UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

ISSN 0103-2569

TESTE DE SOFTWARE E ANÁLISE DE QUALIDADE EM SISTEMAS BIOMÉDICOS - UM MAPEAMENTO SISTEMÁTICO

MISAEL COSTA JÚNIOR
MÁRCIO E. DELAMARO
FÁTIMA L. S. NUNES
RAFAEL A. P. OLIVEIRA

Nº 420

RELATÓRIOS TÉCNICOS

São Carlos – SP
Out./2017

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
DEPARTAMENTO DE SISTEMAS DE COMPUTAÇÃO (SCC)
Caixa Postal 668 – CEP 13560-970 – São Carlos/SP – Fone (16) 3373-9700 – Fax (16) 3371-2238

Teste de software e análise de qualidade em sistemas biomédicos – Um mapeamento sistemático

Relatório Técnico

Aluno: Misael Costa Júnior (misaeljr@usp.br)

Orientador: Prof. Dr. Márcio E. Delamaro (delamaro@icmc.usp.br)

Colaboradores: Prof[a]. Dr[a]. Fátima L. S. Nunes (fatima.nunes@usp.br)

Prof. Dr. Rafael A. P. Oliveira (raoliveira@utfpr.edu.br)

Resumo

Sistemas biomédicos são aqueles que têm a capacidade de apoiar a decisão clínica de profissionais da área médica. É cada vez mais comum que profissionais da área médica tenham o apoio de sistemas biomédicos durante a realização de suas funções de rotina e em tomadas de decisão. No entanto, devido à alta complexidade de dados do domínio de saída de sistemas biomédicos, a aplicação de técnicas e estratégias de teste de software convencionais é limitada, fazendo com que testes manuais e *ad-hoc* sejam realizados. O presente relatório descreve um Mapeamento Sistemático (MS) realizado por meio da análise de 91 estudos publicados entre 1990 e 2016. O relatório relata e discute abordagens, critérios e estratégias de teste de software e análise de qualidade em sistemas biomédicos, sintetizando as evidências disponíveis e identificando as principais questões que envolvem avaliação de qualidade nesse domínio de pesquisa. Para reunir evidências, o processo de condução do MS foi definido baseado em diretrizes existentes. Assim, de acordo com um conjunto de questões de pesquisa, o processo foi realizado em três etapas: planejamento, condução e apresentação dos resultados. O MS fornece uma compreensão estruturada sobre estratégias e abordagens de teste de software e análise de qualidade em sistemas biomédicos. A partir da identificação de evidências, é possível relatar as dificuldades, carências e limitações na validação de sistemas biomédicos. Além disso, uma taxonomia é definida a partir das abordagens de teste de software e análise de qualidade de sistemas biomédicos. O trabalho evidencia a carência de abordagens e estratégias de teste de software e análise de qualidade em sistemas biomédicos, sejam *ad-hoc* ou automatizadas. Assim, o MS estabelece uma base de comparação para abordagens futuras de validação em sistemas biomédicos, configurando sólida contribuição para essa área de pesquisa.

USP – São Carlos
Outubro/2017

Sumário

1 Introdução

1.1 Contextualização e motivação

O desenvolvimento de ferramentas computacionais, atividade cada vez mais complexa, está relacionado ao processo evolutivo pelo qual a computação passou nas últimas décadas. Tecnologias cada vez mais avançadas, linguagens de programação em mais elevado nível de abstração e o surgimento de variados ambientes de execução são alguns dos fatores que contribuem para o aumento da complexidade desses sistemas computacionais.

A aplicação de atividades de Validação, Verificação e Teste de Software (VV&T), nas etapas do processo de desenvolvimento de software, contribui para a garantia de qualidade do software, papel essencial da Engenharia de Software (ES). No entanto, é senso comum entre engenheiros de software e projetistas que duas questões básicas de projeto são intimamente associadas à aplicação do teste de software: (1) custo; e (2) tempo. Em relação ao custo, pode-se afirmar que o emprego do teste traz a necessidade da contratação de equipes especializadas (testadores e projetistas de teste) e aquisição de licenças de ferramentas específicas, elevando significativamente custos de desenvolvimento. De modo similar, dependendo da complexidade do sistema em desenvolvimento, atividades de teste podem demandar no acréscimo de até 50% do tempo total de projeto (Myers et al., 2011). Tal acréscimo do tempo de projeto é consequência do emprego de técnicas e critérios de teste adequados, da reexecução dos dados de testes após cada correção de inconsistências detectadas e, principalmente, da falta de automatização (Bertolino, 2007).

Dependendo do domínio do SUT, a condução de atividades de validação pode se tornar um desafio. Sistemas Texto-Fala (do inglês, *Text-To-Speech* – TTS) (Taylor, 2009), os programas de bio-informática (Chen et al., 2009), os aplicativos baseados em GUI (do inglês, *Graphical User Interface*) (Banerjee et al., 2013), (Banerjee et al., 2003), a maior parte de aplicações CAD (do inglês, *Computer Aided Design*) (Zhang et al., 2008), sistemas estatísticos (Zhou et al., 2004) e sistemas que processem imagens médicas em modelos tridimensionais (Galarreta et al., 2013) são exemplos contemporâneos de sistemas com saídas complexas. Uma grande parcela desses sistemas é composta por programas de apoio à área médica e, geralmente, estratégias de teste para tais sistemas são limitadas pela falta de estratégias sistemáticas. A falta de atividades sistemáticas para a automatização de testes leva à aplicação de testes manuais desempenhadas pelo próprio testador, de modo informal, *ad-hoc* e improdutivo.

Sistemas biomédicos exemplificam um domínio de sistema cujas saídas são em formatos incomuns, tornando as atividades de teste complexas e "escassas". A alta complexidade dos programas nas análises de dados e a consistência na geração de resultados esperados são problemas evidentes nessa área, dificultando a verificação devido à falta de regularidade.

Tais sistemas vêm desempenhando papel fundamental para a evolução da área médica em diversos segmentos.

Devido à alta complexidade de dados do domínio de saída de sistemas biomédicos, a aplicação de técnicas e estratégias convencionais de teste de software é limitada, fazendo com que testes manuais e *ad-hoc* sejam realizados. Dessa forma, não foram encontrados na literatura estudos que indiquem teste sistemático ou análise de qualidade em sistemas biomédicos. Por isso, um Mapeamento Sistemático (do inglês, *Systematic Mapping* – MS) acerca de técnicas e estratégias de teste de software e análise de qualidade de sistemas biomédicos foi conduzido.

De acordo com Kitchenham et al. (2007), um MS parte da realização de pesquisas pré-moldadas para uma Revisão Sistemática (do inglês, *Systematic Review* – RS), mas que, no decorrer da pesquisa, observa-se a baixa relevância de resultados, caracterizando uma granularidade com relação aos possíveis resultados e abrangência na pesquisa. Dessa forma, a principal diferença entre um MS e uma RS está na abrangência da análise dos resultados.

1.2 Objetivos

O MS foi conduzido seguindo as diretrizes propostas por Kitchenham (2004a), Petersen et al. (2015) e Keele (2007). O presente MS tem como objetivo geral *identificar trabalhos na literatura que relatam ou proponham métodos, abordagens e estratégias de teste de software e/ou análise de qualidade em sistemas biomédicos.* Para atingir tal objetivo geral, são definidos os seguintes objetivos específicos:

- identificar métodos, técnicas, critérios, estratégias e abordagens de testes de software e análise de qualidade em sistemas biomédicos;
- definir uma taxonomia de estratégias de teste de software e análise de qualidade em sistemas biomédicos;
- obter uma visão geral dos procedimentos experimentais mais adequados para a validação de sistemas biomédicos;
- relatar as principais limitações e dificuldades na validação de sistemas biomédicos;
- identificar a colaboração entre academia e indústria; e
- avaliar a maturidade da área por meio de dados visuais.

1.3 Organização do relatório

Além dessa seção introdutória, o restante do documento está organizado da seguinte forma:

- A Seção 2 apresenta algumas definições, conceitos e taxonomias relacionadas a teste de software, teste de sistemas com saídas complexas, sistemas biomédicos e teste de sistemas biomédicos. Além disso, a Seção 2 apresenta conceitos relacionados a engenharia de software baseada em evidências;
- A Seção 3 apresenta um mapeamento sistemático acerca de estudos que apresentem métodos ou estratégias de teste de software e avaliação de qualidade em sistemas biomédicos;
- A Seção 4 apresenta uma discussão e análise dos estudos identificados no MS;
- A Seção 5 são apresentadas as lacunas de pesquisa. As lacunas de pesquisa foram definidas de acordo com a análise realizada em cada estudo do MS;
- A Seção 6 apresenta as ameaças à validade identificadas durante a realização do trabalho; e
- A Seção 7 apresenta as considerações finais e trabalhos futuros referentes ao MS.

2 Fundamentação teórica

A presente seção apresenta os conceitos necessários para total compreensão dos temas envolvidos no MS. Assim, essa seção apresenta conceitos associados a teste de software (Seção 2.1), particularidades de estratégias de teste de software para sistemas de saídas complexas (Seção 2.1.1), sistemas biomédicos (Seção 2.2) e estratégias de validação para sistemas biomédicos (Seção 2.2.1). Além disso, são apresentados e discutidos conceitos associados a Engenharia de software baseada em evidências (Seção 2.3).

2.1 Teste de software

Durante o processo de desenvolvimento de um software, o programador pode cometer diversos enganos. Especificação incompleta, erros de escrita de código, escolhas de tecnologias e linguagens de programação são alguns dos motivos que levam o SUT a um estado inconsistente. Outros motivos também podem colaborar para falhas por parte do programador na escrita de um software como, por exemplo, o programador pode ter errado na interpretação de um requisito do sistema e, por isso, o código foi escrito da maneira incorreta. Dessa forma, atividades VV&T são consideradas fundamentais (Delamaro et al., 2016). A Validação assegura que o produto sendo desenvolvido corresponde ao produto correto, conforme os requisitos do usuário, enquanto a Verificação tende a assegurar consistência, completude, corretude do produto em cada fase e entre fases consecutivas do ciclo de vida. Por sua vez, o Teste de Software é um processo em que um programa é executado com o objetivo de revelar erros, aumentando a confiabilidade do sistema.

Atividades VV&T estão diretamente relacionadas com garantia de qualidade, objetivando a detecção de erros em sistemas. Para fins conceituais, a IEEE 610.12 (IEEE, 1990) padroniza termos utilizados na área de Teste de Software: (i) **Engano (*mistake*):** codificação incorreta causada por erro humano (programador); (ii) **Defeito (*fail*):** passo, processo ou definição de dados incorretos, podendo configurar diversos; (iii) **Erros (*error*)**: diferença entre o valor obtido e o esperado; e (iv) **Falha (*failure*)**: parte do código é executada levando o sistema a um estado inesperado, diferente da saída esperada.

Assim, o teste pode ser realizado para avaliação de diversos cenários da construção do software, desde propriedades de visualização do usuário a falhas provenientes da má interpretação da especificação do sistema. A detecção de falhas de modo prematuro permite aos desenvolvedores a realização de correções antes da entrega do produto final, melhorando a qualidade do SUT e evitando desperdício de tempo e dinheiro.

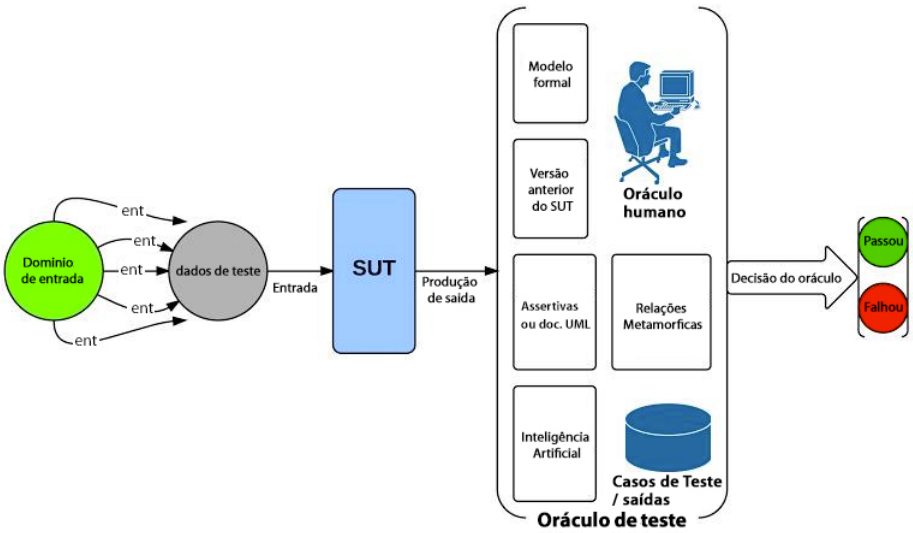
O teste de software é uma atividade dinâmica que consiste em executar um SUT, em um ambiente controlado, para aumentar a confiabilidade que o SUT se comporta conforme sua especificação (Bertolino, 2003). De acordo com Myers (2004), as atividades de teste de software são um processo, ou série de processos, que têm a função de executar um sistema com a intenção de encontrar possíveis erros. De um modo pragmático, as atividades de teste de software consistem em exercitar uma aplicação, observando seus resultados, comparando-os com alguns valores de resultados esperados e relatando suas consequências (Hoffman, 2001). Esse fato faz com que as indústrias de desenvolvimento de software considerem a qualidade um fator essencial. Dependendo do domínio de saída do SUT, a atividade de teste passa a ser uma atividade muito valorizada no processo de desenvolvimento, e isso é predominante em sistemas críticos. Nesse caso, a ausência de atividades de teste sistemáticas podem ser a causa de resultados inesperados.

