforma.

2. Metodologia

A metodologia deste estudo foi delineada de forma a atender aos objetivos específicos, utilizando uma abordagem qualitativa-descritiva baseada em estudo de caso, tendo como objeto a Plataforma SobreVidas - “Câncer de Boca”. O processo foi estruturado em etapas sequenciais, de modo a garantir a rastreabilidade das ações e a reprodutibilidade do estudo. O estudo foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal

de Goiás (UFG), conforme protocolo nº 3.239.242.

2.1. Análise da relevância dos testes de software em saúde

Inicialmente, realizou-se uma revisão bibliográfica direcionada em bases como IEEE *Xplore*, *ACM Digital Library* e *Scielo*. Essa etapa teve como objetivo identificar riscos e implicações decorrentes da ausência de testes em sistemas de saúde. Artigos relacionados a testes de software foram selecionados (*software testing*, *healthcare systems*, *patient safety* e *quality assurance*). Os resultados dessa análise serviram como fundamento teórico para contextualizar a relevância dos testes de software, especialmente no que se refere à segurança do paciente, à precisão diagnóstica e à confiança dos profissionais de saúde.

2.2. Caracterização das metodologias de qualidade

Nessa etapa, procedeu-se à caracterização das metodologias de qualidade adotadas durante o ciclo de desenvolvimento da plataforma. Essa caracterização foi realizada a partir da documentação disponível, incluindo registros de *sprints*, critérios de aceite e *pipelines* de integração contínua. A análise desses registros permitiu sistematizar práticas como TDD, testes automatizados de API e validação de critérios de aceite técnico e de negócio, demonstrando como essas metodologias foram incorporadas de forma efetiva ao processo de desenvolvimento.

2.3. Aplicação de testes no *back-end*

Posteriormente, foram aplicados testes no *back-end* da solução. Para isso, implementaram-se testes unitários e de integração utilizando Python/Pytest e JUnit, enquanto os testes de desempenho e carga foram conduzidos com o apoio do Locust. Além disso, ferramentas de análise estática, como o SonarQube, foram utilizadas para reforçar a qualidade do código. Essa etapa contemplou os casos de uso mais críticos, como a classificação realizada pela inteligência artificial, o funcionamento das APIs de cadastro e *upload* e a anonimização de dados clínicos. Também foram modelados cenários de falha, incluindo tentativas de injeção de SQL e indisponibilidade de serviços externos, assegurando maior robustez à Plataforma.

2.4. Aplicação de testes no *front-end*

Em seguida, foram conduzidos testes no *front-end* da plataforma. Essa verificação contemplou usabilidade, acessibilidade, responsividade e automação funcional. A coleta de dados ocorreu por meio de laboratórios presenciais e remotos realizados com profissionais de saúde, complementados por testes automatizados com o *Cypress*. Os fluxos mais relevantes, como cadastro de pacientes, consulta de resultados e *upload* de imagens, foram priorizados. Os critérios de acessibilidade seguiram as recomendações do W3C (WCAG 2.1), com foco no contraste de cores, na navegação por teclado e na compatibilidade com leitores de tela [World Wide Web Consortium 2018].

2.5. Experiência de testes em ambiente real

Essa etapa consistiu na realização de testes em ambiente real. Para isso, foi conduzido um piloto nos municípios de Firminópolis-GO e São Luís de Montes Belos-GO, durante os meses de julho e agosto de 2025. Essa etapa envolveu laboratórios de usabilidade remotos, realizados com dentistas, enfermeiros, médicos e gestores municipais de saúde. A coleta de dados incluiu observação direta, entrevistas estruturadas e *feedback* qualitativo. A

análise comparativa entre o Produto Mínimo Viável (do inglês, MVP inicial e o protótipo reformulado possibilitou identificar adaptações necessárias para garantir maior aderência ao contexto clínico.

2.6 - Validação e Avaliação do Componente de Inteligência Artificial

A avaliação do componente de inteligência artificial não envolveu o treinamento ou ajuste fino (*fine-tuning*) de modelos proprietários, mas sim a seleção e validação de modelos de linguagem de grande porte (*Large Language Models* – LLMs) previamente treinados e integrados à plataforma. A estratégia adotada foi baseada em uma abordagem de avaliação comparativa e orientada por cenários.

Inicialmente, os modelos candidatos foram pré-selecionados com base em seu desempenho em benchmarks e rankings públicos, específicos para tarefas relevantes aos requisitos do sistema, tais como capacidade de raciocínio geral, acurácia na chamada de ferramentas (*tool-calling*) e correção na passagem de parâmetros durante a invocação dessas ferramentas.

Posteriormente, os modelos selecionados foram submetidos a múltiplas execuções de cenários clínicos e operacionais idênticos, representativos dos casos de uso da plataforma. Esse processo teve como objetivo avaliar a consistência das respostas, a robustez e a estabilidade do comportamento dos modelos ao longo de diferentes execuções, bem como sua capacidade de invocar corretamente as ferramentas apropriadas e retornar saídas estruturadas com dados válidos.

A avaliação concentrou-se na adequação funcional e na confiabilidade do componente de IA dentro do contexto sociotécnico da plataforma de saúde, em vez de métricas tradicionais de aprendizado supervisionado. Essa abordagem está alinhada ao objetivo do sistema de garantir um comportamento confiável de apoio à decisão quando incorporado a um ambiente de software crítico à segurança.

