



XVI Encontro Nacional de Pesquisa em Ciência da Informação (XVI ENANCIB)
ISSN 2177-3688

GT 11 – Informação & Saúde
Comunicação Oral

A INFORMAÇÃO EM ONCOLOGIA NA ERA DO BIG DATA¹

THE INFORMATION IN ONCOLOGY IN THE ERA OF BIG DATA

Amanda Damasceno de Souza, UFMG
amanda@ufmg.br

Fernanda Farinelli, UFMG
fernandafarinelli@eci.ufmg.br

Mauricio Barcelos de Almeida, UFMG
mba@ufmg.br

Sebastião Cabral Filho, Centro de Quimioterapia Antiblastica e Imunoterapia
cabral@cqai.com.br

Resumo: Analisa o fenômeno *Big Data* na área de saúde especificamente na oncologia, chamado neste trabalho de *Big Health Data*. A contribuição do *Big Health Data* para os oncologistas refere-se a prover desde dados sobre o paciente para responder questões sobre efetividade de tratamento e resultados. Neste contexto a padronização e interoperabilidade é requisito essencial. Coloca-se a seguinte questão de pesquisa emerge para ser investigada: de que forma as lacunas de interoperabilidade semântica existentes nas soluções de *Linked Health Data* podem ser contornadas a fim de minimizar a variabilidade dos dados em soluções de *Big Health Data* na área de Oncologia? Foram analisados artigos da área de medicina do PUBMED (2010-2015) e no Scielo (2000 a 2015). Como perspectivas futuras haverá a continuidade do crescimento do *Big data* sobre câncer. O trabalho transdisciplinar será importante para otimização e interoperabilidade entre *Big data* e sistemas de informação em oncologia.

Palavras-chave: *Big Data*. Sistema de informação. Câncer. Tecnologia da informação e comunicação. Profissional da Informação, *Big Health Data*, *Linked Health Data*.

Abstract: Analyzes the phenomenon of Big Data in healthcare specifically in oncology, called this work of Big Health Data. The Big Health Data contribution to oncologists refers to provide data from the patient to answer questions about the effectiveness of treatment and results. In this context the standardization and interoperability is an essential requirement. Place the following research question

¹ O conteúdo textual deste artigo, os nomes e e-mails foram extraídos dos metadados informados e são de total responsabilidade dos autores do trabalho.

emerges to be investigated: how semantic interoperability gaps existing in the Linked Health Data solutions can be circumvented in order to minimize data variability in Big Health Data solutions in oncology area? They analyzed medical area of articles of PubMed (2010-2015) and Scielo (2000-2015). As future prospects will be the continued growth of Big data on cancer. The transdisciplinary work will be important for optimization and interoperabilidade between Big data and information systems in oncology.

Keywords: *Big Data*. Information Storage and Retrieval Systems. Cancer. Information Technology. , *Big Health Data*, *Linked Health Data*.

1 INTRODUÇÃO

Os dados sempre foram usados pela humanidade como insumo para tomada de decisões em geral. Entretanto, nos dias de hoje estamos lidando com volumes de dados cada vez maiores, que são criados em uma velocidade na escala dos segundos e envolve ainda uma variedade de formatos que são disponíveis, como textos, sons, imagens e vídeos. Ao se referir aos dados acompanhados da tríade característica volume, velocidade e variedade, estamos falando de um fenômeno chamado de *Big Data* (BD).

Observa-se hoje que as instituições de saúde atuam em um ambiente que vem sendo moldado por este fenômeno, onde a cada segundo são gerados milhares de registros de saúde do paciente, seja por meio de um exame de laboratório ou de imagem, ou por um registro no prontuário eletrônico. Ao nos referirmos ao fenômeno *Big Data* na área de saúde, neste trabalho adota-se a nomenclatura de *Big Health Data* (BHD).

Um bom exemplo de como a aplicação de tecnologias baseadas no conceito do BHD pode contribuir foi publicada pela *Public Health Watch*². O sistema *HealthMap*³, foi capaz de prever o surto de Ebola na África Ocidental nove dias antes de ser anunciado formalmente pela Organização Mundial da Saúde (OMS). Este sistema identificou a existência de uma “febre hemorrágica desconhecida” no sudeste da Guiné por meio de cruzamento de dados de milhares de sites diferentes, como sites governamentais, redes sociais, sites especializados em doenças infecciosas e outras fontes.

Em um levantamento realizado sobre o uso do BHD na área de Oncologia, verifica-se o trabalho de Meyer e Basch (2015), que relatam sobre a contribuição do BHD para os oncologistas, ao prover desde dados sobre o paciente às práticas essenciais para responder

² Reportagem na íntegra, foi publicada em 10/08/2014, disponível no seguinte endereço de internet: <<https://publichealthwatch.wordpress.com/2014/08/10/how-a-computer-algorithm-predicted-west-africas-ebola-outbreak-before-it-was-announced/>>. Acessada em 26/07/2015.

³ O sistema Healthmap, foi criado por pesquisadores, epidemiologista e desenvolvedores de software, no Boston Children's Hospital em 2006. Disponível no seguinte endereço de internet: <<http://www.healthmap.org/pt/>>. Acessado em 26/07/2015.

questões sobre efetividade de tratamento e resultados a longo prazo. Entretanto, os dados sobre a saúde de um paciente podem se originar de várias fontes e/ou sistemas de informação (SI) e estes SIs podem adotar terminologias diferentes. Por exemplo, na área de Oncologia o *National Cancer Institute* (NCI) dos EUA criou uma terminologia de domínio o *NCI-Thesaurus* com a finalidade de padronizar os termos sobre câncer. Na área médica em geral, existe o SNOMED CT é uma coleção de terminologia médica que busca descrever conceitos para documentação clínica, utilizadas para organizar de forma sistemática e para ser processável por computador (TSATSARONIS et al., 2013). Outra iniciativa é o *Unified Medical Language System* (UMLS) – *Metathesaurus* um vocabulário biomédico que apresenta um conjunto de cerca de 100 fontes de vocabulários (SMITH et al., 2007).

O fato de os SIs não adotarem uma terminologia única é um desafio a ser vencido nas soluções de BHD, pois a falta de interoperabilidade semântica restringe a troca de direta de informações entre SIs. A heterogeneidade semântica introduz mais um ponto de variabilidade de dados no BHD.

Diante do que foi exposto, algumas questões foram identificadas e foram investigadas: 1) Como o fenômeno *Big Health Data* está transformando a tomada de decisão e a rotina médica na Oncologia?; 2) Como a Ciência da Informação pode apoiar os profissionais de saúde, que lidam com Oncologia, no sentido de prover a organização e a recuperação de informações em ambientes distribuídos?.

Desta forma, este artigo visa: i) Identificar o estado da arte das pesquisas sobre *Big Data* na área de Oncologia; ii) Discutir o envolvimento da Ciência da Informação em soluções de organização e recuperação de informação que estão fisicamente distribuídas; e iii) Identificar o que a Ciência da Informação pode realizar para viabilizar o crescimento de solução de *Big Health Data* na área de Oncologia.

