MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação Biologia Computacional e Sistemas

MINERADOR DE EVENTOS ADVERSOS MALÁRICOS NO TWITTER – O CASO DA DOXICICLINA

FELIPE VIEIRA DUVAL

Rio de Janeiro
Maio de 2016

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ
Programa de Pós-Graduação em Biologia Computacional e Sistemas

FELIPE VIEIRA DUVAL

Minerador de eventos adversos maláricos no Twitter – O caso da Doxiciclina

Dissertação apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Mestre em Biologia Computacional e Sistemas.

Orientador: Prof. Dr. Fabrício Alves Barbosa da Silva

RIO DE JANEIRO
Maio de 2016

Ficha catalográfica elaborada pela
Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

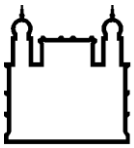
D983 Duval, Felipe Vieira

Minerador de eventos adversos maláricos no Twitter: o caso da doxiciclina / Felipe Vieira Duval. – Rio de Janeiro, 2016.
xv, 80 f. : il. ; 30 cm.

Dissertação (Mestrado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2016.
Bibliografia: f. 55-59

1. Evento adverso. 2. Farmacovigilância. 3. Twitter. 4. Mineração de dados. 5. Big Data. 6. Análise de desproporcionalidade. I. Título.

CDD 615.704



Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: FELIPE VIEIRA DUVAL

MINERADOR DE EVENTOS ADVERSOS MALÁRICOS NO TWITTER – O CASO DA DOXICICLINA

ORIENTADOR: **Prof. Dr. Fabrício Alves Barbosa da Silva**

Aprovada em: _____/_____/_____

EXAMINADORES:

Prof. Marcelo Ribeiro Alves (INI/FIOCRUZ) – Presidente e Revisor
Prof. Maria Clícia Stelling de Castro (UERJ)
Prof. Marcelo Ferreira da Costa Gomes (PROCC/FIOCRUZ)
Prof. Ernesto Caffarena (PROCC/FIOCRUZ) - Suplente
Prof. Kele Teixeira Belloze (CEFET/RJ) - Suplente

Rio de Janeiro, 16 de maio de 2016

Anexar a cópia da Ata que será entregue pela SEAC já assinada.

Agradecimentos

Aos meus pais, Carlos Alberto e Esmeralda, que sempre me deram a base e o incentivo para a continuidade dos meus estudos, por me apoiarem nos momentos difíceis e pela compreensão da minha ausência em vários momentos.

A minha amada namorada, que esteve presente me ajudando e alegrando nos momentos fáceis e difíceis ao longo de toda trajetória desse trabalho. Te amo Amanda!

Ao Fabrício, meu orientador, pela confiança e paciência que teve comigo na realização desse trabalho. Pelas várias reuniões que teve comigo durante o desenvolvimento desse trabalho e principalmente na ajuda que me deu na etapa final.

Aos amigos do PROCC, Rafael, Lucas, Deborah, Janaina, Gisele, Lucianna, André, Vanessa, Thiago, Carlos, Aline e todos os outros, obrigado pela convivência durante meus anos na Fiocruz.

Aos docentes deste programa de pós-graduação, só tenho a agradecer pela vivência que obtive nesses anos de experiência acadêmica e científica onde obtive os créditos necessários para a conclusão da minha dissertação. Não poderia deixar de citar alguns nomes como: Ernesto, Oswaldo, Ana Carolina, Paulo Ricardo, Leonardo e Maurício.

Ao Marcelo Ribeiro, por aceitar fazer a revisão deste trabalho.

Aos membros da banca examinadora, por terem aceitado o convite da avaliação desse trabalho e pela certeza da ajuda crítica e construtiva do mesmo.

À CAPES, pelo investimento através da bolsa de estudos recebida.

À todos aqueles que participaram direta ou indiretamente e que, de alguma forma, contribuíram para minha formação.

Provavelmente eu me esqueci de citar alguém, mas ficam aqui meus sinceros agradecimentos.

“Why so serious?”

The Joker

(Coringa interpretado por Heath Ledger em Batman – O Cavaleiro das Trevas)

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

MINERADOR DE EVENTOS ADVERSOS MALÁRICOS NO TWITTER – O CASO DA DOXICICLINA

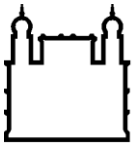
RESUMO

DISSERTAÇÃO DE MESTRADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Felipe Vieira Duval

Durante o período de pós-comercialização, quando medicamentos são usados por grandes populações e por períodos de tempo maiores, eventos adversos inesperados podem ocorrer, o que altera a relação risco-benefício dos medicamentos o suficiente para exigir uma ação regulatória. Eventos adversos no período de pós-comercialização podem requerer um aumento significativo de cuidados de saúde e resultar em danos desnecessários, muitas vezes fatais, aos pacientes. Portanto, a descoberta o quanto antes de eventos adversos no período de pós-comercialização é um objetivo principal do sistema de saúde. Alguns países possuem sistemas de vigilância farmacológica responsáveis pela coleta de relatórios voluntários de eventos adversos na pós-comercialização, mas estudos já demonstraram que com a utilização de redes sociais como o Twitter, pode-se conseguir um número maior de relatórios. O objetivo principal desse projeto é construir um sistema totalmente automatizado que utilize o Twitter como fonte para encontrar eventos adversos novos e já conhecidos e fazer a análise estatística dos dados obtidos. Para tal, foi construído um sistema que coleta, processa, analisa e avalia tweets em busca de eventos adversos, comparando-os com dados da FDA e do padrão de referência construído. Nos resultados obtidos conseguimos encontrar eventos adversos novos e já existentes relacionados ao medicamento doxiciclina o que demonstra que o Twitter pode sim, ser útil para a Farmacovigilância quando utilizado em conjunto com outras fontes de dados.

Palavras Chave: Evento adverso, Farmacovigilância, Twitter, Mineração de dados, Big Data, Análise de desproporcionalidade.



Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

MINING MALARIAL ADVERSE EVENTS ON TWITTER – THE DOXYCYCLINE CASE

ABSTRACT

MASTER DISSERTATION IN COMPUTATIONAL SYSTEMS BIOLOGY

Felipe Vieira Duval

At the post-marketing phase when drugs are used by large populations and for long periods, unexpected adverse events may occur altering the risk-benefit trade off of drugs, sometimes requiring a regulatory action. These events at the post-marketing phase require a significant increase in health care since they result in unnecessary damage, often fatal, to patients. Therefore, the early discovery of adverse events in the post-marketing phase is a primary goal of the health system. Some countries have pharmacovigilance systems used to collect voluntary reports of adverse events from post-marketing, but studies have shown that with the use of social networks like Twitter, you can get a greater number of reports. The main objective of this project is to build a fully automated system that uses Twitter as a source to find new and already known adverse events and do statistical analysis of the data. To achieve this, we build a system that collects, processes, analyzes and evaluates tweets in searching for adverse events and comparing them with the FDA data and gold standard built. New and already known adverse events related to the doxycycline drug were found in the result, which shows that Twitter can indeed be useful for pharmacovigilance when used with other data sources.

Keywords: Adverse event, Pharmacovigilance, Twitter, Data mining, Big Data, Disproportionality analysis.

ÍNDICE

ÍNDICE DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

ADP - Análise de Desproporcionalidade

ADReCS - Adverse Drug Reaction Classification System, na sigla em inglês

AIDS - Acquired Immune Deficiency Syndrome

ANVISA - Agência Nacional de Vigilância Sanitária

API - Application Program Interface, na sigla em inglês

BCPNN - Bayesian Confidence Propagation Neural Network, na sigla em inglês

cTAKES - clinical Text Analysis and Knowledge Extraction System, na sigla em inglês

EA - Evento Adverso

FAERS - FDA Adverse Event Reporting System, na sigla em inglês.

FDA - Food and Drug Administration, na sigla em inglês

IC - Information Component, na sigla em inglês

KDD - Knowlegde Discovery in Databases, na sigla em inglês

LVG - Lexical Variant Generation, na sigla em inglês

MedLEE - Medical Language Extraction and Encoding, na sigla em inglês

MGPS - Multi-item Gamma-Poisson Shrinker, na sigla em inglês

NIH - National Institutes of Health, na sigla em inglês

Notivisa - Sistema de Notificações em Vigilância Sanitária

OMS - Organização Mundial da Saúde

PLN - Processador de Linguagem Natural

PRR - Proportional Reporting Ratios, na sigla em inglês

RAM - Reações Adversas a Medicamentos

RDC - Resolução da Diretoria Colegiada

REST - Representational State Transfer, na sigla em inglês

ROR - Reporting Odd Ratios, na sigla em inglês

RR - Reporting Ratios, na sigla em inglês

SIDER - Side Effect Resource, na sigla em inglês

SNOMED - Systematized Nomenclature of Medicine, na sigla em inglês

SRE - Sistemas de Relato Espontâneo

UFARM - Unidade de Farmacovigilância

UMLS - Unified Medical Language System, na sigla em inglês

1 Introdução

Durante o período de pós-comercialização, quando medicamentos são usados por grandes populações e por períodos de tempo maiores, eventos adversos (EA) podem ocorrer, o que altera a relação risco-benefício dos medicamentos o suficiente para exigir uma ação regulatória. Os eventos adversos são definidos como agravos à saúde de um usuário ou de um paciente que podem surgir durante o tratamento com um produto farmacêutico, podendo ser erros de medicação, desvio de qualidade dos medicamentos, reações adversas a medicamentos (RAM), interações medicamentosas e intoxicações [1].

Os eventos adversos podem ser identificados durante a fase de estudo sobre o medicamento que ocorre antes da comercialização, conhecida como fase clínica. Os testes clínicos com medicamentos ocorrem em três etapas distintas, conhecidas como Fases I, II e III, sendo iniciado com voluntários saudáveis e número limitado de pacientes. Contudo, o número de pacientes submetidos aos estudos nas Fases I a III é limitado e a seleção e tratamento dos pacientes geralmente difere dos utilizados na prática clínica [2]. Além disso, há exclusão de muitos subgrupos importantes da população que são potenciais usuários, e estes estudos são realizados por tempo geralmente curto. Por outro lado, EAs detectados tardiamente no período de pós-comercialização (também conhecido como Fase IV) podem requerer um aumento significativo de cuidados de saúde e resultar em danos desnecessários, muitas vezes fatais, aos pacientes [3]. Portanto, a descoberta, o quanto antes, de EAs no período de pós-comercialização é um objetivo principal do sistema de saúde, e em particular, dos sistemas de vigilância farmacológica.

Métodos computacionais comumente referidos como "detecção de sinais" ou algoritmos de "rastreamento" permitem que os avaliadores de segurança de medicamentos analisem grandes volumes de dados para identificar sinais de risco de potenciais EA, e provaram ser um componente fundamental na Farmacovigilância. Como exemplo, a agência americana de controle de alimentos e medicamentos (*Food and Drug Administration* - FDA) usa rotineiramente um processo de rastreamento de sinais para calcular estatísticas relatando associações para todos os milhões de combinações de medicamentos e eventos em seu sistema de comunicações de eventos adversos [3]. Não obstante, estes sinais por si só não são suficientes para estabelecer uma relação causal, mas são considerados avisos

iniciais que requerem avaliação aprofundada por especialistas para estabelecer a causalidade. Esta nova avaliação tipicamente consiste de um processo complexo em que os avaliadores de segurança de medicamentos analisam informações adicionais, tais como relações temporais, relatos de casos publicados na literatura, plausibilidade biológica e clínica, dados de ensaios clínicos e estudos epidemiológicos em vários bancos de dados relacionados à saúde [3].

Dedicar recursos para a posterior avaliação de cada um dos múltiplos sinais normalmente gerados por algoritmos de detecção não é possível, e recursos desviados para indicações falsas podem inviabilizar um sistema de vigilância farmacológica [4]. Portanto, estratégias automatizadas para reduzir a quantidade de falsas indicações e definir prioridades, de modo a permitir que apenas os candidatos mais promissores sejam avaliados, são imperativas.

Desta forma, o objetivo principal deste trabalho é desenvolver um sistema automatizado para a vigilância farmacológica e que seja capaz de identificar associações novas e já existentes de “medicamento-EA”. Pretendemos utilizar como fonte de dados uma base não convencional devido a maior facilidade e rapidez de acesso aos seus dados. Exemplos de bases não convencionais que vem sendo utilizadas recentemente em vigilância epidemiológica são *logs* de busca [5-7] e redes sociais [8, 9]. Em nosso projeto utilizamos como base o Twitter.

Este capítulo está organizado da seguinte forma: na seção 1.1 temos um embasamento teórico com alguns termos relativos a área farmacológica e na seção 1.2 temos a descrição das fases do ciclo de vida de um medicamento. Temos na seção 1.3 alguns casos de fármacos que foram retirados do mercado. Na seção 1.4 abordamos alguns aspectos técnicos referentes a computação e na seção 1.5 falamos um pouco sobre a rede social Twitter que é utilizada neste trabalho. No item 1.6 apresentamos alguns trabalhos relacionados, e no 1.7 a importância de estudos nessa área.

1.1 Definições em Farmacovigilância

Nesta seção, temos um embasamento teórico com alguns termos relativos à área farmacológica. São definições importantes para o entendimento do trabalho como: fármaco, evento adverso e Farmacovigilância.

1.1.1 Droga

É uma infeliz tradução do inglês "*drug*" que contamina boa parte dos livros textos (traduzidos em português). Em português, temos a palavra "FÁRMACO", muito melhor para distinguir o "princípio ativo" de um medicamento das "drogas ilícitas", como cocaína. Ou seja, quando se vê "*drug*" em inglês, deve se usar "fármaco" em português. Da mesma forma, quando se vê "*drug product*" em inglês, deve se usar "medicamento" em português.

Nota-se que pode haver incoerência de terminologia na própria legislação brasileira quando o termo "droga" coabita com "fármaco", apropriadamente usado na definição de medicamento, como vimos anteriormente. Exemplo disso é a definição de "droga" na RDC nº 135 de 29/05/2003: "Substância ou matéria-prima que tenha finalidade medicamentosa ou sanitária" [10].

1.1.2 Fármaco

De acordo com a Portaria nº 3.916/MS/GM, de 30 de outubro de 1998 do Conselho Federal de Farmácia, fármaco é a substância química que é o princípio ativo do medicamento [11].

Remédio é qualquer substância ou recurso (ex. radioterapia) usado para combater uma moléstia. Apesar de ser muito usado popularmente, deve-se utilizar o termo medicamento quando se deseja falar especificamente de uma formulação ou produto farmacêutico que contém um ou vários princípios ativos denominados fármacos [10].

Neste contexto, vale a pena conferir a definição, mais apropriada, dada pela ANVISA (RDC nº 135) [12] para medicamento: “produto farmacêutico, tecnicamente obtido ou elaborado, com finalidade profilática, curativa, paliativa ou para fins de diagnóstico (Lei nº 5.991, de 17/12/73)”. Ou a definição do Instituto Virtual de

Fármacos do Estado do Rio de Janeiro de que medicamento é uma forma farmacêutica terminada que contém o fármaco, geralmente em associação com adjuvantes farmacotécnicos [10].

1.1.3 Evento Adverso

Segundo a ANVISA [13], evento adverso é qualquer ocorrência médica desfavorável, que ocorra durante o tratamento com um medicamento, mas que não tem necessariamente relação causal com esse tratamento. Para efeitos da Resolução RDC 04/09 [14] considera-se como evento adverso:

- Suspeita de reações adversas a medicamentos (RAM);
- Eventos adversos por desvios da qualidade de medicamentos;
- Eventos adversos decorrentes do uso não aprovado de medicamentos;
- Interações medicamentosas;
- Inefetividade terapêutica, total ou parcial;
- Intoxicações relacionadas a medicamentos;
- Uso abusivo de medicamentos;
- Erros de medicação, potenciais e reais;

São considerados graves os eventos adversos apresentados a seguir:

- Óbito;
- Ameaça à vida: há risco de morte no momento do evento;
- Hospitalização ou prolongamento de hospitalização já existente: hospitalização é um atendimento hospitalar com necessidade de internação. Também inclui um prolongamento da internação devido a um evento adverso;
- Incapacidade significativa ou persistente: é uma interrupção substancial da habilidade de uma pessoa conduzir as funções de sua vida normal;
- Anomalia congênita;
- Evento clinicamente significante: e qualquer evento decorrente do uso de medicamentos que necessitam intervenção médica, a fim de se evitar óbito, risco à vida, incapacidade significativa ou hospitalização;

Para efeitos da Resolução RDC 04/09 [14], considera-se também como

evento adverso grave qualquer suspeita de transmissão de agente infeccioso por meio de um medicamento.

Dentre os eventos adversos citados, o foco da maioria dos trabalhos é a primeira, RAM [15-19]. De acordo com a ANVISA, Reação adversas a medicamento é qualquer resposta prejudicial ou indesejável, não intencional, a um medicamento, que ocorre nas doses usualmente empregadas no homem para profilaxia, diagnóstico, terapia de doença ou para a modificação de funções fisiológicas [13].

Efeito colateral é um efeito diferente do responsável pelo uso terapêutico do fármaco, podendo ser benéfico ou indiferente e não necessariamente adverso, indesejável ("*unwanted side effect*"). O termo antigo *side effect* (i.e. efeito colateral) foi usado de várias maneiras no passado, usualmente para descrever efeitos negativos (não favoráveis), mas também efeitos positivos (favoráveis) [20]. É recomendado que este termo não seja mais usado e que, particularmente, não seja considerado como sinônimo de “reação adversa" [10].

1.1.4 Farmacovigilância

Mesmo com a obrigatoriedade de realização de uma grande quantidade de testes clínicos e laboratoriais com os medicamentos nas fases de pré-comercialização, determinadas Reações Adversas a Medicamentos (RAM) podem não ser previstas. Este fato é devido a sua rara prevalência ou, por ocorrerem exclusivamente em conjunto com outros medicamentos, alimentos ou outras substâncias não testadas nas fases anteriores à comercialização [21]. Embora as reações adversas não sejam desejáveis, há aquelas que são consideradas “aceitáveis”. São essas as reações indesejáveis previstas que podem ser controladas, ou cuja ocorrência é compensada pelo benefício promovido pelo medicamento. Entretanto, algumas reações adversas são tão severas ou graves que impedem a comercialização do medicamento.

Para monitorar a ocorrência dos eventos adversos do uso dos medicamentos, surge a Farmacovigilância, que compreende atividades de detecção, avaliação, compreensão e prevenção dos efeitos adversos ou quaisquer problemas relacionados a medicamentos [1]. Para conseguir esse objetivo, o principal instrumento da Farmacovigilância é a notificação espontânea, documento que informa aos órgãos do governo sobre os Eventos Adversos (EA) que ocorreram por

uso dos medicamentos. Tendo essa informação, os órgãos responsáveis pela Farmacovigilância podem tomar as medidas necessárias sobre cada evento. Uma notificação pode conter informações sobre EA e sobre desvios de qualidade de medicamentos (conhecidos como queixa técnica) [21].

No Brasil, as ações de Farmacovigilância são realizadas de forma compartilhada pelas vigilâncias sanitárias dos estados, municípios e pela Agência Nacional de Vigilância Sanitária - ANVISA [13, 22]. A ANVISA foi criada em 1999, e com ela o Sistema Nacional de Farmacovigilância, gerenciado pela Unidade de Farmacovigilância (UFARM). A UFARM, por sua vez, é integrante da Gerência Geral de Segurança Sanitária de Produtos de Saúde Pós-comercialização. Uma estratégia da UFARM é a descentralização da Farmacovigilância, com foco nas Vigilâncias Sanitárias Estaduais [1].

Até março de 2008, as notificações voluntárias de EA que chegavam à Unidade de Farmacovigilância da ANVISA, através de formulário de notificação disponível no seu sítio eletrônico, eram avaliadas e armazenadas, de forma manual, no banco de dados denominado Bdfarm. Já relatos de EA que chegavam por outros meios eram cadastrados no banco de dados Sisfarmaco [22]. A partir desta data, com o intuito de armazenar informações qualificadas diretamente no banco de dados, foi criado o Sistema de Notificações em Vigilância Sanitária (Notivisa) [23]. Esse banco de dados recebe notificações de eventos adversos de profissionais de saúde ou de usuários cadastrados, por meio de formulários de notificação [13]. Os usuários podem também comunicar eventos adversos ao profissional de saúde ou para a Vigilância Sanitária Local, que deve, por sua vez, repassar essa informação à ANVISA.

1.2 Ciclo de vida de um medicamento

Antes de chegar a sua fase comercial, um medicamento tem que passar por diversos estágios que variam de acordo com o seu país de desenvolvimento. Na Figura 1 temos o exemplo de uma linha do tempo de um medicamento que começa com o estágio de invenção do medicamento e testes laboratoriais e em animais antes de começar com testes em voluntários humanos. Essas fases duram aproximadamente seis anos. Depois disso, começam os testes em humanos (Fases I, II e III), inicialmente com um pequeno grupo de voluntários, sendo aumentado esse

número sucessivamente a cada fase. Após isso temos a sua aprovação pelo governo e finalmente chegamos a fase 4 ou fase de pós-comercialização.

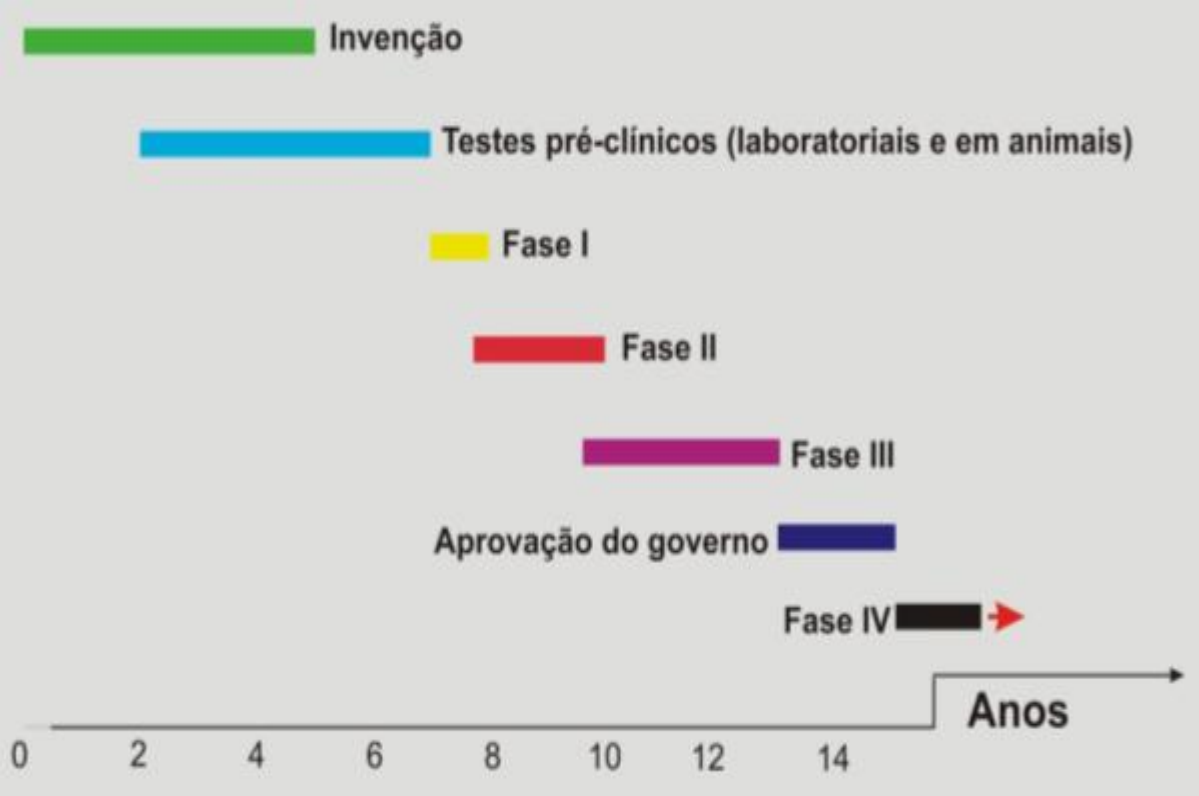


Figura 1. Linha do tempo de um medicamento.
Adaptado de: http://www.pfizer.com.br/sobre-a-pfizer/industria-farmaceutica/pesquisa-cl%C3%ADnica

1.2.1 Invenção e Testes pré-clínicos

A etapa de invenção inclui todas as atividades necessárias para a identificação e a validação do alvo terapêutico e das moléculas candidatas. Segundo Berkowitz *et al*. [24], as moléculas candidatas a fármacos são descobertas, em sua maioria, por meio de:

- Identificação e elucidação de um alvo para o composto;
- Planejamento racional do fármaco, com base no conhecimento dos mecanismos biológicos, estrutura dos receptores e estrutura própria;
- Modificações químicas de uma molécula conhecida;
- Triagem de grande número de produtos naturais à procura de atividade biológica, bancos de entidades químicas previamente descobertas e

grandes bancos de peptídios, ácidos nucléicos e outras moléculas orgânicas;

- Biotecnologia e clonagem utilizando genes para produzir peptídios e proteínas maiores;

Os candidatos a fármacos que passam pelos procedimentos iniciais de triagem e estabelecimento do perfil farmacológico são cuidadosamente avaliados quanto aos riscos potenciais, antes e no decorrer dos testes clínicos. Esses riscos são monitorados por meio de testes pré-clínicos [25]. Os testes pré-clínicos, realizados em laboratório (*in vitro*) ou em animais (*in vivo*), buscam determinar se a molécula é segura e eficaz o suficiente para iniciar os testes em humanos [26].