2.7. Síntese e lições aprendidas

Por fim, foi realizada a síntese dos achados e a definição de lições aprendidas. Essa síntese ocorreu por meio da triangulação entre resultados de testes automatizados, métricas de desempenho e percepções qualitativas relatadas pelos usuários. A análise cruzada desses dados permitiu identificar boas práticas que podem ser replicadas em outros sistemas de saúde. Além disso, foram projetados os próximos passos para a evolução da Plataforma, incluindo a ampliação da automação de testes, a incorporação de práticas de DevSecOps e a expansão para novos municípios.

3. Resultados

Esta seção apresenta os principais achados do estudo, organizados do nível estratégico ao operacional. Inicialmente, descrevemos as estratégias de garantia de qualidade incorporadas ao ciclo de desenvolvimento (3.1), seguidas do ecossistema de ferramentas e das métricas de avaliação (3.2). Em seguida, reportamos os resultados técnicos e a validação em campo (3.3), bem como os impactos práticos na Plataforma (3.4). Detalhamos então os testes do *back-end* (unitários, integração, desempenho/carga e segurança) (3.5) e os testes do *front-end* (usabilidade, responsividade, acessibilidade e automação) (3.6). Por fim, consolidamos os testes com usuários em ambiente real (3.7) e integramos as evidências de capacidade e desempenho obtidas com Locust, Netdata e Gra-

fana (3.8), destacando implicações para *Service Level Agreement/Service Level Objective* (SLO/SLA) e um plano de ação para otimização e escalabilidade.

3.1. Estratégias abordadas

As estratégias de garantia de qualidade foram implementadas de maneira proativa e integrada ao ciclo de desenvolvimento da Plataforma SobreVidas - “Câncer de Boca”, não sendo tratadas como etapa isolada ou meramente finalística. O refinamento sistemático de requisitos com a Faculdade de Odontologia da UFG permitiu alinhar funcionalidades às necessidades reais do rastreamento do câncer de boca, reduzindo ambiguidade e risco de inconsistências. A adoção de TDD [Basit *et al.* 2018] a implementação com testes unitários especificados antes do código; testes automatizados de API foram incorporados à integração contínua para prevenir regressões e preservar os contratos entre *front-end* e *back-end*; e critérios de aceite técnicos e de negócio foram aplicados a cada incremento. Esse conjunto foi reforçado por revisões contínuas de código e documentação, resultando em evolução controlada e auditável, orientada à segurança clínica. A arquitetura lógico-funcional que sustenta essas estratégias é apresentada na Figura 1, onde se visualizam a camada de apresentação, o *gateway* e autenticação, os módulos de negócio, mensageria e persistência, bem como o conector de interoperabilidade HAPI FHIR.

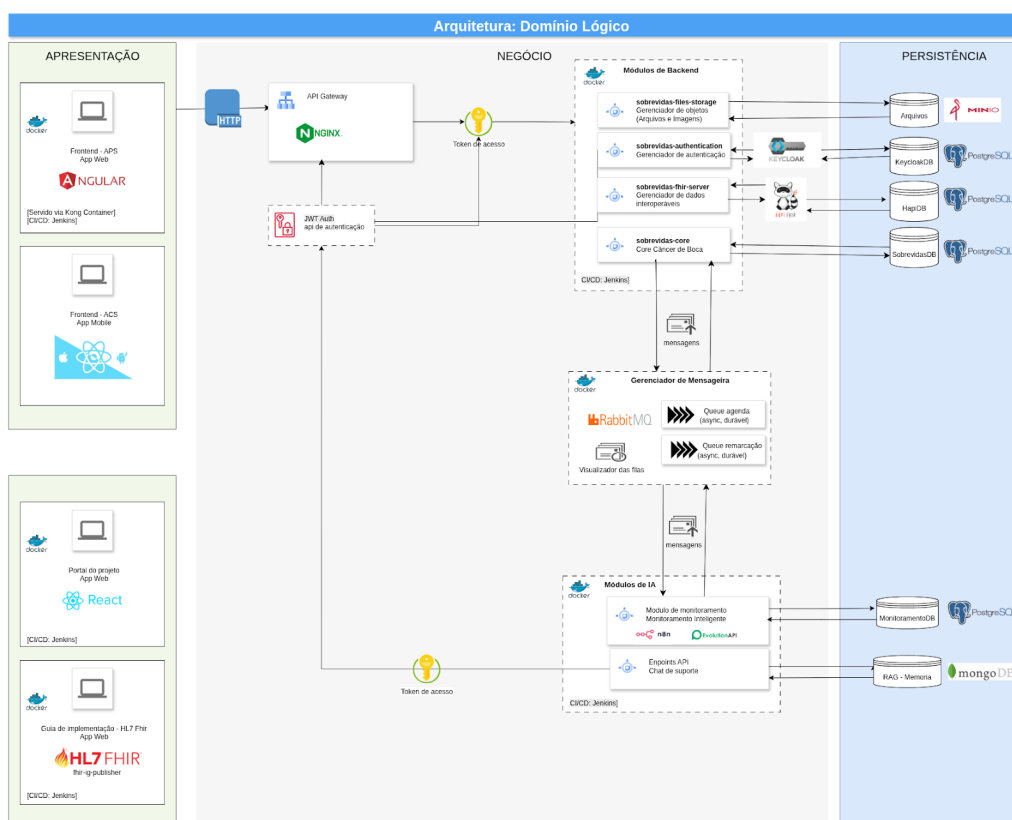


Figura 1. Arquitetura geral e os seus componentes.