Este artigo é resultado de discussões que envolvem duas pesquisas que lidam questões inerentes ao fenômeno BHD: a primeira referente à dissertação de mestrado onde desenvolve uma sistematização de definições formais em ontologias biomédicas aplicadas ao domínio das leucemias mieloides agudas; a segunda referente à tese de doutorado que pesquisa a aplicação de terminologias e ontologias biomédicas para resolver problemas de interoperabilidade semântica em sistemas de *business intelligence* em um ambiente de *Big Data*. Pretende-se com este trabalho gerar subsídios para os pesquisadores envolvidos com esta temática, visando seu aprofundamento e sua aplicação em trabalhos futuros.

1.1 MATERIAIS E MÉTODOS

No campo da Ciência da Informação (CI), este estudo foca na categoria temática "organização do conhecimento" conforme a taxonomia da área de CI, elaborada por Hawkins, Larson, e Caton (2003). Na tradução adaptada pelas autoras Oddone e Gomes (2003), à categoria temática que este estudo se enquadra é "processamento, recuperação e disseminação da informação". No geral, este trabalho é aplicado ao campo da saúde, visando solucionar problemas de organização e recuperação da informação biomédica, mas com as devidas adaptações pode ser expandidos para outros campos.

Este estudo se classifica como uma pesquisa exploratória, pois visa proporcionar aos pesquisadores uma maior familiaridade sobre o tema pesquisado, uma vez que este ainda é pouco conhecido, pouco explorado. Quanto aos procedimentos técnicos, adota-se a pesquisa bibliográfica, pois explora e discute material já publicado (LAKATOS; MARCONI, 1991; GIL, 2002).

Para tal, inicialmente, discutise questões do campo da CI que pode ser explorados na era do BD, revisitam-se aspectos teóricos relevantes para o entendimento da noção de interoperabilidade, ontologias, web semântica, *Linked Data (LD)* e *Big Data (BD)* (seção 2). Em seguida, discorre-se sobre o estado da arte da temática *Big Health Data (BHD)* na área de oncologia (seção 3). Cabe destacar que não se pretende uma revisão exaustiva das iniciativas nacionais e internacionais, mas apenas o necessário para caracterizar o atual estado da arte nacional. Em seguida, apresenta-se os resultados da bibliometria em torno do BD e Câncer e discuti-se sobre questões da interoperabilidade e ontologias em um ambiente de LHD e sobre a variabilidade de dados em soluções de BHD (seção 4). Finalmente, conclui-se que ontologias podem desempenhar um papel relevante na busca pela interoperabilidade em soluções de LD e BD, mas que muito ainda está por ser (seção 5).

2 O PAPEL DA CIÊNCIA DA INFORMAÇÃO NA ERA DO BIG DATA

Para desenvolver esta seção, utilizou-se artigos e livros indexados na ferramenta de busca Google Acadêmico⁴ que indexa, em um único lugar, publicações - artigos, teses, livros, etc. - disponíveis em outras bases. Na busca foram utilizadas as palavras chaves: *Big data*, *Linked Data*, *Big Health Data*, *Linked Health Data*. Além de pesquisas com combinação desta palavras. A escolha dos documentos, baseou-se na ordem de relevância que o Google

⁴ Scholar Google – [<http://scholar.google.com.br/>]

Acadêmico adota⁵ que considera o autor, a publicação do documento, o texto integral de cada artigo, a frequência de citação. Optou-se por selecionar somente documentos de conteúdo teórico, por se tratar de uma pesquisa teórica. Assim, apresenta-se conceitos relacionados às questões de investigação e discute-se o papel da Ciência da Informação na Era do *Big Data*.

2.1 ORGANIZAÇÃO DA INFORMAÇÃO DO CONHECIMENTO

Dentre os estudos encontrados na CI, destacam-se os Sistemas de Organização de Conhecimento (SOC). Nos últimos anos, vários pesquisadores desenvolveram estudos relacionados a estes tópicos, e seus resultados serviram de contribuição a múltiplos campos de estudo. Os SOC's são conhecidos como produtos de representação (teorias, processos e instrumentos) gerados por meio de interpretações relevantes de um mundo escolhido ou domínio de informação, registrados em sistemas de informação e documentos. Os SOC podem contribuir na desambiguação de vocabulário e das estruturas necessárias para se garantir a semântica e compreensão dos termos, podendo assim ser aplicado para solucionar problemas relacionados à interoperabilidade semântica. Seu estudo e desenvolvimento se justificam principalmente em função da web semântica (SOUZA; TUDHOPE; ALMEIDA, 2012).

2.1.1 ontologias e interoperabilidade semântica

No campo da CI, as ontologias são instrumentos de representação semântica, e são estudadas como um tipo de SOC que permite a representação do conhecimento (SOERGEL, 1997 *apud* FARINELLI *et al*, 2013). Segundo Farinelli et al. (2013) nas ontologias formais, os termos e as relações entre os termos do vocabulário aparecem como variáveis, constantes e predicados de uma linguagem lógica formal. São adicionados axiomas lógicos para expressar relações entre entidades, para restringir o significado dos termos nela representados, desta forma aceitam-se certas interpretações e rejeitam-se outras. Assim a semântica de cada termo é definida explicitamente e sem ambiguidades.

A OWL (*Web Ontology Language*) é a linguagem recomendada pelo W3C para definir ontologias. É baseada em RDF e RDF *Schema*, porém mais rica semanticamente por incluir axiomas lógico para definir as relações. O RDF, é uma recomendação do W3C, especifica uma estrutura que permite representar informações na web. RDF possui uma representação através de documentos XML, instrumento que traz grandes facilidades para troca de metadados. As informações são representadas por meio de tripla contendo sujeito, predicado e objeto, todos

⁵Google Acadêmico – “Sobre o Google Acadêmico”. Disponível em: [<https://scholar.google.com.br/intl/pt-BR/scholar/about.html>]. Acesso em: 07/2015.

representados univocamente por uma URI (W3C, 2004, 2015).

No início dos anos 1990, ontologias passaram a ser amplamente aplicadas nas áreas de medicina e biomedicina como forma de estruturar o grande volume de dados gerados. Além disso, inúmeras pesquisas sobre interoperabilidade de SIs a partir de ontologias vêm sendo realizadas (SIMON; SMITH; 2004 apud FARINELLI et al., 2013).

No trabalho de Farinelli et al. (2013), o termo interoperabilidade é definido como “a capacidade que um sistema possui de compartilhar e trocar informações e aplicações com outro sistema”. Considerando a amplitude de abrangência do termo interoperabilidade e sua definição, Ukoln (2005) e Miller (2000) se complementam, ao propor uma subdivisão a noção de interoperabilidade para tratar as diferentes formas. Não desconsiderando a importância das demais, este trabalho foca na interoperabilidade semântica. Assim, a interoperabilidade semântica refere-se a capacidade dos SIs para comunicar informações sendo que o SI receptor está habilitado para interpretar corretamente – com o mesmo sentido – tal informação, conforme previsto pelo SI transmissor (FARINELLI et al., 2013).

Considere dados originados de duas fontes de informação SI 1 e SI 2, cada uma com seu próprio vocabulário (A e B), durante a comunicação entre estas fontes ou sistemas, ocorre um problema semântico. A fonte 1 não conhece o vocabulário B assim como a fonte 2 não conhece o vocabulário A. Isto porque os SIs adotam terminologias distintas que podem apresentar diferenças como pode ser observado no estudo de Bodenreider (2008).

