UNIVERSIDADE DE SÃO PAULO

FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

DEPARTAMENTO DE MEDICINA SOCIAL

PROGRAMA DE PÓS-GRADUAÇÃO EM SAÚDE PÚBLICA

ESTUDO DE MODELO DE PREDIÇÃO DE ABANDONO AO TRATAMENTO DA TUBERCULOSE (TB)

VERENA HOKINO YAMAGUTI

ORIENTADOR: PROF. DR. ANTÔNIO RUFFINO-NETTO

CO-ORIENTADOR: PROF. DR. RUI PEDRO CHARTERS LOPES RIJO

Ribeirão Preto - SP

2023

UNIVERSIDADE DE SÃO PAULO

FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

DEPARTAMENTO DE MEDICINA SOCIAL

PROGRAMA DE PÓS-GRADUAÇÃO EM SAÚDE PÚBLICA

ESTUDO DE MODELO DE PREDIÇÃO DE ABANDONO AO TRATAMENTO DA TUBERCULOSE (TB)

VERENA HOKINO YAMAGUTI

Tese apresentada ao Programa de Pós-Graduação em Saúde Pública da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Doutora em Saúde Pública.

Orientador: Prof. Dr. Antônio Ruffino-Netto

Co-orientador: Prof. Dr. Rui Pedro Charters Lopes Rijo

Ribeirão Preto - SP

2023

Yamaguti, Verena Hokino

Estudo de modelo de predição de abandono ao tratamento da tuberculose (TB). Ribeirão Preto, 2023. 61 p.; 30cm.

Tese de Doutorado apresentada à Faculdade de Medicina de Ribeirão Preto/USP.

Orientador: Prof. Dr. Antônio Ruffino-Netto

1. Tuberculose. 2. Abandono. 3. Modelo de predição. 4. Seleção de atributos. 5. CART. 6. Agrupamento hierárquico.

UNIVERSIDADE DE SÃO PAULO

FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

DEPARTAMENTO DE MEDICINA SOCIAL

PROGRAMA DE PÓS-GRADUAÇÃO EM SAÚDE PÚBLICA

ESTUDO DE MODELO DE PREDIÇÃO DE ABANDONO AO TRATAMENTO DA TUBERCULOSE (TB)

VERENA HOKINO YAMAGUTI

Tese apresentada ao Programa de Pós-Graduação em Saúde Pública da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Doutora em Saúde Pública.

Aprovado em_____.

Membros da Banca:

_____

Membro do Programa

_____

Membro Externo ao Programa

_____

Membro Externo à FMRP

Ribeirão Preto - SP

2023

Ao meu padrasto, com amor, dedicação e muita saudade...

Ao meu marido por sempre acreditar e ser minha maior rede de apoio

E a todos amigos e familiares que sempre torcem por mim.

AGRADECIMENTO

Ao meu padrasto Horácio Moribe, por todos os ensinamentos, incentivos, exemplos e amor deixados, para que hoje eu me tornasse a pessoa que sou.

À minha irmã Jacqueline Hokino Yamaguti, por ser minha melhor amiga, por toda confiança depositada em mim, pela força e carinho que me fazem seguir em frente.

Ao meu marido Rômulo, por estar sempre ao meu lado, dando o maior apoio, amor e carinho. Só tenho a agradecer por toda sua paciência e por acreditar e apoiar todos os meus sonhos e por fazer dos meus sonhos os seus também.

Aos meus pais, e aos meus familiares, que sempre me apoiaram e torceram pelo meu sucesso.

Ao meu orientador Prof Dr. Antonio Ruffino-Netto, por aceitar a me orientar, por toda atenção e paciência em compartilhar seus conhecimentos e sanar minhas dúvidas. Tenho sorte em ter a oportunidade de aprender a cada dia com essa pessoa incrível e que é um grande exemplo de vida, de pessoa e de profissional.

Ao meu coorientador Prof Dr. Rui Rijo, por me orientar com o planejamento das atividades, o que fez com que eu evoluísse muito rápido durante o doutorado e por sempre ser muito atencioso e estar disposto a ajudar, com muita alegria e energia.

A todos meus amigos e aqueles que sempre me apoiam, sou muito grata por ter cada um de vocês na minha vida.

À Universidade de São Paulo por oferecer ensino gratuito e de qualidade, por me acolher desde a graduação e abrir várias portas.

Às agências de fomento à pesquisa que financiaram o presente trabalho: Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processos 2018/23963-2 e 2021/08341-8. “O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001”.

Ao Centro de Vigilância Epidemiológica da Secretaria Estadual De Saúde pela autorização do uso dos dados para o desenvolvimento do presente trabalho.

“Se não der para vencer pelo talento, triunfe pelo esforço.

Vai na raça, ninguém segura uma pessoa altamente determinada!

Se der certo, você vence. Se der errado, você aprende.

Dos dois jeitos você ganha algo!”

Caio Carneiro

SUMÁRIO

ABREVIATURAS E SIGLAS

BRICS – *Brazil, Russia, India, China and South Africa*

CART – *Classification And Regression Trees*

CEP – Comitê de Ética em Pesquisa

COVID-19 – Coronavírus

DOTS – Estratégia de Tratamento Diretamente Observado (*Directly Observed Treatment Short-course*)

HC – Hospital das Clínicas

HIV – Vírus da Imunodeficiência Humana (*Human Imunodeficiency Vírus*)

ICC – Índice de Comorbidade de Charlson

MS – Ministério da Saúde

OMS – Organização Mundial da Saúde

PEP – Prontuário Eletrônico do Paciente

PNCT – Plano Nacional de Controle da Tuberculose

RNA – Redes Neurais Artificiais

ROC - *Receiver Operating Characteristic*

SINAN – Sistema de Informação de Agravos de Notificação

SIS – Sistema de Informação de Saúde

SITETB – Sistema de Informação de Tratamentos Especiais de Tuberculose

SR – Sintomáticos Respiratórios

SUS – Sistema Único de Saúde

SVM – *Support Vector Machine*

TB – Tuberculose

TB-HIV – Coinfecção por *Mycobacterium tuberculosis* e HIV

TBWEB – Sistema de Notificação e Monitoramento de Casos de Tuberculose no Estado de São Paulo

TDO – Tratamento Diretamente Observado

VPN – Valor Preditivo Negativo

VPP – Valor Preditivo Positivo

RESUMO

YAMAGUTI, V. H. **Estudo de modelo de predição de abandono ao tratamento da tuberculose (TB)**. 2023. 61 p. Tese de Doutorado. Faculdade de Medicina de Ribeirão Preto – Universidade de São Paulo, Ribeirão Preto, SP, Brasil, 2023.