3.2. Ferramentas utilizadas e métricas

O ecossistema de ferramentas viabilizou a aplicação das estratégias em cada camada do sistema. JUnit e Pytest foram empregados na execução de testes unitários e de integração;

SonarQube forneceu relatórios de qualidade com meta mínima de 80% de cobertura e ausência de alertas críticos; e Locust embasou os cenários de desempenho e carga. As principais métricas monitoradas incluíram tempos de resposta (mediana e p_{95}), *throughput* em requisições por segundo, taxa de falhas, cobertura de testes e volume de execuções automatizadas. A Figura 2 demonstra que, no período de teste, não houve pressão por recursos (load baixo, PSI=0%, I/O e rede ociosos), descartando infraestrutura como causa de degradação de desempenho. Esse diagnóstico se detalha nos demais painéis do Netdata (Figura 3, Figura 4, Figura 5, Figura 6, Figura 7, Figura 8 e Figura 9), todos apontando folga de recursos e estabilidade do SO durante as campanhas.

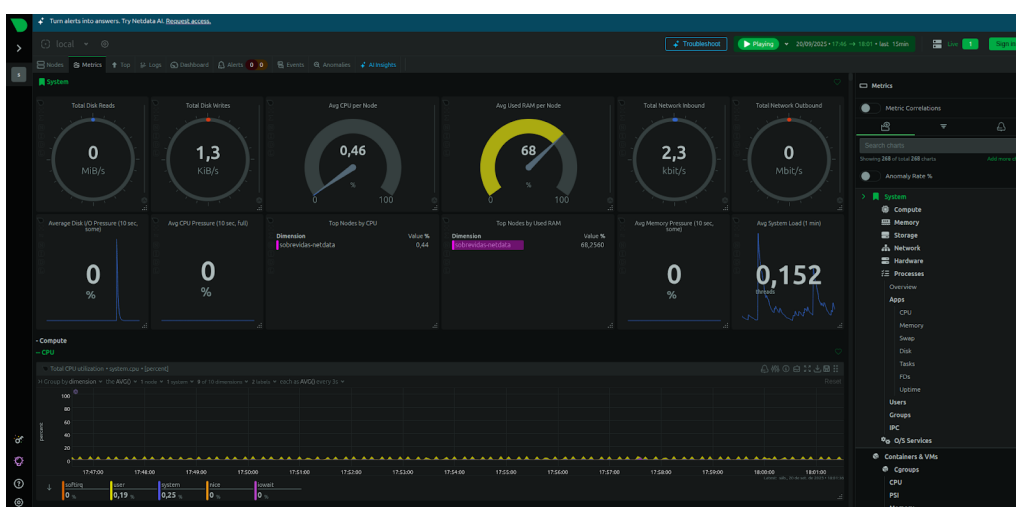


Figura 2. System Overview (Visão Geral da Plataforma).

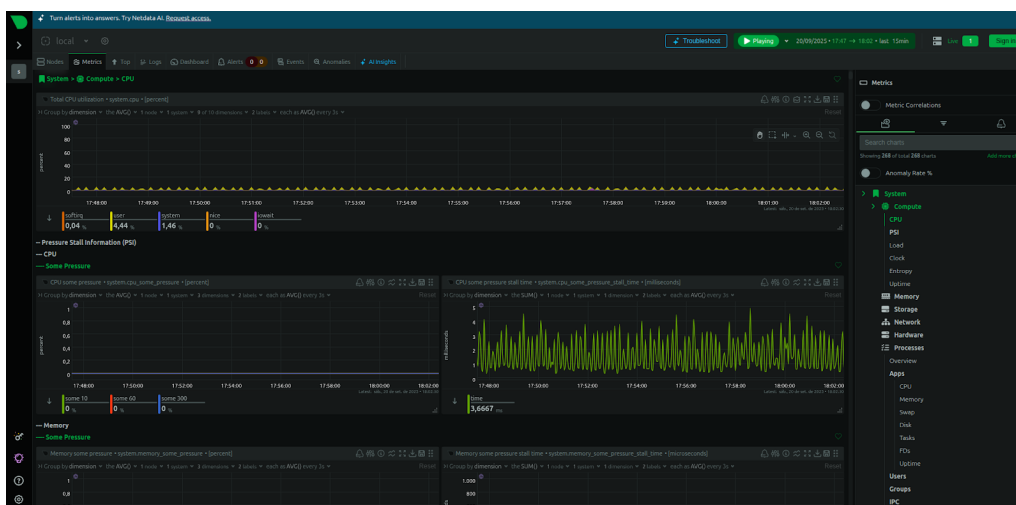


Figura 3. Compute CPU (Uso do Processador).



Figura 4. Memory(Usu da Memória RAM).