O objetivo da ontologia é criar um padrão semântico que seja capaz de integrar os vocabulários A e B em um vocabulário C, de forma que as fontes 1 e 2 sejam capazes de compreender uma a outra. As ontologias são instrumentos aptos a especificar explicitamente a semântica de termos pertencentes a diferentes domínios. O problema de falta de interoperabilidade semântica que entre dois sistemas pode ser tratado a partir de ontologias (FARINELLI et al., 2013).

2.2 DA WEB SEMÂNTICA AO LINKED DATA

A Web Semântica pode ser entendida como uma extensão da Web tradicional que estrutura o significado do conteúdo da Web de forma clara e bem definida, permitindo aos computadores interagir entre eles trocando informações (BERNERS-LEE; HENDLER; LASSILA, 2001).

Uma das maneiras para se publicar dados na web de forma estruturada é pela adoção dos princípios chamados *Linked Data* (LD - dados ligados). O conceito LD surgiu da ideia de Tim Berners-Lee em interligar dados na web ao invés de documentos. Em suma, LD é uma forma de

publicar dados na web de forma estruturada, de modo que uma pessoa ou máquina possa explorar estes dados (BERNERS-LEE, 2006). LD é um conceito que surge a partir da web semântica e se refere a um conjunto de melhores práticas para publicação e conexão de dados estruturados na Web (BIZER; HEATH; BERNERS-LEE, 2009).

Este conjunto de melhores práticas tornaram-se conhecidos como os princípios do LD. Estes princípios são os seguintes: i) Definir URIs como nomes para as coisas que se deseja representar; ii) Definir URIs HTTP de modo que as pessoas procurem por esses nomes em URI; iii) Quando alguém procurar por um URI, fornecer informações úteis, usando os padrões (RDF*, SPARQL); e iv) Incluir links para outros URIs a fim de que eles possam descobrir mais coisas. O LD ainda propõe o uso de *Resource Description Framework* (RDF) como mecanismo padrão para interligar os dados. Onde seus URIs identificam unicamente qualquer tipo de objeto ou conceito (BERNERS-LEE, 2006; HEATH; BIZER, 2011).

A idéia de interligação da proposta LD permite aos usuários da web navegar entre diferentes fontes. Além disso, as ferramentas de busca ficam aptas a indexar a web e fornecer recursos de pesquisa mais sofisticados sobre o conteúdo rastreado (HEATH; BIZER, 2011).

Neste contexto de LD ainda é relevante acrescentar dois novos conceitos: Dados Abertos e *Linked Open Data* (LOD). O conceito de dados abertos remete a ideia de conteúdo aberto ou disponível para todos. LOD é LD que foi lançado sob uma licença aberta, o que não impede a sua reutilização de graça, incluindo aqui o princípio de conteúdo aberto (BERNERS-LEE, 2006).

Ou seja, são dados interligados que se encontram disponíveis livremente na web, LOD é uma tendência para os acervos de dados, em qualquer campo. Ocorre, que o fato dos dados virem de fontes diversas, estas fontes podem adotar diferentes terminologias, gerando assim problemas de interoperabilidade semântica nas soluções de LD ou LOD.

2.2.1 Lacunas de interoperabilidade semântica nas soluções de Linked Data

As melhores práticas do LD são fundamentais para conectar o conteúdo de diferentes fontes em um único espaço de informação global (HEATH; BIZER, 2011). Entretanto, este conteúdo, muitas vezes definido como uma ontologia consiste apenas em representações relacionadas à semântica formal, como à Teoria dos Modelos e à Teoria dos Conceitos. Estas representações são válidas para CI, mas não são adequadas para uso em computadores e sistemas automatizados, que necessitam de representações mais formais (SOUZA *et al*, 2013).

Um dos grandes desafios a ser enfrentado para possibilitar uma maior interligação no LD, se encontra na capacidade de interoperabilidade semântica entre as entidades representadas

e publicadas. No que tange o uso de ontologias para melhorar a interoperabilidade semântica entre os dados em um ambiente de LD, considerando o ambiente composto por múltiplos SI e consequentemente múltiplas terminologias, o problema de heterogeneidade semântica pode ser solucionado de duas formas: por meio de uma solução que adota ontologias formais ou por uma solução que adota ontologias informais (FARINELLI et al., 2015, 2013).

No primeiro caso, são adicionados axiomas lógicos aos vocabulários representados com o objetivo de definir relações entre entidades nele representadas. Tais axiomas restringi o significado dos termos onde aceitam-se certas interpretações e rejeitam-se outras. Isto possibilita a eliminação automatizada de ambiguidade entre os termos. No segundo caso, a eliminação da ambiguidade é limitada, pois o fato do vocabulário ter sido definido sem o uso dos axiomas impede a restrição do o significado dos termos de forma a rejeitar ou aceitar certas interpretações, em um ambiente computacional (FARINELLI et al., 2013).

2.3 BIG DATA

Atualmente estamos cercados por uma grande quantidade de dados e informação, que apoiam nossa tomada de decisões. São dados sobre nosso dia a dia, registros sobre educação, produção de bens e serviços, investimentos e impostos governamentais, estatísticas sobre a economia, dados sobre o consumo, e dados da área de saúde (HEATH; BIZER, 2011).

Este fenômeno originalmente caracterizado pelo volume de dados, pela velocidade de geração de dados e pelas variedades de formato que os dados são disponibilizados é conhecido como *Big Data (BD)* (TAURION, 2011). Além do Volume, Velocidade e Variedade, o BD recentemente incluiu mais duas importantes características, Veracidade e Valor. Tais características são conhecidos como os 5 Vs do BD (TAURION, 2011).

O volume refere-se a grande quantidade de dados e informações que é criada a cada dia e que precisamos utilizar, a velocidade se refere aotempo que é necessário processar a agir diante da quantidade de dados. Precisamos agir praticamente em tempo real sobre este imenso volume de dados. A variedade refere-se ao fato de que os dados hoje são disponibilizados em formatos variáveis, podem ser dados estruturados como os existentes em bancos de dados ou não estruturados como textos, emails, posts de blogs e redes sociais, sons, vídeos, imagens, dados de GPS, páginas web, sensores, etiquetas RFID, etc. Entende-se como veracidade a confiabilidade que os dados proporcionam, ou seja, os dados devem expressar a realidade e devem ser consistentes. A última característica, valor, refere-se à aplicabilidade ou utilidade dos dados ao negócio, ou seja, os dados devem agregar valor ao negócio (TAURION, 2011).

O fenômeno BD surgiu, graças ao desenvolvimento das tecnologias de informação, que não só incrementou as fontes geradoras de dados, mas também proporcionou novas formas de armazenar, representar, organizar, compartilhar e analisar os dados.

2.3.1 Combinando Linked Data e Big Data

Refletindo sobre as dimensões que envolvem o BD e a perspectiva do uso do LD para promover melhorias nos serviços de saúde prestados, Hitzler e Janawicz (2013) indicam que o LD reduz a variabilidade do BD utilizando como base um dos princípios do LD que é o uso de um Identificador Uniforme de Recursos (URI, *Uniform Resource Identifier*) como mecanismo global e único de identificação de uma entidade. Entretanto, um desafio que deve ser enfrentado para possibilitar o LD, esta exatamente relacionado à interoperabilidade semântica.