A tuberculose (TB) está entre as principais causas de mortes por doenças transmissíveis no mundo e até a pandemia do coronavírus (COVID-19), a TB era a principal causa de morte no mundo por doenças infectocontagiosas, estando acima do vírus da imunodeficiência humana (HIV). Para o controle efetivo da doença no Brasil, a abordagem mais significativa, vem sendo utilizada é a Estratégia do Tratamento Diretamente Observado (TDO). Contudo ainda são notificados números significativos de casos. A utilização de um sistema de informação tem como objetivo melhorar a qualidade do planejamento, implementação do programa de combate de TB, controle do tratamento e gerenciamento das informações. Por isso, é decisivo dotar tais sistemas com capacidades que ajudem os profissionais de saúde a gerirem os escassos recursos disponíveis e focar seus esforços de forma adequada a cada caso. Dessa forma, o objetivo geral desse projeto é desenvolver um modelo de predição de abandono ao TDO dos pacientes com TB. Isto possibilitaria a identificação de casos de abandono do tratamento com antecedência e redirecionamento de recursos para melhorar a adesão desses casos (para o tratamento), reduzindo o abandono e a taxa de infecção por bacilos resistentes. Inicialmente (Artigo 1), analisamos diferentes bases de dados comparando-as quanto a completude e confiabilidade das informações de 50 tratamentos em 2 ambulatórios no estado de São Paulo. As bases de dados utilizadas para coleta das informações foram: 1) SISTb; 2) Hygia; 3) Prontuários locais dos ambulatórios; 4) Prontuários eletrônicos do Hospital das Clínicas; e 5) TBWEB. A base de dados que apresentou os melhores resultados para completude e confiabilidade foi o TBWEB. Em seguida (Artigo 2), foi realizada uma seleção dos atributos mais relacionados ao desfecho do tratamento da TB na base de dados TBWEB. Posteriormente (Artigo 3 e 4), foram desenvolvidos dois modelos utilizando diferentes métodos para predição do desfecho do tratamento da TB. Em primeiro momento, foi utilizado o modelo CART (Artigo 3) e posteriormente foi feito um classificador utilizando como base os agrupamentos gerados por um modelo de classificação não supervisionada hierárquica (Artigo 4). Ambos os modelos desenvolvidos apresentaram uma performance similar ou melhor a outros trabalhos encontrados na literatura. Além disso, ambos os modelos se destacam por possuírem uma representação visual simples, possibilitando que equipes de saúde possam utilizá-los para identificar padrões durante o tratamento de TB.

Palavras-chave: tuberculose, abandono, modelo de predição, seleção de atributos, CART, agrupamento hierárquico.

ABSTRACT

YAMAGUTI, V. H. **Study of a prediction model for tuberculosis treatment loss to follow up**. 2023. 61 p. Doctoral thesis. Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, SP. Brazil, 2023.

Tuberculosis is among the leading causes of death from infectious diseases. Tuberculosis was the leading cause of deaths by infectious disease until the beginning of the coronavirus pandemics. For effective control of the disease in Brazil, the Directly Observed Treatment Strategy is used. However, there is a significant number of new cases reported. The use of information system aims to improve the quality of planning, program implementation, treatment control and information management. It is therefore crucial to provide such systems with capabilities that help health professionals to manage the resources available and focus their efforts on cases that require their attention. Thus, the general objective of this project is to carry out a study for a prediction model to the tuberculosis treatment abandonment. This would provide a way to predict treatment loss to follow-up and redirect resources in advance to improve the adherence of these cases, reducing the treatment loss to follow-up rate and the number of infections by resistant bacilli. Initially (Article 1), we analyzed different databases comparing them for the completeness and reliability of the data from 50 treatments in 2 ambulatories in the state of São Paulo. The databases used to collect the information were: 1) SISTb; 2) Hygia; 3) Local ambulatory medical records; 4) Clinical Hospital electronic medical records; and 5) TBWEB. The database which presented the best completeness and reliability scores was TBWEB. Subsequently (Article 2), the selection of the most related attributes to the tuberculosis treatment outcome was performed. Later (Article 3 and 4), we developed two different models for tuberculosis prediction. First (Article 3), we developed a classifier using CART model. Later (Article 4), we developed a prediction classifier from the clusters generated by a hierarchical clustering model. Both developed models showed similar performance when compared to other models found in the literature. Besides, our models are highlighted by generating human comprehensible models that can be used by healthcare professionals through tuberculosis treatment to identify undesired outcomes patterns.

Keywords: tuberculosis, treatment loss to follow-up, prediction model, selection of attributes, CART, hierarchical clustering.

1. INTRODUÇÃO

A tuberculose (TB) é uma das principais causas de mortes por doenças transmissíveis no mundo. Até a pandemia do coronavírus (COVID-19), TB era a principal causa de morte no mundo por doenças infectocontagiosas, estando acima do vírus da imunodeficiência humana (HIV) (WHO, 2022). Segundo a Organização Mundial da Saúde (OMS), é estimado que aproximadamente um quarto da população mundial foi infectada pelo bacilo. A pandemia do COVID-19 continua a ter um grande impacto negativo no controle da doença, isto é, no tratamento da TB. Os progressos feitos até o ano de 2019 diminuíram, estagnaram ou foram perdidos, e os principais objetivos globais para a TB não estão sendo alcançados.