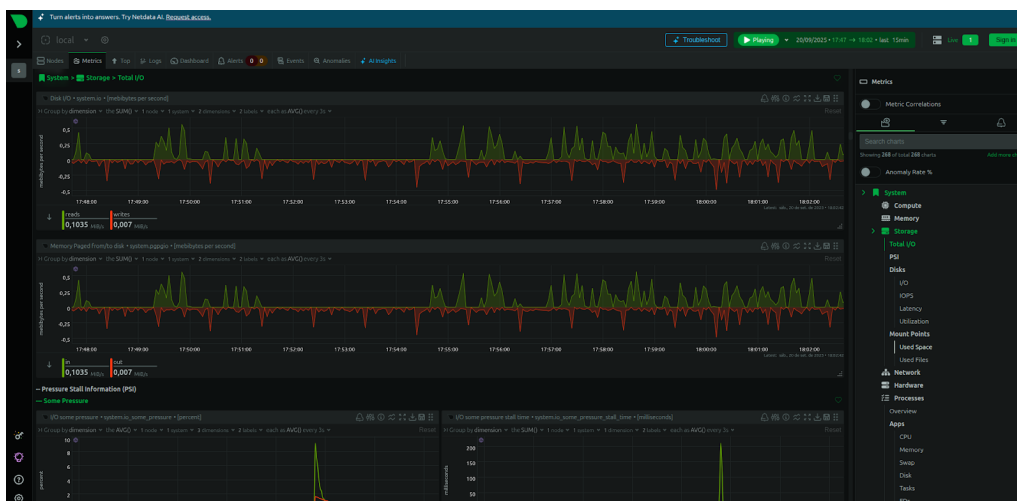


Figura 5. Storage I/O (Acesso ao Disco).

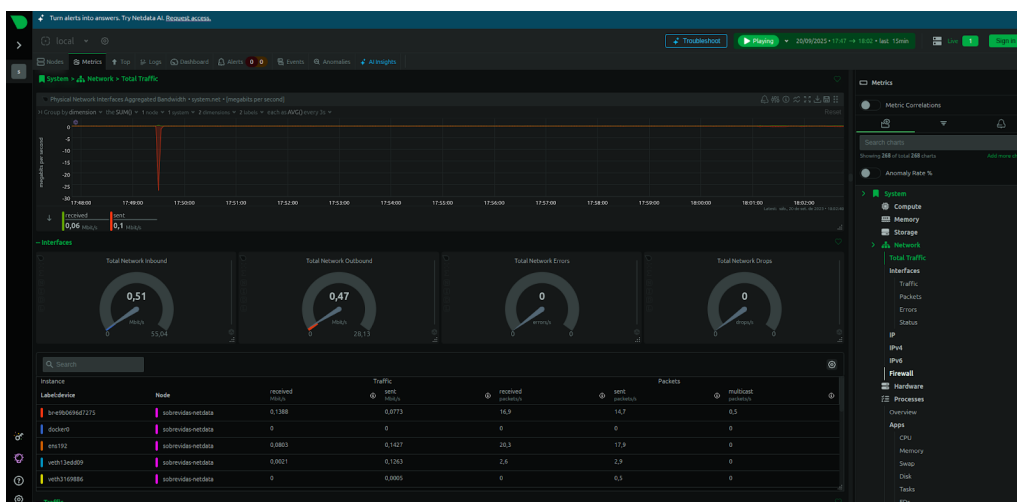


Figura 6. Network Traffic (Tráfego de Rede)

3.3. Resultados técnicos e validação em campo

A testagem identificou e corrigiu falhas críticas como respostas HTTP 500, retornos HTTP 200 sem corpo e inconsistências de *payloads*, por meio de ajustes no código, documentação e critérios de aceitação. Nos testes de carga com Locust observou-se estabilidade até cerca de 25 usuários concorrentes, com *throughput* próximo de 10 requisições por segundo, mediana de resposta de 1,5 a 2,5 segundos e p_{95} entre 3 e 4,5 segundos. A partir de aproximadamente 55 usuários emergiu um ponto de saturação, caracterizado por crescimento abrupto da latência (p_{50} de 6 a 9 segundos e p_{95} de 15 a 17 segundos), estagnação do *throughput* e aparecimento de falhas. Essas curvas estão sintetizadas na Figura 10, evidenciando um teto de processamento compatível com contenção lógica interna (aplicação/BD) e não com limites de máquina, conclusão corroborada pelos painéis do Netdata citados antes.