Dentre os paradigmas de realização de atividades de teste, destacam-se: (i) Teste manual e (ii) Teste automatizado. Em relação ao teste manual, uma pessoa (testador) alimenta o sistema em teste com entradas, avaliando os resultados do processamento de tais entradas nesse sistema (Hoffman, 2001). O teste automatizado é realizado com o apoio de ferramentas desenvolvidas com o objetivo de contribuir e automatizar na busca de qualidade de sistemas em desenvolvimento (Rafi et al., 2012).

A Figura 1 apresenta uma estrutura genérica associada ao fluxo de tarefas de teste de software. Tal estrutura contempla diversas etapas que podem variar dependendo da natureza do sistema em teste. Entretanto, o fluxo do teste habitualmente se inicia a partir da análise do "domínio de entradas" do SUT, que representa todos os possíveis dados de entrada. Por meio dessa análise, o projetista do teste seleciona algumas entradas específicas que irão compor os "dados de teste". Em geral, um projetista de teste seleciona tais dados ("*ent*") cuidadosamente a partir de algum critério de alguma técnica de teste.

Essa seleção de dados de teste é fundamental para a produtividade do teste, haja vista que muitas vezes o domínio de entradas do SUT é infinito e o teste do sistema com todos os elementos do SUT é impraticável (Myers et al., 2011).

Figura 1: Fluxo genérico de teste de software



Em seguida, o SUT é executado a partir das entradas "*ent*", produzindo uma saída qualquer comumente chamada de "saída real". É exatamente nessa etapa do teste que agem os "oráculos de teste" – responsáveis por avaliar as saídas do SUT diante de alguma referência. A Figura 1 apresenta diferentes recursos técnicos que podem ser explorados como fonte de informação para os oráculos: (1) modelos formais executáveis que representem o comportamento do SUT (Yin et al., 2012); (2) versões anteriores confiáveis do SUT que possam ser uma fonte segura acerca dos resultados de teste (Yu et al., 2013); (3) assertivas previamente computadas (Cheon, 2007); (4) documentações UML (do inglês, *Unified Modeling Language*) (Briand e Labiche, 2002); (4) relações metamórficas (Asrafi et al., 2011); (5) inteligência artificial (Frounchi et al., 2011); (6) saídas esperadas de casos de teste previamente computados (Javed et al., 2007); e (7) o próprio testador verificando saídas e comportamentos do SUT, configurando um oráculo humano (McMinn et al., 2010).

2.1.1 Teste de sistemas com saídas complexas

Domínios de saída complexa limitam, severamente, a aplicabilidade das atividades de Teste de Software automatizado (Chan e Tse, 2013). Devido à complexidade associada com os dados processados pelo SUT, técnicas de testes tradicionais são ineficientes e,

muitas vezes, impraticáveis (Oliveira et al., 2014). Assim, estratégias para a automação de testes em sistemas com saídas não-triviais são um desafio. Além disso, a maior parte das abordagens existentes são definidas para um domínio específico, tendo pouca ou nenhuma padronização. Nesse contexto, ao testar os sistemas de saídas complexas, os testadores geralmente "ignoram" etapas de testes tradicionais, como a análise de domínio de entrada do SUT para projetar um conjunto de casos de testes válidos, prevendo os efeitos das entradas por meio de saídas explícitas, mitigando o compartilhamento de conhecimento neste campo. A complexidade associada à saída do SUT leva ao emprego de estratégias subjetivas e *ad-hoc* de teste (Hinterleitner et al., 2011). De um modo geral, um ser humano exercita o SUT, observando as suas saídas e preenchendo um protocolo de aspectos, tornando as atividades de teste difícil, tediosa e custosa.

Entre os sistemas com saídas complexas e cujas saídas são em formatos incomuns, destacam-se os sistemas biomédicos. É cada vez mais comum profissionais da área médica precisarem de recursos tecnológicos que os auxiliem na tomada de decisão. Nesse cenário, a indústria de cuidados médicos necessita da tecnologia para realização de muitas de suas funções, que vão desde a gestão de informações hospitalares (Kawamura et al., 2014) a sistemas de monitoramento remoto de pacientes em tratamento (Postolache et al., 2012). Consequentemente, é essencial que esses sistemas apresentem resultados consistentes e com qualidade, facilitando o trabalho dos profissionais e oferecendo assistência com qualidade aos pacientes.

2.2 Sistemas biomédicos

Sistemas biomédicos são aqueles que têm a capacidade de apoiar a decisão clínica de profissionais da área médica (Jetley et al., 2006). Sistemas de diagnóstico de imagens de ultrassonografia (Filho et al., 2014; Jiang et al., 2011; Wang et al., 2014), sistemas de Processamento de Linguagem Natural (PLN) (Cohen et al., 2013; Zheng et al., 2015), Sistemas de Informações Hospitalares (IH) (Kawamura et al., 2014; Paech e Wetter, 2008) e sistemas de diagnóstico radiológico (Ward et al., 2009) são exemplos de sistemas que apoiam profissionais da área médica em funções clínicas essenciais. Sistemas biomédicos englobam um conjunto de características fundamentais que necessitam de verificação e validação para oferecer qualidade ao sistema como um todo.

A integração da tecnologia na medicina tem evoluído como uma nova área de pesquisa na última década. O surgimento dessa integração é, em parte, devido aos inúmeros desafios relacionados à prática da medicina ao longo dos anos. Trabalhos médicos mais eficientes, migração de bases de dados do papel para ambientes digitais, redução de erros em diagnósticos e gerenciamento de informações relacionadas aos pacientes são alguns dos desafios enfrentados pelos profissionais da área médica.

Duas características comuns em sistemas biomédicos é quanto à necessidade da geração dos resultados esperados e a alta complexidade dos programas de análise de dados. Falhas em sistemas biomédicos são geralmente deterministas e podem não ser facilmente reproduzidas. Adicionalmente, são poucos os erros que ocorrem inesperadamente sem apresentar causas especificas e que se manifestam devido a uma combinação rara de eventos e/ou uma sequência de eventos e condições contributivas (Jetley et al., 2006).

Existem diversas iniciativas que impulsionam a adoção da tecnologia da informação em áreas da saúde, embora havendo necessidades a serem cumpridas como, por exemplo: diminuir custos, melhorar a segurança do paciente e melhorar a qualidade dos cuidados médicos (Hoyt et al., 2008). Dessa forma, é essencial o surgimento de novas iniciativas de busquem uma melhor integração da tecnologia da informação em áreas da saúde, impulsionando qualidade e segurança ao paciente e ao profissional da área médica. De acordo com Hoyt et al. (2008) e Siddiqi et al. (2009), os principais objetivos da informática médica são:

- Reduzir os erros de diagnóstico médico e litígios resultantes. O desenvolvimento de estratégias em informatizar atividades de análise de diagnóstico tendem a auxiliar o profissional da área médica em tomar melhores decisões baseadas no diagnóstico do paciente;
- Melhorar a comunicação entre agentes, instituições de saúde e pacientes. Banco de dados compartilhados e maior rapidez no envio da informação médica são algumas das alternativas que visam tornar a decisão médica mais ágil e, de certa forma, trazendo maiores benefícios ao paciente;
- Melhorar a qualidade dos cuidados médicos. O desenvolvimento de aplicações que automatizam a decisão do profissional da área médica torna o processo mais ágil e aumenta a qualidade no diagnóstico;
- Padronizar as informações geradas por diferentes médicos e instituições de saúde. A integração da informática da medicina vai tornar possível atividades que visam compartilhar informações entre profissionais da área médica e instituições de saúde;
- Melhorar a produtividade do clínico. Um dos fatores que mais prejudica a agilidade em resultados de diagnósticos médicos está relacionado a falta de produtividade na realização de atividade ou na geração de resultados de atividades médicas; e
- Proteger a privacidade e garantir a segurança dos pacientes, médicos e instituições. A integração da informática com a medicina torna as informações de cada paciente, como também de instituições de saúde, mais seguras, tornando possível o compartilhamento apenas de informações de interesse e importantes para o andamento de algum procedimento médico.

2.2.1 Teste de sistemas biomédicos

A difusão da computação nas diversas áreas de ciências biomédicas, especialmente na medicina, trouxe uma crescente agilidade no desenvolvimento de atividades médicas. No entanto, tal integração aumentou a complexidade do software biomédico, tornando um grande desafio garantir a confiabilidade na implementação de sistemas nesse domínio (Kamali et al., 2015).

Wallace e Kuhn (2001) realizam uma análise de falhas em dispositivos médicos que podem não ter causado nenhuma morte ou lesão, mas foram causas de inúmeros *recall's* por parte dos fabricantes em relação ao software dos dispositivos médicos nos últimos 15 anos. Assim, é realizada uma classificação das falhas e de seus respectivos métodos de prevenção. Segundo Wallace e Kuhn (2001), diversas falhas foram detectadas nas quais envolviam procedimentos tradicionais de desenvolvimento e garantia de qualidade em software como, por exemplo: para práticas de desenvolvimento e manutenção foi recomendado especificação de requisitos, rastreabilidade dos artefatos de desenvolvimento, software de gerenciamento de configuração, análise na mudança de impacto, atividades de treinamento e etc.

Koru et al. (2007) apresenta um levantamento a partir de desenvolvedores de software no domínio de OS (do inglês, *Open-Source*) biomédicos para entender as práticas de controle de qualidade e as estratégias e métodos utilizados durante o desenvolvimento dos softwares. Dessa forma, as categorias de práticas de revisões e testes que são, efetivamente, utilizadas em projetos OS biomédicos foram analisadas. Assim os autores constataram que atividades de revisões de código em pares são pouco utilizadas e, quando utilizadas, não ocorrem de forma sistemática. As revisões em pares são um meio eficaz de encontrar erros no software ou enganos por parte do programador, embora sejam complementares às atividades de teste (Koru et al., 2007).

A literatura não relata estratégias sobre padronização de abordagens e métodos sistemáticos de teste de software e análise de qualidade em sistemas biomédicos e, quando descritas, demandam mais recursos e tempo que as abordagens tradicionais, tornando um desafio garantir a qualidade nesse domínio de sistema. Dessa forma, é essencial que atividades VV&T sejam realizadas de forma planejada e sistemática, oferecendo qualidade em sistemas que apoiem decisões clínicas.

2.3 Engenharia de software baseada em evidências

Uma área de pesquisa amadurece devido ao aumento acentuado no número de relatórios e resultados disponibilizados, e torna-se importante para resumir e fornecer uma visão geral de um determinado tópico de pesquisa (Petersen et al., 2008, 2015). Dessa

forma, diversas metodologias especificas foram propostas objetivando a busca por estudos secundários. Kitchenham (2004b) apresenta uma tendência que busca um foco maior sobre novos métodos de investigação, empíricos e sistemáticos, favorecendo a ES. A Engenharia de Software baseada em Evidências (do inglês, *Evidence-Based Software Engineering* – EBSE) visa aplicar uma abordagem baseada em evidências para a pesquisa e a prática de ES (Kitchenham et al., 2004).

Pesquisa e prática baseada em evidências é um termo que inicialmente foi desenvolvido na medicina, percebendo-se que a opinião de especialistas com base no conselho médico não era tão confiável quanto o conselho com base na acumulação de resultados de experimentos científicos (Kitchenham et al., 2004). Nesse cenário, diversas disciplinas foram beneficiadas em adotarem práticas semelhantes como a psiquiatria, enfermagem, política social e educação.

Para fins conceituais, a evidência é definida como uma síntese dos estudos científicos de melhor qualidade sobre um tema ou questão específica de pesquisa. Vale ressaltar que, com o amadurecimento de uma área de pesquisa, o número de estudos e resultados cresce significativamente (Petersen et al., 2008, 2015). Um dos principais métodos da EBSE são as Revisões Sistemáticas da Literatura (do inglês, *Systematic Literature Review*), classificadas como estudos secundários, já que, dependem dos estudos primários utilizados para revelar evidências e construir o conhecimento. Sendo assim, a literatura diferencia as SLR's em dois tipos: As Revisões Sistemáticas da Literatura (RSL); e os estudos de MS (Petticrew e Roberts, 2008).

Uma RSL é um meio de identificar, avaliar e interpretar todas as pesquisas disponíveis relevantes para uma determinada questão de pesquisa ou tópico de uma área, ou fenômeno de interesse (Kitchenham, 2004a). Dessa forma, os estudos relevantes a serem considerados em uma Revisão Sistemática são chamados de estudos primários, assim uma Revisão Sistemática é chamada de estudo secundário.

Por sua vez, um MS é uma vertente de SLR, na qual se realiza uma revisão mais ampla dos estudos primários, objetivando identificar quais evidências estão disponíveis, bem como identificar lacunas no conjunto dos estudos primários onde seja direcionado o foco de revisões sistemáticas futuras e identificar áreas onde mais estudos primários precisam ser conduzidos (Kitchenham, 2004b). Um MS fornece uma estrutura do tipo de relatórios de pesquisa e resultados que têm sido publicados objetivando categorizá-los. Em geral, fornece um resumo visual, do mapa, dos resultados obtidos. Em contraste ao método de RSL, que além da identificação visa também analisar, avaliar e interpretar todas as pesquisas disponíveis para uma questão determinada, o MS fornece uma abordagem mais abrangente e concisa em relação aos estudos primários existentes para o tema investigado.

Petersen et al. (2008) apresenta o processo do MS dividido em três fases: (i) Planejamento; (ii) Condução; e (iii) Apresentação. Na fase de **Planejamento**, os objetivos e as questões de pesquisa são definidos. Além disso, é nesta fase que ocorre a definição do protocolo de pesquisa, mecanismos que definem a proposta e procedimentos de uma revisão, definição dos critérios de inclusão e exclusão dos artigos, critérios de qualidade, meio pelo qual irá auxiliar o pesquisador na interpretação dos resultados, etc.