1.2.2 Ensaios Clínicos – Fases I, II e III

Os testes clínicos, por sua vez, têm como objetivo obter evidências quanto à segurança e à eficácia do uso do produto por seres humanos, sendo divididos em três etapas básicas (fases I, II e III). O custo total e o tempo necessário são crescentes a cada etapa, principalmente em função da ampliação do tamanho da amostra de voluntários. Entretanto, as primeiras fases envolvem um maior desafio tecnológico, principalmente as fases I e II, que são onde se definem a dosagem segura do novo medicamento e a avaliação inicial da eficácia de sua ação [27].

A fase I tem como objetivo principal verificar a tolerância e a segurança do novo medicamento. Durante essa fase, várias dosagens do medicamento em estudo são administradas a um pequeno número de voluntários (20-100), normalmente sadios, sob supervisão de um investigador, para avaliação da sua ação metabólica e farmacológica [28]. Essa fase necessita de um treinamento específico do investigador para identificar e manejar Evento Adverso (EA), juntamente com uma infraestrutura dedicada, pois, para o acompanhamento dos voluntários são necessários exames diferentes dos disponíveis na rede assistencial. É alta a complexidade de elaboração de protocolo clínico de fase I dada a dificuldade de determinar a causalidade dos EA [29].

O objetivo da fase II é verificar se o medicamento é efetivo a curto prazo no tratamento da doença-alvo, gerando informações sobre segurança, efeitos adversos e potenciais riscos. Nessa fase, busca-se determinar quais as doses mais efetivas, além do método e da frequência mais apropriados de administração de acordo com

a velocidade de liberação necessária do medicamento no organismo [29]. É nesta fase que os estudos são realizados com um número maior de voluntários (100-300) e, normalmente, estes possuem a doença-alvo. É objetivo dessa fase estabelecer as relações dose-resposta, visando obter dados para seguir à fase III [27].

A fase III tem como objetivo testar de forma mais ampla a segurança e a eficácia do medicamento. Os testes duram em média cinco anos e podem envolver milhares de voluntários (1.000-3.000) e em vários locais, a depender da incidência da doença-alvo e do tipo de substância em teste [28]. A maior parte da informação incluída nas bulas de medicamentos e avaliação de risco-benefício de um novo medicamento costumam ser fornecidas pelos testes da fase III [29].

Nos testes clínicos, os dados são analisados para identificar eventos adversos que ocorrerem a taxas significativamente maiores com o medicamento de interesse. Normalmente, os dados de todos os ensaios são reunidos em uma análise de segurança global. Os ensaios clínicos irão identificar a maioria das reações adversas mais comuns, mas muitas vezes têm limitações importantes, como [30]:

- O número de pacientes estudados geralmente não é suficiente para identificar Reações Adversas a Medicamentos (RAM) raras, mas graves;
- A duração do acompanhamento geralmente é curta, durando semanas ou meses ao invés de anos;
- Seleção de pacientes - aqueles com maior risco de RAM são, por muitas vezes, excluídos;
- As condições artificiais - os pacientes são susceptíveis a serem mais estreitamente monitorados do que na vida real;

1.2.3 Ensaios Clínicos - Fase IV (Farmacovigilância)

Também chamada de Farmacovigilância, a fase IV é de responsabilidade da agência reguladora e é o estudo do uso real do fármaco na prática médica. Podem ser descobertos nessa fase novos efeitos tóxicos ou terapêuticos e efeitos a longo prazo ou raros que não puderam ser detectados com os pequenos grupos de indivíduos das fases anteriores [26].

Durante essa fase, as informações podem ser obtidas através de notificações voluntárias pelos profissionais de saúde e pacientes [13]. Dentre as informações contidas nessas notificações encontram-se, segundo o Manual de Orientação em

Pesquisa Clínica e Farmacovigilância [27]:

- Dados do paciente: sexo, idade, história médica;
- Dados do medicamento ou dos medicamentos suspeitos: nome do medicamento e/ou substância ativa, apresentação utilizada, dose, data de início do uso, se foi interrompido o uso e se foi reintroduzido;
- Dados do evento: intensidade e gravidade do evento, data de início, como evoluiu, como foi tratado, dados de exames laboratoriais, etc;

Essa fase é de suma importância para que fatalidades e outros eventos adversos graves não venham a ocorrer como é descrito na próxima seção.

1.3 Casos de Insucesso de Fármacos

As agências de vigilância sanitária são responsáveis pelo monitoramento dos medicamentos presentes no mercado e possuem a autoridade de introduzir e/ou retirar um medicamento de circulação, de forma a obter qualidade, eficácia e segurança [26]. Nos EUA a retirada de fármacos do mercado tem ocorrido desde 1937 por consequência do uso de um xarope de sulfanilamida contendo como veículo o dietilenoglicol, que causou a morte de mais 100 pessoas [31]. A quantidade de fármacos retirados pela FDA nas últimas décadas tem aumentado por razões de segurança [26].

São vários os casos na literatura onde medicamentos precisaram ser retirados do mercado pelas agências reguladoras, dentre eles, um dos primeiros foi o medicamento talidomida prescrito como um sedativo e anti-naúseas, indicado para o alívio de enjoos matinais, comum em gestantes. A talidomida foi responsável pelo nascimento de milhares de crianças com deformações congênitas (focomelia), atribuído posteriormente ao seu perfil teratogênico [32].

No final dos anos 1950, havia pouca, ou nenhuma, regulamentação de medicamentos fora dos EUA (onde a talidomida não foi comercializada), e seus testes e desenvolvimentos foram feitos quase que inteiramente pelas companhias farmacêuticas. No caso da talidomida, reivindicações de segurança na gravidez foram feitas sem justificativas e o seu uso como um sedativo foi dirigido a mulheres grávidas. O medicamento acabou por se mostrar teratogênico, produzindo uma variedade de defeitos congênitos, em particular defeitos nos membros conhecido como focomelia. Em todo o mundo, cerca de 10.000 fetos foram afetados,

especialmente na Alemanha, onde o medicamento foi comercializado pela primeira vez. Como a focomelia era uma anomalia congênita muito rara, a existência de um grande aumento na sua incidência não passou despercebida na Alemanha, mas a causa foi inicialmente atribuída a um fator ambiental. Em 1961, uma série de três casos associados à talidomida foi relatada na revista The Lancet. O problema foi finalmente reconhecido e o medicamento foi retirado de venda [30].

Outro exemplo é o do anti-histamínico terfenadina que em alguns casos causava arritmia cardíaca potencialmente fatal. Foi descoberto somente após sua aprovação, onde fármacos inibidores da enzima citocromo P450, como por exemplo, eritromicina ou cetoconazol, poderiam ser os responsáveis pelo aparecimento de arritmias quando utilizados em associação com a terfenadina [33].

Esses casos, entre muitos outros não citados aqui, demonstram o quão importante é a tarefa das agências de vigilância sanitária em fiscalizar um medicamento após a sua aprovação. As principais lições aprendidas com casos assim foram:

- A necessidade de testes adequados com os medicamentos antes da comercialização;
- A necessidade de regulamentação dos medicamentos pelo governo;
- A necessidade de sistemas para identificar os efeitos adversos de medicamentos;
- A relação de potencial entre reivindicações de *marketing* e segurança;
- Evitar o uso desnecessário de medicamentos na gravidez;
- Alguns riscos podem ser minimizados com sucesso, como a má-formação fetal;

1.4 Aspectos Técnicos em Computação

Nesta seção, temos um embasamento teórico de alguns tópicos relativos à área da Computação. São definições importantes utilizadas na metodologia deste trabalho como: Mineração de Dados, Mineração de textos e Big Data e Processador de Linguagem Natural.

1.4.1 Mineração de Dados

A Mineração de Dados [34] é o processo de mineração ou descobrimento de novas informações na forma de padrões e regras a partir de grandes bases de dados. A Mineração de Dados ajuda certos tipos de decisão, extraindo padrões que não seriam encontrados utilizando apenas consultas à base. Essa combinação é a pedra fundamental na construção de Sistemas de Apoio à Decisão.

Apesar de ser a etapa mais importante, a Mineração de Dados faz parte de um processo maior denominado KDD (*Knowledge Discovery in Databases*), responsável pelo descobrimento de conhecimento em bases de dados.

Existem inúmeras tarefas de Mineração de Dados que podem ser classificadas em duas categorias: a Mineração preditiva, que pode mostrar como certos atributos vão se comportar no futuro a partir de dados da base; e, a Mineração descritiva, que utiliza padrões e regras para descrever os dados, indicando características importantes [35].

Dentre algumas das tarefas de Mineração de Dados temos: regras de associação, classificação, padrões de sequência, padrões em séries temporais, clusterização e mineração de textos [36].

1.4.2 Mineração de Textos

A mineração de textos consiste em usar técnicas para recuperar informação textual, extrair informação, bem como processar linguagem natural com os algoritmos e métodos de descoberta de conhecimento, mineração de dados e aprendizado de máquina [37].

Extrair informação de textos é uma forma de análise de linguagem natural e está se tornando uma tecnologia central para diminuir a distância entre um texto não estruturado e conhecimento formal expresso em ontologias. Os métodos de processamento de linguagem natural fornecem a fundamentação para as investigações no campo da mineração de textos biomédicos [38].

A mineração de texto abrange um vasto campo de abordagens teóricas e métodos cujo objeto comum de entrada de informação é o texto. Isso permite várias definições, que vão desde uma extensão de mineração de dados clássica de textos, às formulações sofisticadas aplicadas às grandes coleções *online* para descoberta

de novos fatos e tendências sobre o próprio mundo [39].

1.4.3 Processamento de Linguagem Natural

O processamento de linguagem natural (PLN) é uma área de pesquisa e aplicação que explora como computadores podem ser usados para compreender e manipular texto em linguagem natural para fazer coisas úteis [40].

PLN é uma das principais técnicas para mineração de textos. Por meio de conhecimentos da área de linguística, o PLN consegue aproveitar ao máximo o conteúdo do texto, extraindo entidades, seus relacionamentos, encontrando sinônimos, fazendo a correção de palavras escritas erradas e ainda desambiguizando-as [41]. Os fundamentos do PNL encontram-se numa série de disciplinas como [40]: Computação e Ciências da Informação, Linguística, Matemática, Engenharia Elétrica e Eletrônica, Inteligência Artificial e Robótica, Psicologia, etc.

1.4.4 Big Data

Big data é um termo usado para descrever conjuntos de dados cujo processamento por sistemas de gerenciamento de bancos de dados convencionais é problemático, devido a qualquer combinação de (a) seu tamanho (Volume); (b) sua frequência de atualização (Velocidade); ou (c) sua diversidade (Variedade) [42].

A multiplicidade de dispositivos e a capacidade destes dispositivos interagirem em rede estão promovendo uma verdadeira inundação de dados. Cada um de nós carrega junto de si um celular que, agindo como um sensor, pode obter informação de localização das pessoas e permitir a realização de negócios direcionados. Ao levarmos em consideração que o mundo tem cerca de 7 bilhões de habitantes e que aproximadamente 6 bilhões possuem celulares, pensem no Volume e na Variedade de dados que pode ser gerado, captado, processado, (re)utilizado e entregue [43].

A melhoria dos canais de transmissão, com redes em fibra ótica e emissores de sinais de alta capacidade, o uso de satélites, o uso de outras bandas para a telefonia celular, as comunicações em tempo real para controle de processos na internet, os *workflows* científicos com processamento paralelo e *cluster* de processamentos vem possibilitando atingir uma maior velocidade para troca de

dados e informação [44].

O que é considerado *Big Data* varia de acordo com as capacidades dos usuários e suas ferramentas. Assim, o que é considerado "grande" em um ano se tornará comum em anos posteriores.

Em nosso projeto utilizamos conceitos de Big Data dado ao grande volume e velocidade de dados do Twitter considerando que estamos realizando a coleta de vários medicamentos e doenças apesar de termos focado em apenas um neste trabalho (doxiciclina).

1.5 Twitter

Twitter [45] é um serviço de rede social *on-line* que permite o envio de mensagens de até 140 caracteres chamados *tweets*. Criado no ano de 2006 por Jack Dorsey, Evan Williams, Biz Stone e Noah Glass, o Twitter rapidamente tornou-se popular no mundo todo.

Um *tweet* pode ser lido por qualquer outro usuário do Twitter. Usuários podem seguir as pessoas que lhe interessam e, nesse caso, são notificados quando essa pessoa posta um *tweet* novo. Um usuário que está sendo seguido por outro usuário não tem necessariamente que retribuir, seguindo-o de volta.

Hoje em dia as pessoas estão cada vez mais contando suas vidas em redes sociais como Facebook e Twitter. Um dos temas frequentes é medicamentos. Existem até redes sociais específicas para falar sobre medicamentos, como PatientsLikeMe [46] e AskaPatient [47]. O uso de redes sociais tornou-se tão importante que a indústria Farmacêutica Britânica fez um guia [48] com sugestões para as empresas monitorarem Evento Adverso (EA) e reclamações de produtos nas redes sociais.

O Twitter está entre os dez *sites* mais visitados no mundo todo, possui mais de 1 bilhão de usuários e cresce a um número estimado de 135 mil usuários por dia, gerando quase 60 milhões de *tweets* por dia [49, 50]. Devido ao seu grande volume de dados gerado e ao livre acesso às suas discussões, o Twitter é uma fonte valiosa para pesquisadores [19].

1.5.1 Revendedores de dados

Para se ter acesso a todos os dados do Twitter é necessário ter acesso ao seu *firehose* [51], que em inglês significa a mangueira utilizada pelos bombeiros, por representar um alto fluxo de dados. A principal diferença entre o *firehose* e a Streaming API é que com o *firehose* temos garantia de coleta de 100% dos *tweets* que buscamos. A quantidade dos tweets retornados pela Streaming API pode chegar até 1% de todos os tweets públicos que estão sendo gerados, logo o volume de dados dependerá da quantidade de palavras selecionadas e da quantidade de tweets totais gerados no momento. Exemplificando, se num dado segundo houve 100 mil tweets, o sistema poderia ter acesso até mil desses tweets.

São poucos os que possuem acesso a esse *firehose*, dentre eles alguns revendedores de dados. Dentre eles temos o Topsy [52], Gnip [53], Datasift [54] e Dataminr [55]. A maioria deles tem APIs REST para acesso aos seus dados. Para obter os dados de um revendedor, é necessária uma licença.

O Topsy foi o primeiro revendedor de dados do Twitter e alegava ter todos os *tweets* desde 2006. O Topsy foi comprado pela Apple em 2013 e atualmente seu *site* não existe mais [56].

O Gnip possui dados do Twitter e de outras redes sociais como Tumblr, WordPress, Foursquare, Disqus, IntenseDebate, StockTwits, e GetGlue. Ele foi recentemente comprado pelo próprio Twitter [57].

O DataSift tem dados de muitas redes sociais e possui uma API muito bem documentada para os desenvolvedores. Ele oferece um período de teste gratuito de sete dias, sendo pago depois desse período [54].

O Dataminr, usando algoritmos proprietários, analisa instantaneamente todos os *tweets* públicos e outros dados publicamente disponíveis em sinais acionáveis, descobrindo informações em tempo real para os clientes em Finanças, Setor Público e Notícias [55].

1.6 Trabalhos relacionados

Vários estudos têm sido publicados sobre o tema de mineração de mensagens em redes sociais para extração de informações relacionadas à saúde. Além disso, outros estudos têm sido propostos para extrair informações das

mensagens contidas no Twitter. Contudo, ainda existem poucos estudos sobre a mineração de mensagens do Twitter para encontrar Eventos Adversos (EA) [58].

A maioria dos trabalhos anteriores de mineração de texto relacionados com a Farmacovigilância está focada em registros eletrônicos de saúde, e em relatos de casos médicos [59, 60]. Harpaz *et al*. [61] fornecem um estudo aprofundado sobre as abordagens existentes para a fase de pós-comercialização, explorando vários recursos, tais como registros eletrônicos de saúde e sistemas de relato espontâneo de EA à medicamentos. As redes sociais vêm sendo utilizadas para esse propósito recentemente. Leaman *et al*. [62] analisaram os comentários de usuários em redes sociais e demonstraram que estes contêm informações sobre medicamentos que podem ser extraídas para posterior análise. Para tal, os autores utilizaram um sistema baseado em um conjunto de palavras e em algumas regras para a extração de conceitos relacionados a Reações Adversas a Medicamentos (RAM). Nikfarjam & Gonzalez [63] propuseram uma técnica baseada em padrões com base na mineração de regras de associação, que extrai menções de RAM baseado nos padrões de linguagem utilizados pelos pacientes nas redes sociais para expressar RAM. Em um estudo recente, Yates & Goharian [64] analisaram o valor dos comentários de usuários em revelar EA desconhecidos avaliando as RAM extraídos com a base de dados SIDER [65], que contém informações sobre os EA conhecidos. Há estudos semelhantes para extração de menções de RAM automática a partir de discussões on-line de pacientes [64, 66, 67].

Muitos estudos têm sido publicados sobre a mineração de mensagens do Twitter para extração de informações relacionadas à saúde. Cobb et. al. [68] analisaram várias redes sociais, incluindo Twitter, para estudar como essas plataformas podem facilitar a cessação do tabagismo. Prier et. al. [69] realizaram um estudo empírico para explorar os *tweets* relacionados com o tabaco para a identificação de tópicos relacionados com a saúde. Paul et. al. [70] propuseram um modelo de análise para a mineração de temas de saúde pública no Twitter. Além disso, vários outros modelos analíticos têm sido propostos para extrair as mais variadas informações das mensagens do Twitter, variando desde previsão de resultados de votos de eleições [71] à estudos sobre padrões globais de humor [72].

A grande maioria das pesquisas utilizam redes sociais que já são voltadas para a área médica, e as que utilizam o Twitter buscam outras informações que não EA. Alguns estudos recentes utilizaram o Twitter para essa função [17, 18, 58] e

mostraram que o uso de seus *tweets* pode levar a farmacovigilância em tempo real. Freifeld [17] utilizou o Twitter para avaliar o nível de concordância entre os *tweets* com menções de EA (Proto-AE) e relatórios espontâneos de EA da FDA (FAERS). Nesse estudo foram coletados 6,9 milhões de *tweets* com nomes de medicamentos, dos quais 4401 foram identificados como Proto-AEs. Eles mostraram que o Twitter teve quase três vezes mais Proto-AEs que relatórios da FDA [17]. Rachel *et al*. [18] criou um *corpus* (grande coleção de textos) de 10.822 *tweets* anotados manualmente que podem ser usados para treinar ferramentas automatizadas para minerar RAM no Twitter.

As pesquisas que buscam EA no Twitter geralmente coletam dados de alguns poucos meses para encontrar RAM já conhecidos, usam uma ou nenhuma ontologia para fazer isso e tem etapas manuais do seu *pipeline*. Em nosso trabalho, criamos um *pipeline* automático para coleta, armazenamento e processamento de *tweets* que utiliza uma ontologia completa e totalmente voltada para a busca de RAM.

1.7 Justificativa

Esta dissertação tem o objetivo de auxiliar o sistema de vigilância farmacológica coordenado pela ANVISA, acelerando a identificação de potenciais associações entre medicamentos e eventos adversos. Consequentemente, relações causais podem ser estabelecidas mais rapidamente e medicamentos que tiverem uma nova Reação Adversa a Medicamento (RAM) comprovada e que forem consideradas graves, poderão ser retiradas mais rapidamente do mercado, beneficiando grandes populações de pacientes. Em outros casos, poderia ocorrer uma revisão da prescrição. Esta dissertação se caracteriza por grande impacto social e grande relevância no contexto brasileiro, uma vez que os recursos dedicados à vigilância farmacológica no Brasil são limitados, assim como o seu alcance. Os resultados da pesquisa descrita nesta dissertação permitirão a otimização dos recursos disponíveis e tornarão a vigilância farmacológica mais eficiente.

Atualmente existem pouquíssimos sistemas que realizam essa tarefa de buscar Evento Adverso (EA), e os que existem são ou utilizam softwares proprietários. Estes sistemas geralmente atuam de forma específica, atuando apenas sobre alguns eventos adversos e utilizando um determinado formato de

relatório médico [73]. Não existe software livre que realize tal tarefa como é apresentado neste trabalho, e a sua criação é de grande importância para o âmbito da Farmacovigilância, tanto nacionalmente como internacionalmente.

1.8 Organização do Trabalho

Esta dissertação de mestrado foi elaborada no formato de 6 capítulos que estão dispostos da seguinte maneira. No Capítulo 2 estão definidos os objetivos do presente trabalho. A descrição detalhada dos materiais e dos métodos empregados na construção do *pipeline* é feita no Capítulo 3. O Capítulo 4 corresponde à apresentação dos resultados, sob a forma de evidências para a confirmação das questões da pesquisa e dos objetivos alcançados. A discussão dos resultados obtidos é feita no Capítulo 5. No Capítulo 6 são detalhadas as conclusões do trabalho e sugestões de trabalhos futuros para a continuidade da pesquisa.

2 Objetivos

2.1 Objetivo Geral

O objetivo principal deste trabalho é desenvolver o Tweet Adverse Event Miner (TweetAEMiner), um sistema automatizado de vigilância farmacológica que será capaz de processar um grande volume de dados do Twitter. Este sistema será capaz de identificar associações novas (não conhecidas previamente) e já existentes entre medicamentos e eventos adversos. O TweetAEMiner será executado regularmente em recursos computacionais disponibilizados pela FIOCRUZ e parceiros, e servirá de apoio às redes de vigilância farmacológica implantadas no Brasil.