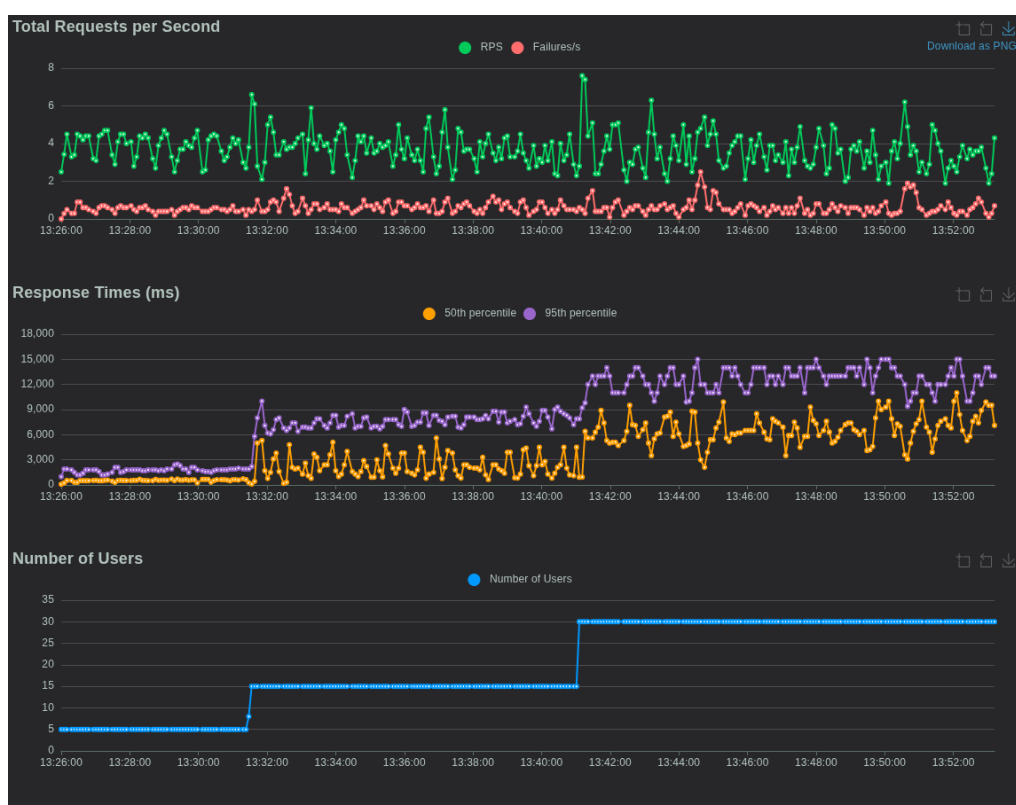


Figura 10. Curvas de RPS, latência p_{50}/p_{95} , falhas e usuários.

A validação em campo, realizada em pilotos nos municípios de Firminópolis e São Luís de Montes Belos, confirmou a estabilidade funcional e apontou melhorias de experiência: simplificação de fluxos, padronização terminológica e realce de alertas clínicos. Essas mudanças foram incorporadas em ciclos curtos, elevando a aderência do protótipo reformulado às rotinas dos profissionais de saúde.

3.4. Impactos práticos na Plataforma

A aplicação de SQA reduziu a probabilidade de falhas que possam comprometer a segurança do paciente e a precisão diagnóstica, reforçando a confiabilidade operacional. O cumprimento à Lei Geral de Proteção de Dados (LGPD) foi atendido por anonimização,

criptografia em repouso e em trânsito e controle de acesso baseado em papéis, apoiado por trilhas de auditoria. Do ponto de vista de experiência, os testes de usabilidade e acessibilidade se traduziram em maior fluidez operacional e melhor adoção por equipes multiprofissionais. A priorização de otimizações ficou evidenciada no ranqueamento de endpoints exibido no Grafana, com destaque para rotas de acompanhamento e decisão clínica, conforme Figura 11 e Figura 12.

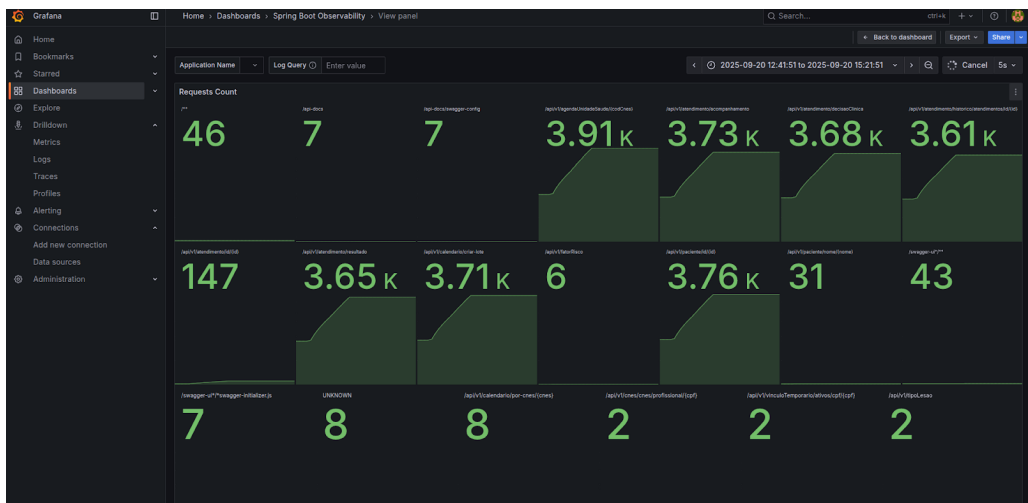


Figura 11. Identificação do gargalo no Grafana (volume por endpoint).



Figura 12. Taxa de requisições ao longo do tempo.

3.5. Testes de *Back-End*

O *back-end* concentrou a execução de algoritmos de inteligência artificial, a manipulação de dados sensíveis e a integração com serviços externos. Testes unitários em funções críticas preservaram a estabilidade de sensibilidade e especificidade dos modelos de IA, além de garantir contratos e consistência de *payloads* nas APIs de cadastro, *upload* e retorno de análises. A integração foi verificada quanto à consistência transacional do PostgreSQL e à resiliência a falhas externas por meio de estratégias como *retry*, *fallback*,

timeouts e *circuit breaker*. Em cenários ponta a ponta, do cadastro à análise por IA e encaminhamento clínico, observou-se robustez na orquestração. Em desempenho e carga, a plataforma apresentou capacidade próxima de 10 requisições por segundo por réplica, mantendo $p_{95} \leq 5$ segundos; acima disso ocorreu degradação típica de gargalo lógico, coerente com a Figura 10 e com os painéis de folga de infraestrutura das Figuras 2 a 9. Essas evidências sustentam as recomendações de otimizar consultas SQL nas rotas de maior custo e ajustar a política do pool de conexões antes de ampliar o número de réplicas.