A fase de **Condução ou Execução** do mapeamento objetiva gerar os resultados finais e artefatos intermediários, tais como: o registro inicial de busca, a lista de publicações selecionadas, os registros das avaliações de qualidade e os dados extraídos para cada uma das publicações selecionadas. Assim, é a fase na qual ocorre a coleta e análise dos estudos primários, ou seja, publicações e outras fontes de estudo relacionadas à questão de pesquisa. Nesse cenário, esta fase pode resumir tópicos relacionados a: (i) identificação da pesquisa; (ii) seleção dos estudos primários; (iii) avaliação de qualidade; (iv) extração e monitoração dos dados; e (v) síntese dos dados.

Por fim, na fase de **Apresentação** dos resultados é realizada uma análise e descrição do MS em referência aos resultados obtidos. Os resultados dos estudos preliminares que satisfazem o objetivo da avaliação são extraídos e sintetizados. Nessa etapa também consiste, de acordo com a análise e síntese dos dados, na escrita do relatório do MS. Kitchenham (2004a) estabelece recomendações acerca da estrutura e dos conteúdos que possam ser inseridos em um relatório técnico.

3 Mapeamento sistemático

O presente MS foi conduzido seguindo as diretrizes propostas por Keele (2007); Kitchenham (2004a); Petersen et al. (2008, 2015). O MS foi conduzido durante o período de cinco meses (Agosto/2015 a Dezembro/2015) e atualizado durante o período de quatro meses (Setembro/2016 a Dezembro/2016). A partir da identificação dos estudos mais relevantes, espera-se apresentar e discutir as principais práticas, limitações e dificuldades em atividades de análise, validação e controle de qualidade em sistemas biomédicos.

Na literatura, encontram-se trabalhos relacionados a diversos domínios de sistemas biomédicos. Dessa forma, um MS promove a compreensão estruturada de abordagens e estratégias para atividades de teste de software e análises de qualidade em sistemas biomédicos. A partir da identificação de evidências, é possível relatar as dificuldades e limitações identificadas na validação de sistemas biomédicos. Para aumentar a replicabilidade do estudo, uma planilha online foi disponibilizada na *Web* apresentando informações específicas de cada estudo obtido por meio do MS[1].

[1] https://goo.gl/2pwp87

O MS foi realizado sendo conduzido seguindo três etapas: (i) planejamento, (ii) condução e (iii) apresentação dos resultados. A seguir, tais etapas são apresentadas e a metodologia utilizada na condução do MS é descrita.

3.1 Planejamento

O planejamento do MS foi realizado de acordo com as diretrizes descritas em (Keele, 2007; Kitchenham, 2004a; Petersen et al., 2008, 2015). A presente seção apresenta os principais tópicos relacionados à fase de planejamento do MS.

3.1.1 Objetivos

O objetivo geral do MS é identificar estudos na literatura que relatam ou propõem métodos, abordagens ou estratégias de teste de software e/ou análise de qualidade em sistemas biomédicos. Para atingir tal objetivo geral, são definidos os seguintes objetivos específicos:

- identificar métodos, técnicas, critérios, estratégias e abordagens de testes de software e análise de qualidade em sistemas biomédicos;
- definir uma taxonomia de estratégias de teste de software e análise de qualidade em sistemas biomédicos;
- obter uma visão geral dos procedimentos experimentais mais adequados para a validação de sistemas biomédicos;
- relatar as principais limitações e dificuldades na validação de sistemas biomédicos;
- identificar a colaboração entre academia e indústria; e
- avaliar a maturidade da área por meio de dados visuais.

3.1.2 Questões de pesquisa

Como uma estratégia de seleção e busca de possíveis estudos que se adequam aos objetivos deste MS, as seguintes questões de pesquisa (QP's) foram formuladas com base em seis QP's principais:

- **QP 1**: De acordo com a literatura, existem técnicas, critérios e abordagens de teste de software e análise de qualidade que são usados em sistemas biomédicos?
  - **QP 1.1**: Existem estratégias *ad-hoc* ou informais de teste de software e análise de qualidade aplicadas em sistemas biomédicos?
  - **QP 1.2**: Existem estratégias sistematizadas e/ou automatizadas, incluindo o uso de ferramentas e *frameworks*, de teste de software e análise de qualidade aplicadas em sistemas biomédicos?
- **QP 2**: Quais procedimentos experimentais são mais utilizados e adequados para a validação de sistemas biomédicos?

  - **QP 2.1**: As validações ocorrem em sistemas biomédicos reais ou *toys*?
  - **QP 2.2**: Quais são os domínios dos sistemas biomédicos envolvidos?
- **QP 3**: Quais estratégias e abordagens, *ad-hoc* ou sistematizadas, de teste de software e análise de qualidade, são mais utilizadas para validar domínios de sistemas biomédicos?
- **QP 4**: Existe colaboração entre Indústria e Academia no desenvolvimento de estratégias de validação para sistemas biomédicos?
- **QP 5**: Quais são os desafios e limitações de validação mencionados nos estudos primários?
- **QP 6**: Qual é o impacto da pesquisa relacionada à validação de sistemas biomédicos?
  - **QP 6.1**: Qual a localização das universidades e institutos de pesquisa que mais pesquisam na área têm o maior impacto de pesquisas relacionadas à validação de sistemas biomédicos?
  - **QP 6.2**: Quais são os meios de divulgação utilizados para disseminar as estratégias desenvolvidas para validar domínios de sistemas biomédicos?

Com base em Kitchenham (2004a) e Keele (2007), um critério é indicado para estruturar as questões de pesquisa, nomeado de PICO (C) (do inglês, *Population, Intervention, Comparison, Outcome, Context*). PICO (C) é um critério proposto por Petticrew e Roberts (2006) utilizado para estender as orientações médicas tradicionais, fornecendo maior compreensão na especificidade das questões de pesquisa. Assim, *Population* (População) corresponde ao grupo a ser observado no estudo. *Intervention* (Intervenção) indica o que deve ser observado na pesquisa, em uma dada população. *Comparison* (Comparação) é o conjunto de dados iniciais de que o pesquisador já possui. Outcome (Resultados) representam o que é esperado para chegar no final da revisão sistemática. Por fim, *Context* (Aplicação) indica que tipo de profissional e áreas serão beneficiadas com os resultados da revisão sistemática.

Assim, as QP's podem ser estruturadas da seguinte maneira:

- População: pesquisas sobre teste de software e análise de qualidade;
- Intervenção: métodos, técnicas, critérios, estratégias, abordagens e procedimentos experimentais, em geral, utilizados em teste de software e análise de qualidade;
- Comparação: conjunto de trabalhos relacionados a teste de software e trabalhos que relatam o desenvolvimento de sistemas biomédicos;
- Resultados: técnicas, critérios, estratégias e abordagens de teste de software ou análise de qualidade, sistemáticas ou informais, em sistemas biomédicos e procedimentos experimentais adequados para validar sistemas biomédicos; e
- Aplicação: pesquisadores e desenvolvedores envolvidos com o desenvolvimento de sistemas biomédicos.

3.1.3 Estratégia de seleção de estudos primários

Dadas as questões de pesquisa, decidiu-se, então por um processo de seleção dos estudos primários realizado em duas etapas: (i) busca automática e (ii) busca manual. A busca automatizada consiste do uso de `Strings` de busca em pesquisas nas bases indexadas de artigos científicos. Em complemento ao processo automatizado, a busca manual está relacionada à realização do processo de busca em referência dos estudos primários, *Snowballing* (análise de referências bibliográficas de estudos selecionados), e estudos indicados por especialistas.
A seguir, as etapas para busca e seleção dos estudos primários são descritas:

- Bibliotecas de pesquisa: a Tabela 1 apresenta as bibliotecas de pesquisa utilizadas para busca de estudos primários do MS. A escolha das bibliotecas de pesquisa está relacionada à experiência dos pesquisadores envolvidos no MS;

Tabela 1: Bibliotecas de pesquisa utilizadas no MS.

Bibliotecas	**Endereço**
ACM Digital Library	http://dl.acm.org/
IEEE Xplore	http://ieeexplore.ieee.org/
Compendex	http://www.engineeringvillage.com/
Scopus	http://www.scopus.com/

- Idioma dos estudos: optou-se por não considerar no MS estudos que estejam escritos em idiomas não-alfanuméricos (japonês, chinês, mandarim, candii e etc). A escolha do idioma dos estudos está relacionada a experiência dos pesquisadores envolvidos na condução do MS;
- Palavras-chave: as palavras-chave utilizadas para a busca dos estudos primários foram elaboradas a partir dos cinco principais termos da pesquisa: critérios ou abordagens, teste de software ou análise de qualidade, e sistemas biomédicos. A Figura 2 apresenta a `String` de busca definida a parte das palavras-chave;
  - **Critérios**: *technique, techniques, method, criterion, strategy e approach*;
  - **Teste de software e análise de qualidade**: *"software testing" e "software quality"*; e
  - **Sistemas biomédicos**: *biomedics, biomedical, "medical software", "medical systems" e "medical devices".*

Figura 2: String de busca do MS.

(“technique” OR “techniques” OR “method” OR “criterion” OR “strategy” OR “approach”) AND (“software testing” OR “software quality”) AND (“biomedics” OR “biomedical” OR “medical software” OR “medical systems” OR “medical devices”)

- Trabalhos relacionados: após a seleção dos estudos primários, foi realizada uma pesquisa nos trabalhos relacionados a estes estudos. De acordo com Kitchenham (2004a), a prática do *Snowballing*, com a busca de estudos nas referências dos estudos primários, pode “revelar” mais artigos relevantes; e
- Estudos sugeridos/indicados por especialistas: estudos sugeridos/indicados por especialistas, no domínio de teste de software, foram considerados como fonte de pesquisa no MS. Embora a escolha dessa opção possa causar um possível viés, é necessário considerar os estudos sugeridos/indicados objetivando evitar a perda de estudos que possam ser boa evidência para o MS. Além disso, tais estudos foram utilizados como base para calibração da String de busca.

3.1.4 Critérios e procedimentos para seleção dos estudos primários

Um passo importante na seleção dos estudos primários do MS é a definição dos critérios de inclusão (CI) e dos critérios de exclusão (CE). A definição dos critérios de seleção é uma tarefa essencial para assegurar a viabilidade da pesquisa (Kitchenham, 2004a). Dessa forma, os seguintes critérios de seleção foram definidos:

- **Critérios de inclusão:**
  - *CI1*: Qualquer estudo primário que aborde ou discuta (direta ou indiretamente) métodos ou abordagens de teste de software ou análise de qualidade em sistemas biomédicos;
  - *CI2*: Qualquer estudo primário que relate ou proponha um novo método ou abordagem, seja *ad-hoc* ou automatizada, de teste de sistemas biomédicos ou análise de qualidade;
  - *CI3*: Qualquer estudo primário que apresente ou discuta procedimentos experimentais mais adequados para validar sistemas biomédicos; e
  - *CI4*: Estudos primários cujo manuscrito esteja completamente disponível na *Web*.
- **Critérios de exclusão:**

- *CE1*: Estudos que lidam ou mencionam sistemas biomédicos, porém não apresentam avaliações formais ou informais da qualidade apresentada por eles;
- *CE2*: Estudos escritos em idiomas não-alfanuméricos (japônes, chinês, mandarim, candii e etc) ou em idiomas não conhecidos pelos autores (polonês, alemão, etc);
- *CE3*: A versão completa do manuscrito não está disponível na *Web* (após solicitar o autor) ou não foi publicada;
- *CE4*: A pesquisa é uma versão anterior de um estudo mais completo. Ex: *conference paper* que foi expandido para *journal paper*; e
- *CE5*: O estudo é uma monografia, tese, dissertação ou um trabalho de conclusão de curso.

3.1.5 Processo de seleção dos estudos primários

O processo de seleção de estudos primários foi dividido em três etapas:

- **Seleção preliminar**:
  - Inicialmente a *String* de busca foi customizada e executada em cada biblioteca de pesquisa, assim como apresentado na Tabela 1;
  - Durante o processo de aplicação da *String* de busca, uma avaliação da precisão da *String* foi realizada por supervisão de especialistas em revisões bibliográficas;
  - Realização de uma análise com base no Título, *Abstract* e *Keywords* de cada artigo por meio dos CI e CE, assim como definidos na Seção 3.1.4. Dessa forma, foi possível classificar um estudo como relevante ou não-relevante; e
  - Os artigos retornados foram inseridos e gerenciados por meio da ferramenta BibDesk[2].
- **Primeira seleção**:
  - Os estudos primários foram analisados por meio de uma leitura completa, aplicando os CI e CE;
  - Após a leitura completa, os estudos foram classificados como relevantes ou não relevantes. Os estudos primários classificados como relevantes, incluídos no MS, passaram por um processo de extração de dados. Os estudos classificados como

[2]O seguinte link se refere ao software de gerenciamento de referências BibDesk, OS, disponível para MAC OS X: http://bibdesk.sourceforge.net/

não relevantes, foram documentados e uma justificativa do motivo do estudo não ter sido incluído no MS foi descrita. A Tabela 2 apresenta as informações extraídas durante a análise de cada estudo primário; e

– Os estudos classificados como relevantes, incluídos, e os estudos classificados como não relevantes, não incluídos, foram documentados e disponibilizados por meio de uma planilha online.

Tabela 2: Informações extraídas durante a análise dos estudos primários.