2.2 Objetivos Específicos

- Desenvolver um sistema que seja capaz de monitorar tanto doenças quanto medicamentos selecionados previamente, com a capacidade de processar um grande número de alvos.
- Aplicar técnicas de mineração de texto para extração de Evento Adverso (EA).
- Criar um padrão de referência “medicamento-evento adverso”, a ser utilizado no processo de avaliação deste sistema e em outras pesquisas.
- Criar um coletor de dados para o Twitter.
- Analisar e avaliar os dados de um medicamento de uma doença para teste do sistema.

3 Materiais e Métodos

Nesse capítulo é descrito o *pipeline* do sistema que é apresentado na Figura 2. São descritas suas quatro principais partes: extração, processamento, análise e avaliação dos dados, juntamente com o banco de dados que utilizamos para armazenar os *tweets* e o padrão de referência criado.

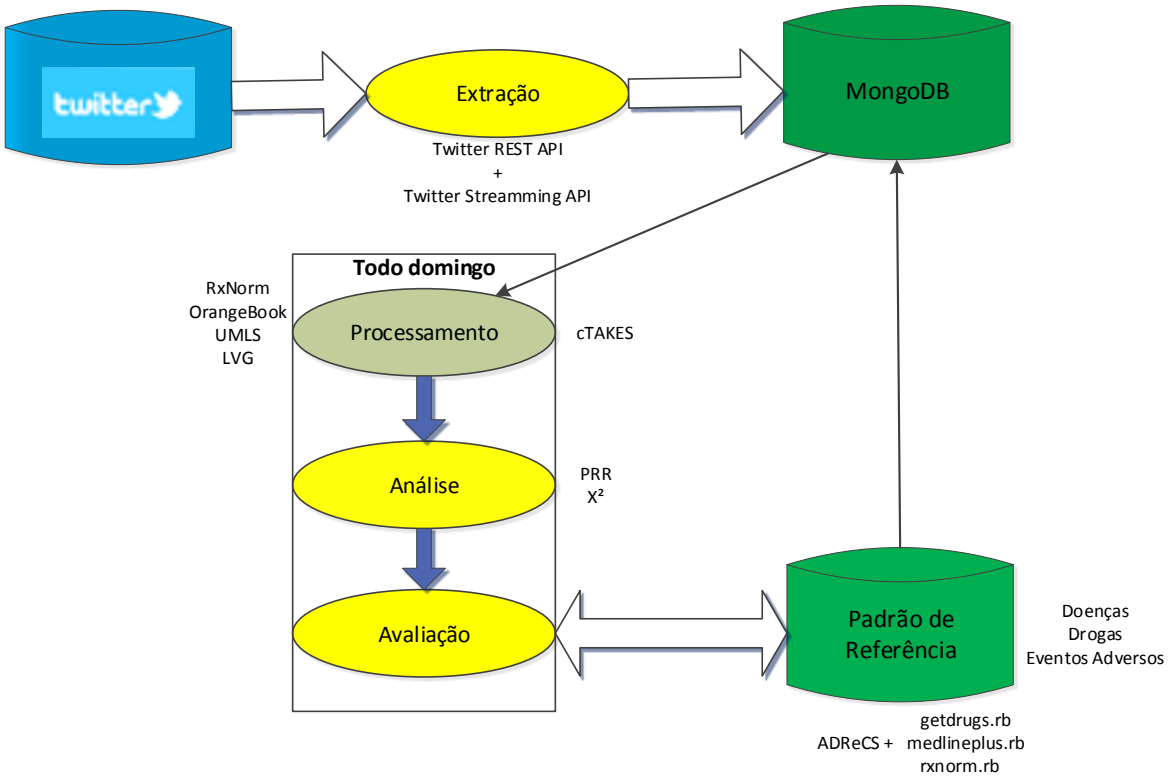


Figura 2. "TweetAEMiner": Visão geral

O TweetAEMiner realiza continuamente a coleta de *tweets* utilizando as APIs do Twitter com palavras pré-determinadas (doenças ou medicamentos). Esses *tweets* são armazenados em nossa base de dados. Periodicamente o sistema inicia o processamento e a análise desses *tweets*. Atualmente o sistema está configurado para rodar aos domingos, pois é quando começa uma nova semana epidemiológica, mas isso pode ser alterado. No processamento é utilizado um Processador de Linguagem Natural (PLN) e, com a saída desse processamento, é feita a análise dos dados. Por último é feita uma avaliação dos resultados obtidos com o padrão de referência.

O sistema gera uma lista de sinais específicos, que são avaliados tendo-se como base um padrão de referência. Um sinal corresponde a uma associação “medicamento-EA” identificada pelo *pipeline*.

Todas as etapas do *pipeline* são executadas em um servidor com a seguinte configuração: processador Intel Xeon com 4 núcleos; placa gráfica Nvidia GTX 770; 32 gigabytes de memória RAM; sistema operacional Xubuntu 14.04.2 – LTS.

3.1 Padrão de Referência e MongoDB

Para construir a nossa base padrão de referência, utilizamos principalmente o ADReCS (*Adverse Drug Reaction Classification System*) [74], uma ontologia para termos de reações adversas que usa fontes como DailyMed [75], SIDER2 [65], *US Food and Drug Administration* (FDA), UMLS (*Unified Medical Language System*) [76] e DrugBank [77]. Acrescentamos a essa ontologia: o relacionamento entre doenças e seus medicamentos utilizando o site “www.drugs.com” (Anexo A); algumas reações adversas encontradas com o *Connect Service MedlinePlus Web* [78] (Anexo B); e, sugestões de escritas para algumas palavras utilizando a RxNorm API RESTful [79] (Anexo C). Com essas fontes, criamos uma base com as doenças estudadas, os medicamentos utilizados em seus tratamentos e os efeitos adversos de cada uma destas doenças. Um exemplo das principais coleções do nosso padrão de referência (DISEASES, DRUGS e ADRS) é mostrado na Figura 3. Todas essas coleções ficam na base de dados *ontology* do nosso padrão de referência.

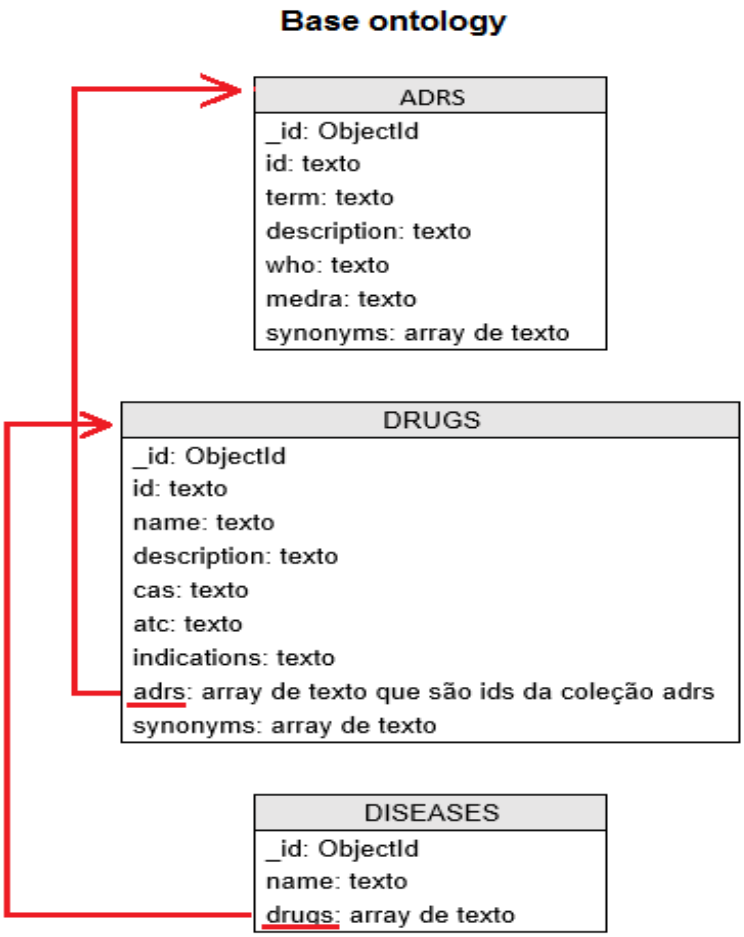


Figura 3. Modelo da ontologia base do Padrão de Referência. As setas em vermelho indicam que os campos são identificadores de outras coleções.

Na coleção DRUGS é armazenado somente o ADREeCS_ID dos Eventos Adversos (EA) no campo “adrs”, o nome e a frequência são obtidos nos documentos da coleção ADRS. A Tabela 1 mostra um exemplo de alguns dos EA armazenados relacionados ao medicamento doxiciclina. Todos os 126 EA da doxiciclina presentes em nosso padrão de referência podem ser vistos na tabela do Anexo D.

Tabela 1. Alguns EA da doxiciclina presentes no padrão de referência

ADR Term[(a)]	ADReCS_ID[(b)]	Frequência[(c)]
Abdominal discomfort	07.01.06.001	-
Abdominal distension	07.01.04.001	-
Abdominal pain	07.01.05.002	-
Abdominal pain upper	07.01.05.003	-
Abscess	11.01.08.001	-
Anaemia	01.03.02.001	-
Anaphylactic reaction	10.01.02.001; 24.06.03.006	-
Anaphylactic shock	10.01.02.002; 24.06.02.004	-
Angioedema	10.01.05.009; 23.04.01.001	-
Anorexia	08.01.09.025; 14.03.01.001	-
Anxiety	19.06.02.002	-
Aphthous stomatitis	07.05.06.001	-
Arthralgia	15.01.02.001	6.00%
Back pain	15.03.04.005	3.00%
Benign intracranial hypertension	17.07.02.001	-
Blood pressure increased	13.14.03.005	-
Blood urea increased	13.13.01.006	-
Bowel discomfort	07.01.06.006	infrequent
Bronchitis	11.01.09.001; 22.07.01.001	3.00%
Candidiasis	11.03.03.001	-
Cough	22.02.03.001	4.00%
Decreased appetite	08.01.09.028; 14.03.01.005	-
Dermatitis	23.03.04.002	-
Dermatitis exfoliative	10.01.01.004; 23.03.07.001	rare
Diarrhoea	07.02.01.001	infrequent
Discomfort	08.01.08.003	-
Dry mouth	07.06.01.002	-
Dysmenorrhoea	21.01.01.002	4.00%
Dyspepsia	07.01.02.001	6.00%
Dysphagia	07.01.06.003	-

[(a)]Eventos Adversos; [(b)]Identificadores que estão associados a esse EA no ADReCS; [(c)]Frequência com que o EA ocorre quando pode ser obtido, caso contrário é mostrado o símbolo "-".

Caso o nosso processamento encontre alguma relação entre um medicamento e um evento adverso que não esteja nessa base, isso é visto como um indício de um possível novo evento adverso.

Tanto o padrão de referência como os dados coletados são armazenados em nossa base de dados. Devido à enorme quantidade de dados que são coletados diariamente, é necessário a utilização de um banco de dados voltado para

manipulação de *Big Data* como explicado na seção 1.4.4. Por esse motivo, o sistema utiliza como base de dados um noSQL DB. Em nosso caso específico utilizamos o mongoDB [80].

3.2 Extração

O Twitter possui duas APIs para coleta de *tweets*, a REST API [81] e a Streaming API [82], ambas tem em comum a filtragem de palavras e algumas limitações. As duas API's permitem apenas o acesso a *tweets* recentes, portanto, estamos construindo uma base de dados de Farmacovigilância que vai ser útil para pesquisas futuras.

Temos coletados *tweets* relacionados a doenças negligenciadas desde o início de 2014. Todos os *tweets* coletados são armazenados no banco de dados mongoDB. Eles são armazenados em semanas epidemiológicas para uma melhor análise. Por exemplo, todos os *tweets* de malária coletados na quinta semana epidemiológica de 2014 foram armazenados em uma coleção com o rótulo "malaria_2014_5". A Figura 4 demonstra esse armazenamento, juntamente com os campos que cada documento possui. Iniciamos com uma abordagem de coleta de *tweets* por doença, mas depois também coletamos pelo nome do medicamento.

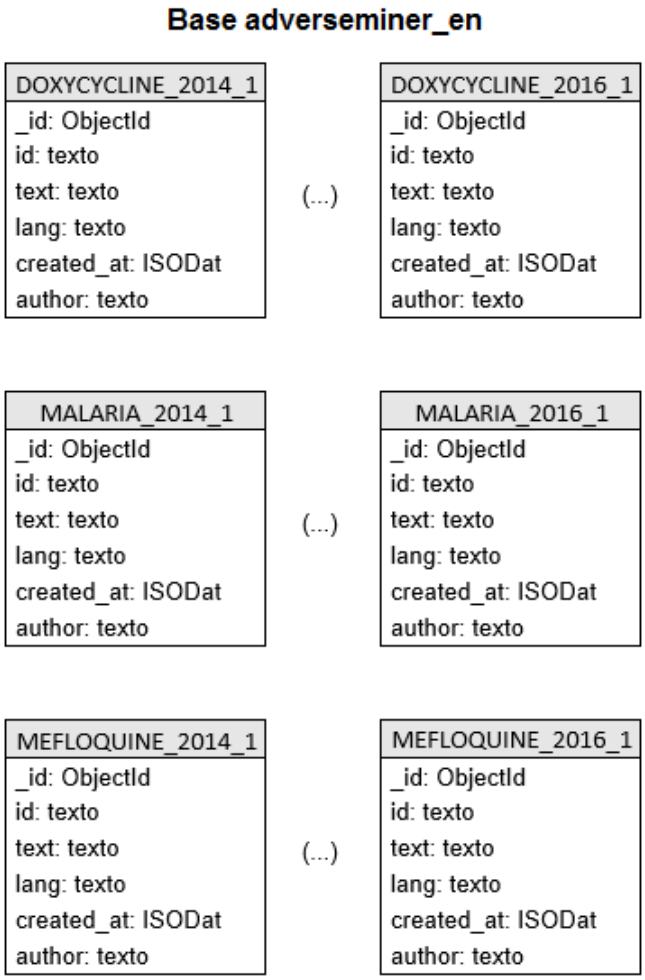


Figura 4. Armazenamento por semana epidemiológica. O campo "id" é o identificador do tweet, o campo "text" é onde se encontra o texto do tweet, o campo "lang" diz em qual lingua está o tweet, o campo "created_at", a data de criação, o campo "author", o autor e o campo "_id" é um identificador criado pelo próprio mongoDB.

Até o momento, estão sendo processados apenas os tweets em inglês, que se encontram em nossa base "adverseminer_en". Isso se deve ao fato da ontologia criada para o nosso padrão de referência ser constituída somente com palavras da língua inglesa. *Tweets* em português também estão sendo coletados para pesquisas futuras, quando tivermos uma ontologia para a língua portuguesa.

A REST API fornece os métodos GET e POST para ler e escrever vários recursos do Twitter. Ela retorna apenas os *tweets* recentes (6-9 dias de *tweets*) e permite apenas 450 pedidos a cada 15 minutos. Esta API é utilizada quando ocorre algum problema com a API Streaming. Para usar a REST API fizemos um programa em linguagem Ruby, que armazena todos os *tweets* em arquivo (Anexo E) e posteriormente, no nosso banco de dados.

A Streaming API permite o acesso a uma conexão contínua com o Twitter permitindo obter *tweets* com as palavras selecionadas desde o início da conexão. A quantidade dos *tweets* retornados pode chegar até 1% de todos os *tweets* públicos que estão sendo gerados, logo o volume de dados dependerá da quantidade de palavras selecionadas e da quantidade de *tweets* totais gerados no momento. Exemplificando, se num dado segundo houve 100 mil *tweets*, o sistema poderia ter acesso a até mil desses *tweets*.

Criamos um programa em linguagem Ruby para utilizar esta API. Ele recebe como entrada as palavras buscadas e, como saída, ele armazena os *tweets* de notícias no banco de dados (Anexo F). Toda hora é verificado se o programa está rodando e, caso não esteja, é iniciado a REST API para coletar os dados do tempo que o sistema ficou inativo e posteriormente a Streaming API para reestabelecer a conexão contínua.

Como uma abordagem inicial, começamos a coletar os *tweets* relacionados a doenças negligenciadas, como a malária, dengue, doença de chagas, tuberculose e leishmaniose [83]. Posteriormente, nossas consultas foram expandidas para outras doenças, incluindo também as não negligenciadas, como a AIDS. No momento, estamos coletando chikungunya, ebola, esquistossomose, filariose, cisticercose, hanseníase, elefantíase, varicela, esporotricose, sífilis, toxoplasmose, gonorreia e H1N6, como palavras a serem usadas na busca de *tweets*.

Nesta primeira abordagem descobrimos que malária era a doença com mais *tweets*. Apesar de algumas destas doenças não terem ainda um medicamento associado, seus *tweets* podem ser úteis em outros projetos, como por exemplo, projetos de estudo epidemiológico.

A coleta de *tweets* por doença poderia omitir alguns *tweets*, pois não é sempre que a pessoa que utiliza um medicamento menciona também a sua doença. Por isso começamos a coletar os *tweets* pelo nome de medicamentos, porém, o Twitter limita o número de palavras que podem ser buscadas. Portanto, usar todos os nomes de medicamentos relacionados com todas as doenças não seria viável. Por isso, coletamos apenas medicamentos contra a malária, que é a doença negligenciada com mais *tweets*.

Para obter todos os nomes de medicamentos relacionados com a malária, usamos o *site* “www.drugs.com”. No *site* é possível encontrar nomes de medicamentos comerciais e genéricos. Foi feito um programa que, tendo como

entrada o nome de uma determinada doença, retorna os medicamentos relacionados à mesma (Anexo A). Para a doença malária, utilizamos os seguintes nomes comerciais de medicamentos: Plaquenil, Malarone, Doryx, Lariam, Daraprim, Aralen, Fansidar, Morgidox, Ocudox e Oraxyl. E, como nomes genéricos: Atovaquone, Proguanil, Doyicycline, Mefloquine, Pyrimethamine, Sulfadoxine, Hydroxychloroquine, Chloroquine e Primaquine.

3.3 Processamento

Depois de recolher os *tweets*, usamos um processador de linguagem natural (PLN) para processá-los. Existem algumas ferramentas utilizadas no campo da medicina, como Medlee [84], cTAKES [85] e MetaMap [86]. Medlee é o mais conhecido PLN mas não é mais gratuito. Por isso escolhemos o cTAKES da *Apache Software Foundation*.

cTAKES é um PLN de código aberto usado para extrair informações a partir de texto livre. Ele utiliza diferentes fontes como o RxNorm [87], SNOMED (*Systematized Nomenclature of Medicine*) [88], Orange Book [89], UMLS [76] e LVG (*Lexical Variant Generation*) [90].

Para melhor aproveitamento do cTAKES, um ID de usuário e senha UMLS são necessários, que podem ser adquiridos gratuitamente no NIH (*U.S. National Library of Medicine / National Institutes of Health*) [91]. Para usar o cTAKES, baixamos o seu código fonte e seus bancos de dados e fizemos um programa que processa os *tweets* armazenados retornando as doenças, os medicamentos e as reações adversas associadas, bem como outras informações médicas.

Um exemplo da utilização do cTAKES através de sua ferramenta visual é mostrado na Figura 5. Nela podemos verificar as anotações que o cTAKES realiza.

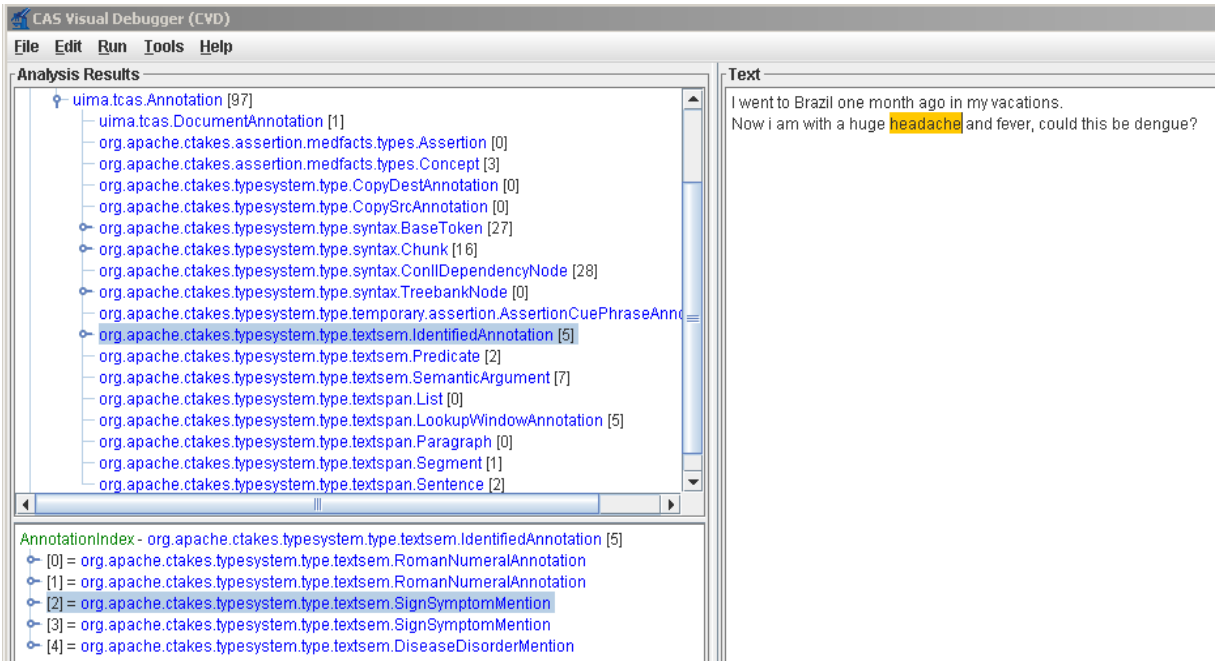


Figura 5. Exemplo de anotação do cTAKES. Destacado em amarelo está o EA headache (dor de cabeça) que é classificado pelo cTAKES como "SignSymptomMention". O outro "SignSymptomMention" que aparece na lista é referente ao EA fever (febre). A classificação "DiseaseDisorderMention" é referente ao termo dengue. Todos esses termos fazem parte da lista de "IdentifiedAnnotation" do cTAKES, que são os termos encontrados em seu processamento.

3.4 Análise

Após o processamento, utilizamos uma medida de análise de desproporcionalidade para os dados a serem analisados. Esta medida foi usada para classificar pares "Medicamento-evento adverso" identificados na etapa de processamento anterior. Observe que o método de análise pode variar em função dos dados que forem processados. Sistemas de relato espontâneo (SRE) baseados em Reação Adversa a Medicamento (RAM) realizam a detecção de sinais mais frequentemente utilizando medidas de desproporcionalidade.

Não há consenso sobre qual é a melhor abordagem, frequentista ou Bayesiana [92]. Do ponto de vista do cenário internacional, ambas as abordagens são utilizadas. A FDA (*Food and Drug Administration*), agência de vigilância dos Estados Unidos, utiliza o MGPS (*Multi-item Gamma-Poisson Shrinker*) [61], um método Bayesiano. Na UE (União Europeia), é utilizado o método frequentista PRR (*Proportional Reporting Ratios*) pelo seu SRE, EudraVigilance [93]. A Organização Mundial de Saúde, por sua vez, utiliza o BCPNN [61], que é uma versão Bayesiana do IC (*Information Component*).