Esta heterogeneidade de termos e significados pode ser vista em uma solução de BD como a variabilidade. No estudo de Farinelli et al. (2013) esta heterogeneidade semântica é vista como um problema de interoperabilidade semântica entre SIs. Os autores analisam sua solução através da especificação dos vocabulários subjacentes aos SIs sem ambiguidade. Assim, concluem que as ontologias podem ser aplicadas na busca pela interoperabilidade entre SI.

3 INFORMAÇÃO E BIG DATA EM ONCOLOGIA: ESTADO DA ARTE

Para desenvolver esta seção, utilizou-se a análise de conteúdo como modo de organizar e categorizar a produção científica. Foram analisados artigos extraídos da base indexadora PUBMED na área de Medicina em um período de cinco anos (2010-2015) e na base Scielo por um período maior (2000 a 2015) devido à carência sobre estudos relacionados à informação sobre câncer. Descritores utilizados: *Big data* e Câncer. Foram utilizados os seguintes filtros na pesquisa no PUBMED: publicados nos últimos 5 anos, somente em: Humanos, nos idiomas: inglês, alemão, português, espanhol, com o tema: Câncer. No Scielo a pesquisa realizada utilizou a busca avançada com os termos: (*big data*) AND (*cancer*). O critério de inclusão dos artigos foi à frequência dos termos no título: *Big data* e Câncer.

3.1 ORGANIZAÇÃO E REPRESENTAÇÃO DA INFORMAÇÃO NA SAÚDE

A área de saúde é constituída por uma grande quantidade e diversidade de informações provenientes de diversas fontes. Fonte de informação é um documento, dado ou registro capaz de prover aos usuários de serviços de informação, informações que quando acessadas sejam aptas a responder a suas necessidades. Um desafio na área da saúde é a recuperação da

informação devido ao volume de dados disponíveis e a diversidade de fontes de informação existentes (OLIVEIRA; ALMEIDA; QUINTELA, 2013).

Ao se pensar nas informações necessárias para a prática médica, o prontuário médico é considerado um dos mais importantes, pois é o registro do atendimento do paciente. Originalmente estes prontuários eram registrados em papel e mantidos em diversos formatos e conteúdos. Tal prática resultava em duplicação de informações, falta de padrão para troca de informações, extravio dos documentos, dentre outros problemas (MASSAD et al., 2003).

Os dados em oncologia para a prática do oncologista e atenção ao paciente com câncer podem ser obtidos em Prontuários, Registros Hospitalares de câncer, terminologias, bases de dados, etc. A área de saúde também apresenta vários recursos informacionais para organizar e disponibilizar o conhecimento, bases e dados em saúde, terminologias médicas, ontologias biomédicas, entre outros. Recursos esses que também apresentam informações e dados sobre o Câncer. A seguir serão apresentados alguns desses recursos.

Schulz e López-Garcia (2015) descrevem os diversos tipos de sistemas de organização do conhecimento da área médica, que busca o desenvolvimento e aperfeiçoamento de recursos de informação e dados mais que qualquer outra área o conhecimento:

- Tesouros: descrevem termos especializados existentes e utiliza relações semânticas como a sinonímia. Exemplo: MeSH.
- Classificações: Exemplo: Classificação Internacional de Doenças (CID 10); Classificação Internacional de Doenças para Oncologia (CID-O).
- Nomenclaturas: são padronizações de terminologias. Exemplo: SNOMED.
- Ontologias formais: são sistemas de classificação que descrevem classes de objetos do domínio usando a lógica formal. Elas são geralmente ligadas à terminologia em sentido estrito, o que pode ser visto como uma ponte para as unidades de representação ontológica primárias independentes da língua. Exemplo: OBO *foundry*⁶.

3.1.1 Prontuário eletrônico do paciente e Registro Eletrônico de Saúde

O prontuário do paciente é o documento que reúne as informações oriundas do atendimento a um paciente, necessárias para garantir a continuidade do tratamento prestados a ele. Este documento é tido como meio de comunicação entre os membros da equipe de saúde responsável pelo atendimento na instituição onde o paciente está recebendo cuidados. O prontuário subsidia a continuidade e evolução dos cuidados de saúde, abrange informações

⁶ <http://www.obofoundry.org/>

como procedimentos resultam em melhoria ou não do problema que originou a busca pelo atendimento, a identificação de novos problemas de saúde e as condutas diagnósticas e terapêuticas.

A implementação de Prontuário Eletrônico do Paciente (PEP) no Brasil ainda representa um grande desafio para devido à estrutura necessária de informática e os requisitos padrões de normas e certificações estabelecidas pelo Conselho Federal de Medicina (CFM) e pela Sociedade Brasileira de Informática em Saúde (SBIS) em relação aos critérios de segurança da informação. O PEP é necessário para possibilitar a rapidez, qualidade da informação clínica, confiabilidade dos dados e assim viabilizar a gerência de dados para: assistência ao paciente, pesquisas clínicas, criação de indicadores administrativos e gerenciar processos de avaliação da qualidade hospitalar. A tecnologia beneficia a oncologia ao melhorar o tempo de resposta e a precisão da informação, bem como a qualidade do tratamento do câncer (SOUZA, 2013).

Nos EUA, em 2006 a *American Society of Clinical Oncology* (ASCO) em conjunto com a *Clinical Practice Committee, Information Technology Committee, Cancer Research Committee*, e *The Quality Advisory Group*, criaram o *Electronic Health Record Working Group* (EHR WG). Um dos objetivos foi fornecer aos membros ASCO os conhecimentos necessários para orientar a seleção e implementação de PEP. Visou-se ainda promover o desenvolvimento de funcionalidades dentro do PEP para apoiar o tratamento do câncer.

Iniciativas como essas buscam a inovação na oncologia além de potencializar o uso de grandes quantidades de dados para direcionar o futuro do tratamento do câncer (YU, 2014). No Brasil o prontuário em papel ainda é amplamente usado, em muitos casos este prontuário é digitalizado, o que dificulta a compreensão e gestão de informações do paciente (CHIAVEGATTO FILHO, 2015).

Com o desenvolvimento das tecnologias da informação e comunicação (TIC), surgiu a possibilidade de compartilhar as informações de saúde entre instituições de saúde distintas. Assim, o PEP (de uso exclusivo e interno de uma instituição de saúde) em conjunto com as TICs deram origem ao Registro Eletrônico de Saúde (RES). O RES traz o conceito de compartilhamento de informações atrelado a ele, podendo ser usado por instituições distintas e fisicamente distribuídas (CFM;SBIS, 2012).

3.1.2 Registro hospitalar de câncer

No Brasil a sistematização de dados sobre o câncer só é possível através do Registro Hospitalar de Câncer (RHC). Segundo o INCA (2010) os RHCs são sistemas de coleta,

armazenamento, processamento e análise de dados, por localização primária do tumor e comportamento dessa patologia, em uma área geográfica específica ou unidade hospitalar. Os RHCs são SI [...] instalados em hospitais gerais ou especializados em oncologia, com o objetivo de coletar dados referentes ao diagnóstico, tratamento e evolução dos casos de neoplasia maligna atendidos nessas instituições, sejam públicas, privadas, filantrópicas ou universitárias (INCA, 2010).