Informações extraídas
Identificação do artigo [Título, Autores, biblioteca de publicação, Ano, Tipo de publicação e País]
Proposta do estudo [Objetivo/Propósito do estudo]
Ferramentas de suporte [Ferramentas de suporte utilizadas/mencionadas no estudo]
Domínio de aplicação do sistema [Área de aplicação do sistema utilizado/mencionado no estudo]
Abordagem do domínio de teste de software [Automatizado e/ou Manual]
Características relatadas no estudo [Usabilidade, Adaptabilidade, Efetividade de comunicação, etc]
Origem do estudo [Academia/Indústria]
Tipo de sistema testado [Sistema real ou Protótipo]
Fase de teste de software citada no estudo [Unidade, Integração, Sistema e/ou Regressão]
Trabalhos futuros [Os autores mencionam trabalhos futuros?]
Estratégia do estudo [Estudo empírico, Estudo de caso ou Prova de Conceito]
Metodologia aplicada [Metodologia aplicada no estudo]
Limitações e dificuldades [Dificuldades mencionadas no estudo pelos autores]

- ***Snowballing***: o termo *Snowballing* se refere ao processo de análise das listas de referência de estudos primários, podendo ser utilizado para inclusão e seleção de novos estudos na SLR (Kitchenham et al., 2007). Segundo Kitchenham et al. (2007) e Kitchenham (2004b), a prática do *Snowballing* aumenta as possibilidades de "revelar" mais estudos relevantes para a pesquisa. Dessa forma, as seguintes atividades foram realizadas objetivando encontrar mais estudos relevantes nas referências dos estudos primários:

  – Após a seleção dos estudos primários, foi realizada uma pesquisa nas referências dos estudos, objetivando encontrar mais estudos relevantes para o MS;

  – Os CI e CE foram aplicados da mesma forma durante essa etapa e as atividades envolvidas durante essa seleção envolveu uma busca preliminar com base no Título e *Abstract*. Por fim, os estudos foram avaliados com base na leitura completa; e

  – A seleção dos artigos nessa etapa foi realizada com a supervisão de um especialista em RSL, objetivando mitigar possíveis vieses.

3.2 Condução

A fase de condução ou execução de uma RSL objetiva gerar os resultados finais e artefatos intermediários, tais como: identificação das fontes de pesquisa, seleção dos

trabalhos, extração dos dados e monitoramento do progresso e síntese dos dados. Assim, é a fase na qual ocorre a coleta e análise dos estudos primários, ou seja, publicações e outras fontes de estudo relacionadas às QPs (Kitchenham, 2004a). A presente seção apresenta os principais tópicos relacionados à fase de condução ou execução do MS.

3.2.1 Seleção preliminar

Nessa etapa, foi realizado o processo de customização e calibração da *String* de busca para aplicação em cada biblioteca de pesquisa, resultando em um conjunto de estudos primários. A etapa de seleção preliminar é constituída pelas seguintes atividades: construção da *String* de busca, pesquisa de estudos em quatro bases de pesquisa e eliminação de estudos. Documentos retornados que tinham como objetivo promover eventos científicos não foram contabilizados.

String de busca:

De acordo com o processo de definição dos termos de pesquisa e seleção dos sinônimos, foi definida uma String de busca a ser aplicada a cada uma das bases de pesquisa, objetivando o retorno de estudos relevantes para o MS. Assim, para cada base em particular foi aplicado um processo de customização e calibração da *String*. Para todos os termos principais e relacionados, foram adicionados operadores lógicos OR e AND para a combinação dos termos.

O processo de aplicação da *String* nas bibliotecas foi precedido por uma processo de customização, uma vez que cada biblioteca possui particularidades na definição da *String*. A seguir são discutidos mais detalhes sobre o processo de aplicação da *String* nas bibliotecas de pesquisa.

Aplicação da *String* nas bibliotecas de pesquisa:

- **ACM Digital Library**: A busca na ACM Digital Library foi realizada durante os meses de Agosto/2015 a Setembro/2015 e Setembro/2016 a Outubro/2016, retornando 18 estudos ao total. A *String* de busca foi customizada para se adequar ao padrão exigido por cada base, e a pesquisa foi realizada com base no Título, *Abstract* e *Keywords*. O processo de customização da *String* para a ACM é uma atividade mais trabalhosa em relação às demais bases, exigindo mais atenção na customização da *String*. A *String* com o número de estudos retornados pode ser visualizada com mais detalhes na Tabela 3.

  Após obter o resultado da pesquisa, foi realizado o download do arquivo do tipo *bibtex* contendo o Titulo e *Abstract* de cada artigo, por meio da funcionalidade de exportação disponível na ACM.

Tabela 3: String de busca com o respectivo número de estudos retornados.

String	**Retorno**
Title:((technique or techniques or method or criterion or strategy or approach) and ("software testing" or "software test" or "software quality") and (biomedics or biomedical or "medical software" or "medical systems" or "medical devices")) or **Keywords:**((technique or techniques or method or criterion or strategy or approach) and ("software testing" or "software test" or "software quality") and (biomedics or biomedical or "medical software" or "medical systems" or "medical devices")) or **Abstract:**((technique or techniques or method or criterion or strategy or approach) and ("software testing" or "software test" or "software quality") and (biomedics or biomedical or "medical software" or "medical systems" or "medical devices"))	18

- **IEEE Xplore**: A busca na IEEE Xplore foi realizada entre os meses de Agosto/2015 a Setembro/2015 e Setembro/2016 a Outubro/2016, retornando 69 estudos no total. Assim como demonstrado na Tabela 5, a String foi dividida em três partes para que a pesquisa fosse realizada com base no Titulo, *Abstract* e *Keywords* dos artigos.

Tabela 4: String de busca e o número de estudos retornados na IEEE Xplore.

String	**Retorno**
(((((“Document Title”: technique OR techniques OR method OR criterion OR strategy OR approach) AND (“Document Title”: “software testing” OR “software test” OR “software quality”) AND (“Document Title”: biomedics OR biomedical OR “medical software” OR “medical systems” OR “medical devices”)))))	66
(((((“Abstract”: technique OR techniques OR method OR criterion OR strategy OR approach) AND (Abstract: “software testing” OR “software test” OR “software quality”) AND (Abstract: biomedics OR biomedical OR “medical software” OR “medical systems” OR “medical devices”)))))	73
(((((“Keywords”: technique OR techniques OR method OR criterion OR strategy OR approach) AND (“Keywords”: “software testing” OR “software test” OR “software quality”) AND (“Keywords”: biomedics OR biomedical OR “medical software” OR “medical systems” OR “medical devices”)))))	66

Após obter o resultado da pesquisa, foi realizado o download do arquivo do tipo *bibtex* contendo o Título e *Abstract* de cada artigo, por meio da funcionalidade de exportação disponível na IEEE. Em seguida, os artigos foram agrupados, excluindo estudos duplicados de cada retorno obtido.

- **Scopus**: A busca na Scopus foi realizada entre os meses de Agosto/2015 a Setembro/2015 e Setembro/2016 a Outubro/2016, retornando 157 estudos no total. A pesquisa foi realizada com base no Titulo, *Abstract* e *Keywords* dos artigos. A String pode ser analisada com mais detalhes na Tabela 6.

Tabela 5: String de busca e o número de estudos retornados na Scopus.

String	**Retorno**
TITLE-ABS-KEY ((technique OR techniques OR method OR criterion OR strategy OR approach) AND ("software testing" OR "software test" OR "software quality") AND (biomedics OR biomedical OR "medical software" OR "medical systems" OR "medical devices"))	157

Após obter o resultado da pesquisa, foi realizado o download do arquivo do tipo *bibtex* contendo o Título e *Abstract* de cada artigo, por meio da funcionalidade de exportação disponível na Scopus. Em seguida, os artigos foram agrupados, excluindo estudos duplicados de cada retorno.

- **Compendex**: A busca na Compendex foi realizada entre os meses de Agosto/2015 a Setembro/2015 e Setembro/2016 a Outubro/2016, retornando 122 estudos no total. A pesquisa foi realizada com base no *Titulo*, *Abstract* e *Keywords* dos artigos. A String pode ser analisada com mais detalhes na Tabela 7.

Tabela 6: String de busca e o número de estudos retornados na Compendex.

String	**Retorno**
(((technique or techniques or method or criterion or strategy or approach) and ("software testing" or "software test" or "software quality") and (biomedics or biomedical or bioinformatics or "medical software" or "medical systems" or "medical devices")) WN KY)	122

Assim como nas demais bases, após obter o resultado da pesquisa, foi realizado o download do arquivo do tipo *bibtex* contendo o Título e *Abstract* de cada artigo, por meio da funcionalidade de exportação disponível na Compendex. Em seguida, os artigos foram agrupados, excluindo estudos duplicados de cada retorno.

Após a busca em cada uma das bases, foram obtidos 366 estudos no total, considerando os artigos duplicados nas diferentes bases. Durante a seleção preliminar, foram eliminados: estudos duplicados, artigos que não tinham as palavras-chaves no Título ou *Abstract* e artigos em idiomas não-alfanuméricos (japônes, chinês, mandarim, candii e etc). Além

disso, o *Abstract* de cada estudo foi lido e os critérios de seleção definidos foram aplicados. Um total de 13 estudos não estavam disponibilizados na Web e, mesmo sem sucesso, houve uma tentativa de contato com os respectivos autores com o objetivo de obter o estudo para leitura e análise. A Tabela 8 sumariza o número de artigos selecionados e não-selecionados durante as atividades desenvolvidas nessa fase.

Tabela 7: Relação entre número de artigos selecionados e eliminados.

Fonte	Selecionados	Não-Selecionados
Scopus	24	133
IEEE Xplorer	35	34
ACM Digital	3	15
Compendex	77	45
Total	139	227

3.2.2 Primeira seleção

A presente seção apresenta os resultados obtidos durante a primeira seleção do MS. Após a seleção preliminar, foi realizada uma leitura completa de todos os estudos obtidos e, em seguida, classificados como *relevantes* (incluídos) ou *não-relevantes* (excluídos) com base na aplicação dos critérios de seleção. Dessa forma, 82 estudos foram considerados relevantes para o MS. A Tabela 9 sumariza o número de artigos selecionados e não-selecionados durante essa fase.

Tabela 8: Relação entre artigos selecionados e eliminados.

Fonte	Selecionados	Não-selecionados
Scopus	13	11
IEEE Xplorer	17	18
ACM Digital	2	1
Compendex	50	27
Total	82	57

3.2.3 Snowballing e indicação de especialista

Após a seleção dos estudos primários, foi realizada uma busca manual nas referências de cada estudo em busca de novas contribuições relevantes aos objetivos do MS. Dessa forma, 17 estudos foram selecionados e, em seguida, foi realizada a aplicação dos critérios

de seleção. A princípio, foi realizada a exclusão por duplicidade em relação aos estudos já selecionados pela busca automatizada e, em seguida, foi aplicado um primeiro filtro com base na leitura do Título e *Abstract* de cada estudo. Por fim, foi realizada uma leitura completa de cada estudo com objetivo de selecionar os estudos mais relevantes com base nos critérios de seleção definidos no MS. A Tabela 10 sumariza os estudos selecionados após a realização do *Snowballing*.

Tabela 9: Contribuições por *Snowballing*.

Autor	Contribuição
Madsen EL	(Madsen, 2000)
Gibson NM et al.	(Gibson et al., 2001)
Barbosa, P.E.S. et al.	(Barbosa et al., 2013)
Majma, Negar et al.	(Majma e Babamir, 2014)
Torfeh, Tarraf, et al.	(Torfeh et al., 2007)

As contribuições por *Snowballing* foram adicionadas à base de artigos obtidos por meio da seleção preliminar e primeira seleção. Os estudos obtidos por meio do *Snowballing* e por indicação de especialista foram adicionados em uma categoria renomeada como Misc (*Miscellaneous*). Dessa forma, a Tabela 11 apresenta o número de artigos resultantes após a adição dos estudos obtidos pelo *Snowballing*. Quatro estudos indicados por especialistas, Goh e Lee (2007), Delamaro et al. (2013), Kamali et al. (2015) e Giannoulatou et al. (2014) foram adicionados na mesma categoria dos estudos selecionados por *Snowballing*. Importante ressaltar que tais estudos foram base para auxilio na elaboração da *String* de busca. No total, 91 estudos resultaram no final do processo de seleção.

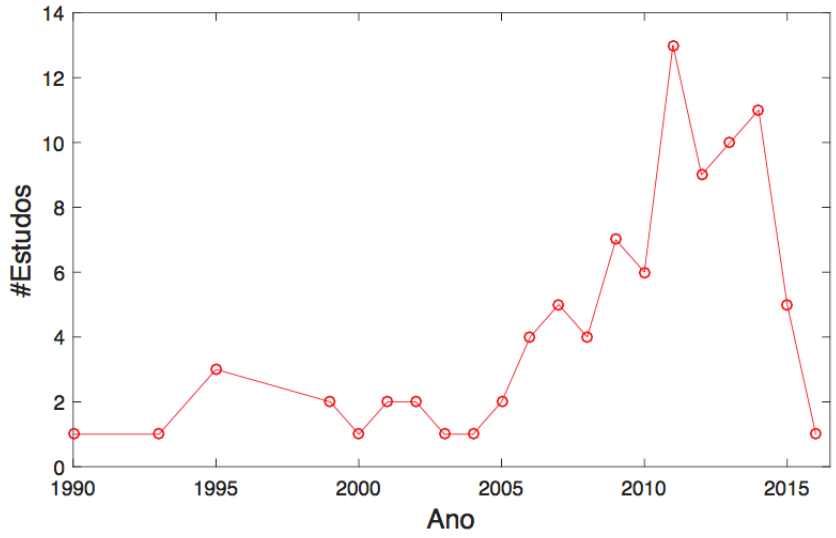
Tabela 10: Relação entre artigos selecionados e eliminados.

Fonte	Selecionados	Não-selecionados
Scopus	13	11
IEEE Xplorer	17	18
ACM Digital	2	1
Compendex	50	27
Misc	9	12
Total	91	69

4 Discussão e análise dos resultados

Por meio dos dados coletados, fase de análise e síntese de dados, permitiu uma representação das informações extraídas dos estudos primários com análises qualitativas e quantitativas. A maioria dos trabalhos selecionados são recentes, mostrando que há interesse de pesquisa nesse domínio de pesquisa. A Figura 3 mostra o interesse na área de 1990 a 2016 por meio de um gráfico com o número de estudos publicados a cada ano. Esses aspectos quantitativos mostram uma evolução considerável de pesquisas relacionadas às abordagens de validação em sistemas biomédicos.

Figura 3: Análise de publicações por ano entre 1990 e 2016.



A seguir, são apresentadas as respostas para cada QP do MS com base em dados quantitativos e análises subjetivas.