Neste trabalho utilizamos a abordagem freqüentista PRR (*Proportional Reporting Ratios*). Medidas Bayesianas tendem a produzir valores menos extremos do que PRR quando o número de casos é muito pequeno. No entanto, quando a sensibilidade, especificidade e poder preditivo dessas medidas foram comparadas utilizando dados holandeses em 2002 [94], não foram encontradas diferenças importantes quando pelo menos 3 casos foram relatados. Por esse motivo, damos um enfoque maior na análise com o PRR que já foi utilizada em vários trabalhos para detecção de RAM em SRE [93, 95, 96] e é uma das principais medidas utilizadas pela UE. Juntamente com o PRR, foi calculado o seu intervalo de confiança de 95% e também realizado o teste χ^2 para validação dos sinais gerados, da mesma maneira que é realizada pelo EudraVigilance [93].

No contexto de notificação de RAM espontânea, um sinal é normalmente uma série de casos de suspeita de RAM semelhantes, reportados em relação a um determinado medicamento. Com exceção de certos tipos de eventos, que são particularmente importantes e susceptíveis de serem relacionados com o medicamento (por exemplo, anafilaxia), um único caso não é geralmente suficiente para elevar um sinal. O número mínimo de casos necessários a serem considerados geralmente são três [30]. Quando a suspeita de RAM é uma doença que é rara em uma população em geral (por exemplo, anemia aplástica, necrose epidérmica tóxica), um número muito pequeno de casos, associado com um único fármaco, não deve ser um fenômeno do acaso, mesmo se o medicamento tivesse sido amplamente usado.

Os dados obtidos com a análise foram comparados com dados da FDA que foram obtidos com o openFDA [97].

3.4.1 Análise de desproporcionalidade

Métodos de análise de desproporcionalidade (ADP) em vigilância farmacológica representam a classe principal de métodos analíticos para a análise de dados de sistemas de relato espontâneo (SRE) [61]. SRE são relatórios que compreendem um ou mais medicamentos, de um ou mais eventos adversos (EAs), e possivelmente, alguns dados demográficos de base.

Estes métodos identificam associações relevantes em bases de dados de SRE, com foco em projeções de baixa dimensionalidade dos dados, mais

especificamente tabelas de contingência 2x2. Tanto a FDA (*Food and Drug Administration*) como a OMS utilizam métodos de ADP para achar essas associações [61]. A Tabela 2 mostra uma tabela típica utilizada para o cálculo de diversas medidas de associação onde:

- A letra “a” representa no nosso caso a quantidade de *tweets* que contém o medicamento i e o EA j;
- A letra “b” representa a quantidade de *tweets* que contém o medicamento i mas não contém o EA j;
- A letra “c” representa a quantidade de *tweets* que não contém o medicamento i mas contém o EA j;
- A letra “d” representa a quantidade de *tweets* que não contém o medicamento i nem o EA j;

Tabela 2. Tabela de contingência 2x2

	EA j = Sim	**EA j = Não**	**Total**
Medicamento i **= Sim**	a=13	b=2250	n=a+b=2263
Medicamento i **= Não**	c=251	d=1632315	c+d=1632566
Total	m=a+c=264	b+d=1634565	t=a+b+c+d=1634829

A tarefa básica de um método ADP é a classificação das tabelas em ordem de “interesse”. Diferentes métodos ADP focam em diferentes medidas estatísticas de associação como a sua medida de "interesse". A Tabela 3 apresenta as fórmulas para as diferentes medidas de associação mais comumente usadas, juntamente com a sua interpretação probabilística, onde "¬ medicamento" denota os relatórios que não incluem o medicamento alvo.

Tabela 3. Medidas comuns de associação em análises de SRE

Medida da Associação	Formula	Valor	Interpretação probabilística
RRR – Relative Reporting Ratio	$\frac{t \times a}{m \times n}$	**35.57355**	$\frac{Pr\,(ea\,\vert medicamento)}{Pr\,(ea)}$
PRR – Proportional Reporting Ratio	$\frac{(a \times (t-n))}{c \times n}$	**37.36421**	$\frac{Pr\,(ea\,\vert medicamento)}{Pr\,(ea\,\vert \neg\, medicamento)}$
ROR – Reporting Odds Ratio	$\frac{a \times d}{c \times b}$	**37.57431**	$\frac{((Pr\,(ea\,\vert medicamento)}{Pr\,(\neg\, ea\,\vert medicamento))}{\frac{(Pr\,(ea\,\vert \neg\, medicamento)}{Pr\,(\neg\, ea\,\vert \neg medicamento))}}$
CI - Componente de informação	$\log_2(RRR)$	**5.15273**	$\log_2 \frac{Pr\,(ea\,\vert medicamento)}{Pr\,(ea)}$

Um medicamento em particular que mais provavelmente causa um EA específico que qualquer outro, normalmente terá o valor da medida de associação mais elevado. Se um EA e um medicamento são estocasticamente independentes, o valor da medida de associação receberá o valor igual a 1. Como normalmente cada EA de um medicamento individual ocorre numa proporção pequena do total de notificações, geralmente a << b ou a << c e c << d, na prática estas medidas tendem a ter valores, bem como interpretações, idênticos. Um valor de 3 indica que há três vezes mais notificações envolvendo o par medicamento/EA do que esperado se não houvesse associação entre os dois [95].

Apesar de nosso sistema estar utilizando *tweets* ao invés de relatórios espontâneos, procuramos filtrar nossos *tweets* para que tenham no mínimo um medicamento e um AE, descartando *tweets* que não os tenham, uma abordagem parecida com os Proto-AE de Freifeld [17].

Devido ao fato de nossa ontologia possuir mais de 8 mil EA sem contar os seus sinônimos, os *tweets* não puderam ser coletados usando a abordagem baseada nos EA, do inglês, *event-based approach* [98]. Nesta abordagem, um conjunto de eventos específicos são inspecionados por sua associação com todos os medicamentos possíveis, ou seja, deveríamos coletar *tweets* com os 8 mil EA e neles buscar por medicamentos.

Para o cálculo das medidas, utilizamos uma abordagem que se baseia na seleção de medicamentos, do inglês, *drug-based approach* [98]. Essa abordagem foi escolhida por não sabermos o número de *tweets* com determinado EA e também por esta ter como base a quantidade de *tweets* com EA e os medicamentos relacionados à doença alvo. Com essa abordagem, é mais apropriado considerar um *tweet* com o

nome do medicamento do que coletar um *tweet* qualquer que pode não estar relacionado a medicamentos.

3.5 Avaliação

O sistema verifica se na análise dos dados houve algum sinal (uma associação “medicamento-EA”) utilizando a medida de desproporcionalidade calculada, PRR, em conjunto com seu intervalo de confiança de 95% e também com a utilização do teste χ^2.

3.5.1 Intervalo de confiança de 95% do PRR

O PRR, por se tratar de um método muito sensível, pode gerar muitos falsos positivos, especialmente se o número de notificações for baixo. Para reduzir isso, utilizamos os mesmos critérios utilizados pelo EudraVigilance para definir um sinal de desproporcionalidade para determinado medicamento/EA. Um desses critérios é o cálculo do intervalo de confiança de 95%.

Intervalos de confiança são usados para indicar a confiabilidade de uma estimativa. Por exemplo, um intervalo de confiança pode ser usado para descrever o quanto os resultados de uma pesquisa são confiáveis. Sendo todas as estimativas iguais, uma pesquisa que resulte num intervalo de confiança pequeno é mais confiável do que uma que resulte num intervalo de confiança maior.

O intervalo de confiança de 95% para o logaritmo neperiano de PRR é estimado como $\mathbf{ln}(\mathbf{PRR}) \pm \mathbf{1.96se}$, onde “se” (*standard error*) é o erro padrão da média do logaritmo natural do PRR que pode ser calculado com a seguinte formula [94, 99]:

$$se = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{(a+b)} - \frac{1}{(c+d)}}$$

Onde as letras a, b, c, e d seguem a descrição apresentada na Tabela 2.

Utilizando uma propriedade logarítmica, temos que:

$$IC\ 95\%\ do\ PRR = \frac{PRR}{\exp\ (1.96se)}, PRR\ \times \exp(1.96se).$$

Se o PRR for apresentado com intervalo de confiança de 95%, será considerado como um sinal de desproporcionalidade quando:

- Limite inferior do intervalo >= 1
- Número de casos >= 3

3.5.2 Teste χ^2 (qui-quadrado)

A estatística do χ^2 é um teste inexato resultante da aproximação normal da distribuição de Poisson. É um teste de independência de variáveis categóricas e utilizado como uma medida alternativa da heterogeneidade da tabela de contingencia construída com um medicamento P e um evento adverso E [93]. Em tabelas de contingência 2x2, o χ^2 pode ser calculado com a seguinte fórmula [100]:

$$\chi^2 = \frac{t\ (ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Onde as letras a, b, c, e d seguem a descrição apresentada na Tabela 2.

Se o PRR for apresentado com o χ^2, será considerado como um sinal de desproporcionalidade quando:

- PRR >= 2;
- χ^2 >= 4;
- Número de casos >= 3;

Com um χ^2 >= 4, a probabilidade de esse resultado ter ocorrido por acaso é menor que 5%, já que 4 é maior que o 3,841 da tabela de distribuição do χ^2 com um grau de liberdade, o que significaria uma probabilidade das 2 variáveis da tabela de contingência serem independentes, ser menor que 5%.

A EudraVigilance utiliza o χ^2 como uma medida alternativa de detecção de um sinal em conjunto com o PRR.

4 Resultados

Um dos principais resultados do nosso trabalho foi a construção de uma ferramenta automática para a coleta e análise de Eventos Adversos (EA) no Twitter, pois a aquisição de *tweets* antigos só é possível mediante pagamento a revendedores como Gnip [53], Datasift [54] e Topsy [52]. Nossa ferramenta foi feita de maneira a possibilitar a portabilidade do *pipeline* para outros tipos de textos com o mínimo de esforço possível.

No início do projeto fizemos com ajuda do Topsy [52] algumas consultas por *tweets* relacionados a algumas doenças negligenciadas desde a criação do Twitter em 2006. Até 2008 haviam poucos *tweets* por mês (menos de 200), por isso, decidimos contabilizar apenas os *tweets relacionados* a partir de Janeiro de 2008 até junho 2014. Procuramos por tuberculose (196.790 *tweets*), doença de Chagas (19.999 *tweets*), leishmania (53.338 *tweets*), dengue (3.587.284 *tweets*) e malaria (2.161.169 *tweets*). Dengue e malária foram aquelas com maior quantidade de *tweets*, como pode ser visto na Figura 6.

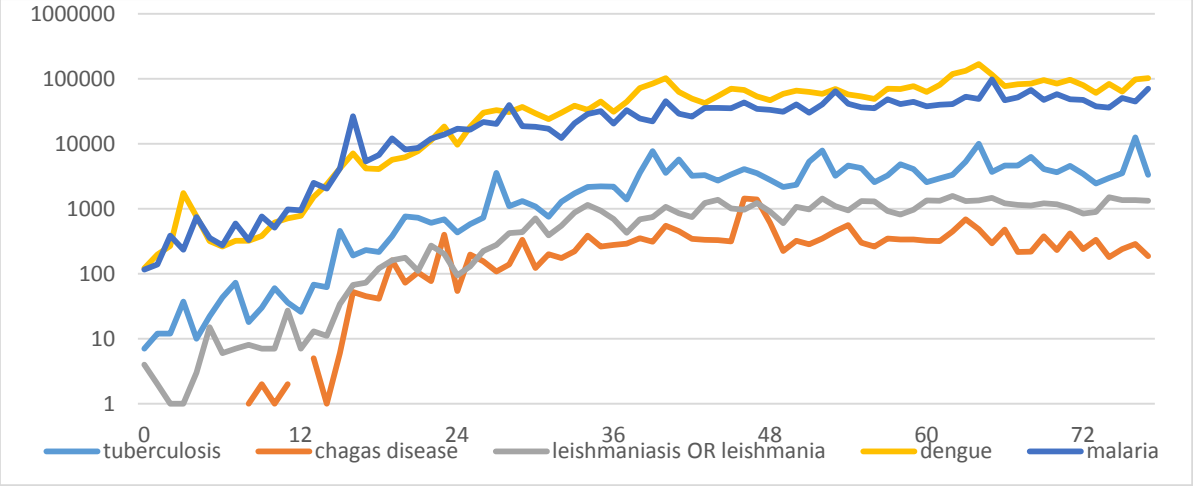


Figura 6. Tweets com doenças desde 2008. No eixo X temos as semanas epidemiológicas e no eixo Y o número de tweets utilizando escala logarítmica na base 10.

Por não existir remédio para a dengue, que não para o tratamento de seus sintomas, nós optamos por fazer um estudo mais aprofundado com malária, que foi a segunda doença com mais *tweets*. Porém, pela dengue ser uma das doenças mais prevalentes no Brasil, nós estudamos seus *tweets* um pouco mais e encontramos o seguinte gráfico mostrado na Figura 7 abaixo.

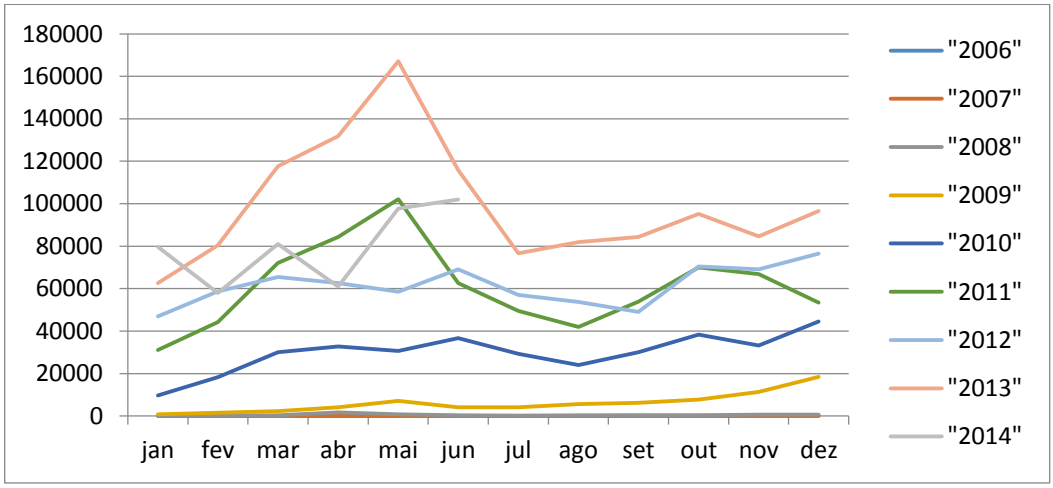


Figura 7. Número de Tweets com dengue entre os anos de 2006 e 2014. No eixo X temos os meses do ano e no eixo Y o número de tweets.

Como pode ser visto na Figura 7, o pico de *tweets* de dengue ocorre próximo aos meses de pico de notificações da doença [101]. Apesar da dengue não possuir nenhum medicamento atualmente, seus dados poderão ser úteis posteriormente para um estudo específico, de maneira similar a feita no Dengue Trends [7] do Google, e por esse motivo, seus *tweets* continuam sendo coletados.

Coletamos *tweets* relacionados a diversas doenças durante todo o ano de 2014, focando nossa análise de desproporcionalidade em fármacos para a malária, como dito anteriormente. Dentre os 19 medicamentos para a malária do nosso padrão de referência, aqueles que apresentaram a maior quantidade de *tweets* em 2014 foram, em primeiro lugar, a doxiciclina, com 14333 *tweets*; em segundo lugar, a cloroquina, com 2912 *tweets*; e em terceiro lugar, mefloquina, com 1312 *tweets* (Tabela 4). Os demais medicamentos possuíam menos de 1000 *tweets* no ano de 2014. Utilizando-se o medicamento doxiciclina juntamente com os seus sinônimos vibramycin (8551), oxytetracycline (353), monodox (133), oracea (57), vibra-tabs (7), doxytetracycline (1), supracyclin (1), doxy-caps (0) e doxycyclinum (0), a quantidade de *tweets* aumentou para mais de 23.000. Dentre os eventos adversos que mais ocorreram nos *tweets* com doxiciclina destacam-se *pain* (dor) e *cough* (tosse), que são dois eventos adversos já conhecidos desse medicamento e estão entre os sintomas com maiores frequências conhecida no ADReCS (6% e 4%), considerando *pain* (dor) como subgrupo de *arthralgia* (artralgia).

Tabela 4. Números de tweets com medicamentos maláricos em 2014

Medicamento[(a)]	Num[(b)]
Morgidox	0
Ocudox	0
Oraxyl	0
Daraprim	35
sulfadoxine	61
proguanil	98
Aralen	122
Doryx	173
atovaquone	191
Fansidar	193
pyrimethamine	216
primaquine	359
Lariam	671
hydroxychloroquine	819
Malarone	890
Plaquenil	982
mefloquine	1312
chloroquine	2912
doxycycline	14333

[(a)]Nomes de medicamentos maláricos; [(b)]Quantidade de *tweets* encontrados com este medicamento.

Por doxiciclina ter sido o medicamento com mais *tweets*, optamos por focar nossa análise no mesmo. Alguns dos *tweets* encontrados em nossa coleta associados a doxiciclina, são exemplificados abaixo:

- *"So with Doxycycline I get indigestion, bloating, nausea, chest pain and other unmentionable symptoms. FUN STUFF."*
- *"I took a doxycycline before bed 11/3 (sinus infection) and woke up to a burning chest/throat...my esophogas is killing me - help!"*
- *"I'm dosed up with doxycycline, pholcodine linctus and hot milk and honey, I hope to all the gods I don't have another coughing fit"*
- *"Advice-Don't go out in the sun on Doxycycline, causes extreme sun sensitivity & painful tingley on fire hands"*

Na Figura 8, podemos ver um gráfico de todos os *tweets* coletados que contém doxiciclina ou seus sinônimos e de relatórios da FDA com doxiciclina, que foram obtidos com o openFDA [97].

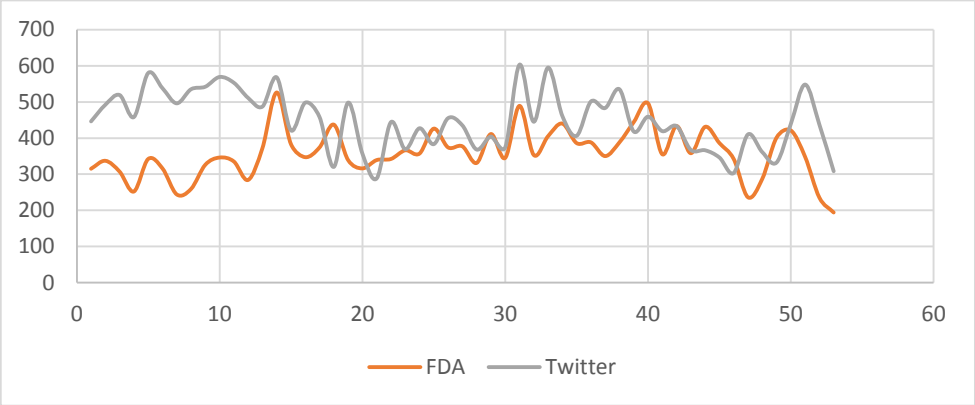


Figura 8. Tweets X Relatórios com doxiciclina por semana epidemiológica em 2014. Os números no eixo X representam as semanas epidemiológicas e no eixo Y, tweets e relatórios do Twitter e FDA, respectivamente.

Como pode ser visto no gráfico (Figura 8), na maioria das semanas epidemiológicas houve mais *tweets* que relatórios do FDA. O total de *tweets* foi 23.460 e de relatórios foi de 18.845 no ano de 2014.

Nas próximas seções temos os resultados para as Análises dos dados do Twitter e do FDA.

4.1 Análise dos dados do Twitter

No cálculo da análise de desproporcionalidade foram considerados somente os *tweets* de medicamentos maláricos que tivessem algum Evento Adverso (EA). Utilizamos também todos os sinônimos de ADR existentes no ADReCS na contagem para a construção das tabelas de contingência.

Na Tabela 5 temos o relatório do PRR para o medicamento doxiciclina com os EA já conhecidos do medicamento no padrão de referência e que possuíam ao menos um *tweet*.

Tabela 5. Relatório do PRR para EA do medicamento doxiciclina (Twitter). Quando é detectado um sinal pelo χ2, a célula é preenchida de vermelha e, de laranja, quando detectado um sinal pelo intervalo de confiança de 95% do PRR. A coluna "FDA"é preenchida de verde quando o sinal tiver ocorrido no Twitter e na FDA.

Reação Adversa[a]	PRR(-)[b]	PRR[c]	PRR(+)[d]	X²[e]	#Tweets[f]	FDA[g]
Abdominal Discomfort	Não Calculado	99,9	Não Calculado	2,356	11	
Abdominal Distension	Não Calculado	99,9	Não Calculado	1,071	5	
Abdominal Pain Upper	Não Calculado	99,9	Não Calculado	6,434	30	
Abscess	Não Calculado	99,9	Não Calculado	0,428	2	
Anaemia	0,197	1,634	13,568	0,166	6	
Anaphylactic Reaction	0,038	0,272	1,933	1,529	2	SIM
Angioedema	0,108	0,233	0,504	12,807	12	
Anorexia	0,017	0,272	4,353	0,764	1	
Anxiety	1,812	4,466	11,007	10,022	82	SIM
Aphthous Stomatitis	1,079	4,493	18,716	4,035	33	
Arthralgia	0,314	0,953	2,894	0,006	14	
Back Pain	0,427	0,657	1,012	2,897	70	
Blood Pressure Increased	Não Calculado	99,9	Não Calculado	2,356	11	
Bronchitis	0,113	0,272	0,654	7,651	10	
Candidiasis	Não Calculado	99,9	Não Calculado	2,999	14	
Cough	0,567	1,634	4,706	0,664	24	
Decreased Appetite	Não Calculado	99,9	Não Calculado	0,428	2	
Dermatitis	Não Calculado	99,9	Não Calculado	0,214	1	SIM
Diarrhoea	0,214	0,681	2,169	0,336	10	
Discomfort	0,017	0,272	4,353	0,764	1	SIM
Dyspepsia	Não Calculado	99,9	Não Calculado	0,642	3	
Dysphagia	0,055	0,272	1,349	2,293	3	
Ear Infection	0,113	0,272	0,654	7,651	10	
Emotional Distress	Não Calculado	99,9	Não Calculado	0,214	1	SIM
Fungal Infection	1,417	4,539	14,543	6,159	50	
Gingivitis	Não Calculado	99,9	Não Calculado	0,214	1	
Haemolytic Anaemia	0,017	0,272	4,353	0,764	1	
Headache	0,165	0,327	0,648	8,937	18	
Hypersensitivity	0,482	0,754	1,179	1,211	72	
Hypertension	Não Calculado	99,9	Não Calculado	1,713	8	
Infection	2,664	4,341	7,076	32,958	271	
Inflammation	Não Calculado	99,9	Não Calculado	1,499	7	
Influenza	0,229	0,256	0,285	528,852	557	
Injury	0,172	0,363	0,767	6,032	16	SIM
Insomnia	0,088	0,182	0,377	20,92	12	
Intracranial Pressure Increase	Não Calculado	99,9	Não Calculado	0,214	1	
Leukopenia	Não Calculado	99,9	Não Calculado	0,428	2	
Malaise	Não Calculado	99,9	Não Calculado	0,856	4	SIM
Muscle Spasms	Não Calculado	99,9	Não Calculado	2,356	11	SIM
Myalgia	0,085	0,817	7,852	0,024	3	
Nasal Congestion	Não Calculado	99,9	Não Calculado	0,214	1	
Nasopharyngitis	0,009	0,091	0,872	5,37	1	
Nausea	0,943	3,949	16,54	3,245	29	
Oedema	0,009	0,091	0,872	5,37	1	
Oesophageal Ulcer	Não Calculado	99,9	Não Calculado	0,642	3	SIM
Oesophagitis	Não Calculado	99,9	Não Calculado	0,642	3	
Oropharyngeal Pain	0,039	0,163	0,683	6,316	3	
Pain	1,556	2,465	3,905	12,485	181	
Photosensitivity Reaction	Não Calculado	99,9	Não Calculado	1,928	9	SIM

Reação Adversa[a]	PRR(-)[b]	PRR[c]	PRR(+)[d]	X^2[e]	#Tweets[f]	FDA[g]
Pigmentation Disorder	0,049	0,545	6,005	0,199	2	
Rash	0,974	2,451	6,17	3,048	45	
Rhinorrhoea	Não Calculado	99,9	Não Calculado	0,214	1	
Sinusitis	0,172	0,272	0,432	27,638	36	
Stevens-Johnson Syndrome	0,036	0,091	0,229	32,26	6	
Stomatitis	Não Calculado	99,9	Não Calculado	0,428	2	
Swelling	1,383	10,076	73,414	6,272	37	
Tension	Não Calculado	99,9	Não Calculado	4,716	22	
Thrombocytopenia	Não Calculado	99,9	Não Calculado	0,428	2	
Tooth Abscess	0,038	0,272	1,933	1,529	2	
Toothache	Não Calculado	99,9	Não Calculado	1,499	7	
Ulcer	Não Calculado	99,9	Não Calculado	3,643	17	
Urticaria	0,064	0,117	0,213	55,349	15	
Vomiting	Não Calculado	99,9	Não Calculado	3,428	16	SIM

[a]Nomes dos EA; [b]Limite inferior do intervalo de confiança de 95% do PRR; [c]Valor do PRR para o EA; [d]Limite superior do intervalo de confiança de 95% do PRR; [e]Valor do teste χ^2; [f]Número de *tweets*; [g] Mostra se houve sinal desse EA na FDA no mesmo período de 2014.