O progresso global para o combate à doença depende principalmente de avanços na prevenção e no cuidado ao paciente em todos os países que apresentam uma alta carga de TB, entre eles o Brasil (MINISTÉRIO DA SAÚDE, 2019) que ocupa a 20ª posição quanto à carga da doença (TB) e a 19ª no que se refere à taxa de co-infecção de tuberculose-HIV (TB-HIV) no mundo. O Brasil reporta 0,9% dos casos mundiais de TB, 33% dos casos das Américas e em conjunto com os demais países que compõem o BRICS (*Brazil, Russia, India, China and South Africa*) cerca de 50% dos casos de TB no mundo e utilizando mais de 90% dos recursos necessários de fontes domésticas de financiamento para ações de controle da doença (MINISTÉRIO DA SAÚDE, 2019).

Reduções no número de diagnósticos notificados de TB em 2020 e 2021 sugerem que o número de pessoas com TB não diagnosticada ou tratada aumentou. Dessa forma, ocorreu-se um aumento do número de mortes e de infecções. No mundo, a taxa de mortalidade por TB aumentou entre os anos de 2019 e 2021, revertendo anos de queda que ocorreram entre os anos de 2005 e 2019. Em 2021, estima-se que ocorreram 10,6 milhões de casos, 1,4 milhões de óbitos no mundo entre pessoas HIV-negativas e 187.000 óbitos entre pessoas soropositivas (WHO, 2022) em decorrência da TB.

A TB é hoje considerada uma doença prioritária, estabelecendo-se uma linha de ação através do Plano Nacional de Controle da Tuberculose (PNCT). A equipe do

PNCT está em constante ampliação e qualificação, adotando cada um dos elementos da estratégia do *Stop TB*, a começar pelo seu primeiro elemento que trata da implementação do Tratamento Diretamente Observado (TDO).

A TB é uma doença curável em praticamente 100% dos casos novos, sensíveis aos medicamentos anti-TB, desde que obedecidos os princípios básicos da terapia medicamentosa e a adequada operacionalização do tratamento que deve ser feito diariamente por um longo período de tempo, de (no mínimo) 6 meses (e em alguns casos, até 2 anos) sem interrupções. Portanto, a adesão do paciente ao tratamento apresenta um grande impacto sobre o sucesso do mesmo.

O tratamento requer o uso diário de quatro drogas diferentes. No entanto, os efeitos colaterais dessas drogas e do desaparecimento dos sintomas após o início do tratamento, estão entre as principais causas do abandono e consequentemente, o aumento da transmissão da doença e a resistência do bacilo à medicação (MINISTÉRIO DA SAÚDE, 2019).

Em 1996, o Ministério da Saúde (MS) lançou o Plano Emergencial, recomendando a implantação da Estratégia de Tratamento Diretamente Observado (DOTS) para o controle da TB no país, sendo formalmente oficializado em 1999 por intermédio do PNCT.

“Em outubro de 1998, lançou-se o Plano Nacional de Controle da Tuberculose (MINISTÉRIO DA SAÚDE, 1996), com as seguintes diretrizes gerais: 1) o MS é responsável pelo estabelecimento das normas; 2) a aquisição e abastecimento de medicamentos; 3) referência laboratorial e de tratamento; 4) coordenação do sistema de informações; 5) apoio aos Estados e Municípios; 6) articulação intersetorial visando maximizar os resultados de políticas públicas. Reconhece que a condição essencial é a articulação e a complementaridade de ações dos três níveis de gestão do SUS (União, Estados e Municípios); envolver obrigatoriamente a participação social e organizações não governamentais. Detecção e diagnóstico feitos fundamentalmente através da baciloscopia em todos os sintomáticos respiratórios (SR) e contatos. Disponibilizar tuberculostáticos, incluindo um estoque estratégico, assegurar tratamento supervisionado e vigilância da resistência das drogas. Prover um sistema de informações de acordo com recomendações da OMS. O Plano introduz duas inovações: o tratamento supervisionado e a instituição de um bônus de R$ 150,00 e de R$ 100,00 para cada caso de doente de tuberculose diagnosticado,

tratado e curado, respectivamente, se foi utilizado ou não o tratamento supervisionado. O repasse desse bônus é feito automaticamente por ocasião da notificação da alta curado do paciente. Fica suprimida toda e qualquer burocracia de assinaturas de convênio para esses repasses. Observa-se, então, uma substituição da administração burocrática por uma administração gerencial" (RUFFINO-NETTO, 2002).

O DOTS é uma das principais contribuições para que o país atinja a meta de cura de 85% dos doentes e a diminuição da taxa de abandono para 5%, evitando o surgimento de bacilos resistentes e possibilitando um efetivo controle da TB no país (MINISTÉRIO DA SAÚDE, 2017).

Esta estratégia é composta por cinco componentes: 1) Diagnóstico por qualidade garantida por baciloscopia de escarro; 2) TDO e monitorado quanto sua evolução; 3) Fornecimento regular e ininterrupto de medicamentos; 4) Sistema de registro padronizado e relatórios que assegure a avaliação do tratamento; 5) Compromisso político e financeiro sustentado para controlar a TB colocando como prioridade entre as políticas públicas (OPAS; OMS, 1997).

A associação medicamentosa adequada, as doses corretas e o uso por tempo suficiente são os princípios básicos para o tratamento adequado, evitando a persistência bacteriana e o desenvolvimento de resistência aos fármacos, assegurando, assim, a cura do paciente. A esses princípios soma-se o TDO como estratégia fundamental para assegurar a cura do doente. O tratamento é desenvolvido sob regime ambulatorial e é diretamente observado a menos que haja indicação em sentido contrário.

O TDO consiste na observação da ingestão dos medicamentos, preferencialmente todos os dias da semana durante a fase inicial e no mínimo três vezes por semana na fase de manutenção do tratamento, administrado por profissionais de saúde ou eventualmente por outra pessoa, desde que devidamente capacitada e sob monitoramento do enfermeiro. Nos finais de semana e feriados os medicamentos são auto administrados.

Verifica-se, empiricamente, que numa primeira fase os doentes apresentam adesão ao tratamento, especialmente nos primeiros meses, durante os quais apresentam os sintomas da doença e enfrentam forte desconforto físico. Neste período, os doentes realizam a medicação e apresentam-se nas consultas de

acompanhamento. Uma vez ultrapassada esta fase (de mal-estar físico e os sintomas tornam-se menos acentuados, dando uma percepção visível de melhora, embora, infelizmente não estejam curados) a tendência é de abandonarem o tratamento.