3.6. Testes de *Front-End*

A interface, ponto de contato com os profissionais, foi avaliada quanto à clareza informacional, carga cognitiva e fluidez de navegação, com sessões remotas e presenciais. A análise levou à simplificação de linguagem e à redução de etapas em tarefas críticas, mantendo consistência visual em múltiplos dispositivos e sistemas operacionais. A verificação de acessibilidade, pautada nas diretrizes WCAG 2.1, contemplou contraste, tipografia, hierarquia e compatibilidade com leitores de tela, ampliando a inclusão digital. A automação com *Cypress* consolidou rotinas de regressão em fluxos de login, cadastro e visualização de resultados; desafios do *Document Object Model* (DOM) foram mitigados com identificadores estáveis para teste, aumentando a confiabilidade das execuções.

3.7. Testes com usuários em ambiente real

Os laboratórios e treinamentos conduzidos entre julho e agosto de 2025, nos municípios de Firminópolis e São Luís de Montes Belos, envolveram médicos, cirurgiões-dentistas e gestores. O MVP inicial mostrou excesso de etapas, que foi corrigido no protótipo com redução de cliques, padronização de terminologia e destaque a alertas clínicos. Observou-se preferência por sessões presenciais, mais ricas para coleta de *feedback*. Também se identificou a necessidade de protocolos operacionais para as sessões, incluindo preparação antecipada do ambiente, cadastro prévio de usuários, definição de casos de uso com antecedência e registro da velocidade de rede, garantindo padronização e repetibilidade dos ensaios.

3.8. Integração das evidências de capacidade e desempenho

A integração de telemetria e experimentos confirma que a arquitetura modular em contêineres está preparada para escalar. A Figura 1 apresenta os componentes principais da API Gateway (NGINX), autenticação stateless (Keycloak/JWT), módulos de *back-end* em Docker, mensageria em RabbitMQ com DLQ e *parking-lot*, PostgreSQL como fonte de verdade transacional, MinIO para objetos binários, e HAPI FHIR com ETL diário para interoperabilidade. A Figura 2 e os painéis 4 a 10 demonstram folga de CPU, memória, I/O e rede, ausência de filas e estabilidade do sistema operacional, descartando gargalos de *hardware* durante os ensaios. A Figura 10 evidencia o ponto de saturação na aplicação quando a carga ultrapassa a região estável, com latências p_{50}/p_{95} em escalada e *throughput* estagnado; e as Figuras 11 e 12 orientam a priorização de otimizações por *endpoint* e por dinâmica temporal de carga. Em termos de SLO/SLA, os achados apontam que uma réplica sustenta aproximadamente 10 requisições por segundo mantendo $p_{95} < 5$ s; para a Região Oeste II de Goiás (com 737 profissionais), um pico plausível de 10% de simultaneidade em uso intensivo projeta $\sim 14,8$ req/s, justificando duas réplicas de *back-end* após otimizações de código e de consultas SQL. O plano de ação recomendado inclui tornar operações longas assíncronas, otimizar índices e planos de consulta nas rotas críticas,

monitorar a utilização do pool de conexões e configurar autoescalonamento orientado por SLOs.

4. Discussão

Os resultados obtidos indicam que a qualidade em sistemas de saúde não é um atributo alcançado apenas ao final do desenvolvimento, mas sim um produto do processo quando práticas de SQA são incorporadas desde a concepção. Por adotar TDD, automação de testes de API, critérios de aceitação e *quality gates* na integração contínua, a Plataforma SobreVidas internalizou uma lógica de “qualidade por desenho”, reduzindo o risco de regressões e falhas críticas que, em contexto clínico, podem impactar segurança do paciente e confiança profissional. Essa abordagem convergente com modelos como o ISO/IEC 25002:2024 sustenta que atributos de qualidade (confiabilidade, segurança, usabilidade, eficiência de desempenho) emergem quando o ciclo de engenharia é orientado por evidências e por validação contínua [ISO/IEC 2024].

A análise integrada de desempenho e capacidade reforça a centralidade de métricas operacionais para informar decisões de arquitetura em saúde digital. A identificação de uma região estável por réplica (~ 10 req/s com $p_{95} < 5$ s), seguida de saturação lógica com o aumento da concorrência, evidencia que gargalos residem principalmente em camadas de aplicação e banco de dados (consultas e concorrência/*pool*), e não na infraestrutura. Na prática, isso desloca o foco de solução: otimizações de consultas, índices e desenho de *endpoints*, além da conversão assíncrona de operações longas, devem preceder a escala horizontal. A tradução dessas evidências em SLOs explícitos viabilizou autoescalonamento e trade-offs de custo mais racionais para implantações no SUS. Ao traduzir essas evidências em SLOs explícitos e acionáveis (por exemplo, metas de p_{95} por domínio funcional), viabiliza-se autoescalonamento guiado por objetivos e *trade-offs* de custo-desempenho mais racionais, um ponto-chave para implantações regionais no SUS.

Sob a ótica da transformação digital em saúde, os achados demonstram que SQA não é apenas um mecanismo técnico; é um instrumento de governança que habilita inovação responsável. A interoperabilidade baseada em HL7® FHIR® mostrou-se estruturante para garantir continuidade informacional entre serviços e níveis de atenção [Walonoski et al. 2018]. Contudo, a mera adoção do padrão não elimina riscos: sem testes sistemáticos de contratos, semântica e integridade, a interoperabilidade pode tornar-se um vetor de falhas. Os nossos resultados, ao priorizarem validação de payloads, consistência transacional e observabilidade ponta a ponta, indicam que interoperabilidade confiável depende tanto de padrões quanto de SQA efetivo. Em paralelo, a IA incorporada a fluxos críticos (triagem e apoio à decisão) só se converteu em valor clínico por estar acoplada a testes unitários e de integração orientados a sensibilidade/especificidade, avaliação de drift funcional e salvaguardas de segurança e privacidade previstas na LGPD, endossando a tese de que IA clínica requer engenharia de qualidade contínua para garantir acurácia, auditabilidade e aceitação pelos usuários [Mehraeen et al. 2025].