QP 1: De acordo com a literatura, existem técnicas, critérios e abordagens de teste de software e análise de qualidade que são usados em sistemas biomédicos?

Identificaram-se estratégias e abordagens utilizadas para validar domínios de sistemas biomédicos. As estratégias de validação identificadas foram classificadas em categorias de acordo com o processo utilizado para validar o domínio em questão: (i) *ad-hoc* e informal ou (ii) sistemática e automatizada. Estratégias *ad-hoc* e informais referem-se a estratégias de validação manual, em que o processo não requer formas sistemáticas de validação e não requer o uso de ferramentas que automatizam o processo de validação. Estratégias sistemáticas e automatizadas referem-se à aplicação de processos sistemáticos no desempenho da atividade de teste de software ou análise de qualidade na validação de um sistema. Geralmente, abordagens sistemáticas usam ferramentas que automatizam a validação do sistema.

Entre os estudos primários identificados, 98% (89/91) aplicam ou discutem estratégias de validação que utilizam abordagens manuais ou automatizadas. Além disso, 2% (2/91) dos estudos primários discutem apenas o estado da arte do domínio de pesquisa. Essa questão principal visa discutir as duas estratégias de validação, analisando os procedimentos aplicados no processo de validação no campo de pesquisa de sistemas biomédicos. Em seguida, evidências sobre as QPs específicas dessa QP geral são discutidas.

QP 1.1: Existem estratégias *ad-hoc* ou informais de teste de software e análise de qualidade aplicadas em sistemas biomédicos?

19% (18/91) dos estudos primários realizam processos informais ou *ad-hoc* para a condução de atividade de teste de software ou análise de qualidade de um domínio de sistema biomédico. Uma análise realizada nos estudos que utilizam abordagens *ad-hoc* evidenciou o uso de modelos de desenvolvimento e práticas para garantir a qualidade no desenvolvimento de sistemas biomédicos. A Tabela 11 apresenta algumas dessas abordagens identificadas nos estudos primários.

Tabela 11: Validação *ad-hoc* e não-sistemática em sistemas biomédicos: abordagens, ferramentas e modelos.

ID	Abordagem	Domínio do sistema
1	MIL-STD-498 e linguagem de programação Ada	Software médico em geral
2	Processo de Planejamento de produto (PPP)	
3	Padrões ISO 13485 e FDA CFR 820	
4	Análise de falhas em sistemas médicos	
5	Padrão IEC 62304	
6	Práticas de Medicina Familiar (FMF)	
7	Qualidade, Prazo e Confiabilidade (QPC	
8	Med-Adept	
9	Modelo de qualidade de segurança	
10	Qualidade do equipamento de diagnóstico radiológico	Equipamento de diagnóstico radiológico
11	Engenharia de requisitos de qualidade orientada para o mal uso (ERQOMU)	Sistema de aconselhamento sobre drogas
12	Ferramenta de Pesquisa Médica (FPM)	Equipamento de eletrocardiograma
13	ISOgrayTM	Sistemas de simulação virtual
14	OpenDMAP	Aplicações em mineração de texto
15	Formulário Web	Software biomédico *open-source*

Com base nas abordagens apresentadas na Tabela 11, é possível notar uma tendência para proposições de abordagens para o software médico em geral, em detrimento a sistemas de software específicos. Por outro lado, para sistemas que processam imagens, simulações virtuais e outras aplicações de resultados complexos, notaram-se abordagens desenvolvidas especificamente para tais domínios.

QP 1.2: Existem estratégias sistematizadas e/ou automatizadas, incluindo o uso de ferramentas e *frameworks*, de teste de software e análise de qualidade aplicadas em sistemas biomédicos?

As estratégias sistemáticas e automatizadas se referem a um conjunto de atividades que visam à validação de um domínio específico de sistema biomédico com o uso de ferramentas ou *frameworks* que automatizem o processo de validação. As ferramentas e *frameworks* são essenciais para garantir qualidade e a produtividade no processo de validação. Assim, 62% (57/91) dos estudos aplicam procedimentos sistemáticos e automatizados durante a validação de um domínio específico do sistema biomédico. Uma análise realizada em estudos que mencionam ou realizam processos de validação sistemáticos em sistemas biomédicos propiciou a categorização das abordagens e estratégias.

As ferramentas e *frameworks* identificados na validação dos sistemas biomédicos tinham como principal objetivo automatizar a atividade de validação do sistema em análise, oferecendo mais produtividade e eficiência. Nesse cenário, foi possível evidenciar o desenvolvimento de ferramentas e *frameworks* para validar, especificamente, domínios de sistemas biomédicos. Além disso, para determinados domínios de sistemas biomédicos, sistemas que processem ou gerem imagens médicas, a geração de dados de teste são mais complexas. Desse modo, o uso de ferramentas para validar tais sistemas é essencial. A Tabela 12 apresenta a abordagem utilizada na validação do respectivo domínio de sistema biomédico. Tais abordagens incluem o uso de ferramentas e *frameworks*, objetivando a automatização e eficácia na condução da validação do respectivo domínio de sistema biomédico.

QP 2: Quais procedimentos experimentais são mais utilizados e adequados para a validação de sistemas biomédicos?

Realizou-se uma análise para verificar os principais procedimentos experimentais aplicados no processo de validação de domínios de sistemas biomédicos. Para tanto, observou-se a natureza do estudo, prático ou teórico, o tipo de sistema biomédico, sistema real ou *toy*, e o domínio de sistema biomédico a ser validado. Inicialmente, foi realizada uma análise para verificar o número de estudos práticos comparados a estudos teóricos. Estudos práticos apresentam alguma atividade prática durante o processo de validação. Tais atividades práticas, envolvem a condução de um estudo de caso demonstrando a validação de domínio de sistema biomédico em específico. Estudos teóricos descrevem uma área ou tópico de pesquisa, mas não demonstração a condução de um, por exemplo, estudo de caso. Os autores consideraram estudos de caso (44/91), experiências controladas e análises empíricas em geral (45/91) durante essa análise. A Figura 4 apresenta um gráfico de pizza com uma comparação do número de estudos práticos e teóricos incluídos no MS.

Foi analisada a origem dos domínios de sistemas biomédicos identificados – reais ou *toys* – e suas respectivas predominâncias em pesquisas da área biomédica. Além disso, os autores apresentam uma análise comparativa dos domínios dos sistemas biomédicos de 1990 a 2016.

Tabela 12: Validação sistemática e automatizada em sistemas biomédicos: ferramentas, software e *frameworks*.

ID	Abordagem	Domínio do sistema
1	DEVSimPy Environment	Sistemas GUI
2	Sikuli	
3	Robot Framework	
4	O-FIm (Oracle for Images)	
5	NeuroSista	Imagens médicas
6	Projeto Humano Visível	
7	*Medical Imaging Interaction Toolkit* (MITK)	
8	Cardiac T2*	
9	QUALIMAGIQ	
10	Visualization Toolkit (VTK)	
11	CBMC model checker	Sistemas médicos embarcados
12	SATABS Model Checker	
13	NuSMV2	
14	Device Coordination Framework (DCF)	
15	HYRES	
16	SATs	
17	Dynamic Symbolic Execution (DSE)	Banco de dados biomédicos
18	Mock Object Based Test Generation for Database Applications (MODA)	
19	EvIdent	
20	Labview Environment	Software biomédico
21	OntoReTest	
22	BedMaster	
23	Schedulable Online Testing (SOTF)	
24	MuJava	
25	QuickCheck	
26	Control Quality for Ultrasound B-mode equipment (CQUS)	Exames de ultrassonografia
27	GPU-Based Framework	
28	JMP Statistical Discovery Software	
29	OntoReTest	Dispositivos médicos
30	NUnit	
31	Turf - EHR Usability Toolkit	Processamento de Linguagem Natural (PLN)
32	NIH Image	Exames de raio-x
33	MATLAB	Electrocardiogram
34	The Generic PCA Reference Model	Bomba de Infusão

QP 2.1: As validações ocorrem em sistemas biomédicos reais ou *toys*?

Identificaram-se 74 estudos que apresentam estratégias de validação com programas reais ou *toy* explicitamente. 95% (70/74) usam programas reais, 4% (3/74) exploram programas *toy* e 1% (1/74) dos estudos usam programas *toy* e reais. A Figura 5 ilustra a comparação entre o impacto do uso de programas *toy* e reais. Além disso, a Figura 6 ilustra uma análise que diferencia o número de estudos com programas *toy* e reais, ano a ano.

QP 2.2: Quais são os domínios dos sistemas biomédicos envolvidos?

Realizou-se uma análise com o objetivo de verificar os principais domínios de sistemas biomédicos utilizados por pesquisadores e profissionais no processo de validação. A Tabela

Figura 4: Estudos práticos vs estudos teóricos.

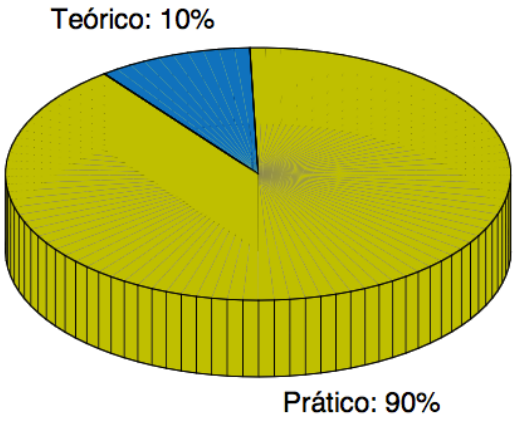


Figura 5: Programas *toy* versus programas reais.

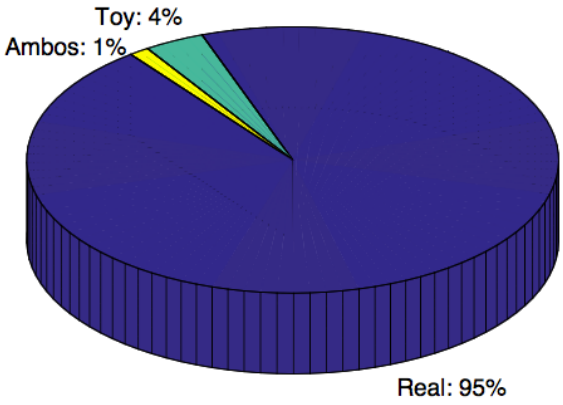
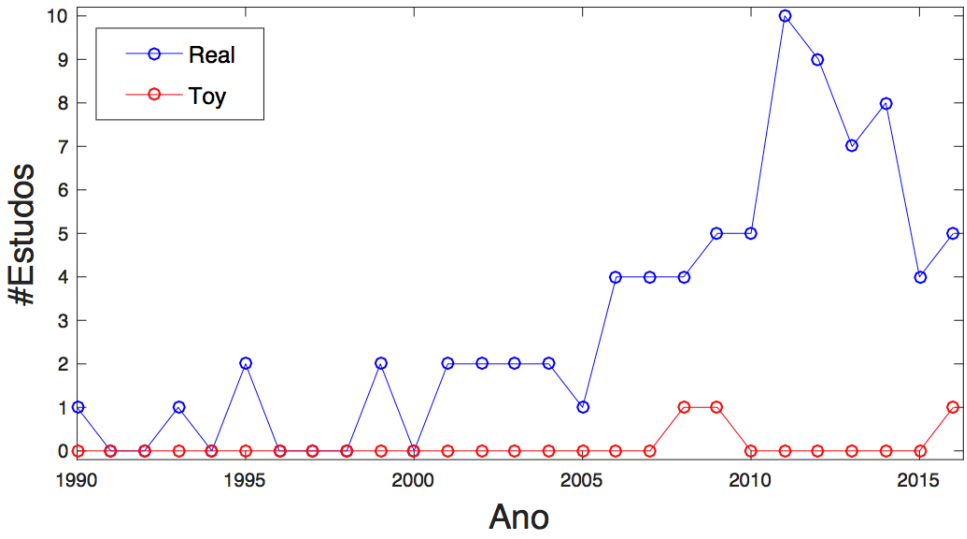


Figura 6: Programas *toy* versus programas reais: análise ano por ano.



13 mostra uma categorização dos domínios dos sistemas biomédicos. Os sistemas foram classificados em 19 categorias distintas: (1) sistemas que processam imagens médicas; (2) sistemas de Processamento de Linguagem Natural (PLN); (3) sistemas de diagnóstico de eletroencefalograma (EEG); (4) sistemas urológicos; (5) dispositivos de detecção de Respiração Cheyne-Stokes (RCS); (6) olfatômetro contínuo de sistemas de monitorização respiratória; (7) dispositivo clínica de detecção de neuropatias diabéticas; (8) Sistemas de Informação Hospitalar (SIH); (9) sistema de monitoramento remoto; (10) sistemas de aconselhamento de drogas; (11) dispositivo de marcapasso; (12) sistemas de análise de dados biomédicos; (13) bombas de infusão; (14) sistemas de telemedicina; (15) dispositivo de assistência ventricular; (16) sistemas de código aberto biomédico; (17) Sistemas de Computação Móvel (SMDCSes); (18) sistemas de exame de eletrocardiograma; e (19) propostas gerais.

Tabela 13: Número de domínio de sistema por categoria.