Para o controle efetivo do tratamento da TB e outras atividades que envolvem a doença, é necessário a utilização de sistemas de informação que visem o gerenciamento dos casos, a fim de facilitar as atividades dos profissionais de saúde e gestores, além de garantir a qualidade dos dados, evitando informações incompletas e incorretas, de modo a ajudar no planejamento e tomada de decisões que possam melhorar o amparo ao paciente, bem como prevenir a doença.

Existe uma estreita relação entre os diferentes sistemas de informação e o sucesso das operações em hospitais (GONÇALVES et al., 2007; GE et al., 2012; LOURENÇO et al., 2014; MARTINHO; RIJO; NUNES, 2015). Diferentes sistemas de informação foram desenvolvidos visando auxiliar no acompanhamento dos pacientes e no gerenciamento de seus dados de saúde, facilitando as atividades de diversos profissionais de saúde. No entanto, notamos disparidades nas informações fornecidas pelos diferentes sistemas. Essas disparidades podem impactar negativamente nos processos de planejamento e tomada de decisão em saúde que afetam diretamente o apoio dado aos pacientes e a prevenção de doenças (CREPALDI et al., 2014).

A OMS define o Sistema de Informação de Saúde (SIS) como um mecanismo para coleta, processamento, análise e transmissão de informações necessárias para planejar, organizar, operar e avaliar os serviços de saúde (LIPPEVELD; SAUERBORN; BODART, 2000). Considera-se que a transformação de dados em informação requer, além da análise, disseminação e inclusão de recomendações de ação. Um SIS deve ser usado para gerenciar as informações que os profissionais de saúde precisam para realizar atividades com eficácia e eficiência, facilitando a comunicação, integrando informações, coordenando ações entre vários membros da equipe profissional de atendimento, e disponibilizando recursos para apoio financeiro e administrativo (SHORTLIFFE; PERREAULT, 2001).

Existem várias iniciativas de SIS que ajudam na gestão do tratamento da TB. No Brasil, existem diversas ferramentas que permitem o registro dos dados dos pacientes com TB, como o Sistema de Informação de Agravos de Notificação (SINAN), no qual são investigados os casos de doenças incluídas na lista nacional

de doenças de notificação compulsória; o Sistema Especial de Informação sobre o Tratamento da Tuberculose (SITETB), no qual é possível notificar, acompanhar e monitorar os resultados dos casos especiais de TB; e o Sistema de Notificação e Monitoramento de Casos de Tuberculose no Estado de São Paulo (TBWEB). Esses sistemas estão focados na notificação de doenças e no acompanhamento do tratamento por meio de uma compilação de informações. A equipe de saúde também utiliza outros sistemas ao longo do tratamento, como sistemas para solicitação e recebimento de resultados de exames, registros médicos eletrônicos e registros médicos em papel.

A utilização de sistemas de informação tem como objetivo melhorar a qualidade do planejamento, implementação do programa, e portanto, melhorar o controle do tratamento e gerenciamento das informações relacionados aos pacientes de TB, contribuindo com gestores e profissionais de saúde na assistência contínua e efetivo manejo clínico da doença em todos os estágios para garantir a adesão ao tratamento.

Neste contexto é fundamental (e de bom senso) dotar tais sistemas de informação com capacidades que ajudem os profissionais de saúde a gerirem os escassos recursos materiais e humanos disponíveis para a demanda existente, de modo a concentrarem mais atenção e esforços nos doentes que apresentam maior probabilidade de abandonarem o tratamento após os 3 meses iniciais.

CHRISTENSEN e SMITH (1995) realizaram um estudo mostrando a relação entre diferentes aspectos pessoais dos pacientes com sua adesão ao tratamento. Acreditamos que com o uso de um modelo semelhante seríamos capazes de classificar um doente entre os grupos, possibilitando o foco de esforços de ação nos doentes que possuem maior tendência a abandonar o tratamento, em detrimento dos doentes que, por sua iniciativa, iriam continuar o mesmo. Conseguindo-se prever o comportamento dos doentes, seria possível direcionar o esforço das equipes no trabalho de campo e concentrar os esforços naqueles doentes mais propensos a abandonar o tratamento, aumentando-se assim a adesão ao tratamento e, portanto, aumentando o percentual de cura, diminuindo o percentual de abandonos, recidivas e a propagação da doença.

2. OBJETIVOS

2.1 Objetivo Geral

O objetivo geral deste projeto é desenvolver um modelo de predição de abandono do tratamento da TB, possa ser implementado utilizando dados extraídos a partir de sistemas de informação de saúde para gestão de TB. Dessa forma, desenvolvendo um modelo de fácil compreensão humana que possa ser interpretado por profissionais de saúde.

2.2 Objetivos Específicos

Os objetivos específicos deste projeto podem ser sumarizados por:

- Fazer uma revisão dos modelos e técnicas existentes para predição de abandono e sua acurácia (Seção 3);
- Analisar a qualidade de dados em várias bases de dados de pacientes de TB para ser utilizada no estudo (Artigo 1);
- Avaliar a importância de diversas variáveis e do Índice de Comorbidade de Charlson para predição do desfecho do tratamento da TB (Artigo 2);
- Propor um modelo/estratégia de atenção para prever o abandono de doentes de TB (Artigo 3 e Artigo 4).

3. TRABALHOS CORRELATOS

3.1 Delineamento de estudos correlatos

Nesta seção são analisados diferentes estudos correlatos encontrados na literatura vigente. Estes têm como principal objetivo o levantamento de fatores preditivos da TB e a revisão dos modelos e técnicas existentes para predição de abandono e sua acurácia. Diversos dos trabalhos apresentados na literatura vigente são dados como estudos descritivos que realizam uma análise das principais variáveis que se relacionam com o tratamento da TB.