Os dados coletados e consolidados provenientes dos RHC de todo o Brasil são disponíveis pelo Módulo Integrador dos Registros Hospitalares de Câncer IntegradorRHC⁷, desenvolvido pelo INCA. Para a coleta e gestão dos prontuários, foi desenvolvido o SisRHC, um sistema que abrange desde a captação, análise, pesquisa, tabulações e exportações dos dados. É um sistema informatizado via internet, “*para o envio, a consolidação, o acompanhamento e a análise dos dados nacionais dos RHC brasileiros - Integrador RHC*” (INCA, 2010).

A coleta de dados pelo RHC possibilita os profissionais de saúde à compreensão e o enfrentamento da incidência de câncer de forma mais eficaz (CDC, 2015). Esta coleta só é possível através de prontuários que segundo Souza (2013) o prontuário é “*fonte de informação no tratamento do câncer, fonte de ensino e pesquisa para o conhecimento e notificação dos casos e características da enfermidade*”. Na oncologia RHC serve também para a pesquisa clínica e é uma fonte para dados estatísticos do INCA de incidência, prevalência e mortalidade do câncer.

A qualidade das informações do prontuário do paciente é fator fundamental para qualidade do RHC. No futuro, a ligação entre os RHCs e PEPs com outras fontes de dados sobre câncer podem permitir a transferência mais transparente e uma entrada de dados mais precisa do que a entrada de dados feita de forma manual, proporcionando assim ferramentas eletrônicas mais eficientes e livres para expandir a quantidade de dados capturada, bem como aumentar a qualidade desses dados (PEZZI, 2014).

3.2 BIG DATA E IMPLICAÇÕES PARA A PRÁTICA DO ONCOLOGISTA

Soluções para área de saúde baseadas no conceito de BD podem ser usadas na medicina de precisão, nos registros eletrônicos de saúde e internet das coisas. Observa-se que seu uso proporcionará retornos substanciais à área de saúde, tanto em termo de tempo, dinheiro e vidas quanto em termos de oportunidades para o avanço do conhecimento em saúde

⁷ Disponível em: <https://irhc.inca.gov.br/RHCNet/>.

(CHIAVEGATTO FILHO, 2015). Na era do BD a capacidade de criação, armazenamento, compartilhamento e manipulação de dados irá crescer exponencialmente. Grandes quantidades de dados sobre pacientes com câncer são utilizados, por exemplos, em estudos epidemiológicos como o realizado por Medeiros et al., (2015) que analisou os dados de aproximadamente 8336 pacientes com Leucemia Mieloide Aguda (LMA) nos Estados Unidos.

Estudos como esse são importantes por analisar uma grande quantidade de dados na busca por características semelhantes em determinada população e assim realizar estimativas de incidência e contribuir com dados para o planejamento da assistência e campanhas de detecção precoce do câncer. O desafio desta era tecnológica será possibilitar que dados, sistemas de organização do conhecimento e infra-estrutura tecnológica trabalhem juntos para viabilizar a pesquisa e tratamento do paciente com câncer.

Meyer e Basch (2015) analisam que para a prática do oncologista a infra-estrutura de *BD* contribui para a tomada de decisão em relação ao tratamento do paciente que demonstra ser cada vez mais complexa devido necessidade de uma medicina personalizada juntamente com o ritmo acelerado do desenvolvimento científico. A qualidade dos dados ainda é um tema a ser amplamente discutido (PEZZI, 2014).

3.2.1 Volume dos dados na área de Oncologia

A adoção dos RES vem gerando enormes volumes de dados. Um estudo realizado pela *American Hospital Association* mostrou que nos Estados Unidos, a adoção de RESs dobrou de 2009 a 2011, graças aos investimentos em TI na saúde (MURDOCH; DETSKY, 2013). Na área de oncologia uma das principais fontes de informações sobre o paciente e a doença é o prontuário do paciente. Com os dados do prontuário é realizado o RHC que fornece uma quantidade de dados crescente sobre como os cânceres são tratados para propósitos de pesquisa clínica em oncologia (PEZZI, 2014).

Devido aos altos índices de incidência do câncer no Brasil e no mundo, os dados tendem a crescer exponencialmente. O INCA estimou para o ano de 2015, a ocorrência de aproximadamente 576 mil casos novos de câncer. O mais incidente na população brasileira será o câncer de pele do tipo não melanoma (182 mil casos novos), seguido pelos tumores de próstata (69 mil), mama feminina (57 mil), cólon e reto (33 mil), pulmão (27 mil), estômago (20 mil) e colo do útero (15 mil). Para os homens, os tipos mais incidentes serão os cânceres de próstata, pulmão, cólon e reto, estômago e cavidade oral. Entre as mulheres, será mama, cólon e reto, colo do útero, pulmão e glândula tireoide.

O INCA (2014) relata ainda que em 2030, serão 21,4 milhões de casos novos de câncer e 13,2 milhões de mortes por câncer, em função do aumento e envelhecimento da população e também em função da redução da mortalidade infantil e das mortes por doenças infecciosas nos países em desenvolvimento. Meyer e Basch (2015) relatam que como perspectivas futuras haverá a continuidade do crescimento de dados e informação sobre o câncer, bem como a estrutura tecnológica e sistemas de informação em saúde.

3.2.2 Variabilidade dos dados na área de Oncologia

Na oncologia é necessário manter uma infinidade de dados, como: prevenção, estimativas de câncer, perfil epidemiológico, terminologias e classificações (CID-O, *French-American-British* – FAB, CICI) estadiamento de tumores (TNM). Esses dados são obtidos por meio da: história do paciente, diagnóstico, hemograma, amostras da medula óssea (aspiração e biópsia); análise citogenética; exames específicos como imunofenotipagem, FISH, HER2; além de dados sobre tratamento (quimioterapia, radioterapia – doses, data, etc), transplante de células tronco; terapias alvo específicas (também chamadas de anticorpos monoclonais), cuidados paliativos, hormonioterapia e por último, medicamentos de suporte, prognóstico, sobrevida, mortalidade, informações genéticas, etc. (NATIONAL CANCER INSTITUTE, 2013).

Esses dados podem ser encontrados em uma infinidade de formatos: formatos textuais estruturados ou livres (não-estruturados); bases de dados estruturadas, gráficos, imagens, vídeos, áudios, sinais, dados categóricos, dados numéricos escalares, ordinais, dados lógicos, descritivos, códigos, e outros (SABBATINI, 1994). Assim irão produzir milhares de informações e artigos na área de oncologia.

Uma análise bibliométrica sobre a representatividade do câncer na literatura médica no ano de 2007, relatou que artigos sobre oncologia constituíram cerca de 25% de todas as publicações entre as 20 revistas médicas com os maiores fatores de impacto (GLYNN et al., 2010).

Estima-se que 80% dos dados de saúde que são criados se encontram disponíveis em formato não-estruturado, grande parte dos quais é baseado em texto (MURDOCH; DETSKY, 2013). Um desafio ainda é a pouca utilização de padrões de dados uniformes para registros médicos eletrônicos existentes. Sugere-se que é preciso ter a capacidade de processamento de dados em qualquer formato (SCHNEEWEISS, 2014).

3.2.3 Oportunidades ao profissional da informação

Os estudos que utilizam dados do BHD necessitam mais do que o simples acesso a esses grandes volumes de dados, necessitam de recursos humanos e tecnológicos na integração de sistemas. Isso exige o trabalho colaborativo transdisciplinar do profissional da informação com profissionais da TI (MEYER et al., 2014).