Como pode ser visto na Tabela 3, quando a letra “c” da tabela de contingência 2x2 é zero, não é possível calcular o PRR, por esse motivo é atribuído arbitrariamente “99,9” na coluna “PRR” da Tabela 5 para refletir a presença de um possível sinal. Nesses casos, os limites do intervalo de confiança não são calculados como podem ser visto nas colunas “PRR(-) e PRR(+)”. Lembrando que a letra “a” significa o número de *tweets* com o medicamento alvo e o EA em análise e a letra “c”, o número de *tweets* com o EA mas que não possuem o medicamento alvo. Logo, se “a” é maior que zero e “c” é zero, significa que o EA em análise está relacionado unicamente com o medicamento alvo na base.

Os intervalos de confiança de 95% do PRR dos Eventos Adversos (EA) que geraram algum sinal no Twitter podem ser visto na Figura 9. Nessa figura foi removido o EA *swelling* (inchaço) para uma melhor visualização dos outros, pois este possui um limite superior muito alto.

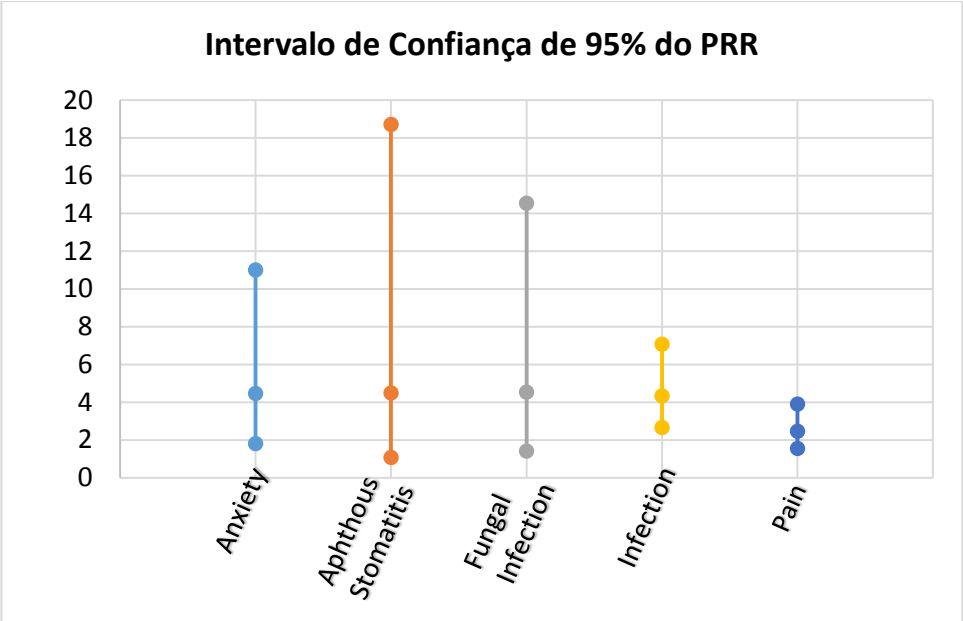


Figura 9. Intervalo de Confiança de 95% do PRR. Para cada um dos EA temos no eixo Y os pontos com os valores do PRR e de seus limites (inferior e superior) do Intervalo de Confiança de 95%

Os EA não relacionados ao medicamento doxiciclina e que tiveram algum sinal são exibidos na Tabela 6.

Tabela 6. Relatório do PRR para EA não relacionados ao medicamento doxiciclina que geraram sinais (Twitter).

Reação Adversa[a]	PRR(-)[b]	PRR[c]	PRR(+)[d]	X²[e]	#Tweets[f]	FDA[g]
Alopecia	Não Calculado	99,9	Não Calculado	33,5190	155	
Rosacea	Não Calculado	99,9	Não Calculado	5,7890	27	SIM

[a]Nomes dos EA; [b]Limite inferior do intervalo de confiança de 95% do PRR; [c]Valor do PRR para o EA; [d]Limite superior do intervalo de confiança de 95% do PRR; [e]Valor do teste χ^2; [f]Número de *tweets*; [g] Mostra se houve sinal desse EA na FDA no mesmo período de 2014.

Foram detectados sinais para dois possíveis novos EA: alopecia (alopécia) e rosacea (rosácea). Alopécia tem como definição a ausência de cabelo em áreas onde é normalmente presente, tendo como sinônimo a palavra calvície. Já rosácea é uma doença cutânea que ocorre principalmente em convexidades da parte central da face como testa, bochecha, nariz e queixo. É caracterizada por rubor, eritema, edema, pápulas e sintomas oculares. Pode ocorrer em qualquer idade, mas normalmente com mais frequência após os 30 anos de idade. Ambos também aparecem no openFDA no mesmo período como pode ser visto na Tabela 7. No openFDA são relatados mais de 200 EA, que podem ser vistos no Anexo G.

Tabela 7. Comparativo entre quantidade de EA encontrados nos tweets e nos relatórios para o medicamento doxiciclina no ano de 2014.

EA[a]	NumTweets[b]	NumRelatoriosFDA[c]
Abdominal Discomfort	11	21
Abdominal Distension	5	10
Abdominal Pain Upper	30	32
Abscess	2	-
Alopecia	155	18
Anaemia	6	33
Anaphylactic Reaction	2	12
Angioedema	12	-
Anorexia	1	-
Anxiety	82	86
Aphthous Stomatitis	33	-
Arthralgia	14	48
Back Pain	70	29
Blood Pressure Increased	11	16
Bronchitis	10	33
Candidiasis	14	-
Cough	24	48
Decreased Appetite	2	36
Dermatitis	1	11
Diarrhoea	10	96
Discomfort	1	17
Dyspepsia	3	11
Dysphagia	3	17
Ear Infection	10	-
Emotional Distress	1	47
Fungal Infection	50	-
Gingivitis	1	-
Haemolytic Anaemia	1	-
Headache	18	119
Hypersensitivity	72	29
Hypertension	8	22
Infection	271	19
Inflammation	7	14
Influenza	557	16
Injury	16	54
Insomnia	12	29
Intracranial Pressure Increase	1	-
Leukopenia	2	-
Malaise	4	91
Muscle Spasms	11	42
Myalgia	3	32
Nasal Congestion	1	-
Nasopharyngitis	1	20
Nausea	29	200
Oedema	1	12
Oesophageal Ulcer	3	18
Oesophagitis	3	-
Oropharyngeal Pain	3	23
Pain	181	122
Photosensitivity Reaction	9	18
Pigmentation Disorder	2	-

EA[(a)]	NumTweets[(b)]	NumRelatoriosFDA[(c)]
Rash	45	90
Rhinorrhoea	1	15
Rosace	27	9
Sinusitis	36	18
Stevens-Johnson Syndrome	6	-
Stomatitis	2	-
Swelling	37	9
Tension	22	-
Thrombocytopenia	2	12
Tooth Abscess	2	-
Toothache	7	-
Ulcer	17	-
Urticaria	15	47
Vomiting	16	137

[(a)]Nomes dos EA; [(b)]Quantidades de *tweets*; [(c)]Quantidade de relatórios da FDA, mostrando o sinal "-" quando não houve nenhum.

4.2 Análise dos dados da FDA

Para podermos comparar a quantidade de *tweets* com EA coletados com a quantidade de relatórios do FDA no mesmo período, realizamos a coleta e análise dos dados da FDA com dois programas que fizemos, que utilizam dados da FDA obtidos através do *site* http://www.researchae.com/ .

O primeiro programa, "getEAfromFDADrug.rb" (Anexo H), retorna os EA do medicamento alvo ordenados alfabeticamente e também o número de ocorrências nos relatórios. Considerando como exemplo a doxiciclina, a Figura 10 mostra os primeiros EA retornados.

```
> ruby getEAfromFDADrug.rb Doxycycline.html
ABDOMINAL DISCOMFORT    21
ABDOMINAL DISTENSION    10
ABDOMINAL PAIN  75
ABDOMINAL PAIN LOWER    13
ABDOMINAL PAIN UPPER    32
ABNORMAL BEHAVIOUR  9
ABORTION INCOMPLETE 22
ACNE    12
ALOPECIA    18
ANAEMIA 33
ANAPHYLACTIC REACTION   12
ANXIETY 80
APHAGIA 11
ARRHYTHMIA  16
ARTHRALGIA  48
ASTHENIA    52
ASTHMA  16
ATRIAL FIBRILLATION 11
BACK PAIN   28
BALANCE DISORDER    18
BLISTER 21
BLOOD ALKALINE PHOSPHATASE INCREASED    10
BLOOD CREATININE INCREASED  9
BLOOD GLUCOSE INCREASED 10
BLOOD PRESSURE DECREASED    9
BLOOD PRESSURE INCREASED    15
BRADYCARDIA 16
BRONCHITIS  32
BRONCHOSPASM    9
CELLULITIS  17
CEREBROVASCULAR ACCIDENT    9
CHEST DISCOMFORT    20
CHEST PAIN  46
CHILLS  28
CHLAMYDIAL INFECTION    16
CHOLELITHIASIS  10
CHRONIC OBSTRUCTIVE PULMONARY DISEASE   24
```

Figura 10. Exemplo do resultado do programa getEAfromFDADrug.rb para o medicamento doxiciclina. É exibido o EA do medicamento doxiciclina e sua quantidade.

O segundo programa, “analiseDadosFDA.rb” (Anexo I), faz a análise dos dados de um medicamento. Para isso também é necessário saber quantidade de EA dos outros medicamentos maláricos, pois seus valores são necessários ao cálculo da tabela de contingência para cada EA. A Figura 11 mostra um exemplo de resultado para o medicamento doxiciclina.

```
> ruby analiseDadosFDA.rb Doxycycline.html,Aralen.html,chloroquine.html,
Doryx.html,Fansidar.html,Lariam.html,mefloquine.html,Ocudox.html,Plaquenil.html,
proguanil.html,sulfadoxine.html,atovaquone.html,Daraprim.html,hydroxychloroquine.html,
Malarone.html,Morgidox.html,Oraxyl.html,primaquine.html,pyrimethamine.html
ABDOMINAL DISCOMFORT    0.726   1.184   1.93    0.095   21
ABDOMINAL DISTENSION    0.421   0.833   1.651   0.057   10
ABDOMINAL PAIN  1.385   1.82    2.392   3.917   75
ABDOMINAL PAIN LOWER    Nao cal 99.9    Nao cal 10.315  13
ABDOMINAL PAIN UPPER    0.677   0.997   1.47    0.0 32
ABNORMAL BEHAVIOUR  2.657   8.626   28.002  3.864   9
ABORTION INCOMPLETE Nao cal 99.9    Nao cal 17.462  22
ACNE    0.683   1.314   2.531   0.139   12
ALOPECIA    0.435   0.719   1.188   0.346   18
ANAEMIA 0.748   1.1 1.618   0.049   33
ANAPHYLACTIC REACTION   5.15    23.003  102.754 7.468   12
ANXIETY 1.884   2.494   3.3 9.033   80
APHAGIA 4.675   21.086  95.108  6.702   11
ARRHYTHMIA  Nao cal 99.9    Nao cal 12.697  16
ARTHRALGIA  0.289   0.388   0.522   8.849   48
ASTHENIA    0.692   0.936   1.266   0.038   52
ASTHMA  0.571   0.989   1.713   0.0 16
ATRIAL FIBRILLATION 0.423   0.811   1.553   0.083   11
BACK PAIN   0.419   0.624   0.93    1.135   28
BALANCE DISORDER    0.649   1.095   1.848   0.024   18
BLISTER 1.244   2.119   3.607   1.656   21
BLOOD ALKALINE PHOSPHATASE INCREASED    2.622   7.668   22.425  4.0 10
BLOOD CREATININE INCREASED  2.657   8.626   28.002  3.864   9
BLOOD GLUCOSE INCREASED 0.688   1.42    2.932   0.188   10
BLOOD PRESSURE DECREASED    0.905   2.03    4.551   0.636   9
BLOOD PRESSURE INCREASED    0.374   0.646   1.116   0.516   15
BRADYCARDIA 2.14    4.382   8.972   4.035   16
BRONCHITIS  0.519   0.757   1.105   0.433   32
```

Figura 11. Exemplo do resultado do programa analiseDadosFDA.rb. Os valores exibidos após o EA são respectivamente PRR(-), PRR, PRR(+), χ^2 e número de relatórios. Da mesma maneira como na análise do tweets, quando a letra "c" da tabela de contingência é zero, não é possível calcular o PRR, por esse motivo é atribuído o valor "99,9" seus limites aparecem como "Nao cal", que significa não calculado.

Dos 19 medicamentos maláricos contidos na Tabela 4, apenas Oraxyl não possuía nenhum relatório no ano de 2014, como pode ser visto na Tabela 8. Como os relatórios são voltados para a detecção de EA, é normal que sua análise tenha um grande número de sinais. A doxiciclina teve um total de 138 sinais de EA que são mostrados na Tabela 9. Essa tabela foi criada utilizando os resultados dos programas de coleta e análise dos dados da FDA.

Tabela 8. Quantidade de EA nos relatórios com medicamentos maláricos em 2014.

Medicamentos[a]	Num[b]
Oraxyl	0
Primaquine	24
Fansidar	34
Sulfadoxine	36
Aralen	48
Lariam	110
Daraprim	128
Pyrimethamine	198
Mefloquine	319
Malarone	385
Proguanil	429
Morgidox	533
Ocudox	533
Chloroquine	621
Doryx	640
Atovaquone	1040
Doxycycline	6079
Plaquenil	7664
Hydroxychloroquine	10564

[a]Nomes de medicamentos maláricos; [b]Quantidade de relatórios encontrados com este medicamento.

Tabela 9. Sinais detectados na análise dos dados da FDA para o medicamento doxiciclina. Quando é detectado um sinal pelo χ2, a célula é preenchida de vermelha e, de laranja, quando detectado um sinal pelo intervalo de confiança de 95% do PRR.

EA[a]	PRR(-)[b]	PRR[c]	PRR(+)[d]	X²[e]	#Relatórios[f]
ABDOMINAL PAIN	1,385	1,820	2,392	3,917	75
ABDOMINAL PAIN LOWER	Não calculado	99,900	Não calculado	10,315	13
ABNORMAL BEHAVIOUR	2,657	8,626	28,002	3,864	9
ABORTION INCOMPLETE	Não calculado	99,900	Não calculado	17,462	22
ANAPHYLACTIC REACTION	5,150	23,003	102,754	7,468	12
ANXIETY	1,884	2,494	3,300	9,033	80
APHAGIA	4,675	21,086	95,108	6,702	11
ARRHYTHMIA	Não calculado	99,900	Não calculado	12,697	16
BLISTER	1,244	2,119	3,607	1,656	21
BLOOD ALKALINE PHOSPHATASE INCREASED	2,622	7,668	22,425	4,000	10
BLOOD CREATININE INCREASED	2,657	8,626	28,002	3,864	9
BRADYCARDIA	2,140	4,382	8,972	4,035	16
BRONCHOSPASM	Não calculado	99,900	Não calculado	7,140	9
CHEST PAIN	1,047	1,470	2,062	1,038	46
CHLAMYDIAL INFECTION	Não calculado	99,900	Não calculado	12,697	16
CHOLELITHIASIS	4,909	38,339	299,444	6,842	10
CHRONIC OBSTRUCTIVE PULMONARY DISEASE	1,906	3,286	5,664	4,260	24
CLOSTRIDIUM DIFFICILE INFECTION	1,674	3,560	7,570	2,569	13
COAGULOPATHY	Não calculado	99,900	Não calculado	10,315	13
COMPLETED SUICIDE	3,536	6,255	11,065	10,778	31

EA[a]	PRR(-)[b]	PRR[c]	PRR(+)[d]	X²[e]	#Relatórios[f]
CONDITION AGGRAVATED	1,356	1,819	2,440	3,386	65
CONSTIPATION	1,229	1,917	2,991	1,759	29
CRYING	4,675	21,086	95,108	6,702	11
DEATH	1,636	2,403	3,531	4,393	42
DEEP VEIN THROMBOSIS	1,489	2,487	4,153	2,683	24
DEHYDRATION	1,283	2,163	3,644	1,823	22
DEPRESSED MOOD	4,359	8,215	15,483	12,533	30
DEPRESSION	1,425	1,971	2,725	3,613	55
DERMATITIS	3,358	10,543	33,099	5,244	11
DEVICE DISLOCATION	Não calculado	99,900	Não calculado	14,285	18
DEVICE ISSUE	Não calculado	99,900	Não calculado	15,873	20
DIPLOPIA	8,963	38,339	163,982	13,688	20
DISCOMFORT	2,059	4,073	8,058	3,959	17
DISEASE PROGRESSION	4,736	16,613	58,282	7,403	13
DISINHIBITION	2,323	6,390	17,574	3,529	10
DIZZINESS	1,098	1,384	1,744	1,572	96
DRUG ADMINISTRATION ERROR	2,657	8,626	28,002	3,864	9
DRUG ERUPTION	3,518	12,780	46,422	5,186	10
DRUG HYPERSENSITIVITY	2,147	2,579	3,097	22,716	187
DRUG INTERACTION	1,630	2,340	3,360	4,653	47
DRUG-INDUCED LIVER INJURY	1,381	3,195	7,391	1,703	10
DYSARTHRIA	4,120	9,858	23,593	8,306	18
DYSPNOEA EXERTIONAL	3,728	17,252	79,829	5,185	9
DYSSTASIA	4,675	21,086	95,108	6,702	11
DYSURIA	4,372	34,505	272,305	6,059	9
ELECTROCARDIOGRAM QT PROLONGED	3,518	12,780	46,422	5,186	10
EMOTIONAL DISTRESS	12,110	30,671	77,680	26,432	40
EPISTAXIS	1,562	2,875	5,295	2,606	18
ERYTHEMA	1,878	2,475	3,263	9,130	82
EYE IRRITATION	5,150	23,003	102,754	7,468	12
FACTITIOUS DISORDER	Não calculado	99,900	Não calculado	7,140	9
FEAR	1,067	2,465	5,691	0,987	9
FEBRILE NEUTROPENIA	2,314	6,901	20,584	3,357	9
FEELING HOT	6,184	14,240	32,792	14,061	26
FLUSHING	2,019	3,229	5,163	5,541	32
FOETAL EXPOSURE DURING PREGNANCY	1,117	2,147	4,128	1,139	14
GASTRIC ULCER	2,048	5,751	16,151	2,924	9
GENERAL PHYSICAL HEALTH DETERIORATION	1,490	2,519	4,260	2,636	23
GENITAL HAEMORRHAGE	Não calculado	99,900	Não calculado	7,934	10
HAEMATEMESIS	7,531	32,588	141,013	11,342	17
HAEMORRHAGE	3,538	5,388	8,205	16,031	52
HALLUCINATION	1,293	2,706	5,663	1,567	12
HEPATIC FAILURE	2,048	5,751	16,151	2,924	9
HYPERAESTHESIA	6,101	26,837	118,053	9,010	14
HYPERSENSITIVITY	1,109	1,731	2,703	1,237	28
HYPOAESTHESIA ORAL	4,201	19,169	87,466	5,940	10
HYPONATRAEMIA	4,606	9,585	19,945	11,375	25
HYPOXIA	5,150	23,003	102,754	7,468	12
INJURY	22,369	92,013	378,475	35,818	48
JARISCH-HERXHEIMER REACTION	3,518	12,780	46,422	5,186	10
JAUNDICE	2,776	5,112	9,412	7,048	24
LETHARGY	1,739	3,408	6,679	2,985	16
LIP SWELLING	Não calculado	99,900	Não calculado	10,315	13
LIPASE INCREASED	4,329	15,335	54,327	6,657	12