Muitas das variáveis analisadas pelos autores remetem as características sociais que os pacientes portadores da TB possuem (GALDÓS TANGÜIS et al., 2000; DO PRADO et al., 2011; KALHORI; ZENG, 2013; SILVA; ANJOS; NOGUEIRA, 2014; HARLING et al., 2017). Além disso, devido aos diferentes níveis socioeconômicos apresentados por bairros em um mesmo município, alguns autores analisam a região de residência e/ou a região de tratamento dos doentes como parte de fatores preditivos para o abandono do tratamento (GALDÓS TANGÜIS et al., 2000; KALHORI; ZENG, 2013; HARLING et al., 2017).

Outras variáveis normalmente analisadas relacionam as comorbidades como HIV, diabetes, tabagismo e idade, e variáveis relacionadas ao histórico do tratamento de TB do paciente.

SILVA; ANJOS; NOGUEIRA (2014) desenvolveram um estudo descritivo, de abordagem quantitativa. Sua base de dados conta com registros de casos de TB notificados no banco de dados do SINAN. Os dados são referentes ao município de João Pessoa do estado de Paraíba, Brasil; no período compreendido entre janeiro de 2001 e dezembro de 2008. O estudo utilizou inicialmente um total de 5.164 observações, com as variáveis: sexo, idade, cor, escolaridade, tipo de entrada, tipo de tratamento e situação de encerramento.

HARLING et al. (2017) utilizam dados provenientes de tratamentos de TB da cidade de Fortaleza no estado do Ceará, Brasil; entre os anos 2007 à 2014. Os dados utilizados foram extraídos do banco de dados do SINAN. O estudo também se destaca por realizar um mapeamento geográfico dos casos para com a região dos centros de notificação. Dessa forma, realizando uma análise geográfica da correlação das diferentes variáveis e do abandono do tratamento da TB. O estudo realizou uma análise em 117 bairros que notificaram um total de 12,352 casos de TB que ocorreram num intervalo de 9 anos.

As variáveis utilizadas por HARLING et al. (2017) buscam descrever tantas características clínicas como sociais da doença. Como variáveis temos: ano de notificação do caso; idade; sexo; etnia; escolaridade; estado da gravidez; histórico de HIV; estado da diabetes; alcoolismo; existência outras condições que agravam a TB; se a infecção da TB ocorreu em local de trabalho; local de infecção; resultados de exames médicos; resultados da cultura de escarro positivo; utilização do TDO e o seu histórico ao longo do tratamento; localização da casa do paciente; e local do centro de tratamento utilizado pelo paciente.

Os dados utilizados nas análises de KALHORI; ZENG (2013) foram coletados por médicos, enfermeiros e outros profissionais de saúde em diferentes centros de tratamento de TB no Irã no transcorrer do ano de 2005. O trabalho realizou a análise de 6.450 resultados de tratamentos de TB. KALHORI; ZENG (2013) reportam inicialmente mais de 35 variáveis. Posteriormente, selecionando somente algumas variáveis independentes que se mostraram correlacionadas com o desfecho do tratamento. A correlação entre as variáveis é medida por meio de uma correlação bivariada ($p < 0.05$). As variáveis selecionadas por esse método foram: sexo; idade; peso; nacionalidade; bairro de residência; participação no sistema prisional; tipo do caso; tipo de tratamento; classificação da TB; reincidência de TB; diabetes; HIV; duração do tratamento; baixo peso corporal; toxicodependência; atividade sexual de risco.

GALDÓS TANGÜIS et al. (2000) estudaram fatores preditivos para o abandono do tratamento de TB entre pacientes infectados pelo HIV em Barcelona, Espanha entre os anos de 1987-1996. Os dados utilizados no estudo foram todos retirados do Sistema de Vigilância Epidemiológica do Programa de Controle e Prevenção da Tuberculose de Barcelona. Dentre as variáveis utilizadas por GALDÓS TANGÜIS et

al. (2000) temos: sexo; faixa etária do paciente (dividida em grupos de idade de 15-29, 30-39, 40+); bairro de residência; alcoolismo; grupo de transmissão de HIV; o fato do paciente ser sem-teto; histórico prisional; histórico de tratamento de TB; e ano do tratamento.

ARROYO et. al (2019) utilizam dados extraídos do TBWEB entre os anos 2006 e 2015 para desenvolver um modelo de predição para os desfechos do tratamento de casos diagnosticados de TB multidroga resistente. Um total de 16 atributos foram utilizados que descrevem características sociodemográficas do paciente e clínico-operacional do tratamento.

APUNIKE et. Al (2020) também fazem o uso de dados extraídos do TBWEB. Contudo, o fazem com o intuito de correlacionar uma série de eventos clínicos cronológicos a um desfecho do tratamento. Um total de 16 atributos que descrevem o histórico diagnóstico e a evolução do esquema medicamentoso ao longo do tratamento do paciente são utilizados. Dados demográficos não são utilizados para o desenvolvimento do modelo de aprendizado não supervisionado.

3.2 Métodos e modelos preditivos de abandono

Nesta seção são analisados os métodos aplicados nos estudos descritos anteriormente, discutindo seus principais pontos. Foram considerados tantos métodos para descrever fatores preditivos do abandono de TB, como métodos que definem modelos preditivos utilizando técnicas estatísticas e computacionais de aprendizado de máquina.

GALDÓS TANGÜIS et al. (2000) realizaram em seu estudo análises em níveis bivariados e multivariados. A análise em nível bivariado é feita utilizando um teste Qui-quadrado para variáveis qualitativas. As associações entre as variáveis são medidas pelo *odds ratio* com intervalos de confiança. Em nível multivariado, somente variáveis que mostraram associação estatística significativa ($p < 0.05$) foram analisadas. Estas foram analisadas por meio de um modelo de regressão logística ajustado pelo teste de Hosmer-Lemeshow.

Em nível bivariado GALDÓS TANGÜIS et al. (2000) observaram que existem maiores probabilidades de abandono entre pessoas residentes de bairros de baixo

nível socioeconômico, alcoólatras, usuários de drogas intravenosas, sem-tetos, com um histórico prisional ou de tratamento de TB. Em nível multivariado, viver em bairros de baixo nível socioeconômico, ser um sem-teto e ter um histórico de tratamento de TB são fatores de risco para o abandono do tratamento.