O bibliotecário, por exemplo, que atua na Oncologia integra a equipe clínica multiprofissional com o papel de procurar, filtrar e fornecer melhores evidências para decisões clínicas, dando suporte à equipe médica através de levantamentos bibliográficos. (SOUZA; SANTOS, 2013). Assim este profissional, possui a habilidade e o conhecimento para ajudar a equipe médica em oncologia em relação às possibilidades do BHD.

Para se obter ganhos reais com BHD não basta digitalizar os PEP é necessário garantir seu uso integrado, que permita o uso remoto do mesmo prontuário por todas as instituições de saúde (CHIAVEGATTO FILHO, 2015). Esta integração é uma atividade de organização e entendimento da realidade característicos da prática em CI. (SOUZA et al., 2013).

4 RESULTADOS: APRESENTAÇÃO E DISCUSSÃO

Esta seção apresenta os resultados da pesquisa bibliométrica realizada para definir o estado da arte do BD na oncologia. Além disso, discute sobre a aplicação das ontologias formais nas soluções de LD como alternativa para aprimorar a interoperabilidade semântica minimizando a variabilidade dos dados nas soluções de BD aplicados à Oncologia. Adota-se o nome *Linked Health Data* (LHD) para se referir à aplicação de LD na área biomédica.

4.1 BIBLIOMETRIA: BIG DATA E CANCER

Na pesquisa no PUBMED foram encontrados 165 artigos, entretanto somente 9 apresentavam as palavras *Big Data* e câncer ou oncologia no título, relacionados a temática da pesquisa. No Scielo a busca avançada com os termos: (big data) AND (cancer) obteve apenas 4 artigos que não correspondiam ao tema proposto.

QUADRO 1 – Artigos sobre Big data e câncer no PUBMED

TÍTULO	PERIÓDICO	ASSUNTO	ANO DE PUBLICAÇÃO
Cancer_research in the era of next-generation sequencing and_big data_calls for intelligent modeling.	Chin J Cancer	Utilização de algoritmos em <i>big data</i> , <i>machine learning</i> na área de pesquisa sobre câncer.	2015
Artificial Intelligence, Big Data, and Cancer	JAMA Oncol.	the incorporation of large databases using artificial intelligence into large cancer open networks to inform, educate, and help	2015

		cancer treatment and research.	
Big data analysis of treatment patterns and outcomes among elderly acute myeloid leukemia patients in the United States.	Ann Hematol	Análise de dados da base epidemiológica: o SEER. Analisaram dados de pacientes idosos diagnosticados com leucemia mieloide aguda em relação a tratamento e sobrevida.	2015
Big data mining yields novel insights on cancer.	Nat Genet	Considerando os desafios da efetividade da mineração e modelagem usando o 'big data' em novas descobertas biológicas foi realizado um novo estudo sobre <i>Gene Expression Omnibus</i> para fazer novas descobertas sobre genes envolvidos na reparação de danos de DNA e genoma de instabilidade no câncer	2015
Can big data cure cancer?	Fortune	Tema fictício: conta história de adolescentes que construíram um <i>app</i>	2014
Big data for population-based cancer research: the integrated cancer information and surveillance system.	N C Med J	Utilização de informações de base populacional do <i>The Integrated Cancer Information and Surveillance System</i> (ICISS) na pesquisa do câncer, que possibilita o desenvolvimento de sistemas para manipulação de grandes volumes de dados, permitindo os pesquisadores melhor aproveitar o poder do Big Data na promoção da saúde da população.	2014
Collaborative biomedicine in the age of big data: the case of cancer.	J Med Internet Res	Analisa o conceito de Medicine 2.0 relacionado as plataformas de informática e políticas relacionadas com a saúde podem ser utilizada para incentivar a liquidez de dados e inovação, através da visão de um "sistema de saúde aprendizagem rápida" com permissão ao código aberto, a abordagem de base populacional para prevenção e controle do câncer.	2014
Individualizing care for ovarian cancer patients using big data.	J Natl Cancer Inst.	Análise de dados dos estudos publicados sobre câncer de ovário agrupados em uma base de dados e avaliados para prever sobrevida em pacientes com esta neoplasia.	2014
Big data and clinical research in oncology: the good, the bad, the challenges, and the opportunities.	Ann Surg Oncol.	Analisa os principais problemas e perspectivas futuras dos três programas nacionais de registro hospitalar de câncer nacionais dos EUA.	2014
Big data infrastructure for cancer outcomes research: implications for the practicing oncologist	Journal of Oncology Practice	Analisa o termo Big Data no contexto da oncologia, seus desafios, oportunidades e iniciativas em andamento.	2015

Em relação aos artigos encontrados na pesquisa verifica-se a utilização do *Big Health Data* na oncologia relacionado a: *Big data* e *machine learning* em pesquisas de câncer ; utilização de dados da base epidemiológica: o SEER⁸, bases de dados e inteligência artificial no tratamento do câncer; utilização de informações de RHC em relação ao BD na promoção

⁸ Surveillance, Epidemiology, and End Results (SEER)-Medicare database.

da saúde da população; extração de dados em bases de dados na busca por padrões; registro hospitalar de câncer. Pela análise de assunto dos artigos, verificou-se que os desafios do BD na Oncologia se referem à necessidade de integração de dados, modelagem conceitual, mineração de dados, inteligência artificial relacionados à pesquisa e tratamento do câncer. Entretanto o tema ontologia para promover a interoperabilidade, ainda não aparece neste contexto.

4.2 A CI EM FAVOR DO BIG HEALTH DATA

Em um ambiente de *Linked Health Data* (LHD), a ideia é interligar dados médicos oriundos de diversas fontes. Estas fontes podem ser SIs distintos de uma mesma instituição de saúde, ou no caso de um ambiente de *Linked Open Health Data* (LOD), os dados pode ser oriundos de SIs de instituições distintas. Pelo que foi apresentado na seção 2.2.1, um desafio nestes ambientes é sobre a semântica ou o significado dos termos definidos em cada SIs, uma vez que eles podem trabalhar com terminologias distintas.

Considerando a existencia de três SI – sistema A, sistema B e sistema C – sendo que o sistema A utiliza o *NCI-Thesaurus*, o sistema B utiliza o SNOMED CT e o sistema C utiliza o UMLS. Cada um destes SI podem utilizar um termo diferente para se referir a entidades que possuem o mesmo significado. Podem ainda utilizar o mesmo termo, mas para se referir a entidades distintas, que possuem significados diferentes. Este desafio de lidar com semânticas distintas torna-se ainda mais complexo quando o formato dos dados também é distinto, como no caso dos dados de saúde, que podem assumir formatos e tipos de dados distintos (seção 3.3.2).

Para lidar com estas dificuldades, o LD estabelece princípios e melhores práticas de forma à estrutura os dados em um esquema padronizado. Para isso pode-se trabalhar com as definições em ontologias. As ontologias podem ajudar a aumentar a semântica associada à representação de conteúdos de fontes e formatos diversas na web, possibilitando o entendimento homogêneo dos conceitos de diferentes domínios. Mas não basta simplesmente representar os termos ou conteúdos por meio de uma ontologia, esta deve ser uma ontologia formal, onde define-se as relações entre entidades por meio de axiomas, permitindo assim restringir o significado dos termos e a eliminação automatizada de ambiguidade (seção 2.1.1).