EA[a]	PRR(-)[b]	PRR[c]	PRR(+)[d]	X²[e]	#Relatórios[f]
LIVER FUNCTION TEST ABNORMAL	1,162	2,023	3,525	1,334	19
LYME DISEASE	2,314	6,901	20,584	3,357	9
MALAISE	1,074	1,364	1,732	1,349	90
MEDICAL DEVICE DISCOMFORT	Não calculado	99,900	Não calculado	7,934	10
MENORRHAGIA	Não calculado	99,900	Não calculado	9,521	12
MENSTRUATION IRREGULAR	3,728	17,252	79,829	5,185	9
MENTAL STATUS CHANGES	Não calculado	99,900	Não calculado	8,728	11
MOOD SWINGS	2,121	5,272	13,100	3,319	11
MUSCLE ATROPHY	3,358	10,543	33,099	5,244	11
MUSCLE SPASMS	1,188	1,700	2,431	1,783	43
MYALGIA	1,057	1,606	2,440	1,037	31
NAUSEA	1,309	1,540	1,811	5,663	198
NEURALGIA	1,294	2,949	6,722	1,507	10
NIGHT SWEATS	1,522	3,834	9,654	1,951	9
NO THERAPEUTIC RESPONSE	4,675	21,086	95,108	6,702	11
OESOPHAGEAL ULCER	4,120	9,858	23,593	8,306	18
OSTEOMYELITIS	1,212	2,875	6,821	1,302	9
PAIN OF SKIN	2,385	5,367	12,078	4,292	14
PALLOR	Não calculado	99,900	Não calculado	16,668	21
PANCREATITIS ACUTE	6,556	15,974	38,922	14,066	25
PATHOGEN RESISTANCE	Não calculado	99,900	Não calculado	14,285	18
PELVIC PAIN	Não calculado	99,900	Não calculado	12,697	16
PHOTOSENSITIVITY REACTION	2,965	6,274	13,276	6,268	18
PLEURAL EFFUSION	1,163	2,481	5,293	1,222	11
POLLAKIURIA	1,732	4,260	10,479	2,446	10
PREGNANCY	5,551	19,169	66,195	8,912	15
PRODUCT USE ISSUE	4,675	21,086	95,108	6,702	11
PROTEINURIA	Não calculado	99,900	Não calculado	6,347	8
PRURITUS	1,148	1,524	2,025	1,776	66
PULMONARY EMBOLISM	2,142	3,263	4,969	7,036	40
QUALITY OF LIFE DECREASED	3,837	30,671	245,185	5,280	8
RASH	1,135	1,448	1,847	1,852	88
RASH ERYTHEMATOUS	1,080	1,868	3,230	1,068	19
RASH GENERALISED	1,133	1,969	3,421	1,240	19
RASH MACULAR	6,198	18,211	53,511	11,128	19
RASH PRURITIC	1,275	2,176	3,714	1,765	21
RENAL APLASIA	Não calculado	99,900	Não calculado	6,347	8
RESPIRATORY DISORDER	6,521	49,840	380,929	9,198	13
RESPIRATORY FAILURE	1,044	1,862	3,322	0,947	17
RHABDOMYOLYSIS	3,257	15,335	72,199	4,437	8
ROSACEA	Não calculado	99,900	Não calculado	6,347	8
SHOCK	3,115	11,502	42,472	4,465	9
SKIN BURNING SENSATION	5,631	11,928	25,263	14,124	28
SKIN EXFOLIATION	2,031	3,834	7,237	4,122	19
SKIN WARM	Não calculado	99,900	Não calculado	6,347	8
SLOW RESPONSE TO STIMULI	Não calculado	99,900	Não calculado	8,728	11
SOCIAL AVOIDANT BEHAVIOUR	3,243	9,201	26,108	5,340	12
SUICIDAL IDEATION	4,278	7,029	11,549	16,651	44
SWELLING FACE	1,042	1,721	2,844	0,952	22
SYNCOPE	2,641	4,382	7,270	8,079	32
TACHYCARDIA	4,480	9,062	18,329	11,483	26
TENDONITIS	1,690	3,383	6,770	2,770	15
TINNITUS	1,211	2,157	3,839	1,481	18
TREATMENT FAILURE	7,627	22,045	63,720	14,175	23

EA[a]	PRR(-)[b]	PRR[c]	PRR(+)[d]	X²[e]	#Relatórios[f]
UPPER GASTROINTESTINAL HAEMORRHAGE	5,446	42,172	326,596	7,626	11
URINARY INCONTINENCE	4,317	9,372	20,343	9,892	22
URTICARIA	1,307	1,856	2,637	2,545	46
UTERINE PERFORATION	Não calculado	99,900	Não calculado	22,229	28
VENTRICULAR TACHYCARDIA	Não calculado	99,900	Não calculado	11,109	14
VERTIGO	1,067	2,465	5,691	0,987	9
VITH NERVE PARALYSIS	5,150	23,003	102,754	7,468	12
VOMITING	1,558	1,910	2,342	8,256	136
WHEEZING	1,258	2,396	4,565	1,556	15
WITHDRAWAL SYNDROME	3,837	30,671	245,185	5,280	8

[a]Nomes dos EA; [b]Limite inferior do intervalo de confiança de 95% do PRR; [c]Valor do PRR para o EA; [d]Limite superior do intervalo de confiança de 95% do PRR; [e]Valor do teste χ^2; [f]Número de relatórios.

5 Discussão

Com o intuito de construir um sistema capaz de coletar, armazenar e processar *tweets* relacionados a medicamentos, implementamos inicialmente um coletor utilizando as API's do próprio Twitter (*REST API* e *Streaming API*). Para que tivéssemos um panorama sobre os *tweets* relacionados a medicamentos, realizamos algumas consultas pelas doenças negligenciadas tuberculose, doença de Chagas, leishmaniose, dengue e malária como mostrado na Figura 6. A doença com o maior número de *tweets* foi a dengue, mas por essa não possuir medicamentos para o seu tratamento, optamos por focar nosso estudo teste da ferramenta na doença malária e seus medicamentos, pois esta era a segunda doença com maior quantidade de *tweets*.

Para podermos realizar nossa análise, foi necessária a criação de um padrão de referência de doenças, os medicamentos para seu tratamento e os EA que estes podem vir a causar. Para a construção desse padrão de referência utilizamos o ADReCS, o *site* "www.drugs.com", MedlinePlus, RxNorm API RESTful e outras bases médicas. Esse padrão foi essencial para a realização da etapa de avaliação dos dados e também para escolha dos medicamentos.

Nossa análise do Twitter foi feita com base nos *tweets* que o nosso sistema coletou durante todo o período de 2014. São *tweets* com 19 medicamentos relacionados à malária. Dentre esses medicamentos, alguns não apresentaram qualquer *tweet* como Morgidox, Ocudox e Oraxyl. Além desses, outros não apresentaram uma quantidade significativa para que se fosse feita alguma análise ou, então, não possuía qualquer Evento Adverso (EA). O maior número de *tweets* retornados foi do medicamento doxiciclina (14333 *tweets* sem incluir medicamentos similares) como mostrado na Tabela 4.

Na análise do Twitter, detectamos sinais para oito EA já conhecidos da doxiciclina: Abdominal Pain Upper (dor abdominal superior), Anxiety (ansiedade), Aphthous Stomatitis (estomatite aftosa), Fungal Infection (infecção por fungos), Infection (infecção), Pain (dor), Swelling (inchaço) e Tension (tensão). Além desses foram detectados dois EA ainda não relacionados: Alopecia (alopécia) e Rosacea (rosácea). Dos oito EA detectados, apenas Anxiety (ansiedade) também foi detectado na análise de dados da FDA.

Para uma melhor comparação entre os sinais detectados em cada uma das duas análises consideramos separadamente 3 tipos de sinais:

- **Tipo A**: gerados pelo critério do intervalo de confiança do PRR, ou seja, quando o limite inferior do intervalo de confiança de 95% do PRR for maior ou igual a 1 e a quantidade de *tweets*/relatórios for maior ou igual a 3;
- **Tipo B**: gerados pelo critério do χ^2, ou seja, PRR maior ou igual a 2 e χ^2 maior ou igual a 4 e a quantidade de *tweets*/relatórios for maior ou igual a 3;
- **Tipo C**: quando ocorreram os sinais do tipo A e B;

Não houve sinais do tipo A gerados pelo Twitter. A FDA gerou um total de 51 sinais do tipo A, dos quais 40 não se encontram no padrão de referência. Os 11 EA dos sinais que já estavam no padrão de referência são *Abdominal pain* (dor abdominal)*, Discomfort* (desconforto)*, Hypersensitivity* (hipersensibilidade)*, Malaise* (mal-estar)*, Muscle spasms* (espasmos musculares)*, Myalgia* (mialgia)*, Nausea* (náusea)*, Rash* (erupção cutânea)*, Rash erythematous* (exantema eritematoso)*, Urticaria* (urticaria) *e Vomiting* (vomito).

Foram gerados 2 sinais do tipo B pelo Twitter, para os eventos adversos *Abdominal Pain Upper* (dor abdominal superior) e *Tension* (tensão), ambos presentes no padrão de referência. E também foram gerados outros 2 sinais do tipo B que não estão no padrão de referência para os EA *Alopecia* (alopécia) e *Rosacea* (rosácea). Dentre estes sinais, apenas *Rosacea* (rosácea) também ocorreu na FDA que teve um total de 24 sinais do tipo B, dos quais, apenas *Menorrhagia* (menorragia) se encontra no padrão de referência.

O Twitter gerou um total de 6 sinais do tipo C para os EA: *Anxiety* (ansiedade)*, Aphthous stomatitis (estomatite aftsoa), Fungal infection* (infeção fúngica)*, Infection* (infecção)*, Pain* (dor) *e Swelling* (inchaço). Todos estes estão presentes no padrão de referência de EA para doxiciclina. Destes sinais, apenas *Anxiety* (ansiedade) ocorreu na FDA que teve um total 63 sinais, dos quais, oito já estavam presentes no padrão de referência: *Anaphylactic reaction* (reação anafilática)*, Anxiety* (ansiedade)*, Dermatitis* (dermatite)*, Emotional distress* (angustia emocional)*, Injury* (lesão)*, Oesophageal ulcer* (úlcera de esôfago)*, Photosensitivity reaction* (reação de fotossensibilidade) *e Rash maculo-papular* (erupção maculopaular) e outros 55 sinais que não se encontravam no padrão de referência.

Comparando-se as Tabelas 5 e 7 verificamos que existem três EA já presentes no padrão de referência e que geraram sinais somente no Twitter pois não houve relatórios da FDA com os mesmos, são eles: *Aphthous stomatitis (estomatite aftsoa), Fungal infection* (infeção fúngica) e *Tension* (tensão), o que demonstra que EA que não aparecem nos relatórios poderiam ser detectados no Twitter, já que estes também são EA de doxiciclina.

Pesquisando sobre os dois EA que não estavam no padrão de referência (rosácea e alopécia) e que foram detectados pelo Twitter, verificamos que os mesmos também aparecem nos relatórios do FDA do mesmo período. Encontramos tantos relatos de que a doxiciclina poderia causar a calvície como de que também poderia evitá-la. Sobre o EA rosácea, a grande maioria de *tweets* e de relatos que pesquisamos falavam que o medicamento era utilizado para o seu tratamento e não o responsável pela sua causa [102], como pode ser visto em alguns dos *tweets*:

- *There is evidence for the use of topical metronidazole and azelaic acid, as well as doxycycline, to treat rosacea. http://t.co/v18dYEPlLy*
- *Was prescribed doxycycline for the conjunctivitis which is often given to treat rosacea so...two birds with one stone, I guess?*

Isto é algo que precisa ainda ser refinado em nossa ferramenta. Distinguir de forma mais eficaz quando um medicamento é citado como causa ou como prevenção.

Tanto alopécia como rosácea aparecem nos relatórios da FDA, contudo apenas rosácea também gerou um sinal na análise de dados da FDA. Isso é mais um indício que a utilização de múltiplas fontes de dados traz uma maior sensibilidade, pois se considerarmos apenas eventos raros, a análise de dados de múltiplas fontes é necessária para se conseguir o poder estatístico e a heterogeneidade populacional necessárias para detectar diferenças da efetividade de drogas em subpopulações, levando-se em conta diferenças genéticas, étnicas e clinicas [103].

O fato de alopécia não estar no padrão de referência significa que ela pode ser um potencial novo EA. Além disso, esse sinal não foi detectado pela FDA, mas somente pelo Twitter, indicando que o Twitter poderia detectar sinais que escapariam de outras fontes.

Todos os resultados das análises são como o próprio nome diz, sinais, e não afirmações sobre nenhuma relação de causa e efeito entre o medicamento e o EA.

Tais afirmações de maneira alguma podem vir a ser feitas de forma automática, sendo possíveis somente por especialistas que viriam a se utilizar de tais sinais, como avisos iniciais para ajudarem sua avaliação mais aprofundada.

Lembramos que os valores de PRR e χ^2 são medidas de associação e não de causalidade e, por isso, alguns eventos podem não ter gerado sinais apesar de serem eventos já relacionados com os medicamentos analisados, e isso ocorre tanto na análise do Twitter como na da FDA. Nenhuma das duas análises gerou sinais para todos os EA já existentes no padrão de referência.

Com base nos resultados obtidos, a resposta para a hipótese inicial de que o Twitter pode ser utilizado como fonte para a Farmacovigilância é positiva. O Twitter é útil para a Farmacovigilância sob algumas condições, não como uma fonte de dados isolada, mas como uma fonte complementar de dados, pois como foi verificado o Twitter foi capaz de gerar tanto sinais novos como os que já existem no padrão de referência que, em algumas situações não foram encontrados nos dados da FDA.

Uma crença emergente na pesquisa em Farmacovigilância é que a combinação de informações de múltiplas fontes de dados pode levar a descoberta mais eficaz e precisa de Eventos Adversos (EAs) [3]. Dependendo das fontes de dados utilizadas e do modo como elas são combinadas, acredita-se que o sistema resultante poderia levar ao aumento da significância estatística dos resultados ou facilitaria novas descobertas que não são possíveis utilizando fontes de dados isoladas. Esta hipótese já foi preliminarmente confirmada recentemente [3], mas novos estudos são necessários. A utilização de múltiplos *pipelines* como o mostrado na Figura 2, com as etapas de processamento, avaliação e análise dos dados, cada um com fontes de dados diferentes, seria uma ótima maneira de corroborar tal hipótese.

Outro fator importante é a disponibilidade dos dados. Com o Twitter temos um acesso em tempo real para análise dos dados, enquanto que as redes de vigilância farmacológica costumam demorar a disponibilizar os seus dados. A FDA, por exemplo, disponibiliza os seus dados por trimestre, mas não necessariamente os dados se tornam público após decorridos 3 meses. Geralmente os dados relativos aos meses de janeiro, fevereiro e março de um ano só vêm a se tornar públicos depois da metade do mesmo.

Portanto, sim, o Twitter é útil para a Farmacovigilância, com ele foram detectados EA que não se encontravam no padrão de referência (alopecia e

rosacea) e dentre estes, alopecia não se encontrava nos sinais gerados pela FDA. Contudo, mais análises devem ser feitas para corroborar esses resultados, análises com outros medicamentos e outros períodos de tempo. A inclusão de novas fontes de dados também seria uma importante forma de validação desses resultados.

6. Conclusões

Neste trabalho, nós criamos um sistema automatizado de coleta e análise de *tweets* com o intuito de identificar eventos adversos novos e já existentes, para verificar se sua utilização pode ser útil como fonte complementar para uso em Farmacovigilância. Desde o início do nosso trabalho, outras pesquisas também exploraram essa ideia, realizando a coleta de dados por um período determinado, utilizando apenas uma ou nenhuma ontologia para fazer isso e com várias etapas manuais em seus *pipelines*.

Realizamos uma análise de desproporcionalidade em cima dos dados utilizando abordagens frequentistas de maneira a identificar sinais de eventos adversos. Os resultados obtidos demonstraram que o sistema conseguiu alcançar o seu objetivo, identificando tanto eventos adversos já existentes quanto novos.

Uma das principais contribuições desse projeto foi a construção de uma base de doenças, medicamentos e eventos adversos para a utilização do sistema no processamento de *tweets*. Tal base poderá ser utilizada com qualquer outro tipo de texto, tendo em vista que o Twitter tem a escrita livre.

O Twitter pode vir a ser considerado como uma fonte de dados válida para a Farmacovigilância desde que seja feito um estudo prévio para verificar a existência de volume de dados para as doenças/medicamentos que serão analisadas, e desde que o Twitter não seja utilizado de forma isolada mas de forma complementar com outras fontes de dados.

Como trabalhos futuros de projeto, temos a tradução da base para a língua portuguesa para utilização em um projeto de pesquisa associado ao convênio INI/PROCC, onde dados de relatórios médicos serão submetidos a um *pipeline* parecido ao do Twitter. Outro importante fator para que o sistema não fique focado exclusivamente em dados do Twitter é por este ser uma rede social e, como tal, correr o risco de um dia acabar como acontecido com o Orkut. De acordo com notícias recentes, o Twitter vem perdendo o seu ritmo de crescimento [104]. Melhorias também precisam e devem ser feitas continuamente no *pipeline* do sistema, tendo em vista que este se utiliza de várias API's de terceiros e estas podem ter mudanças, o que implicaria em modificações no sistema.

7. Referências

1. Mendes, M., et al., *História da farmacovigilância no Brasil.* Rev Bras Farm, 2008. **89**: p. 246-251.
2. Venulet, J. and M. Ten Ham, *Methods for monitoring and documenting adverse drug reactions.* International journal of clinical pharmacology and therapeutics, 1996. **34**(3): p. 112.
3. Harpaz, R., et al., *Combining signals from spontaneous reports and electronic health records for detection of adverse drug reactions.* Journal of the American Medical Informatics Association, 2013. **20**(3): p. 413-419.
4. Hauben, M. and A. Bate, *Data mining in drug safety: side effects of drugs essay.* Side effects of drugs annual, 2007. **29**: p. xxxiii-xlvi.
5. Ginsberg, J., et al., *Detecting influenza epidemics using search engine query data.* Nature, 2009. **457**(7232): p. 1012-1014.
6. Carneiro, H.A. and E. Mylonakis, *Google trends: a web-based tool for real-time surveillance of disease outbreaks.* Clinical infectious diseases, 2009. **49**(10): p. 1557-1564.
7. Gluskin, R.T., et al., *Evaluation of Internet-based dengue query data: Google Dengue Trends.* PLoS Negl Trop Dis, 2014. **8**(2): p. e2713.
8. Signorini, A., A.M. Segre, and P.M. Polgreen, *The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic.* PloS one, 2011. **6**(5): p. e19467.
9. Lampos, V. and N. Cristianini, *Nowcasting events from the social web with statistical learning.* ACM Transactions on Intelligent Systems and Technology (TIST), 2012. **3**(4): p. 72.
10. IVF. *Instituto Virtual de Fármacos do Estado do Rio de Janeiro. 2006. IVFRJ On Line*. [acessado em agosto de 2016]; 13ª:[Disponível em: http://www.ivfrj.ccsdecania.ufrj.br/ivfonline/edicao_0013/terminologia.html.
11. BRASIL. *Ministério da Saúde. Portaria nº 3.916/GM, de 30 de outubro de 1998. Diário Oficial da União, Brasília, nº 215-E, 10 nov. 1998a. Seção 1, p. 18-22*. [acessado em agosto de 2016]; Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/gm/1998/prt3916_30_10_1998.html.
12. BRASIL, *Resolução - RDC n.º 135, de 29 de maio de 2003. Aprova Regulamento Técnico para Medicamentos Genéricos.*, B. Diário Oficial [da] República Federativa do Brasil, Editor. 2003.
13. *ANVISA*. [acessado em agosto de 2016]; Disponível em: http://portal.anvisa.gov.br/wps/content/Anvisa+Portal/Anvisa/Pos+-+Comercializacao+-+Pos+-+Uso/Farmacovigilancia.
14. BRASIL. *Ministério da Saúde - Agência Nacional de Vigilância Sanitária - Resolução RDC 04/09*. [acessado em agosto de 2016]; Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/anvisa/2009/res0004_10_02_2009.html.
15. Patki, A., et al., *Mining adverse drug reaction signals from social media: going beyond extraction.* Proceedings of BioLinkSig, 2014. **2014**.
16. Sampathkumar, H., X.-w. Chen, and B. Luo, *Mining adverse drug reactions from online healthcare forums using hidden Markov model.* BMC medical informatics and decision making, 2014. **14**(1): p. 91.
17. Freifeld, C.C., et al., *Digital Drug Safety Surveillance: Monitoring Pharmaceutical Products in Twitter.* Drug Safety, 2014. **37**(5): p. 343-350.
18. Rachel Ginn, P.P., Azadeh Nikfarjam, MS, Apurv Patki, Karen O'Connor, Abeed Sarker, PhD, Karen Smith, PhD, Graciela Gonzalez, PhD, *Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark.* 2014.
19. O'Connor, K., et al. *Pharmacovigilance on Twitter? Mining Tweets for adverse drug reactions*. in *AMIA Annual Symposium Proceedings*. 2014. American Medical Informatics Association.

20. Guideline, I.H.T., *Clinical safety data management: definitions and standards for expedited reporting.* Recommended for Adoption at Step, 1994. **4**.
21. *Portal da Saúde RJ*. [acessado em agosto de 2016]; Disponível em: http://www.saude.rj.gov.br/vigilancia-em-saude/99-vigilancia-sanitaria/314-farmacovigilancia.html.
22. Balbino, E.E. and M.F. Dias, *Farmacovigilância: um passo em direção ao uso racional de plantas medicinais e fitoterápicos.* Rev bras farmacogn, 2010. **20**(6): p. 992-1000.
23. NOTIVISA. *Sistema de Notificações em Vigilância Sanitária*. [acessado em agosto de 2016]; Disponível em: http://www.anvisa.gov.br/hotsite/notivisa/index.htm.
24. Berkowitz, B. and B. Katzung, *Avaliação básica e clínica de novas drogas.* Katzung BG. Farmacologia básica e clínica 8a ed. Rio de Janeiro: Guanabara Koogan, 2003: p. 54-63.
25. NASCIUTTI, P.R., *Desenvolvimento de Novos Fármacos - Seminário apresentado junto à Disciplina Seminários Aplicados do Programa de Pós-Graduação em Ciência Animal da Escola de Veterinária e Zootecnia da Universidade Federal de Goiás.* 2012.
26. Ferreira, F.G., et al., *Fármacos: do desenvolvimento à retirada do mercado.* Revista Eletrônica de Farmácia, 2009. **6**(1).
27. Pestana, J.O.M., M. Castro, and W. Pereira, *Pesquisa clínica e farmacovigilância.* Prática Hospitalar, 2006. **44**.
28. Guerrero, G.A.M. and M. Lorenzana-Jiménez, *Las fases en el desarrollo de nuevos medicamentos.* Rev Fac Med UNAM, 2009. **52**(6).
29. Gomes, R.d.P., et al., *Ensaios clínicos no Brasil: competitividade internacional e desafios.* BNDES Setorial, n. 36, set. 2012, p. 45-84, 2012.
30. Waller, P., *An introduction to pharmacovigilance*. 2011: John Wiley & Sons.
31. Junod, S.W., *FDA and clinical drug trials: a short history.* US Food and Drug Administration, 2013.
32. Lima, L.M., C.A.M. Fraga, and E.J. Barreiro, *O renascimento de um fármaco: talidomida.* Química Nova, 2001. **24**(5): p. 683-688.
33. Honig, P.K., et al., *Terfenadine-ketoconazole interaction: pharmacokinetic and electrocardiographic consequences.* Jama, 1993. **269**(12): p. 1513-1518.
34. Elmasri, R. and S.B. Navathe, *Fundamentals of Database Systems, 1989.* Redwood City, Calif.: Benjamin/Cummings. **802**.
35. Han, J., M. Kamber, and J. Pei, *Data mining: concepts and techniques: concepts and techniques*. 2011: Elsevier.
36. da Costa Côrtes, S., R.M. Porcaro, and S. Lifschitz, *Mineração de dados-funcionalidades, técnicas e abordagens*. 2002: PUC.
37. Hotho, A., A. Nürnberger, and G. Paaß. *A Brief Survey of Text Mining*. in *Ldv Forum*. 2005.
38. Zweigenbaum, P., et al., *Frontiers of biomedical text mining: current progress.* Briefings in bioinformatics, 2007. **8**(5): p. 358-375.
39. Hearst, M.A. *Untangling text data mining*. in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. 1999. Association for Computational Linguistics.
40. Chowdhury, G.G., *Natural language processing.* Annual review of information science and technology, 2003. **37**(1): p. 51-89.
41. Aranha, C. and E. Passos, *A tecnologia de mineração de textos.* Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi: 10.5329/RESI, 2006. **5**(2).
42. Hay, S.I., et al., *Big data opportunities for global infectious disease surveillance.* PLoS medicine, 2013. **10**(4): p. e1001413.
43. Ribeiro, C.J.S., *Big Data: os novos desafios para o profissional da informação.* Informação & Tecnologia, 2014. **1**(1): p. 96-105.
44. MATTOSO, M., *Scientific Workflows and Big Data. Palestra apresentada no 1o. EMC Summer School on Big Data*. 2013, EMC/NCE/UFRJ. Rio de Janeiro.
45. *Twitter*. [acessado em agosto de 2016]; Disponível em: http://www.twitter.com.