ID	**Domínio do sistema**	#	%
(1)	Imagens médicas (ultra-som, radiologia, radiografia, ressonância magnética, tomografia, raios-x, sistemas CAD)	22	24
(2)	Processamento de Linguagem Natural (PLN)	3	3,2
(3)	Sistemas de diagnóstico de eletroencefalograma (EEG)	1	1
(4)	Sistema urológico	1	1
(5)	Detecção de respiração de Cheyne-Stokes (CSB)	1	1
(6)	Sistema Contínuo de Monitoramento da Respiração do Olfatômetro (CRO)	1	
(7)	Detecção de Neuropatia Clínica por Diabética	1	1
(8)	Sistema de Informação Hospitalar (SIH)	1	
(9)	Sistema de monitoramento remoto	1	1
(10)	Sistemas de consultoria de drogas	1	1
(11)	Dispositivo de marcapasso	5	5,5
(12)	Sistemas de análise de dados biomédicos	3	3,2
(13)	Bombas de infusão	5	5,5
(14)	Sistema de telemedicina	1	1
(15)	Dispositivo de assistência Ventricular	1	1
(16)	Sistemas biomédicos *open-source*	1	1
(17)	Sistemas médicos de computação móvel (SMDCSes)	1	1
(18)	Sistemas de exame de eletrocardiograma	2	2,1
(19)	Propostas gerais	39	42,8

Considerando o escopo do presente trabalho de mestrado, foram avaliados os processos de validação, verificação e teste de sistemas que manipulam imagens sintéticas tridimensionais de redes vasculares. Dentre o número de domínios de sistemas que processam imagens médicas (22/91), apenas um estudo relata alguma abordagem de validação em imagens de redes vasculares. Errico et al. (2008) apresenta um método eficiente para análise de imagens intravasculares objetivando capturar mecanismos complexos e mudanças relacionadas ao diagnóstico precoce de CAV (do inglês, *Cardiac Allograft Vasculopathy*) na geometria coronária. Para alcançar tal objetivo, os autores compararam medidas

obtidas por meio de duas técnicas e avaliaram a variabilidade interobservador da técnica do software, repetindo, aleatoriamente, remotamente. Concluiu-se que uma análise de software com base na detecção semi-automatizada de bordas da imagem proporcionou uma melhor precisão e repetibilidade de medições em relação a um método manual.

QP 3: Quais estratégias e abordagens, *ad-hoc* ou sistematizadas, de teste de software e análise de qualidade são mais usadas para validar domínios de sistemas biomédicos?

Os estudos selecionados foram classificados em duas categorias principais, de acordo com a estratégia de validação discutida: (i) análise de qualidade; e (ii) teste de software. A análise de qualidade refere-se a estudos que focam a abordagem utilizada para avaliar a qualidade do sistema biomédico sem utilizar abordagens específicas da área de teste de software. Além disso, nessa categoria, é predominante o uso de padrões de qualidade na validação do respectivo domínio de sistema biomédico.

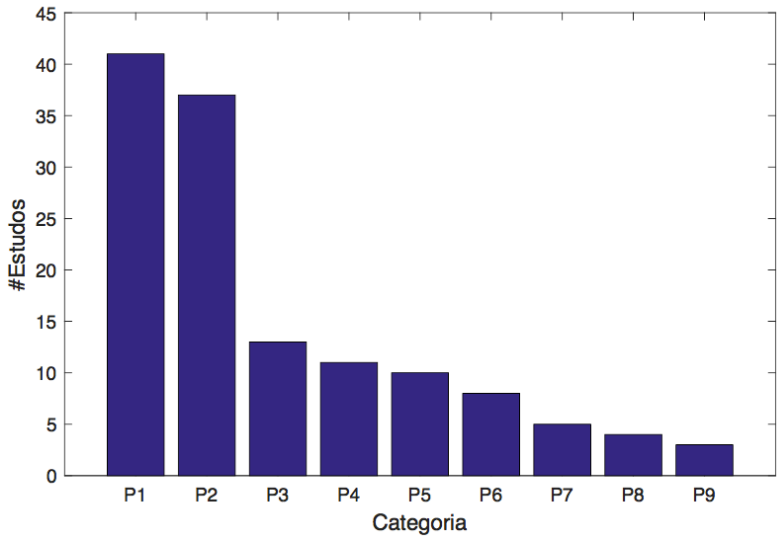
Por outro lado, a categoria de teste de software está relacionada a estudos que utilizam técnicas, critérios e estratégias de teste de software fundamentados na ES. Observou-se uma predominância do uso de abordagens que avalia o uso do sistema por meio do usuário (Teste de usabilidade), estratégias que verificam a saída do sistema com base na especificação (Teste funcional), estratégias que avaliam a cobertura da implementação do sistema por meio do código-fonte (Teste estrutural), estratégias que avaliam a qualidade do sistema durante a operação (Teste de operação), estratégias que validam o sistema com base na especificação formal do software (Teste formal), etc. A Tabela 14 mostra as diferentes categorias com uma descrição que distingue cada abordagem.

Tabela 14: Categorias de abordagens de análise de qualidade e teste de software.

ID	Categoria	Descrição
P1	Análise de qualidade	Estratégias que envolvem garantia de qualidade por meio de metodologias de desenvolvimento de software, padrões de qualidade, etc.
P2	Teste funcional	Estratégias que avaliam o comportamento do sistema de acordo com as especificações.
P3	Teste estrutural	Estratégias que avaliam o software através do código-fonte.
P4	Teste baseado em defeitos	Estratégias que alteram trechos no código fonte para comparar com o programa original.
P5	Teste formal	Estratégias que avaliam o sistema de acordo com a especificação formal.
P6	Teste de usabilidade	Estratégias que avaliam o comportamento do sistema através do usuário.
P7	Teste de unidade	Avaliações em módulos do sistema.
P8	Teste de integração	Estratégia em que os módulos do sistema são testados em grupos.
P9	Teste de regressão	Estratégias que realizam os testes após correções.

A análise e a classificação realizadas durante a leitura dos estudos primários permitiram a identificação de nove abordagens distintas de validação: (i) abordagens que avaliam a qualidade do software (37/91), (ii) abordagens que utilizam testes funcionais (41/91), (iii) abordagens que utilizam testes estruturais (8/91), (iv) abordagens que utilizam conceitos de teste baseado em erro (4/91), (v) abordagens que utilizam a especificação formal ou o modelo formal do software (10/91), (vi) teste de usabilidade (11/91), (vii) testes de unidade (13/91), (viii) testes de integração (5/91) e (ix) teste de regressão (3/91). Importante ressaltar que às abordagens que avaliam a qualidade do software demonstram ou propõem o uso de padrões de qualidade e metodologias de desenvolvimento de software como uma solução de garantia de qualidade. Além disso, nessa categoria, os autores não utilizam técnicas e estratégias da área de teste de software. Com base na Tabela 14 as abordagens utilizadas em cada estudo primário foram identificadas e analisadas. A Figura 7 apresenta o número de abordagens identificadas nos estudos primários.

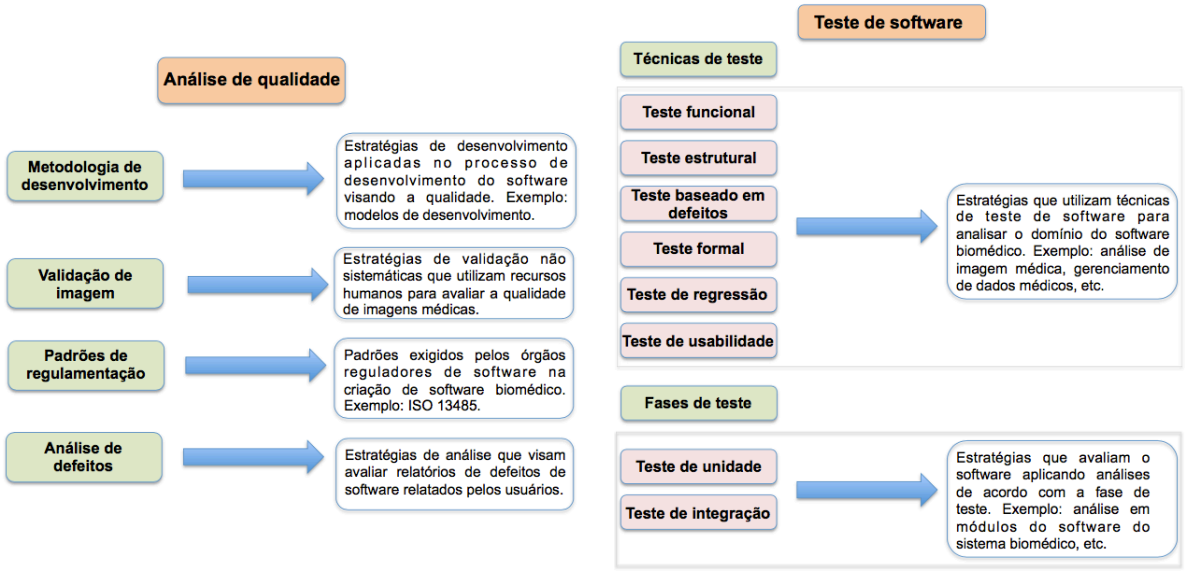
Figura 7: *Bar chart* de abordagens de validação de sistemas biomédicos.



A Tabela 14 apresenta as diferentes categorias de abordagens de validação identificadas durante a análise dos estudos primários do MS. A partir da definição dessas abordagens, definiu-se uma taxonomia que é essencial para apresentar as abordagens de análise de qualidade e teste de software prevalecentes em diferentes domínios de sistemas biomédicos. A Figura 8 apresenta a taxonomia definida com base na análise dos estudos primários.

Durante a análise dos estudos primários, observou-se que alguns estudos não mencionavam claramente a estratégia, o critério ou a abordagem do teste de software ou análise de qualidade utilizada. Assim, identificaram-se as características da abordagem ou cenário de utilização da estratégia de validação com base nas diretrizes da ES. A definição da

Figura 8: Taxonomia de abordagens de validação de sistemas biomédicos.



taxonomia foi realizada identificando a abordagem utilizada na validação do domínio do sistema biomédico e o contexto do processo de validação utilizado.

QP 4: Existe colaboração entre Indústria e Academia no desenvolvimento de estratégias de validação para sistemas biomédicos?

Uma das análises realizadas no MS foi a avaliação da participação da indústria e academia no desenvolvimento de abordagens de validação em sistemas biomédicos. Classificaram-se os estudos em três categorias: (i) Academia; (ii) Indústria; e (iii) Colaboração. Estudos com origem na Academia são aqueles cujo autores têm alguma filiação com universidades. Por outro lado, estudos classificados como provenientes da Indústria, fazem referência a institutos de investigação e pesquisa. Além disso, em alguns estudos, um ou mais autores tinham uma cooperação entre a academia e a indústria.

Durante essa análise, os autores identificaram que 81% (74/91) dos estudos primários são de origem da Academia, 7% (6/91) envolvem colaboração entre Academia e Indústria, e 12% (11/91) são de origem da indústria. A Figura 9 compara as três classificações por meio de um gráfico de pizza. Além disso, um gráfico de linha com uma comparação ano a ano entre as três classificações é apresentado na Figura 10. De acordo com as análises, há uma maior proporção de estudos provenientes da academia, principalmente entre 2009 e 2016. Por sua vez, os estudos da Indústria têm uma maior proporção nos últimos anos. Finalmente, estudos que envolvem colaboração entre Academia e Indústria ainda são pouco explorados, identificados apenas entre 2009 e 2015.

QP 5: Quais são os desafios e limitações de validação mencionados nos estudos?

Identificaram-se características relacionadas a desafios e limitações na validação de domínios de sistemas biomédicos. Nesse contexto, os autores analisaram as caracterís-

Figura 9: Origem de abordagens de validação em sistemas biomédicos: academia, indústria e colaboração.

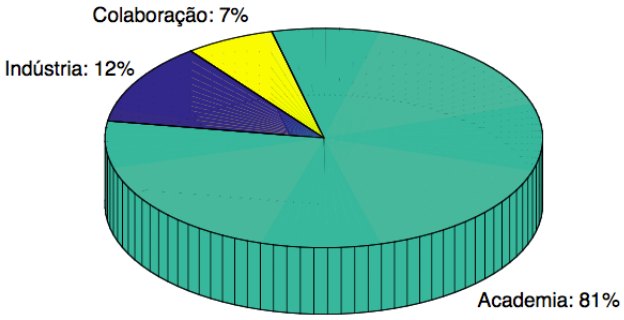
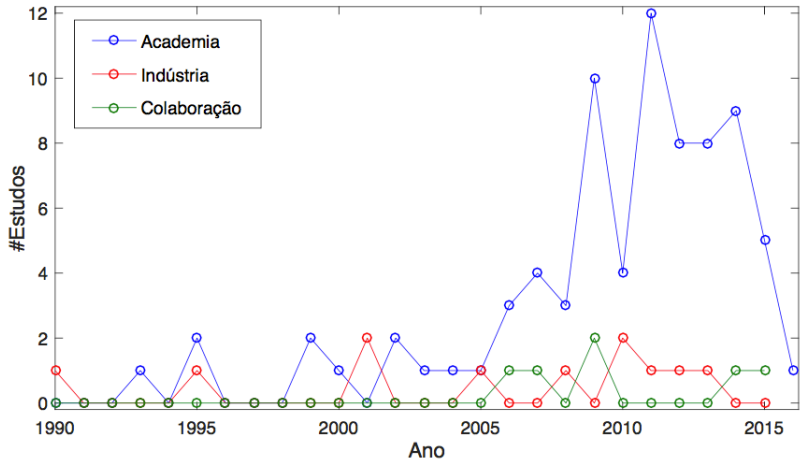


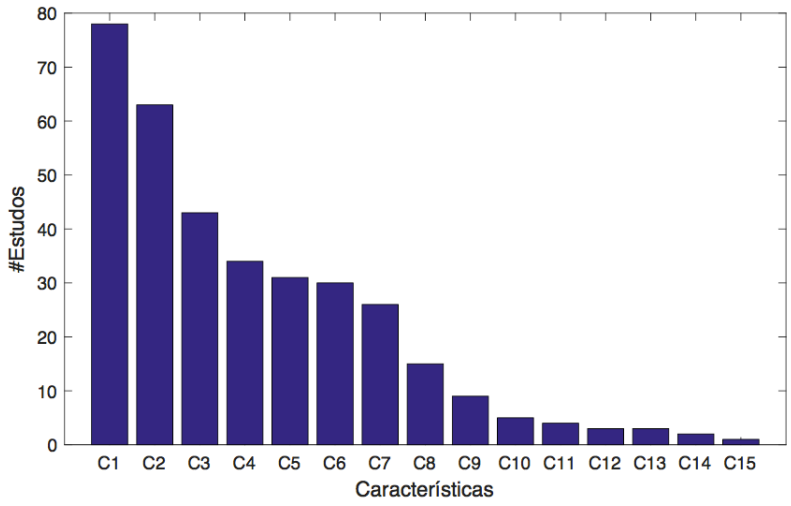
Figura 10: Academia vs Indústria: análise ano por ano.



ticas mencionadas nos estudos primários, verificando as mais frequentes no processo de validação. É essencial identificar as características como uma oportunidade de avaliar os desafios de validação nessa área de pesquisa.