HARLING et al. (2017) também realizaram uma análise descritiva estatística de seus dados identificando os principais fatores preditivos do abandono do tratamento de TB. Contudo, HARLING et al. (2017) estendem sua análise em nível geográfico de maneira muito mais refinada, realizando o mapeamento dos casos de TB e potenciais fatores preditivos através dos bairros. Agrupamentos com altas ou baixas taxas de TB são identificados por meio da análise Local Moran's I. Posteriormente estes são ajustados para comparação usando o método de Benjamini-Hochberg.

Modelos espaciais de auto regressão são utilizados para identificar fatores associados às taxas de infecção de TB. Modelos bivariados são usados para analisar a correlação entre cada possível preditor e a taxa de incidência, e modelos multivariados são construídos a partir de todos os fatores que mostram uma correlação significativa em seu modelo bivariado. Por fim, HARLING et al. (2017) conduzem uma análise hierárquica usando um modelo de regressão logística mista de dois níveis para identificar possíveis fatores preditivos para o abandono do tratamento da TB.

HARLING et al. (2017) mostram uma correlação espacial significante da incidência de TB, identificando alguns agrupamentos significativos. Em seus modelos bivariados HARLING et al. (2017) identificam que o número de casos de TB era maior estatisticamente em bairros com famílias mais numerosas, baixo salário, analfabetismo, eletricidade de difícil acesso, maiores taxas de homicídio e uma maior parcela de sua população de indivíduos negros e pardos.

Quanto ao abandono do tratamento, HARLING et al. (2017) indicam que as fatores que aumentam de maneira significante a chance de abandono são: idade entre 20-29 anos; sexo masculino; notificação tardia; etnia negra ou amarela; HIV positivo; alcoolismo; e outras condições agravantes e comorbidades (e.g.: tabagismo, doenças cardiovasculares, câncer, uso de drogas, etc.). Por outro lado, fatores que caracterizam a redução da chance de abandono são: ensino secundário ou superior completo, gravidez, diabetes e ter TB extrapulmonar. Muitos desses fatores também

são notificados por GALDÓS TANGÜIS et al. (2000) e reforçam sua importância para determinar o nível de adesão do paciente ao tratamento.

SILVA; ANJOS; NOGUEIRA (2014) utilizam um modelo de regressão logística binária que posteriormente é ajustado utilizando o método *Backward*, e pelo uso da matriz de confusão e o gráfico *Receiver Operating Characteristic* (ROC). Inicialmente, SILVA; ANJOS; NOGUEIRA (2014) ajustaram o modelo para todas as variáveis disponíveis, posteriormente, removeram as variáveis não significativas uma a uma, até que se obtivesse o modelo que melhor representasse a situação do encerramento do tratamento. Em seu modelo final é possível observar que entre as variáveis que aumentam significativamente a chance de abandono do tratamento temos: ensino médio incompleto, etnia não-caucasiana e reingresso de tratamento da TB.

O modelo, quando ajustado pelo critério de Youden, apresenta uma área sobre a curva ROC com o valor de 0,722, um Valor Preditivo Positivo (VPP) de 0,4694 e um Valor Preditivo Negativo (VPN) de 0,8914. Apesar de apresentar um desempenho moderado, o modelo gera uma grande quantidade de falsos positivos, o que gera um baixo valor de VPP. No trabalho desenvolvido por SILVA; ANJOS; NOGUEIRA (2014) é possível evidenciar que isso ocorre em decorrência de dois fatores: 1) o não pareamento 1-1 dos casos de abandono e cura estão gerando um viés; 2) o conjunto de dados não segue uma distribuição linear, sendo dessa forma, modelos como o de regressão logística incapazes de descrever corretamente os dados.

KALHORI; ZENG (2013) comparam os resultados de seis técnicas de aprendizado de máquinas para predição do abandono do tratamento de TB. As técnicas utilizadas são: árvores de decisão, redes neurais artificiais (RNA), regressão logística, funções de base radiais, redes Bayesianas, e *Support Vector Machine* (SVM). Cada um dos modelos é construído utilizando uma segmentação da base inicial que conta com 4515 casos. KALHORI; ZENG (2013) também realizam uma seleção de atributos a partir de um teste de significância Kolmogorov-Smirnov. Posteriormente, esses modelos são testados contra uma base de 1935 casos. As métricas utilizadas para comparar os diferentes modelos foram acurácia, *F-measure*, e sensibilidade. KALHORI; ZENG (2013) apontam que a técnica que produziu o melhor modelo foi a árvore de decisão C4.5.

ARROYO et. al (2019) desenvolve seu modelo preditivo baseado em um modelo logístico que resulta em um nomograma que pode ser utilizado como uma

representação visual do modelo. Este modelo foi então avaliado com relação a sua acurácia e ROC.

APUNIKE et. Al (2020) produz um modelo de agrupamento não hierárquico, que apesar de um modelo de aprendizado não supervisionado, pode ser estendido para produzir um classificador preditivo. Cada grupo produzido pelo modelo é relacionado a um desfecho de tratamento e isso pode ser utilizado para predizer qual será o desfecho final de um tratamento ainda em desenvolvimento.

3.3 Considerações finais

Neste capítulo, foram apresentados diferentes trabalhos que objetivavam a definição de um modelo para a predição do abandono de TB. Como foi observado muitos dos autores se utilizam de bases de informações utilizadas para o monitoramento de TB em nível nacional (KALHORI; ZENG, 2013; SILVA; ANJOS; NOGUEIRA, 2014; HARLING et al., 2017) ou regional (GALDÓS TANGÜIS et al., 2000). A partir dessas bases os estudos se desenvolvem para detecção de variáveis que representem possíveis fatores de predição. Esse processo é realizado quase que exclusivamente pelos autores pelo uso de modelos bivariados e multivariados.

Muitos dos fatores preditivos encontrados para o abandono do tratamento se correlacionam a situação socioeconômica do paciente (GALDÓS TANGÜIS et al., 2000; KALHORI; ZENG, 2013; SILVA; ANJOS; NOGUEIRA, 2014; HARLING et al., 2017; ARROYO et al., 2019). Pacientes que possuem um nível socioeconômico menor tendem a abandonar mais facilmente o tratamento. Variáveis como educação, salário, analfabetismo e etnia são citadas repetidamente por diferentes autores. Outros fatores preditivos encontrados pelos autores normalmente se referem ao histórico de tratamento do paciente. Pacientes que possuem um histórico de abandono do tratamento possuem chances maiores de voltarem a abandonar o tratamento.