46. *PatientsLikeMe*. [acessado em agosto de 2016]; Disponível em: http://www.patientslikeme.com/.
47. *AskaPatient*. [acessado em agosto de 2016]; Disponível em: http://www.askapatient.com/.
48. Industry, T.A.o.t.B.P. *Guidance notes on the management of adverse events and product complaints from digital media*. [acessado em agosto de 2016]; Disponível em: http://www.abpi.org.uk/our-work/library/guidelines/Documents/ABPI%20Guidance%20on%20PV%20and%20Digital%20Media.pdf.
49. *Twitter Statistics | Statistics Brain*. [acessado em agosto de 2016]; Disponível em: http://www.statisticbrain.com/twitter-statistics/.
50. *10 dados sobre o Twitter*. [acessado em agosto de 2016]; Disponível em: https://www.revelabit.com.br/blog/infografico-10-dados-sobre-o-twitter-que-voce-precisa-saber/.
51. *Twitter Firehose*. [acessado em agosto de 2016]; Disponível em: https://dev.twitter.com/streaming/firehose.
52. *Topsy*. [acessado em outubro de 2014]; Disponível em: http://topsy.com/.
53. *Gnip*. [acessado em agosto de 2016]; Disponível em: http://gnip.com/.
54. *Datasift*. [acessado em agosto de 2016]; Disponível em: http://datasift.com/.
55. *Dataminr*. [acessado em agosto de 2016]; Disponível em: http://www.dataminr.com/.
56. *Topsy Closes Down* [acessado em agosto de 2016]; Disponível em: http://techcrunch.com/2015/12/15/rip-topsy/.
57. *Twitter compra Gnip*. 2016; Disponível em: http://www.forbes.com/sites/benkepes/2014/04/15/twitter-buys-gnip-its-all-about-the-data/.
58. Bian, J., U. Topaloglu, and F. Yu. *Towards large-scale twitter mining for drug-related adverse events*. in *Proceedings of the 2012 international workshop on Smart health and wellbeing*. 2012. ACM.
59. Toldo, L., S. Bhattacharya, and H. Gurulingappa. *Automated identification of adverse events from case reports using machine learning*. in *Workshop on Computational Methods in Pharmacovigilance*. 2012.
60. H., R., A., & Toldo, L, *Extraction of Adverse Drug Effects from Medical Case Reports.* 2012.
61. Harpaz, R., et al., *Novel data-mining methodologies for adverse drug event discovery and analysis.* Clinical Pharmacology & Therapeutics, 2012. **91**(6): p. 1010-1021.
62. Leaman, R., et al. *Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks*. in *Proceedings of the 2010 workshop on biomedical natural language processing*. 2010. Association for Computational Linguistics.
63. Nikfarjam, A. and G.H. Gonzalez. *Pattern mining for extraction of mentions of adverse drug reactions from user comments*. in *AMIA Annual Symposium Proceedings*. 2011. American Medical Informatics Association.
64. Yates, A. and N. Goharian, *ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites*, in *Advances in Information Retrieval*. 2013, Springer. p. 816-819.
65. Kuhn, M., et al., *A side effect resource to capture phenotypic effects of drugs.* Molecular systems biology, 2010. **6**(1).
66. Adrian Benton, B., et al., *Identifying potential adverse effects using the web: a new approach to medical hypothesis generation.*
67. Sampathkumar, H., B. Luo, and X.-w. Chen. *Mining Adverse Drug Side-Effects from Online Medical Forums*. in *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*. 2012. IEEE.
68. Cobb, N.K., et al., *Online social networks and smoking cessation: a scientific research agenda.* Journal of Medical Internet Research, 2011. **13**(4).

69. Prier, K.W., et al., *Identifying health-related topics on twitter*, in *Social computing, behavioral-cultural modeling and prediction*. 2011, Springer. p. 18-25.
70. Paul, M.J. and M. Dredze. *You are what you Tweet: Analyzing Twitter for public health*. in *ICWSM*. 2011.
71. Tumasjan, A., et al., *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.* ICWSM, 2010. **10**: p. 178-185.
72. Golder, S.A. and M.W. Macy, *Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures.* Science, 2011. **333**(6051): p. 1878-1881.
73. Melton, G.B. and G. Hripcsak, *Automated detection of adverse events using natural language processing of discharge summaries.* Journal of the American Medical Informatics Association, 2005. **12**(4): p. 448-457.
74. Cai MC, X.Q., Pan YJ, Pan W, Ji N, Li YB, Jin HJ, Liu K, Ji ZL., *ADReCS: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms.* Nucleic Acids Research, 2014.
75. *DaylyMed*. [acessado em agosto de 2016]; Disponível em: http://dailymed.nlm.nih.gov/dailymed/index.cfm.
76. Bodenreider, O., *The unified medical language system (UMLS): integrating biomedical terminology.* Nucleic acids research, 2004. **32**(suppl 1): p. D267-D270.
77. Law, V., et al., *DrugBank 4.0: shedding new light on drug metabolism.* Nucleic acids research, 2014. **42**(D1): p. D1091-D1097.
78. *MedlinePlus*. [acessado em agosto de 2016]; Disponível em: http://www.nlm.nih.gov/medlineplus/connect.
79. *RXNorm API*. [acessado em agosto de 2016]; Disponível em: http://mor.nlm.nih.gov/download/rxnav/RxNormAPIREST.html.
80. *MongoDB*. Disponível em: https://www.mongodb.org.
81. *Twitter REST API*. [acessado em agosto de 2016]; Disponível em: https://dev.twitter.com/rest/public.
82. *Twitter Streaming API*. [acessado em agosto de 2016]; Disponível em: https://dev.twitter.com/docs/api/streaming.
83. Felipe Duval, E.C., Oswaldo Cruz and Fabrício Silva, *Mining for adverse drug events on twitter*, in *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. 2014.
84. Friedman, C., et al., *Natural language processing in an operational clinical information system.* Natural Language Engineering, 1995. **1**(01): p. 83-108.
85. Savova, G.K., et al., *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.* Journal of the American Medical Informatics Association, 2010. **17**(5): p. 507-513.
86. Aronson, A.R. *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program*. in *Proceedings of the AMIA Symposium*. 2001. American Medical Informatics Association.
87. *RxNorm*. [acessado em agosto de 2016]; Disponível em: http://www.nlm.nih.gov/research/umls/rxnorm/.
88. Côté, R.A. and C.o.A. Pathologists, *SNOMED International: the systematized nomenclature of human and veterinary medicine*. Vol. 3. 1993: College of American Pathologists.
89. *Orange Book: Approved Drug Products with Therapeutic Equivalence Evaluations*. [acessado em agosto de 2016]; Disponível em: http://www.accessdata.fda.gov/scripts/cder/ob/.
90. *LVG. Lexical Variant Generation.* [acessado em agosto de 2016]; Disponível em: http://www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_004.htm.
91. *U.S. National Library of Medicine / National Institutes of Health*. [acessado em agosto de 2016]; Disponível em: http://www.nlm.nih.gov/.
92. Klarreich, E., *In search of Bayesian inference.* Communications of the ACM, 2014. **58**(1): p. 21-24.

93. Group, E.E.W., *Guideline on the use of statistical signal detection methods in the Eudravigilance data analysis system*. 2006: European Medicines Agency.
94. Van Puijenbroek, E.P., W.L. Diemont, and K. van Grootheest, *Application of quantitative signal detection in the Dutch spontaneous reporting system for adverse drug reactions.* Drug Safety, 2003. **26**(5): p. 293-301.
95. Dias, P., C.F. Ribeiro, and F.B. Marques, *MEDIDAS DE DESPROPORCIONALIDADE NA DETEÇÃO DE SINAL EM FARMACOVIGILÂNCIA.* Revista Portuguesa de Farmacoterapia, 2014. **6**(1): p. 28-32.
96. Evans, S., P.C. Waller, and S. Davis, *Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports.* Pharmacoepidemiology and drug safety, 2001. **10**(6): p. 483-486.
97. *openFDA*. [acessado em agosto de 2016]; Disponível em: https://open.fda.gov.
98. Trifirò, G., et al., *Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor?* Pharmacoepidemiology and drug safety, 2009. **18**(12): p. 1176-1184.
99. Rothman, K. and S. Greenland, *Introduction to categorical statistics.* Modern epidemiology, 1998. **2**: p. 237-239.
100. Everitt, B.S., *The analysis of contingency tables*. 1992: CRC Press.
101. Teixeira, M.d.G., M.L. Barreto, and Z. Guerra, *Epidemiologia e medidas de prevenção do dengue.* Informe epidemiológico do SUS, 1999. **8**(4): p. 5-33.
102. Valentín, S., et al., *Safety and efficacy of doxycycline in the treatment of rosacea.* Clin Cosmet Investig Dermatol, 2009. **2**: p. 129-140.
103. El Emam, K., et al., *A secure distributed logistic regression protocol for the detection of rare adverse drug events.* Journal of the American Medical Informatics Association, 2013. **20**(3): p. 453-461.
104. *Twitter perde milhões*. [acessado em agosto de 2016]; Disponível em: http://g1.globo.com/economia/noticia/2016/02/twitter-perde-us-521-milhoes-em-2015-e-nao-aumenta-n-de-usuarios.html.

8 Anexos

Anexo A - Arquivo getdrugs.rb

```
1   #!/usr/bin/env ruby
2   # encoding: ISO-8859-1
3   require 'rest_client'
4   require 'hpricot'
5
6   URLDRUGSCOM = 'http://www.drugs.com/condition/'
7
8   SEARCH = 'http://www.drugs.com/search.php?searchterm='
9
10  def searchDrugs(url)
11    begin
12          response = RestClient.get url
13          if (response.code == 200)
14                html = Hpricot(response.to_s)
15                html.search('//*[@id="content"]/div[1]/div').each       do
    |div|
16                      aux =  div.search('/h3/a')
17                      if (aux.inner_text.include?("edication"))
18                            link = aux.attr("href")
19                            return getDrugs(link)
20                      end
21                end
22          else
23                puts "O código de resposta da url "+url+" não foi 200"
24                return false
25          end
26    rescue Exception => e
27          return false
28    end
29  end
30
31  def getDrugs(url)
32    begin
33          response = RestClient.get url
34          if (response.code == 200)
35                html = Hpricot(response.to_s)
36                elements = html.search('//*[@id="conditionbox"]')
37                i = 0
38                elements.search("tr").each do |tr| #drugs
39                      if (i == 0)
40                            i = i+1
41                            next
42                      end
43                      if (tr.search("b")[0])
44                            puts tr.search("b")[0].inner_html
45                      end
46                end
47                if (url.include?("generic"))
48                      return true
49                else
50                      url.sub!(".html", "-generic.html")
51                      return getDrugs(url)
52                end
53          else
54                puts "O código de resposta da url "+url+" não foi 200"
55                return false
```

```
56          end
57    rescue Exception => e
58          return false
59    end
60  end
61
62
63  if (ARGV.length < 1)
64    puts "Você pode chamar esse programa da seguinte maneira:"
65    puts "ruby getdrugs.rb malaria"
66    exit
67  end
68
69  disease = ARGV[0]
70  sleep(1)
71
72  if (!getDrugs(URLDRUGSCOM+disease+"-prophylaxis.html"))
73    if (!getDrugs(URLDRUGSCOM+disease+".html"))
74          if (!searchDrugs(SEARCH+disease))
75                puts "Não foram encontrados medicamentos para #{disease}"
76          end
77    end
78  end
```

Anexo B - Arquivo medlineplus.rb

```
1   #!/usr/bin/env ruby
2   # encoding: ISO-8859-1
3   require 'rest_client'
4   require 'hpricot'
5
6   NDC = '2.16.840.1.113883.6.69'
7   RXCUI = '2.16.840.1.113883.6.88'
8
9   LANGUAGE = 'en'
10  RESPONSETYPE = 'text/xml'
11
12  def getAE(codeSystem, code, print)
13    url                                                                    =
    'http://apps.nlm.nih.gov/medlineplus/services/mpconnect_service.cfm?mai
    nSearchCriteria.v.cs='+codeSystem+'&mainSearchCriteria.v.c='+code+'&mai
    nSearchCriteria.v.dn=&informationRecipient.languageCode.c='+LANGUAGE+'&
    knowledgeResponseType='+RESPONSETYPE
14
15    response = RestClient.get url
16    if (response.code == 200)
17          xml = Hpricot(response.to_s)
18          link = xml.at("link")['href'].to_s
19          response = RestClient.get link
20          if (response.code == 200)
21                xml = Hpricot(response.to_s)
22                if (print)
23                      puts "!!!!!EFEITOS ADVERSOS!!!!!"
24                end
25                xml.search('*[@id="pbody"]/ul[2]/li').each     do     |li|
    #efeitos adversos
26                      if (print)
27                            puts li.at("p").inner_html
28                      end
29                end
30                if (print)
31                      puts ""
32                      puts "!!!!!EFEITOS ADVERSOS GRAVES!!!!!"
33                end
34                xml.search('*[@id="pbody"]/ul[3]/li').each     do     |li|
    #graves
35                      if (print)
36                            puts li.at("p").inner_html
37                      end
38                end
39          else
40                puts "O código de resposta do link "+link+" não foi 200"
41          end
42    else
43          puts "O código de resposta da url "+url+" não foi 200"
44    end
45  end
46
47  def NameToRxcui(name)
48    url                                                                    =
    'http://rxnav.nlm.nih.gov/REST/rxcui?name='+name+'&allsrc=1&search=2'
49    response = RestClient.get url
50
51    if (response.code == 200)
52          xml = Hpricot(response.to_s)
```

```
53          id = xml.at("rxnormid").inner_html
54          return id
55    end
56    puts "ERRO no NameToRxcui de: "+name
57  end
58
59  if (ARGV.length < 2)
60    puts "Você pode chamar esse programa da seguinte maneira:"
61    puts "ruby medlineplus.rb type=print code=ndc 00456140501"
62    puts "ruby medlineplus.rb type=ndcfile <arquivo>"
63    puts "ruby medlineplus.rb type=rxcuifile <arquivo>"
64    puts "ruby medlineplus.rb type=namesfile <arquivo>"
65    exit
66  end
67
68  type = ARGV[0].split("=")[1]
69  case type
70  when "print"
71    aux = ARGV[1].split("=")[1]
72    case aux
73    when "ndc"
74          codeSystem = NDC
75          code = ARGV[2]
76    when "rxcui"
77          codeSystem = RXCUI
78          code = ARGV[2]
79    when "name"
80          codeSystem = RXCUI
81          name = ARGV[2]
82          code = NameToRxcui(name)
83    else
84          puts "Parametro "+ARGV[1].to_s+" desconhecido"
85          exit
86    end
87
88  when "ndcfile"
89    codeSystem = NDC
90  when "rxcuifile"
91    codeSystem = RXCUI
92  when "namesfile"
93  else
94    puts "Parametro "+ARGV[0].to_s+" desconhecido"
95    exit
96  end
97
98  if type == "print"
99    getAE(codeSystem,code,true)
100 end
```

Anexo C - Arquivo rxnorm.rb

```
1   #!/usr/bin/env ruby
2   # encoding: ISO-8859-1
3   require 'rest_client'
4   require 'hpricot'
5
6   BASE = 'http://rxnav.nlm.nih.gov/REST'
7
8   if ARGV[0]== "help"
9     puts "Você pode chamar esse programa da seguinte maneira:"
10    puts "ruby rxnorm.rb help"
11    puts "ruby rxnorm.rb type=version"
12    puts "ruby rxnorm.rb type=name mefloquine"
13    puts "ruby rxnorm.rb type=spellingsuggestions mefloquine"
14    puts "ruby rxnorm.rb type=classes NDFRT PE"
15    exit
16  end
17
18  tipo = ARGV[0].split("=")[1]
19  case tipo
20  when "version"
21    url = BASE+'/version'
22  when "rxcui"
23    busca = ARGV[1]
24    url = BASE+'/rxcui?name='+busca+'&allsrc=1&search=2'
25  when "approximateTerm"
26    busca = ARGV[1]
27    url = BASE+'/approximateTerm?term='+busca
28  when "allrelated"
29    rxcui = ARGV[1]
30    url = BASE+'/rxcui/'+rxcui+'/allrelated'
31  when "drugs"
32    busca = ARGV[1]
33    url = BASE+'/drugs?name='+busca
34  when "spellingsuggestions"
35    busca = ARGV[1]
36    url = BASE+'/spellingsuggestions?name='+busca
37  when "classes"
38    src = ARGV[1]
39    type = ARGV[2] == nil ? "" : ARGV[2]
40    url = BASE+'/classes?src='+src+'&type='+type
41  when "members"
42    id = ARGV[1] #é o id dentra do src que vc esta procurando
43    src = ARGV[2] == nil ? "" : ARGV[2]
44    rela = ARGV[3] == nil ? "" : ARGV[3]
45    direct = ARGV[4] == nil ? "" : ARGV[4]
46    tty = ARGV[5] == nil ? "" : ARGV[5]
47    url = BASE+'/members?id='+id+'&src='+src
48    url = url + (ARGV[3] == nil ? "" :
    '&rela='+rela+'&direct='+direct+'&tty='+tty)
49  else
50    puts "Parametro "+ARGV[0].to_s+" desconhecido"
51    exit
52  end
53
54  response = RestClient.get url
55
56  if (response.code == 200)
57    xml = Hpricot(response.to_s)
58    case tipo
```

```
59    when "rxcui"
60          id = xml.at("rxnormid").inner_html
61          puts id
62    when "spellingsuggestions"
63          xml.search('suggestion').each do |aux|
64                puts aux.inner_html
65          end
66    when "classes"
67          i = 0
68          xml.search('nodename').each do |aux|
69                i=i+1
70          end
71          puts i
72    else
73          puts xml.to_s
74    end
75 else
76    puts "ERRO no GET da URL: "+url
77 end
```

Anexo D – Tabela com todos EA da doxiciclina no padrão de referência.

ADR_Term[a]	ADReCS_ID[b]	Frequência[c]
Abdominal discomfort	07.01.06.001	-
Abdominal distension	07.01.04.001	-
Abdominal pain	07.01.05.002	-
Abdominal pain upper	07.01.05.003	-
Abscess	11.01.08.001	-
Anaemia	01.03.02.001	-
Anaphylactic reaction	10.01.02.001; 24.06.03.006	-
Anaphylactic shock	10.01.02.002; 24.06.02.004	-
Angioedema	10.01.05.009; 23.04.01.001	-
Anorexia	08.01.09.025; 14.03.01.001	-
Anxiety	19.06.02.002	-
Aphthous stomatitis	07.05.06.001	-
Arthralgia	15.01.02.001	6.00%
Back pain	15.03.04.005	3.00%
Benign intracranial hypertension	17.07.02.001	-
Blood pressure increased	13.14.03.005	-
Blood urea increased	13.13.01.006	-
Bowel discomfort	07.01.06.006	infrequent
Bronchitis	11.01.09.001; 22.07.01.001	3.00%
Candidiasis	11.03.03.001	-
Cough	22.02.03.001	4.00%
Decreased appetite	08.01.09.028; 14.03.01.005	-
Dermatitis	23.03.04.002	-
Dermatitis exfoliative	10.01.01.004; 23.03.07.001	rare
Diarrhoea	07.02.01.001	infrequent
Discomfort	08.01.08.003	-
Dry mouth	07.06.01.002	-
Dysmenorrhoea	21.01.01.002	4.00%
Dyspepsia	07.01.02.001	6.00%
Dysphagia	07.01.06.003	-
Ear infection	04.03.01.006; 11.01.05.001	-
Emotional distress	19.04.02.008	-
Enterocolitis	07.08.03.003	-
Eosinophilia	01.02.04.001	-
Epidermal necrosis	23.03.03.035	-
Erythema multiforme	10.01.03.015; 23.03.01.003	-
Essential hypertension	24.08.02.008	-
Feeling abnormal	08.01.09.014	-
Fistula	15.03.02.001	-
Fontanelle bulging	17.02.05.006; 18.04.04.001	-
Fungal infection	11.03.05.001	-
Gastric disorder	07.11.01.003	-
Gastrointestinal pain	07.01.05.005	-
Gingival bleeding	07.09.07.001; 24.07.02.010	-
Gingival pain	07.09.04.001	< 1%
Gingivitis	07.09.03.003; 11.01.04.013	-
Glossitis	07.14.01.001	-
Haemolytic anaemia	01.06.03.002	-
Headache	17.14.01.001	26.00%
Henoch-Schonlein purpura	01.01.04.001; 10.02.02.004; 23.06.01.002; 24.07.06.003	-
Hepatotoxicity	09.01.07.009; 12.03.01.008	rare
Hypersensitivity	10.01.03.003	-
Hypertension	24.08.02.001	-
Ill-defined disorder	08.01.03.049	-
Infection	11.01.08.002	2.00%
Inflammation	08.01.05.007	-
Influenza	11.05.03.001; 22.07.02.001	11.00%
Injury	12.01.08.004	5.00%
Insomnia	17.15.03.002; 19.02.01.002	-
Intracranial pressure increased	17.07.02.002	-
Leukopenia	01.02.02.001	-
Malaise	08.01.01.003	-
Menorrhagia	21.01.03.002	-
Muscle spasms	15.05.03.004	-

ADR_Term[a]	ADReCS_ID[b]	Frequência[c]
Musculoskeletal discomfort	15.03.04.001	-
Musculoskeletal pain	15.03.04.007	-
Myalgia	15.05.02.001	1.00%
Nasal congestion	22.04.04.001	-
Nasopharyngitis	11.01.13.002; 22.07.03.002	22.00%
Nausea	07.01.07.001	8.00%
Neck pain	15.03.04.009	-
Necrotising colitis	07.08.01.013	-
Nephropathy toxic	12.03.01.010; 20.05.03.002	-
Neutropenia	01.02.03.004	-
Oedema	08.01.07.006; 14.05.06.010	-
Oesophageal disorder	07.11.02.001	-
Oesophageal ulcer	07.04.05.002	rare
Oesophagitis	07.08.05.001	rare
Oral pain	07.05.03.002	-
Oropharyngeal discomfort	07.05.05.008; 22.02.05.027	-
Oropharyngeal pain	07.05.05.004; 22.02.05.022	5.00%
Pain	08.01.08.004	4.00%
Pain in jaw	15.02.01.003	-
Pericarditis	02.06.02.001	-
Periodontal disease	07.09.05.006	-
Photosensitivity reaction	23.03.09.003	-
Pigmentation disorder	23.05.03.001	-
Postnasal drip	22.02.05.009	-
Premenstrual syndrome	19.04.02.009; 21.01.01.007	-
Proctitis	07.08.04.001	-
Pulpitis dental	07.09.12.001; 11.01.04.009	-
Purpura	01.01.04.003; 23.06.01.004; 24.07.06.005	-
Rash	23.03.13.001	4.00%
Rash erythematous	23.03.06.003	-
Rash maculo-papular	23.03.13.004	-
Rhinitis	11.01.13.004; 22.07.03.006	-
Rhinorrhoea	22.02.05.010	-
Sensitivity of teeth	07.09.06.003	-
Serum sickness	10.01.03.004; 12.02.08.004	-
Shoulder pain	15.03.04.011	-
Sinus congestion	22.04.06.001	5.00%
Sinus headache	17.14.01.002; 22.02.05.023	-
Sinus operation	25.05.05.001	4.00%
Sinusitis	11.01.13.005; 22.07.03.007	3.00%
Skin infection	11.01.12.003; 23.09.04.002	-
Sore mouth	07.05.03.006	-
Stevens-Johnson syndrome	10.01.03.020; 11.07.01.005; 12.03.01.014; 23.03.01.007	-
Stomatitis	07.05.06.005	-
Swelling	08.01.03.015	-
Systemic lupus erythematosus	10.04.03.004; 15.06.02.003; 23.03.02.006	-
Tenderness	08.01.08.005	-
Tension	19.06.02.005	-
Tension headache	17.14.01.004	-
Thrombocytopenia	01.08.01.002	-
Tooth abscess	07.09.01.003; 11.01.04.003	4.00%
Tooth disorder	07.09.05.001	6.00%
Tooth loss	07.09.09.001	-
Toothache	07.09.06.001	7.00%
Toxic epidermal necrolysis	10.01.01.006; 11.07.01.006; 12.03.01.015; 23.03.01.008	-
Ulcer	08.03.06.001	-
Upper-airway cough syndrome	22.02.05.030	-
Urticaria	10.01.06.001; 23.04.02.001	-
Vascular purpura	01.01.04.007; 23.06.01.008; 24.07.06.011	-
Vomiting	07.01.07.003	-
Vulvovaginal candidiasis	11.03.03.005; 21.14.02.003	-
Vulvovaginal mycotic infection	11.03.05.004; 21.14.02.004	-

[a]Eventos Adversos; [b]Identificadores que estão associados a esse EA no ADReCS; [c]Frequência com que o EA ocorre quanto pode ser obtida, caso contrário é mostrado o símbolo "-".