Do ponto de vista sociotécnico, os pilotos com profissionais de saúde evidenciaram que a adesão clínica não é um efeito colateral automático de melhorias técnicas. Ajustes em fluxos, terminologias e visibilidade de alertas foram determinantes para reduzir a carga cognitiva e aumentar a confiança de uso, especialmente em ambientes com

recursos heterogêneos de conectividade e dispositivos. A presença ativa da equipe técnica nas sessões de usabilidade facilitou ciclos curtos de aprendizado e tradução de *feedback* em requisitos efetivos, reduzindo o hiato entre intenção de design e prática assistencial. Esses achados dialogam com a literatura que defende abordagens iterativas e participativas para a aceitação de tecnologias em saúde, mostrando que qualidade percebida pelos profissionais emerge quando engenharia, clínica e gestão co-produzem a solução [Patel et al. 2022].

Os nossos resultados apresentaram três contribuições principais. Primeiro, demonstram empiricamente que estratégias integradas de teste (*back-end*, *front-end*, segurança, desempenho) melhoram confiabilidade e adoção em uma plataforma de IA aplicada ao câncer de boca. Segundo, oferecem um quadro operacional de capacidade ancorado em dados (curvas de RPS/latência, perfis de *endpoints*, telemetria de *host*), útil para planejamento de expansão com SLO/SLA explícitos e custo previsível. Terceiro, evidenciam que interoperabilidade e IA só escalam com segurança quando governadas por SQA, combinando verificação técnica, compliance e validação em campo.

Algumas limitações devem ser reconhecidas. A avaliação de desempenho concentrou-se em um perfil de carga representativo dos cenários-alvo; variações clínicas e sazonais podem demandar reparametrizações das metas de SLO. A validação em campo ocorreu em dois municípios, o que pode restringir a generalização a contextos com infraestrutura de rede substancialmente diferente. A aferição de métricas algorítmicas de IA foi conduzida sob versões específicas de dados e modelos; futuras atualizações exigem monitoramento contínuo para mitigar drift e assegurar desempenho estável. Além disso, embora a conformidade com a LGPD tenha sido garantida por meio de criptografia e anonimização, testes de segurança de caixa branca e red teaming orientados a contextos clínicos reais merecem aprofundamento sistemático. Por *red teaming*, entenda o exercício adversarial independente que simula ataques reais, de ponta a ponta, para avaliar pessoas, processos e tecnologia, incluindo a capacidade de detecção e resposta.

Em síntese, a junção entre SQA, transformação digital, interoperabilidade e IA demonstrou-se virtuosa: qualidade de software atuou como habilitador sociotécnico que converteu avanços tecnológicos em valor clínico concreto, elevando a confiança e a adoção da Plataforma SobreVidas. Assim, este estudo apresentou um caminho replicável para que soluções de saúde digital avancem com segurança, eficiência e legitimidade clínica, especialmente em políticas públicas voltadas ao diagnóstico precoce do câncer de boca.

5. Considerações Finais

A experiência com a Plataforma SobreVidas - “Câncer de Boca” reforça que a qualidade em sistemas de saúde é resultado de um ciclo contínuo, multidimensional e integrado. A integração de práticas de engenharia de software (testes unitários, de integração, de desempenho/carga e de segurança) com validação centrada no usuário (usabilidade, acessibilidade e testes em campo) mostrou-se determinante para conciliar a robustez técnica e aderência clínica. Ao fazer uso de SQA desde a concepção, com o TDD, automação de API e *quality gates* na integração contínua, o projeto evitou a “inspeção tardia” como único mecanismo de controle, deslocando a ênfase para qualidade por desenho, aspecto crucial quando se lida com a IA, interoperabilidade clínica e dados sensíveis protegidos

pela LGPD.

Do ponto de vista empírico, os resultados de desempenho revelaram uma região estável de operação por réplica em torno de ~ 10 requisições por segundo com latência p_{95} inferior a 5 segundos; acima desse patamar, verificou-se saturação lógica com crescimento acentuado de p_{50}/p_{95} e surgimento de falhas. A telemetria de infraestrutura indicou folga consistente de CPU, memória, I/O e rede, sustentando a hipótese de que os gargalos observados são predominantemente de aplicação e banco de dados (consultas, concorrência e gestão de conexões), e não de *hardware*. Em paralelo, os pilotos com profissionais de saúde evidenciaram ganhos de usabilidade após simplificação de fluxos, padronização terminológica e realce de alertas clínicos, reforçando a importância de ciclos curtos de iteração entre o MVP e protótipo.