Posteriormente, como foi observado, alguns autores ainda buscam o desenvolvimento de modelos que permitam a classificação de novos casos a partir do conjunto de fatores preditivos encontrados. Há uma grande diversidade de técnicas utilizadas para criação de modelos preditivos para o abandono do tratamento

de TB. Essas vão desde a simples regressão logística até as complexas RNA. Além disso, podemos observar técnicas de aprendizado não supervisionado que podem ser utilizadas para gerar classificadores.

4. PUBLICAÇÕES

Durante o período de doutorado direto foram desenvolvidos 4 artigos em sequência. O Artigo 1 consistiu-se na comparação entre as informações de diferentes sistemas (SISTb, Hygia, prontuário eletrônico do paciente (PEP) do Hospital das Clínicas (HC) e prontuário dos ambulatórios locais) com objetivo de verificar a qualidade dos dados. Posteriormente, a mesma comparação foi realizada com dados oriundos do TBWEB, o qual, foi então definido com base de dados de referência e utilizado nos demais estudos, dado a qualidade e completude dos seus dados.

No Artigo 2, efetuamos um estudo caso-controle para identificar variáveis que caracterizassem os desfechos de óbito e cura do tratamento da TB, verificando a correlação da TB com diferentes comorbidades por meio de um estudo caso-controle.

Na sequência, um estudo caso-controle foi desenvolvido para criar o modelo preditivo ao longo do Artigo 3, que utilizou como base os resultados produzidos no Artigo 1 e Artigo 2 para definir a base de dados e atributos que seriam utilizados. No Artigo 4, analisamos o desempenho de um modelo preditivo criado a partir do relacionamento de eventos organizados de forma cronológica existentes na base selecionada pelo Artigo 1 para melhor avaliar se seria possível melhorar a acurácia do modelo desenvolvido no Artigo 3.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 121 (2017) 897–903

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2017, 8-10 November 2017, Barcelona, Spain

Data quality in tuberculosis: the case study of two ambulatories in the state of São Paulo, Brazil

Verena Hokino Yamaguti[a*] , Fernanda Bergamini Vicentine[a], Inacia Bezerra de Lima[c], Laís Zago[c], Lídia Maria Lourençon Rodrigues[a], Domingos Alves[a], Nathalia Yukie Crepaldi[a], Rui Pedro Charters Lopes Rijo[b], Antonio Ruffino-Netto[a]

[a]*Departament of Social Medicine, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil*
[b]*School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal*
[c]*School of Nursing of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brasil*

Abstract

Tuberculosis is the second leading responsible cause of death from infectious diseases. Tuberculosis effective control and related activities depends on the use of different systems that aim to assist and monitor patients through managing their health data, and facilitating the activities of several health professionals. However, we noticed disparities in the information provided by some systems, which can negatively impact planning and decision-making. The study is defined as a quantitative and descriptive study of patients' data during the treatment of tuberculosis **in** two different ambulatories. The data was collected from four different sources, including three information systems and the local patient archive in the ambulatories. Collected data was cleansed and standardized for semantic and structure, which allowed data comparison and analysis for its reliability and completeness. Low reliability scores are due to the absence of a semantic standard and the careless validation on recording of data by the professionals. Therefore, this study was able to effectively detect inconsistencies between the different data sources, stressing the need of health standards for data consistency, interoperability, and promoting data quality.

Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.

Keywords: Tuberculosis; Data quality assessment; Data reliability; Electronic health records;

* Corresponding author. Tel.: +55-16-98807-7010; fax: +55-16-3602-1526.
E-mail address: verena.yamaguti@usp.br

1877-0509
Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.
10.1016/j.procs.2017.11.116

1. Introduction

Tuberculosis is the second leading responsible cause of deaths in Brazil from infectious diseases, falling behind only the human immunodeficiency virus[1]. According to the World Health Organization (WHO), one third of the world population is infected by the bacillus and can develop the disease. In 2014, there were 9.6 million cases and 1.5 million deaths around the world[2]. In the same year, Brazil notified 74 thousand new cases of tuberculosis remaining as one of the 22 leading responsible for 80% of the notified cases around the world. In 2013, it was registered rates for cure, abandonment and death, respectively at 72%, 10% and 8%[2].

Tuberculosis treatment must be done daily for a long period of time, from 6 months to 2 years without any interruptions. Therefore, the success of this treatment highly depends on the patient adherence to it. The treatment requires the daily usage of four different drugs. However, the side effects of these drugs and symptoms disappearance after the treatment begins are between the main causes for the treatment abandonment, and, consequently, the increase disease transmission, and the bacillus resistance to the medication[3].

In 1996, the Brazilian Ministry of Health started an emergency plan, recommending the implementation of the Directly Observed Treatment - Short course (DOTS) strategy to control the tuberculosis in the country. This strategy was later officialized by the National Tuberculosis Control Program and continues to be one of the top priorities for the country reaching a 85% cure rate, and reducing the abandonment rate to less than 5%, allowing the effective control of the tuberculosis in the country[3].

DOTS consists of five components, including political commitment, sputum smear microscopy, directly observed treatment and monitoring the treatment evolution, distribution of medication, and information systems that allows to detect cases, the treatment outcomes and the overall performance of the National Tuberculosis Control Program. This approach forces the monitoring of every tuberculosis patient treatment, in which the medication is done daily at home or in an ambulatory.

There is a close relationship between the different information systems and the success of the operations in hospitals[4,5,6,7]. Different information systems were developed aiming assist in monitoring the patients and managing their health data, facilitating the activities of several health professionals. However, we noticed disparities in the information provided by the different systems. Theses disparities can negatively impact in the planning and decision making processes in healthcare that directly affect the support given to patients and disease prevention[8,9].

We collected data from a group of patients from different sources and described these data for later analysing these disparities. Additionally, we used a set of metrics to describe completeness and reliability of these data between the different sources better describing the disparities between them.