As seguintes características expressadas como desafios foram identificadas: confiabilidade (C1); segurança (C2); eficácia do sistema (C3); defeitos de software (C4); dificuldades no campo de aplicação (C5); esforço humano (C6); exposição ao paciente (C7); gerenciamento de requisitos (C8); gerenciamento de projeto (C9); atividades de V&V (C10); dificuldade de operação do sistema (C11); usabilidade (C12); integração do sistema (C13); interoperabilidade (C14); adaptabilidade (C15). A Figura 11 ilustra as características citadas por meio de um gráfico de barras.

Figura 11: Características relacionadas à análise de qualidade e teste de software em sistemas biomédicos.



A característica de *confiabilidade* e *segurança*, são características que fazem referência à atividades que priorizam manter a qualidade dos sistemas em análise e manter a integridade dos dados após a execução do sistema. A característica de *eficácia do sistema* faz referência a capacidade do sistema em produzir o esperado com eficiência. A característica de *defeitos de software* é citada nos estudos como um desafio na realização das atividades de validação em explorar e notificar defeitos no software. A característica de *dificuldades no campo de aplicação* faz referência nos desafios e limitações a serem considerados no domínio de aplicação do sistema em análise. *Esforço humano* é uma característica citada para fazer referência a participação do ser humano como uma ferramenta de validação do sistema em análise. A característica de *exposição ao paciente* faz referência ao risco em expor o paciente a determinados domínios de sistemas biomédicos. As características de *gerenciamento de requisitos* e *gerenciamento de projeto* são características citadas nos estudos para evidenciar a importância em melhor produzir o software do sistema em análise. A característica de *atividades de V&V* fazem referência a importância na aplicação dessas atividades em sistemas da área médica. Por fim, *dificuldades de operação do sistema*, *usabilidade*, *integração do sistema*, *interoperabilidade* e *adaptabilidade*, são características citadas com uma análise das limitações e desafios dos sistemas após a sua execução.

QP 6: Qual é o impacto da pesquisa relacionada à validação de sistemas biomédicos?

Para essa QP, avaliaram-se os meios utilizados para disseminar a pesquisas que reportam estratégias de validação para sistemas biomédicos. Nesse contexto, os autores avaliaram e discutiram o impacto desses estudos em dois aspectos: (i) países onde os

estudos foram realizados; e (ii) os meios utilizados para disseminar os resultados. A seguir, os dois aspectos são discutidos por meio de QPs mais específicas.

QP 6.1: Quais regiões têm o maior impacto de pesquisas relacionadas à validação de sistemas biomédicos?

Durante a condução do MS, foram identificados os países com maior incidência de estudos relacionados à validação de sistemas biomédicos. Para responder essa QP, foi conduzida uma análise demográfica considerando a filiação do primeiro autor do estudo primário com universidades/faculdades e empresas privadas. É importante observar que os autores identificaram o país onde o estudo foi realizado, não necessariamente a origem do primeiro autor.

A Tabela 15 apresenta os dados coletados por esta análise. Além disso, a Figura 12 apresenta um *map chart* ilustrando os países com estudos publicados relacionados à validação de sistemas biomédicos.

Figura 12: *Map chart* de países com impacto na pesquisa.

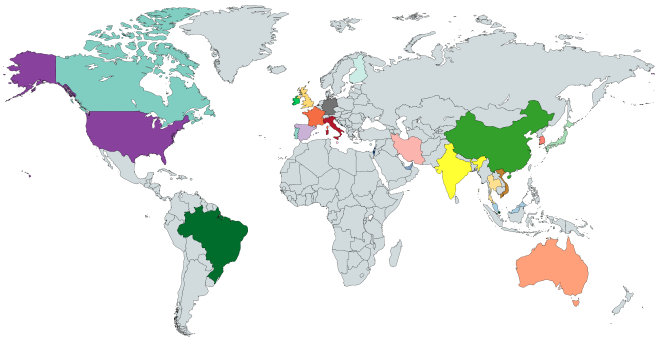


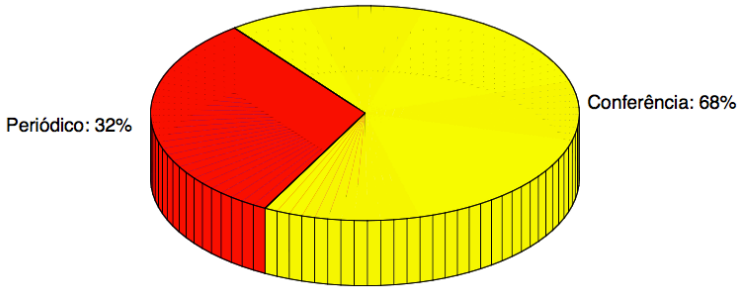
Tabela 15: Análise da divulgação de estudos por país.

País	#	%
USA	28	30,7
CAN	7	7,7
UK	7	7,7
FRA	6	6,6
AUS	5	5,5
CHN	5	5,5
BRA	4	4,4
GER	4	4,4
KOR	3	3,3
IND	3	3,3
THA	2	2,2
ITA	2	2,2
IRI	2	2,2
ESP	2	2,2
Outros	11	12

QP 6.2: Quais são os meios utilizados para disseminar as estratégias desenvolvidas para validar domínios de sistemas biomédicos?

Essa análise tem como objetivo verificar os meios utilizados pelos pesquisadores na disseminação de estudos que relatam estratégias de validação de sistemas biomédicos. Foi realizada uma análise nos canais de divulgação de cada estudo primário. Como resultado, verificou-se que o número de estudos publicados em periódicos 32% (29/91) é duas vezes menor do que os estudos publicados em conferências 68% (62/91). A Figura 13 apresenta uma comparação entre os estudos publicados em periódicos e conferências.

Figura 13: Publicações de abordagens de validação em sistemas biomédicos: periódicos vs conferência.



Com relação aos estudos publicados em periódicos, os autores identificaram que as três fontes mais utilizadas para disseminação das abordagens de teste de software e análise de qualidade são: *Journal Computer*, *Ultrasound in Medicine and Biology* e *Journal Computer Methods and Programs in Biomedicine*. Nos três periódicos citados, foram identificados 20% (6/29) dos estudos, enquanto que 80 % (23/29) dos estudos foram identificados em outros periódicos. A Tabela 16 apresenta a distribuição dos estudos por período.

Tabela 16: Número de estudos publicados por periódico.

Periódico	#	%
Computer Methods and Programs in Biomedicine	2	7
Ultrasound in Medicine and Biology	2	7
Journal Computer Methods and Programs in Biomedicine	2	7
Outros	23	80

Considerando os estudos publicados em conferências, identificou-se um amplo número de conferências em ES, teste de software e informática médica. As principais conferências utilizadas para divulgar estudos sobre garantia de qualidade e estratégias de teste de software aplicadas em sistemas biomédicos foram: CBMS (*Computer-Based Medical Systems*); BMEiCON (*Biomedical Engineering International Conference*); QR2MSE (*Quality, Reliability, Risk, Maintenance, and Safety Engineering*); ICSE (*International Conference on*

Software Engineering); ICBE (*International Conference on Bioinformatics and Biomedical Engineering*); CinC (*Computers in Cardiology*); MIC (*Medical Imaging Conference*); ASS (*Advances in Systems Safety*); ICBBE (*International Conference on Bioinformatics and Biomedical Engineering*); e BMEI (*Biomedical Engineering and Informatics*). A Tabela 17 resume a distribuição de estudos nos procedimentos da conferência.

Tabela 17: Número de estudos publicado por conferência.

Sigla	**Conferência**	#	%
CBMS	Computer-Based Medical Systems	4	6
BMEiCON	Biomedical Engineering International Conference	3	5
QR2MSE	Quality, Reliability, Risk, Maintenance, and Safety Engineering	3	5
ICSE	International Conference on Software Engineering	3	5
ICBE	International Conference on Bioinformatics and Biomedical Engineering	2	3
CinC	Computers in Cardiology	2	3
MIC	Medical Imaging Conference	2	3
ASS	Advances in Systems Safety	2	3
iCBBE	International Conference on Bioinformatics and Biomedical Engineering	2	3
BMEI	Biomedical Engineering and Informatics	2	3
Outros		37	59

5 Lacunas de pesquisa

A partir do MS conduzido, é possível observar que as abordagens, critérios e estratégias de teste de software e análise de qualidade em sistemas biomédicos apontam diversas questões e possibilidades a serem exploradas na literatura. Dessa forma, essa subseção sumariza os *gaps* de pesquisa identificados por meio de dificuldades e ausência de processos de validação em sistemas biomédicos, conforme listado nos tópicos abaixo:

- **Abordagens de validação**: por meio do MS, constatou-se uma maior predominância por estudos que utilizam abordagens de validação em software de dispositivos médicos em geral, como ilustrado na Tabela 11, com ferramentas mais genéricas e utilizáveis em diversos domínios de sistemas biomédicos. Ou seja, há pouca predominância no desenvolvimento de ferramentas de validação para domínios biomédicos em específico;
- **Domínio de sistemas**: a Tabela 11 ilustra os domínios de sistemas biomédicos explorados nos estudos selecionados por meio do MS. Nesse cenário, há pouca predominância em apresentar abordagens de validação para um sistema biomédico em específico, como abordagens e estratégias desenvolvidas para validar um determinado domínio de sistema biomédico. No entanto, para domínios de sistemas biomédicos

em que a saída dos SUTs se referem a imagens, 24% (22/91) exploram sistemas que processam imagens médicas. Tal predominância demonstra um grande interesse da literatura na validação de sistemas biomédicos nesse domínio. Além disso, 90% (20/22) dos estudos sobre imagens mencionavam ou discutiam a complexidade do processo de validação nesse tipo de sistema, sendo indício de que tais sistemas ainda são carentes de processos que validem as saídas produzidas;

- **Abordagens de teste de software**: durante a análise foi observada uma grande predominância do uso de técnicas, critérios ou estratégias de teste de software no processo de validação de sistemas biomédicos, como ilustra a Figura 7. No entanto, poucos estudos deixavam claro qual técnica ou critério de teste de software foi utilizada, o que demonstra a imaturidade de conceitos da ES nesse campo de pesquisa. Dessa forma, em tais situações, os autores do MS analisaram a estratégia utilizada e a classificaram de acordo com as diretrizes da ES e teste de software;
- **Abordagens de análise de qualidade**: assim como houve uma grande predominância no uso de técnicas, critérios e estratégias de teste de software, abordagens que objetivam a garantia de qualidade de sistemas biomédicos também foi bastante explorada nos estudos selecionados, como ilustra na Figura 7. No entanto, foi observado que as abordagens utilizadas nessa categoria com 43% (16/37) são abordagens não sistemáticas ou *ad-hoc*, demonstrando a necessidade de maior esforço no desenvolvimento de abordagens sistemáticas que objetivem a validação de sistemas biomédicos de forma mais produtiva; e
- **Avaliações Experimentais**: a Figura 8 apresenta uma grande predominância de estudos práticos, o que aumenta a relevância das pesquisas realizadas. Nesse cenário, houve um contato com o primeiro autor de um dos estudos do MS que não realiza estudos experimentais. Raoul Jetley, primeiro autor de 2% (2/91) e colaborador em 3% (3/91) dos estudos do MS, foi contatado e o mesmo afirmou a dificuldade na realização de estudos que envolvam a colaboração entre Academia e Indústria. Em um dos estudos de Raoul Jetley, (Jetley et al., 2006), foi realizada uma colaboração com um projeto financiado pela FDA (do inglês, *Food and Drug Administration*) na validação de uma Bomba de Infusão. No entanto, a abordagem apresentada no trabalho não foi validada em sistemas reais, logo não há produto no mercado validado com tal proposta. Esse cenário demonstra a necessidade de amadurecimento da área e de maiores incentivos por parte da indústria.

6 Ameaças à validade

Na presente seção são apresentadas as principais ameaças à validade que podem comprometer a validade do MS. Dessa forma, dentre os pontos a serem levantados, destacam-se:

- **Quantidade de pesquisadores envolvidos**: apenas dois pesquisadores estiveram intensamente envolvidos no processo de busca e seleção dos estudos. Além disso, em situações que resultaram em divergências, um terceiro pesquisador foi envolvido. Nesse cenário, este número pode ser aumentado, a fim de eliminar vieses durante a pesquisa;
- **String de busca**: a string de busca foi definida baseada na experiência e conhecimento dos autores, mas é necessário considerar que pode-se não ter completamente evitado a possibilidade de que alguns termos definidos na string de busca tenham sinônimos que não foram identificados;
- **Quantidade de estudos retornados**: os estudos retornados foram satisfatórios para a pesquisa, ou seja, apresentaram insumos suficientes para a composição do catálogo das abordagens, porém, para melhorar a generalização da mesma, outras fontes de dados poderiam ter sido adicionadas à pesquisa;
- **Seleção dos estudos primários**: não se pode garantir que todos os estudos primários relevantes foram retornados durante o processo de pesquisa e durante a avaliação dos mesmos. Dessa forma, os critérios de qualidade estabelecidos, bem como a atribuição dos *scores*, visa mitigar tal ameaça; e
- **Classificação de abordagens**: alguns estudos não deixavam claro a técnica ou critério de teste de software utilizada na validação da abordagem apresentada. Dessa forma, os autores avaliaram o método utilizado no estudo e o classificaram, seguindo as diretrizes de conceitos sobre Teste de software e Engenharia de Software.