Anexo E - Arquivo rest_tweet.rb

```
1  #!/usr/bin/env ruby
2  # encoding: ISO-8859-1
3  require "rubygems"
4  require "twitter"
5
6  #arquivos de configuração
7  require_relative "config.rb"
8  require_relative "API-keys.rb"
9
10 #The Search API is not complete index of all Tweets, but instead an
   index of recent Tweets. At the moment that index includes between 6-9
   days of Tweets
11
12 if (ARGV[0] == nil)
13   puts "!!!!!Digite qual palavra você deseja procurar nos tweets!!!!!"
14   puts "Ex: ruby rest_tweet.rb malaria"
15   exit
16 else
17   busca = ARGV[0]
18 end
19
20 client = Twitter::REST::Client.new do |config|
21  config.consumer_key = $consumer_key
22  config.consumer_secret = $consumer_secret
23  config.access_token = $access_token
24  config.access_token_secret = $access_token_secret
25 end
26
27 hash = Hash.new
28 hash[:lang] = $lang
29 hash[:result_type] = $result_type
30
31 user = client.user($username)
32 resultado = client.search(busca, hash)
33 puts "Foram encontrados "+resultado.count.to_s+" tweets com "+busca
34 time1 = Time.now.strftime("%d-%m-%Y %H-%M-%S")
35 nomearquivo = "rest_"+time1+"_"+busca+".txt"
36 saida = File.new(nomearquivo, "w+")
37 resultado.each do |tweet|
38   saida.puts  tweet.text
39 end
40 saida.close
41 puts "Arquivo "+nomearquivo+" criado com sucesso"
```

Anexo F - Arquivo tweet_stream.rb

```ruby
#!/usr/bin/env ruby
# encoding: ISO-8859-1
require "rubygems"
require "tweetstream"
require "digest/sha1"
require "mail"
require "mongo"
require "date"
require 'rest_client'

#arquivos de configuração
require_relative "config.rb"
require_relative "API-keys.rb"
require_relative "utils.rb"

now = Time.now
$logger = Utils.createLogger("stream"+now.strftime("%Y_%m_%d")+".log")
$logger.info ("Iniciando stream")

$old_digest = $new_digest = ""
$count = 1 #para contar o número de tentativas após erros
$lastTimeError = now #hora que deu o último erro

$collections = Hash.new
$lasterror = nil

def lerPalavras()
  $collections.clear
  arquivo = File.new($nome_arquivo, "r")
  palavras = Array.new()
  while (linha = arquivo.gets)
        aux = linha.split(',')
        palavras.push(aux[0])
        $collections[aux[0]] = aux[1].strip
  end
  arquivo.close()
  return palavras
end

if (ARGV[0] == nil)
  puts "!!!!!Digite o arquivo com as palavras que você quer procurar no twitter!!!!!"
  puts "Ex: ruby tweet_stream.rb nome-arquivo"
  exit
else
  $nome_arquivo = ARGV[0]
end

TweetStream.configure do |config|
    config.consumer_key       = $consumer_key
    config.consumer_secret    = $consumer_secret
    config.oauth_token        = $access_token
    config.oauth_token_secret = $access_token_secret
    config.auth_method        = :oauth
  end

client = TweetStream::Client.new
begin
  mongoClient = Mongo::Connection.new # defaults to localhost:27017
```

```
59    $mongoDB = mongoClient[$mongo_database]
60    $logger.info ("Conectado ao mongodb")
61
62    hash = Hash.new
63    #hash[:filter_level] = "medium"
64    #hash[:language] = $lang
65    #hash[:locations] = $locations
66
67    while (true)
68          palavras = lerPalavras()
69          $old_digest = Digest::SHA1.hexdigest(File.read($nome_arquivo))
70          data = Time.now()
71
72          $busca = ""
73          queries = Hash.new
74          palavras.each {   |palavra|
75                query = $mongoDB[:queries].find_one( { :text => palavra }
    )
76                if (query == nil)
77                      query = $mongoDB[:queries].insert([8])
78                end
79                queries[palavra] = query
80
81                if ($busca == "")
82                      $busca = palavra
83                else
84                      $busca = $busca + "," + palavra
85                end
86          }
87          hash[:track] = $busca
88          sleep(1)
89          client.on_error do |message|
90                $logger.error("#{$lasterror}ccc")
91                $logger.error("on_error -- #{message}")
92          end.filter(hash) do |object|
93                if object.is_a?(Twitter::NullObject)
94                      $logger.info("Twitter::NullObject")
95                      next
96                end
97                nova_data = Time.now()
98                if ((nova_data - data) > 3600) #se passou mais de 1h
99                      new_digest                                          =
    Digest::SHA1.hexdigest(File.read($nome_arquivo))
100                     if ($old_digest != new_digest)
101                           $logger.info("O arquivo foi modificado")
102                           $old_digest = new_digest
103                           break
104                     else
105                           $logger.info("O arquivo NAO foi modificado")
106                     end
107                     data = nova_data
108               end
109               if object.is_a?(Twitter::Tweet)
110                     #coleção por semana epidemiologica
111                     epiWeek                                             =
    Utils.getEpiWeek(Date.parse(object.created_at.to_s))
112                     queries.keys.each{ |key|
113                           if (object.text.include?(key))
114                                 collection                              =
    $collections[key]+"_"+epiWeek
115                                 puts object.created_at.to_s + " ------ "
    + collection
```

```
116                                 querie                                  =
    $mongoDB[:queries].find_one({:text => key})
117                                 if
    (!querie["epiweeks"].include?(epiWeek))
118                                     $mongoDB[:queries].update(
119                                           {:text => key},
120                                           {:$push   =>   {:epiweeks   =>
    epiWeek}}
121                                     )
122                                 end
123                                 coll = $mongoDB[collection]
124                                 if  (coll.find(  {  :id  =>  object.id  }
    ).count == 0)
125                                     lang  =  if  object.lang.class  ==
    Twitter::NullObject then nil else "#{object.lang}" end
126                                     geo   =   if   object.geo.class   ==
    Twitter::NullObject              then              nil              else
    "#{object.geo.lat};#{object.geo.long}" end
127                                     place  =  if  object.place.class  ==
    Twitter::NullObject then nil else "#{object.place.full_name}" end
128                                     userid  =  if  object.user.id.class
    == Twitter::NullObject then nil else "#{object.user.id}" end
129                                     author                  =         if
    object.user.name.class    ==    Twitter::NullObject    then    nil    else
    "#{object.user.name}" end
130                                     $lasterror                          =
    "#{object.id.class};#{object.text.class};#{object.created_at.class}"
131                                     coll.insert(:id    =>    object.id,
    :text => object.text, :lang => lang, :geo => geo, :place => place,
                                    :created_at     =>     object.created_at,
    :user_id => userid, :author => author , :incomplete => false)
132                                 end
133                             end
134                     }
135                 end
136         end
137         next
138   end
139
140 rescue Twitter::Error::ClientError => exception
141   $logger.info("end of file reached -- recomeçando")
142   $logger.info(exception.message)
143   $logger.info(exception.inspect)
144 rescue Twitter::Error::TooManyRequests => exception
145   $logger.info("Tem    que    esperar    #{exception.rate_limit.reset_in}
    segundos")
146 rescue Twitter::Error => exception
147   $logger.info("Twitter::Error --------> #{exception}")
148 rescue Exception => e
149   $logger.error(e.message)
150   $logger.error(e.backtrace.inspect)
151   $count = $count + 1
152   if ($count < $maxtry)
153         agora = Time.now()
154         if ( (agora - $lastTimeError) > 300 ) #Se tiver passado mais de
    5 minutos desde o último erro, zero as tentativas
155               $count = 1
156         end
157         $logger.info("Irei  tentar  pela  "+$count.to_s+"  vez  daqui  a  1
    minuto")
158         $lastTimeError = agora
159         sleep(60)
```

```
160         retry
161   else
162         $logger.info("Já esgotou o número máximo de tentativas")
163 
164         smtp = Net::SMTP.new('smtp.gmail.com', 587 )
165         smtp.enable_starttls
166         smtp.start('gmail.com', $sender, $pass, :login) do |smtp|
167                     smtp.send_message $message, $sender, $receiver
168         end
169         raise e
170   end
171 end
172 
173 $logger.info("Fechando o arquivo")
174 $logger.close
```

Anexo G – EA dos relatórios da FDA no ano de 2014

Medicamento Doxycycline

EA[a]	NumRelatoriosFDA[b]
ABDOMINAL DISCOMFORT	21
ABDOMINAL DISTENSION	10
ABDOMINAL PAIN	77
ABDOMINAL PAIN LOWER	17
ABDOMINAL PAIN UPPER	32
ABNORMAL BEHAVIOUR	9
ABORTION INCOMPLETE	22
ACNE	12
ACUTE KIDNEY INJURY	11
ALOPECIA	18
ANAEMIA	33
ANAPHYLACTIC REACTION	12
ANHEDONIA	9
ANXIETY	86
APHAGIA	10
ARRHYTHMIA	16
ARTHRALGIA	48
ASTHENIA	53
ASTHMA	16
ATRIAL FIBRILLATION	11
BACK PAIN	29
BALANCE DISORDER	19
BLISTER	21
BLOOD ALKALINE PHOSPHATASE INCREASED	10
BLOOD CREATININE INCREASED	10
BLOOD GLUCOSE INCREASED	11
BLOOD PRESSURE DECREASED	9
BLOOD PRESSURE INCREASED	16
BRADYCARDIA	16
BRONCHITIS	33
BRONCHOSPASM	9
CELLULITIS	17
CEREBROVASCULAR ACCIDENT	10
CHEST DISCOMFORT	20
CHEST PAIN	46
CHILLS	28
CHLAMYDIAL INFECTION	16
CHOLELITHIASIS	10
CHRONIC OBSTRUCTIVE PULMONARY DISEASE	25
CLOSTRIDIUM DIFFICILE INFECTION	13
COAGULOPATHY	13
COMPLETED SUICIDE	31
CONDITION AGGRAVATED	66
CONFUSIONAL STATE	20
CONSTIPATION	29
CONTUSION	15
COUGH	48
CRYING	11
DEATH	42
DECREASED APPETITE	36
DEEP VEIN THROMBOSIS	24
DEHYDRATION	22
DEPRESSED MOOD	31
DEPRESSION	57
DERMATITIS	11
DEVICE DISLOCATION	18
DEVICE ISSUE	22
DIARRHOEA	96
DIPLOPIA	20
DISCOMFORT	17
DISEASE PROGRESSION	13
DISINHIBITION	10

EA[(a)]	NumRelatoriosFDA[(b)]
DIZZINESS	97
DRUG ADMINISTRATION ERROR	9
DRUG DISPENSING ERROR	9
DRUG DOSE OMISSION	17
DRUG ERUPTION	11
DRUG HYPERSENSITIVITY	189
DRUG INEFFECTIVE	115
DRUG INTERACTION	47
DRUG REACTION WITH EOSINOPHILIA AND SYSTEMIC SYMPTOMS	9
DRUG-INDUCED LIVER INJURY	10
DRY EYE	9
DRY MOUTH	14
DRY SKIN	13
DYSARTHRIA	18
DYSPEPSIA	11
DYSPHAGIA	17
DYSPNOEA	99
DYSPNOEA EXERTIONAL	9
DYSSTASIA	11
DYSURIA	9
ELECTROCARDIOGRAM QT PROLONGED	10
EMOTIONAL DISTRESS	47
ENCEPHALOPATHY	9
EPISTAXIS	18
ERYTHEMA	83
EYE IRRITATION	13
FACTITIOUS DISORDER	9
FALL	26
FATIGUE	100
FEAR	10
FEBRILE NEUTROPENIA	9
FEELING ABNORMAL	25
FEELING HOT	26
FLATULENCE	10
FLUSHING	34
FOETAL EXPOSURE DURING PREGNANCY	14
GAIT DISTURBANCE	26
GASTRIC ULCER	9
GASTROINTESTINAL HAEMORRHAGE	16
GENERAL PHYSICAL HEALTH DETERIORATION	26
GENITAL HAEMORRHAGE	13
HAEMATEMESIS	17
HAEMOGLOBIN DECREASED	15
HAEMORRHAGE	53
HALLUCINATION	12
HEADACHE	119
HEPATIC ENZYME INCREASED	10
HEPATIC FAILURE	9
HOT FLUSH	13
HYPERAESTHESIA	14
HYPERHIDROSIS	11
HYPERSENSITIVITY	29
HYPERTENSION	22
HYPOAESTHESIA	14
HYPOAESTHESIA ORAL	10
HYPOKALAEMIA	10
HYPONATRAEMIA	25
HYPOTENSION	33
HYPOXIA	13
INAPPROPRIATE SCHEDULE OF DRUG ADMINISTRATION	17
INCORRECT DOSE ADMINISTERED	12
INFECTION	19
INFLAMMATION	14
INFLUENZA LIKE ILLNESS	16
INJECTION SITE PAIN	25
INJURY	54
INSOMNIA	29

EA[(a)]	NumRelatoriosFDA[(b)]
JARISCH-HERXHEIMER REACTION	10
JAUNDICE	24
JOINT STIFFNESS	10
JOINT SWELLING	10
LETHARGY	16
LIP SWELLING	13
LIPASE INCREASED	12
LIVER FUNCTION TEST ABNORMAL	19
LOCAL SWELLING	9
LOSS OF CONSCIOUSNESS	12
LOWER RESPIRATORY TRACT INFECTION	23
LYME DISEASE	9
LYMPHADENOPATHY	9
MALAISE	91
MEDICAL DEVICE DISCOMFORT	11
MEMORY IMPAIRMENT	11
MENORRHAGIA	13
MENSTRUATION IRREGULAR	9
MENTAL STATUS CHANGES	11
MIGRAINE	13
MOOD SWINGS	11
MUSCLE ATROPHY	11
MUSCLE SPASMS	42
MUSCULAR WEAKNESS	24
MUSCULOSKELETAL PAIN	14
MYALGIA	32
MYOCARDIAL INFARCTION	9
NASOPHARYNGITIS	20
NAUSEA	200
NECK PAIN	10
NEURALGIA	10
NEUTROPENIA	16
NIGHT SWEATS	9
NO THERAPEUTIC RESPONSE	11
OEDEMA	12
OEDEMA PERIPHERAL	15
OESOPHAGEAL ULCER	18
OFF LABEL USE	50
OROPHARYNGEAL PAIN	23
OSTEOMYELITIS	9
PAIN	122
PAIN IN EXTREMITY	55
PAIN OF SKIN	14
PALLOR	21
PALPITATIONS	16
PANCREATITIS	14
PANCREATITIS ACUTE	25
PANIC ATTACK	12
PARAESTHESIA	38
PATHOGEN RESISTANCE	18
PELVIC PAIN	19
PERIPHERAL SWELLING	23
PHOTOSENSITIVITY REACTION	18
PLEURAL EFFUSION	11
PNEUMONIA	41
POLLAKIURIA	10
PREGNANCY	15
PRODUCT QUALITY ISSUE	14
PRODUCT USE ISSUE	12
PRODUCTIVE COUGH	10
PRURITUS	67
PULMONARY EMBOLISM	41
PYREXIA	68
RASH	90
RASH ERYTHEMATOUS	20
RASH GENERALISED	19
RASH MACULAR	19

EA[(a)]	NumRelatoriosFDA[(b)]
RASH PRURITIC	21
RENAL FAILURE	21
RENAL FAILURE ACUTE	16
RESPIRATORY DISORDER	13
RESPIRATORY FAILURE	17
RHEUMATOID ARTHRITIS	9
RHINORRHOEA	15
ROSACEA	9
SCAR	8
SEPSIS	15
SEPTIC SHOCK	10
SHOCK	9
SINUSITIS	18
SKIN BURNING SENSATION	28
SKIN DISCOLOURATION	14
SKIN EXFOLIATION	19
SKIN LESION	12
SKIN ULCER	9
SKIN WARM	8
SLOW RESPONSE TO STIMULI	11
SOCIAL AVOIDANT BEHAVIOUR	12
SOMNOLENCE	25
STAPHYLOCOCCAL INFECTION	15
SUICIDAL IDEATION	45
SWELLING	9
SWELLING FACE	22
SYNCOPE	32
TACHYCARDIA	26
TENDONITIS	15
THERAPEUTIC RESPONSE DECREASED	13
THROAT IRRITATION	8
THROMBOCYTOPENIA	12
THROMBOSIS	18
TINNITUS	18
TOXICITY TO VARIOUS AGENTS	9
TREATMENT FAILURE	23
TREMOR	27
UPPER GASTROINTESTINAL HAEMORRHAGE	11
URINARY INCONTINENCE	22
URINARY TRACT INFECTION	38
URTICARIA	47
UTERINE PERFORATION	31
VAGINAL HAEMORRHAGE	8
VENTRICULAR TACHYCARDIA	14
VERTIGO	9
VISION BLURRED	28
VISUAL IMPAIRMENT	16
VITH NERVE PARALYSIS	12
VOMITING	137
WEIGHT DECREASED	31
WEIGHT INCREASED	19
WHEEZING	15
WHITE BLOOD CELL COUNT DECREASED	12
WITHDRAWAL SYNDROME	8

[(a)]Nomes dos EA; [(b)]Quantidade de relatórios da FDA.

Anexo H - Arquivo getEAfromFDADrug.rb

```
2   # encoding: ISO-8859-1
3
4   require 'rubygems'
5   require 'hpricot'
6
7   if (ARGV.length < 1)
8       puts "Voc□ode chamar esse programa da seguinte maneira:"
9       puts "ruby getEAfromFDADrug.rb <arquivo.html>"
10      exit
11  end
12
13  file = File.open(ARGV[0],"rb")
14  html = Hpricot(file.read)
15  hashEA = Hash.new
16   html.search('*[@id="adverseevents"]/div/div/ul').each do |ul| #eventos
    adversos
17      aux = ul.at("li").inner_html.split('<span class="pull-right">')
18      ea = aux[0]
19      num = aux[1].split(" (")[0]
20      hashEA[ea] = num
21  end
22  hashEA.sort.map do |k,v|
23      puts "#{k}\t#{v}"
24  end
```

Anexo I - Arquivo analiseDadosFDA.rb

```
1       #!/usr/bin/env ruby
2       # encoding: ISO-8859-1
3
4       require 'rubygems'
5       require 'hpricot'
6
7       if (ARGV.length < 1)
8           puts "Voce pode chamar esse programa da seguinte maneira:"
9                                       puts    "ruby    analiseDadosFDA.rb
<droga1.html>,<droga1.htm2>,<droga3.html>,<droga4.html>,<droga5.html>"
10          puts "Sera feita a an□se somente para a droga1"
11          exit
12      end
13
14      arquivos = ARGV[0].split(",")
15      drugXea = Hash.new
16      eaGlobal = Hash.new(0)
17      drugGlobal = Hash.new(0)
18      totalEA = 0
19      first = true
20      firstDrug = nil
21      for arquivo in arquivos
22          file = File.open(arquivo,"rb")
23          drug = arquivo.split(".html")[0]
24          if (first)
25              firstDrug = drug
26              first = false
27          end
28          drugXea[drug] = Hash.new
29          html = Hpricot(file.read)
30            html.search('*[@id="adverseevents"]/div/div/ul').each do |ul|
#eventos adversos
31                  aux = ul.at("li").inner_html.split('<span class="pull-
right">')
32              ea = aux[0]
33              num = aux[1].split(" (")[0].to_i
34              drugXea[drug][ea] = num
35              drugGlobal[drug] = drugGlobal[drug].to_i + num
36              eaGlobal[ea] = eaGlobal[ea].to_i + num
37              totalEA = totalEA + num
38          end
39      end
40
41      drugXea[firstDrug].sort.map do |k,v|
42          a = v.to_i
43          n = drugGlobal[firstDrug].to_i
44          b = n - a
45          m = eaGlobal[k].to_i
46          c = m - a
47          t = totalEA.to_i
48          d = t-a-b-c
49          if (c == 0) #o PRR no foi calculado. Irei setar o PRR para 99.9
50              prr = 99.9
51              prrM = "Nao cal"
52              prrP = "Nao cal"
53          else
54              prr = (a*(t-n))/(c*n).to_f
55                    se  =  Math.sqrt((1/a.to_f)+(1/c.to_f)-(1/(a+b).to_f)-
(1/(c+d).to_f)) #standard error
```

```
56              aux = Math.exp(1.96*se).to_f
57              prrM = prr/aux
58              prrP = prr*aux
59          end
60          qui2 = (n*((a*d-b*c)**2))/((a+b)*(a+c)*(b+d)*(c+d)).to_f
61          if (prr == 99.9)
62                                                                          puts
    "#{k}\t#{prrM}\t#{prr.round(3)}\t#{prrP}\t#{qui2.round(3)}\t#{a}"
63          else
64                                                                          puts
    "#{k}\t#{prrM.round(3)}\t#{prr.round(3)}\t#{prrP.round(3)}\t#{qui2.roun
    d(3)}\t#{a}"
65          end
66      end
67      puts "\n"
68      drugGlobal.sort_by{|key,value| value}.map do |k,v|
69          puts "#{k} #{v}"
70      end
```

9. PUBLICAÇÕES

1 – Artigo aceito no congresso International Conference on knowledge Discovery and Information Retrieval – KDIR
DOI: 10.5220/0005135203540359

2 – Artigo aceito na revista de medicina da Faculdade de Ribeirão Preto (FMRP) - Volume 48 – Suplemento 3 – Novembro/2015 - II Workshop Ibero-Americano de Sistemas Interoperáveis em Saúde.
I.S.S.N. online: 2176-7262