Entretanto, não basta representar os dados em um padrão comum e disponibilizá-los. Para que os dados sejam realmente úteis, é preciso saber capturá-los, processá-los, analisá-los e descartá-los. Assim, uma solução de BHD pode viabilizar a pesquisa e tratamento do paciente com câncer, onde a representação dos dados oncológicos (seção 3.1) é realizada por meio de uma ontologia formal. Os dados podem ser oriundos de um ambiente de LOD interoperável.

No caso, como já existe o SI IntegradorRHC para consolidar os dados sobre ocorrências e tratamento ao câncer no Brasil, este SI poderia evoluir no sentido de um ambiente de LOD.

Quanto mais os dados forem interligados com outros dados, maior será o seu valor e sua utilidade. Além disso, os links têm significado, são links semânticos, expressam a relação entre os recursos linkados. Por meio de SI especializados em BD, seria possível pesquisar este enorme repositório de dados oncológicos a fim de identificar tendências e padrões em diagnósticos e tratamentos, além de subsidiar as pesquisas na área com uma grande quantidade de dados. Tal solução seria favorável para viabilizar e maximizar a pesquisa e tratamento do paciente com câncer.

5 CONSIDERAÇÕES FINAIS

O presente artigo apresentou os conceitos fundamentais de interoperabilidade, ontologia, LD e BD, além de discutir o papel da CI na era do BD. Apresentou os resultados da pesquisa bibliométrica realizada onde buscou-se identificar o estado da arte sobre o BHD no campo da oncologia. Observou-se que o fenômeno BHD tem o potencial de transformar a prática médica na oncologia utilizando informações geradas todos os dias para melhorar a qualidade e a eficiência dos cuidados ao paciente com câncer. Assim, conclui-se que a primeira questão de investigação foi respondida e o primeiro objetivo do artigo alcançado.

Discutiu-se o papel da CI como área de apoio aos profissionais de saúde no que diz respeito às soluções de organização e recuperação de informação. Observou-se que a web semântica e o LD são propostas hábeis para lidar com questões de organização e recuperação de informações em ambientes de fontes de dados distribuídas. Ainda foi discutido sobre o uso das ontologias formais como alternativa a representação da informação a fim de viabilizar a interoperabilidade semântica nos ambientes distribuídos. Assim, conclui-se que a segunda questão de investigação foi respondida e o segundo objetivo do artigo alcançado.

Finalmente, discutiu sobre uma proposta para soluções de BHD na área de Oncologia em um ambiente de LHD semanticamente interoperável com a representação dos dados oncológicos em ontologias formais. Assim, conclui-se que o terceiro objetivo do artigo alcançado.

A respeito das perspectivas da notificação dos casos e câncer no Brasil o INCA (2010) afirma que no futuro, espera-se que haja uso crescente de recursos de informática, interligação de bases de dados, sejam elas do hospital ou externas, com troca automática de informações e implantação de rotinas pré-estabelecidas de convocação e acompanhamento de pacientes, para

controle de doenças crônicas e seguimento. Tudo isto com segurança de acesso, preservação das informações e redução das áreas físicas dos arquivos de prontuários.

No Brasil os desafios são enormes devido a ao fato de muitas instituições de saúde ainda utilizarem o prontuário em papel. O RHC precisa continuar na busca pela qualidade na coleta de seus dados e no futuro otimizar sua coleta através dos PEP. São necessárias mais fontes de BHD direcionada para a oncologia. O sucesso neste contexto dependerá de como oncologistas irão se integrar sistemas de dados. O trabalho transdisciplinar se configura como uma possibilidade para otimização e interoperabilidade entre BHD e SI em oncologia. O papel do profissional da informação será de promover a interoperabilidade entre os sistemas de organização do conhecimento na oncologia e assim potencializar o uso de BHD.

Desta forma, a busca por melhoraria nas coleções de dados assim como a padronização e interoperabilidade é requisito essencial para promover a confiança e transparência em pesquisas clínicas na oncologia.

REFERÊNCIAS

BERNERS-LEE, T. ;HENDER, J. ;LASSILA, O. The semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.

Scientific American, New York, may. 2001. Disponível para comprar em:

<http://www.scientificamerican.com/article/the-semantic-web/>.

BERNERS-LEE, T. Design issues: Linked data. 2006. Disponível em:

<<http://www.w3.org/DesignIssues/LinkedData.html>>. Acesso em: 01 set. 2014.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: Principles and state of the art. In:

World Wide Web Conference. 2008. Disponível em: <<http://goo.gl/a0iKbt>>. Acesso em: 01 set. 2014.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. **Semantic services, interoperability and web applications**: Emerging Concepts, p. 205-227, 2009.

BODENREIDER, O. Comparing SNOMED CT and the NCI Thesaurus through Semantic Web Technologies. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE

REPRESENTATION IN MEDICINE (KR-MED), 3., 2008. **Proceedings...** Disponível em:

<<http://mor.nlm.nih.gov/pubs/pdf/2008-krmed-ob.pdf>>. Acesso em: julho/2015.

BRASIL. Ministério da Saúde. Secretaria de Assistência à Saúde. Instituto Nacional de Câncer. **Registro hospitalar de câncer**: dados dos hospitais do INCA, relatório anual 1994/1998. Rio de Janeiro: INCA, 2004.

CENTERS FOR DISEASE CONTROL AND PREVENTION DIVISION OF CANCER PREVENTION AND CONTROL (CDC). National Program of Cancer Registries (NPCR).

Disponível em: ,<http://www.cdc.gov/cancer/npcr/>>. Acesso em: julho/2015.

CHIAVEGATTO FILHO, A.D.P. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. **Epidemiologia e Serviços de Saúde** [online]. v. 24, n. 2, p. 325-332. 2015.

CONSELHO FEDERAL DE MEDICINA (CFM) & SOCIEDADE BRASILEIRA DE INFORMÁTICA EM SAÚDE (SBIS). Cartilha sobre Prontuário Eletrônico: a Certificação de Sistemas de Registro Eletrônico de Saúde. Fevereiro, 2012. Disponível em: <http://portal.cfm.org.br/crmdigital/Cartilha_SBIS_CFM_Prontuario_Eletronico_fev_2012.pdf>. Acesso em: julho/2015.

FARINELLI, F.; ALMEIDA, M.B.; SOUZA, Y. Linked Health Data: how linked data can help to provide better health decisions Status: In: MEDINFO, 15., 2015. **Proceedings...** Brasil, 2015. (poster).

FARINELLI, F.; SILVA, S.M.; ALMEIDA, M.B. O papel das ontologias na interoperabilidade de sistemas de informação: reflexões na esfera governamental. In: ENANCIB, 14. **Anais...** 2013.

FRIDLEY, B.L.; KOESTLER, D.C.; GODWIN, A.K. Individualizing care for ovarian_cancer_patients using_big data. **The Journal of the National Cancer Institute**, v. 106, n. 5, apr.2014 .

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo, v. 5, 2002.

GLYNN, R. W. et al. Representation of cancer in the medical literature--a bibliometric analysis. **PLoS One**, v. 5, n. 11, p. e13902, 2010.