Esses achados têm implicações diretas para sustentabilidade e escalabilidade em saúde pública. Em cenários de expansão regional, como planejado para a Região Oeste II, estimativas conservadoras de 10% de uso simultâneo em condições intensivas projetam $\sim 14,8$ requisições por segundo, o que demanda ao menos duas réplicas de *back-end* para manter SLOs e SLAs de latência. Antes de simplesmente aumentar réplicas, contudo, os dados sugerem que a priorização de otimizações nos *endpoints* mais custosos, a quebra assíncrona de operações longas, a criação de índices e o ajuste do *pool* de conexões devem preceder a escala horizontal para assegurar eficiência de custos e resiliência operacional. Em termos de governança, consolidar SLOs explícitos (por exemplo, $p_{95} < 3\text{--}5\text{ s}$ por domínio funcional) como gatilhos de autoescalamento e como critérios de aceite no pipeline fortalece a previsibilidade do serviço e alinha engenharia às metas clínicas.

Do lado do processo, emergiram lições com valor de generalização. A antecipação de testes de usabilidade com usuários reais, com a participação ativa da equipe técnica, favoreceu ajustes mais assertivos e redução da carga cognitiva. A padronização de identificadores de teste no *front-end* aumentou a estabilidade da automação de regressão em *Cypress* mesmo em DOMs complexos, reduzindo retrabalho e tempo de ciclo. Tratar desempenho e segurança como requisitos de primeira classe, com métricas e *gates* explícitos, mostrou-se decisivo para prevenir regressões que poderiam afetar a segurança do paciente e a precisão diagnóstica.

Como desdobramentos práticos, recomenda-se ampliar a cobertura de automação para cenários fim a fim que representam os fluxos clínicos completos; institucionalizar verificações de segurança no pipeline segundo práticas de DevSecOps, com *quality gates* mínimos e monitoramento contínuo de superfícies de ataque; e expandir os testes de campo para novos municípios com protocolos operacionais padronizados (preparo antecipado do ambiente, cadastros prévios, definição de casos de uso e telemetria de rede), de modo a avaliar a generalização da solução e sua viabilidade em escala. Em paralelo, orienta-se a implementação gradual de processamento assíncrono nos *endpoints* críticos, a otimização de consultas e índices no PostgreSQL e a ativação de estratégias de autoescalamento orientadas por SLOs. Em conjunto, essas ações fortalecem a maturidade da plataforma para uma implantação regional segura, sustentável e escalável, contribuindo para o diagnóstico precoce do câncer de boca e para a qualificação da saúde pública baseada em evidências.

De forma mais ampla, os resultados obtidos com a Plataforma SobreVidas demonstram como a aplicação de práticas de SQA se integra à agenda contemporânea da transformação digital em saúde. A interoperabilidade viabilizada pelo uso de padrões como o HL7® FHIR® assegurou o compartilhamento seguro e padronizado dos dados clínicos, ampliando o potencial de integração com a Rede Nacional de Dados em Saúde (RNDS). A incorporação de inteligência artificial em fluxos críticos, por sua vez, só se tornou viável e confiável porque foi acompanhada por estratégias sistemáticas de teste, que garantiram métricas de desempenho, segurança e auditabilidade. Assim, o elo entre SQA, transformação digital, interoperabilidade e IA não apenas fortaleceu a maturidade tecnológica da Plataforma, mas também criou condições concretas para sua adoção em escala no SUS, mostrando que inovação responsável em saúde digital depende, de forma indissociável, de qualidade, confiabilidade e confiança clínica.

Referências

- Bates, D. W. and Samal, L. (2018). “Interoperability: What Is It, How Can We Make It Work for Clinicians, and How Should We Measure It in the Future?”. *Health Services Research*, 53(5):3270–3277.
- Bernardo, S., Orviz, P., David, M., Gomes, J., Arce, D., Naranjo, D., Blanquer, I., Campos, I., Moltó, G., and Pina, J. (2024). “Software Quality Assurance as a Service: Encompassing the quality assessment of software and services”. *Future Generation Computer Systems*, 156:254–268.
- Din, M. S. U. S. (2025). “Global digital transformation in healthcare: Impact of connected healthcare systems on patient safety”. *Global Journal on Quality and Safety in Healthcare*, 8(2):60–62.
- Graili, P. and Farhoudi, B. (2025). The intersection of digital health and artificial intelligence: Clearing the cloud of uncertainty. *Digital Health*, 11:20552076251315621.
- International Organization for Standardization (2024). Systems and software engineering — systems and software quality requirements and evaluation (square) — quality model overview and usage. IEC 25002.
- Mehraeen, E., Siami, H., Montazeryan, S., Molavi, R., Feyzabadi, A., Parvizy, I., Masjedlu, Z. A., Dehkalani, M. N., Mahmoudi, S., and Ahmadipour, A. (2025). “Artificial intelligence challenges in the healthcare industry: A systematic review of recent evidence”. *Healthcare Technology Letters*, 12(1):e70017.
- Nan, J. and Xu, L. Q. (2023). “Designing interoperable health care services based on Fast Healthcare Interoperability Resources: Literature review”. *JMIR Medical Informatics*, 11:e44842.
- Obigbesan, O., Graham, K., and Benzies, K. M. (2024). “Software testing of eHealth interventions: Existing practices and the future of an iterative strategy”. *JMIR Nursing*, 7:e56585.
- Ronchieri, E. and Canaparo, M. (2023). “Assessing the impact of software quality models in healthcare software systems, journal = Health Systems”. 12(1):85–97.

Stoumpos, A. I., Kitsios, F., and Talias, M. A. (2023). “Digital transformation in health-care: Technology acceptance and its applications”. *International Journal of Environmental Research and Public Health*, 20(4):3407.

World Wide Web Consortium (2018). Web content accessibility guidelines (wcag) 2.1. Disponível em: <https://www.w3.org/TR/WCAG21/>.