7 Conclusão e trabalhos futuros

Este relatório técnico apresentou os estudos selecionados por meio de um MS realizado com o objetivo de identificar abordagens de análise de qualidade e teste de software de sistemas biomédicos. Por meio da condução do MS, 91 estudos foram identificados na área de teste de software e análise de qualidade em sistemas biomédicos. A partir da definição das QPs, análises quantitativas e qualitativas foram realizadas com o objetivo de sintetizar os principais tópicos de pesquisa explorados na validação de sistemas biomédicos. Como resultado principal, o MS demonstrou que o desenvolvimento de estratégias de validação

para sistemas biomédicos ocorreu com mais frequência recentemente, demonstrando interesse da literatura neste domínio de pesquisa. Outro fator identificado pelos autores diz respeito ao conhecimento de estratégias de validação, da Engenharia de Software, e, principalmente, de teste de software serem pouco disseminados no domínio de sistemas biomédicos.

Os resultados do MS visam disseminar a maturidade das estratégias de validação neste domínio de pesquisa. As limitações e dificuldades identificadas, oferecem lacunas de pesquisa que podem ser exploradas pelos pesquisadores. Além disso, o MS estabelece uma base de comparação para futuras abordagens de validação em sistemas biomédicos, contribuindo de forma sólida para essa área de pesquisa.

Visando aperfeiçoar e abranger a condução do MS, as principais orientações para trabalhos futuros: (i) incluir mais bibliotecas de estudos, incluindo bibliotecas médicas; (ii) expandir a pesquisa de estudos de livros, teses e dissertações; e (iii) analisar cada desafio de pesquisa identificado, avaliando a influência de cada um neste domínio de pesquisa.

Referências

Asrafi, M.; Liu, H.; Kuo, F. C. On testing effectiveness of metamorphic relations: A case study. In: *Proceedings of the 5th International Conference on Secure Software Integration and Reliability Improvement (SSIRI)*, 2011, p. 147–156.

Banerjee, I.; Nguyen, B.; Garousi, V.; Memon, A. Advances in gui testing. Advances in Computers, v. 58, Elsevier, p. 149–201, 2003.

Banerjee, I.; Nguyen, B.; Garousi, V.; Memon, A. Graphical user interface (gui) testing: Systematic mapping and repository. *Information and Software Technology*, p. 1679–1694, 2013.

Barbosa, P. E. S.; Morais, M.; Galdino, K.; Andrade, M. F.; Gomes, L.; Moutinho, F.; de Figueiredo, J. C.; et al. Towards medical device behavioural validation using petri nets. In: *Proceedings of the 26th International Symposium on Computer-Based Medical Systems (CBMS)*, 2013, p. 4–10.

Bertolino, A. Software testing research and practice. In: *Proceedings of the Abstract State Machines*, 2003, p. 1–21.

Bertolino, A. Software testing research: Achievements, challenges, dreams. In: *Proceedings of the Future of Software Engineering*, 2007, p. 85–103.

Briand, L.; Labiche, Y. A uml-based approach to system testing. *Software and Systems Modeling*, p. 10–42, 2002.

Chan, W. K.; Tse, T. H. Oracles are hardly attain'd, and hardly understood: confessions of software testing researchers. In: *Proceedings of the 13th International Conference on Quality Software (QSIC)*, 2013, p. 245–252.

Chen, T. Y.; Ho, J. W. K.; Liu, H.; Xie, X. An innovative approach for testing bioinformatics programs using metamorphic testing. *BMC bioinformatics*, p. 24, 2009.

Cheon, Y. Abstraction in assertion-based test oracles. In: *Proceedings of the 7th International Conference on Quality Software (QSIC)*, 2007, p. 410–414.

Cohen, K. B.; Hunter, L. E.; Palmer, M. Assessment of software testing and quality assurance in natural language processing applications and a linguistically inspired approach to improving it. In: *Proceedings of the Trustworthy Eternal Systems via Evolving Software, Data and Knowledge*, p. 77–90, 2013.

Delamaro, M. E.; Maldonado, J. C.; Jino, M. Capítulo 1 – Conceitos Básicos. In: Delamaro, M. E.; Maldonado, J. C.; Jino, M., eds. *"Introdução ao Teste de Software"*, 2 ed, Rio de Janeiro: Campus, p. 1–7, 2016.

Delamaro, M. E.; Nunes, F. L. S.; Oliveira, R. A. P. Using concepts of content-based image retrieval to implement graphical testing oracles. *Software Testing, Verification and Reliability*, p. 171–198, 2013.

Errico, V. D.; Potena, L.; Fiore, D.; Fabbri, F.; Grigioni, F.; Magnani, G.; Ortolani, P.; Bianchi, I.; Corazza, I.; Zannoli, R.; et al. Reproducibility of ivus measurements in heart transplant recipients: increased quality of data by using dedicated software for image analysis. In: *Proceedings of the Computers in Cardiology*, 2008, p. 537–540.

Filho, A. C. S. S.; Rodrigues, E. P.; Junior, J. E.; Carneiro, A. A. O. A computational tool as support in b-mode ultrasound diagnostic quality control. *Revista Brasileira de Engenharia Biomédica*, p. 402–405, 2014.

Frounchi, K.; Briand, L. C.; Grady, L.; Labiche, Y.; Subramanyan, R. Automating image segmentation verification and validation by learning test oracles. *Information and Software Technology*, p. 1337–1348, 2011.

Galarreta, M. A.; Macedo, M. M. G.; Mekkaoui, C.; Jackowski, M. P. Three-dimensional synthetic blood vessel generation using stochastic L-systems. In: *Proceedings of the Medical Imaging: Image Processing*, 2013, p. 86691I–86691I–6.

Giannoulatou, E.; Park, S.-H.; Humphreys, D. T.; Ho, J. W. Verification and validation of bioinformatics software without a gold standard: a case study of bwa and bowtie. *BMC bioinformatics*, p. S15, 2014.

Gibson, N. M.; Dudley, N. J.; Griffith, K. A computerised quality control testing system for b-mode ultrasound. *Ultrasound in medicine & biology*, p. 1697–1711, 2001.

Goh, O.; Lee, Y. H. Schedulable online testing framework for real-time embedded applications in vm. In: *Proceedings of the Embedded and Ubiquitous Computing*, p. 730–741, 2007.

Hinterleitner, F.; Neitzel, G.; Möller, S.; Norrenbrock, C. An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In: *Proceedings of the Blizzard challenge workshop*, 2011, p. 1.

Hoffman, D. Using oracles in test automation. *Software Quality Methods*, p. 90—-117, 2001.

Hoyt, R. E.; Sutton, M.; Yoshihashi, A. *Medical informatics: Practical guide for the healthcare professional*. Lulu.com, 2008.

IEEE Ieee standard glossary of software engineering terminology. *Office*, p. 1–84, 1990.

Javed, A. Z.; Strooper, P. A.; Watson, G. N. Automated generation of test cases using model-driven architecture. In: *Proceedings of the 2th International Workshop on Automation of Software Test (AST)*, 2007, p. 3–3.

Jetley, R.; Iyer, S. P.; Jones, P. L. A formal methods approach to medical device review. *Computer*, p. 61–67, 2006.

Jiang, F.; Shi, D.; Liu, D. C. Fast adaptive ultrasound speckle reduction with bilateral filter on cuda. In: *Proceedings of the 5th International Conference on Bioinformatics and Biomedical Engineering (iCBBE)*, 2011, p. 1–4.

Kamali, A. H.; Giannoulatou, E.; Chen, T. Y.; Charleston, M. A.; McEwan, A. L.; Ho, J. W. How to test bioinformatics software? *Biophysical Reviews*, p. 343–352, 2015.

Kawamura, T.; Kimura, T.; Tsumoto, S. Data mining-based service quality estimation in hospital information system. In: *Proceedings of the International Conference on Data Mining Workshop (ICDMW)*, 2014, p. 289–295.

Keele, S. Guidelines for performing systematic literature reviews in software engineering. In: *Proceedings of the Technical report, 2.3 EBSE*, 2007.

Kitchenham, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, p. 1–26, 2004a.

Kitchenham, B. *Procedures for Performing Systematic Reviews*. Technical report, Software Engineering Group - Department of Computer Science - Keele University and Empirical Software Engineering - National ICT Australia Ltd, 2004b.

Kitchenham, B.; Charters, S.; Budgen, D.; Brereton, P.; Turner, M.; Linkman, S.; Jørgensen, M.; Mendes, E.; Visaggio, G. *Guidelines for performing systematic literature reviews in software engineering*. Relatório Técnico, Keele University (Software Engineering Group School of Computer Science and Mathematics)and University of Durham (Department of Computer Science), 2007.

Kitchenham, B.; Dyba, T.; Jorgensen, M. Evidence-based software engineering. In: *Proceedings of the 26th international conference on software engineering*, 2004, p. 273–281.

Koru, G.; Emam, E. K.; Neisa, A.; Umarji, M. A survey of quality assurance practices in biomedical open source software projects. *Medical Internet Research*, 2007.

Madsen, E. L. Quality assurance for grey-scale imaging. *Ultrasound in medicine & biology*, p. S48–S50, 2000.

Majma, N.; Babamir, S. M. Specification and verification of medical monitoring system using petri-nets. *Journal of medical signals and sensors*, p. 181, 2014.

McMinn, P.; Stevenson, M.; Harman, M. Reducing qualitative human oracle costs associated with automatically generated test data. In: *Proceedings of the 5th International Workshop on Software Test Output Validation*, 2010, p. 1–4.

Myers, G. J. *The Art of Software Testing*. John Wiley and Sons Inc, 256 p., 2004.

Myers, G. J.; Sandler, C.; Badgett, T. *The art of software testing*. John Wiley & Sons, 2011.

Oliveira, R. A. P.; Memon, A. M.; Gil, V. N.; Nunes, F. L. S.; Delamaro, M. E. An extensible framework to implement test oracles for non-testable programs. In: *Proceedings of the 26th International Conference on Software Engineering and Knowledge Engineering (SEKE 2014)*, 2014, p. 199–204.

Paech, B.; Wetter, T. Rational quality requirements for medical software. In: *Proceedings of the 30th International Conference on Software engineering*, 2008, p. 633–638.

Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. Systematic mapping studies in software engineering. In: *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 2008, p. 71—80.

Petersen, K.; Vakkalanka, S.; Kuzniarz, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, p. 1–18, 2015.

Petticrew, M.; Roberts, H. Systematic reviews in the social sciences: A practical guide. *Malden, MA: Blackwell*, 2006.

Petticrew, M.; Roberts, H. *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons, 2008.

Postolache, O.; Girao, P.; Lunca, E.; Bicleaknu, P.; Andrusca, M. Unobtrusive cardio-respiratory monitoring based on microwave doppler radar. In: *Proceedings of the International Conference and Exposition on Electrical and Power Engineering (EPE)*, 2012, p. 597–600.

Rafi, D. M.; Moses, K. R. K.; Petersen, K.; Mäntylä, M. V. Benefits and limitations of automated software testing: Systematic literature review and practitioner survey. In: *Proceedings of the 7th International Workshop on Automation of Software Test*, 2012, p. 36–42.

Siddiqi, A. A.; Ahmed, M.; Alginahi, Y. M.; Alharby, A. Use of information and mobile computing technologies in healthcare facilities of saudi arabia. In: *Proceedings of the International Conference on Information and Communication Technologies (ICICT)*, 2009, p. 289–294.

Taylor, P. *Text-to-speech synthesis*. Cambridge university press, 2009.

Torfeh, T.; Beaumont, S.; Guédon, J.; Normand, N.; Denis, E. Software tools dedicated for an automatic analysis of the ct scanner quality control images. In: *Proceedings of the Medical Imaging*, 2007, p. 65104G–65104G.

Wallace, D. R.; Kuhn, D. R. Failure modes in medical device software: an analysis of 15 years of recall data. *International Journal of Reliability, Quality and Safety Engineering*, p. 351–371, 2001.

Wang, Y.; Helminen, E.; Jiang, J. Building a virtual breast elastography phantom lab using open source software. In: *Proceedings of the International Ultrasonics Symposium (IUS)*, 2014, p. 1841–1844.

Ward, M. A.; Ofori, E. K.; Scutt, D.; Moores, B. M. Experiences of in-field and remote monitoring of diagnostic radiological quality in ghana using an equipment and patient dosimetry database. In: *Proceedings of the World Congress on Medical Physics and Biomedical Engineering*, 2009, p. 36–39.

Yin, Y.; Liu, B.; Ni, H. Real-time embedded software testing method based on extended finite state machine. *Systems Engineering and Electronics*, p. 276–285, 2012.

Yu, T.; Qu, X.; Acharya, M.; Rothermel, G. Oracle-based regression test selection. In: *Proceedings of the 6th International Conference on Software Testing, Verification and Validation (ICST)*, 2013, p. 292–301.

Zhang, G.; Yan, P.; Zhao, H.; Zhang, X. A computer aided diagnosis system in mammography using artificial neural networks. In: *Proceedings of the International Conference on BioMedical Engineering and Informatics*, 2008, p. 823–826.

Zheng, K.; Vydiswaran, V. G. V.; Liu, Y.; Wang, Y.; Stubbs, A.; Uzuner, O.; Gururaj, A. E.; Bayer, S.; Aberdeen, J.; Rumshisky, A.; et al. Ease of adoption of clinical natural language processing software: An evaluation of five systems. *Biomedical informatics*, p. S189–S196, 2015.

Zhou, Z. Q.; Huang, D. H.; Tse, T. H.; Yang, Z.; Huang, H.; Chen, T. Y. Metamorphic testing and its applications. In: *Proceedings of the 8th International Symposium on Future Software Technology (ISFST 2004)*, 2004, p. 346–351.