GOLBECK, J; et al. **National Cancer Institute's Thesaurus and Ontology** .2003. Disponível em: <<http://www.websemanticsjournal.org/index.php/ps/article/view/27/25>>. Acesso em: 30 out. 2013.

GOMES JUNIOR, S. C. S; ALMEIDA, R. T. Modelo de simulação para estimar a infraestrutura necessária à assistência oncológica no sistema público de saúde. **Revista Panamericana de Salud Pública**, Washington , v. 25, n. 2, p. 113-119, Feb. 2009. Disponível em: <<http://goo.gl/UmuWzm>>. Acesso em: 15 ago. 2015.

HAWKINS, D. T.; LARSON, S.E.; CATON, B. Q. Information science abstracts: tracking the literature of information science. Part 2: a new taxonomy for information science. **Journal of the American Society for Information Science and Technology**, v. 54, n. 8, p. 771-781, 2003.

HEATH, T.; BIZER, C. Linked data: evolving the web into a global data space. **Synthesis lectures on the semantic web: theory and technology**, v. 1, n. 1, p. 1-136, 2011. Disponível em: <<http://goo.gl/t89BbN>>. Acesso em: 15 ago. 2015.

HITZLER, P.; JANOWICZ, K. Linked data, Big Data, and the 4th Paradigm. **Semantic Web**, v. 4, n. 3, p. 233-235, 2013.

INCA. **Registros hospitalares de câncer: planejamento e gestão**. 2. ed. rev. Rio de Janeiro: INCA; 2010.

INCA. **Coordenação de Prevenção e Vigilância Estimativa 2014**: incidência de Câncer no Brasil. Rio de Janeiro: INCA, 2014. 124 p. Disponível em: <<http://www.inca.gov.br/estimativa/2014/estimativa-24042014.pdf>>. Acesso em: ago 2015.

JIANG, P.; LIU, X. S. Big data mining yields novel insights on cancer. **Nature Genetics**.v. 47, n. 2, p. 103-4, Feb. 2015.

KANTARJIAN, H.; YU, P. P. Artificial Intelligence, Big Data, and Cancer. **JAMA Oncology**. 2015.

KLEIN, I. M. Oncology enters the informatics era. **The American Journal of Managed Care**. v.18, n.5 Spec No., p.SP218-21, Dec. 2012.

LAKATOS, E. M.; MARCONI, M. A. **Metodologia do trabalho científico**. São Paulo: Atlas, 1991.

MASSAD, E. MARIN, H. F.; AZEVEDO NETO, R. S.. **O prontuário eletrônico do paciente na assistência, informação e conhecimento médico**. São Paulo: FMUSP/UNIFESP/OPAS, 2003.

MEDEIROS, B.C.; et al. Big data analysis of treatment patterns and outcomes among elderly acute myeloid leukemia patients in the United States. **Annals of Hematology**, v. 94, n. 7, p. 1127-38, jul. 2015.

MEYER, A. M.; BASCH, E. Big data infrastructure for cancer outcomes research: implications for the practicing oncologist. **Journal of Oncology Practice**, v. 11, n. 3, p. 207-8, may 2015.

_____ et al. Big data for population-based cancer research: the integrated cancer information and surveillance system. **North Carolina Medical Journal**, v. 75, n. 4, p. 265-9, Jul./Aug. 2014.

MILLER, P. **Interoperability**: what is it and why should I want it? 2000. Disponível em: <<http://www.ariadne.ac.uk/issue24/interoperability/>>. Acesso em: 05 jun. 2013.

MURDOCH, T. B., DETSKY, A. S. The inevitable application of big data to health care. **The Journal of American Medical Association (Jama)**, v. 309, n. 13 , p. 1351-1352, 2013.

NATIONAL CANCER INSTITUTE (United States). PDQ® - NCI's **Comprehensive Cancer Database**. Disponível em; <<http://www.cancer.gov/cancertopics/pdq>>. Acesso em: 07 fev. 2013.

ODDONE, N.; GOMES, M. Y. F. S. F. Uma nova taxonomia para a ciência da informação. In: ENANCIB, 5., 2003, Belo Horizonte. **Anais...** Belo Horizonte: UFMG, 2003.

OLIVEIRA, J. P; ALMEIDA, M. B.; QUINTELA, E. L. Uma visão geral sobre fontes de informação em saúde. In: CONGRESSO ISKO ESPANHA E PORTUGAL,1., ; CONGRESSO ISKO ESPANHA, 11., 2013. p. 993-1008.

PEZZI, C.M. Big data and clinical research in oncology: the good, the bad, the challenges, and the opportunities. **The Annals of Surgical Oncology**, v. 21, n. 5, p. 1506-7, May 2014.

SABBATINI, R. M. O ensino da informática aplicada à Medicina. Roteiros e bibliografia básica. **Revista Informédica**, v. 2, n. 8, p. 5-12, 1994.

SCHNEEWEISS, S.. Learning from big health care data. **New England Journal of Medicine**, v. 370, n. 23, p. 2161-2163, 2014. Disponível em: <<http://www.nejm.org/doi/pdf/10.1056/NEJMp1401111>>. Acesso em: 13 ago. 2015.

SCHULZ, S.; LÓPEZ-GARCÍA, P. [Big data, medical language and biomedical terminology systems]. **Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz**, v. 58, n. 8, p. 844-52, jun. 2015.

SHAIKH, A.R.; et al. Collaborative biomedicine in the age of big data: the case of cancer. **Journal of Medical Internet Research**, v. 16, n. 4, p. e101, Apr. 2014.

SMITH, B.; et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. **Nature Biotechnology**, v. 25, n. 11, p. 1251-5, nov. 2007.

SOUZA, A. D.; SANTOS, E. L. Atuação do bibliotecário clínico em oncologia. In: DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO, 25., **Anais...** Florianópolis, SC, Brasil, 2013. Disponível em: <<http://goo.gl/YFUM3Z>>. Acesso em: 13 ago. 2015.

SOUZA, R. R. TUDHOPE, D.; ALMEIDA, M. B. Towards a taxonomy of KOS: Dimensions for classifying Knowledge Organization Systems. **Knowledge organization**, v. 39, n. 3, p. 179-192, 2012.

SOUZA, R. R.; ALMEIDA, M. B.; BARACHO, R. M. A.. Ciência da informação em transformação: *Big Data*, nuvens, redes sociais e Web Semântica. **Ciência da Informação**, Brasília, DF, v. 42 n. 2, p. 15-29, maio/ago., 2013. Disponível em: <<http://goo.gl/YFUM3Z>> . Acesso em: 13 ago. 2015.

SOUZA, A.D. **O prontuário do Paciente do Serviço de Oncologia da Santa Casa de Misericórdia de Belo Horizonte**: fonte de informação para melhor atenção no tratamento do paciente com câncer. 54 f. Monografia (Especialização em Gestão Estratégica da Informação) -Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2013.

TAURION, Cezar. **Big data**. Rio de Janeiro: Brasport, 2013.

UKOLN. **Interoperability focus**: looking at interoperability, 2005. Disponível em: <<http://www.ukoln.ac.uk/interop-focus/about/leaflet.html>>. Acesso em: 04 jul. 2013.

YU, P.P. Oncology in the digital health age. **Journal of Oncology Practice**, v. 10, n. 4, p. 227-30, jul. 2014.