2. Material and Methods

The current study is defined as a quantitative descriptive study of patient data during the treatment of tuberculosis in two different ambulatories in the state of São Paulo, Brazil. Since patient healthcare data was part of the material analyzed, the study meet all ethical needs described by resolution 466/12 of the National Health Council[10]. The Clinical Hospital Research Facility and the Ethics Committees in Research of the School of Medicine of Ribeirão Preto approved the project.

2.1. Field Study

Two ambulatories were used in this study. Both provide the treatment for Tuberculosis patients. Ambulatory 1, supports an estimated population of 176,612 individuals, and is associated with the Municipal Health Secretariat and the University of São Paulo, whereas, the Ambulatory 2 supports an estimated population of 106,650 individuals. Additionally, the Ambulatory 1 had a significant number of medicine students and residents, and nursery students. Besides, the ambulatories were used as source of information by many researches associated with the University of São Paulo. We stress that when data was collected a ratio of 48-52% between the patients studied from each ambulatory was used so the final results would not be affected by any bias existing in one the ambulatories population.

2.2. Data collection

Four coordinators collected the tuberculosis patients' data from May/2015 to Jan/2016. Data was collected from 26 patients from Ambulatory 1 and 24 patients for Ambulatory 2. The amount of patients involved relates to the small period of 1 year of the study. The variables collected were: 1) gender; 2) ethnicity; 3) scholar degree; 4) profession; 5) previous treatment; 6) clinical form. Patients selected to this study were enrolled in the tuberculosis treatment, lived in the municipality of Ribeirão Preto and were at least 18 years old. Additionally, all patients who belong to the prison system were excluded from the study. The information would then serve to analyze the data, comparing the reliability and completeness of data in different information systems for the control of tuberculosis.

Data collection process relies on factors that range from the person who is collecting these data to different vulnerability situations that we may encounter in the field and in the tools that are being used to collect the information[11]. Data collected for the selected patients in each ambulatory was from four different sources: 1) SISTb - an information system designed to monitor tuberculosis patients treatment; 2) Local ambulatory medical records; 3) Hygia - an electronic medical record used by the main ambulatories in Ribeirão Preto; 4) HC - Clinical hospital electronic medical record. Variables collected included: gender, ethnicity, scholar degree, profession, previous treatment and tuberculosis clinical form. As we progressed in the data collection, some main limitations were noticed: 1) unavailability of proper infrastructure for the use of the electronic devices in the ambulatories e.g., no Internet access or electric power sources; 2) bureaucratic protocols that needed to be followed to access the patient medical records; 3) local medical records organization system caused problems to find the required data; 4) disparity in the data presented in the different sources would normally require an double check to verify if the data was correct according to its source.

2.3. Data cleaning and standardization

We performed a cleaning and standardization process in the collected data. The cleaning process consisted in the manual verification of data for any erroneous data. Cleaning the data allowed us to identify inconsistent data. Theses inconsistencies were mainly caused by user entry errors, such as the incorrect filling of the patient record in one of the information systems, or errors caused by fatigue of the involved field coordinators. The standardization process was mainly manual due to the lack of any patterns adopted by the information systems in the registry of variables, namely, among others, the clinical form and profession. Other variables were automatically parsed according to the system internal pattern. Moreover, the data standardization performed acted reducing a semantic gap found between the data sources, which would otherwise impair any data further data comparison between the data sources.

2.4. Completeness and reliability scores

After the collected data cleaning and standardization we verified the variables completeness and reliability according to each data source. The completeness (CS) and reliability (RS) scores are described, respectively, by equations (1) and (2).

$$CS = 1 - \frac{n_i}{n} \tag{1}$$

In equation (1), the completeness score (CS) is described as the ratio between the number of instances with no valid information and the total number of instances for the evaluated variable. Instances were considered with no valid information if the data was either not present or found.

$$RS = \frac{\sum_{j \in N} \max_{i \in V} (p_i)}{n} \tag{2}$$

$$p_i = 1 - \frac{v_i}{v} \tag{3}$$

In equation (2), the reliability score (*RS*) is described as the average ratio of all patients' maximum probabilities (p_i) for the set of possible values (V) of the evaluated variable (e.g. the gender variable have 2 possible values, male or female), and the number of data sources (n) that presented valid values for this patient and variable, where N is the whole of patients population and j is a single patient and n is total number of patients in the population. The probability p_i is described as the ratio of the number of occurrences for the value (v_i) and the number of sources of information that presented that value (v).

3. Results

Ambulatory 1 collected data is described on Table 1. We can notice differences in the variables values distribution for almost every variable. Education degree has the most significant discrepancies between the data sources, due to the significant amount of patients with no education in the SISTb data sources. Though other variables had shown differences in their distribution, we can observe that mainly cause to the differences are due to the number of cases that presented no information for the given variable.

Table 1. Collected data for the systems SISTb, Hygia, HC and the local medical records for Ambulatory 1.

		Ambulatory 1							
Variables		**SISTb Reported Cases**		**Local Medical Records Reported Cases**		**Hygia Reported Cases**		**HC Reported Cases**	
		N = 26	%	N = 26	%	N = 26	%	N = 26	%
Gender	Male	15	57.7	14	53.8	16	61.5	11	42.3
	Female	10	38.5	10	38.5	10	38.5	9	34.6
	No information	1	3.8	2	7.7	0	0	6	23.1
Ethnicity	White	15	57.7	7	26.9	2	7.7	12	46.2
	White/Black	4	15.4	4	15.4	2	7.7	5	19.2
	Black	1	3.8	4	15.4	1	3.8	3	11.5
	Yellow	0	0	0	0	0	0	0	0
	No information	6	23.1	11	42.3	21	80.8	6	23.1
Scholar Degree	Never studied	18	69.3	0	0	0	0	1	3.8
	Incomplete fundamental	5	19.3	9	34.7	0	0	4	15.4
	Complete fundamental	1	3.8	2	7.7	0	0	10	38.5
	Incomplete high school	0	0	1	3.8	0	0	1	3.8
	Complete high school	0	0	0	0	0	0	2	7.7
	Complete higher education	0	0	0	0	0	0	0	0
	Incomplete higher education	1	3.8	1	3.8	0	0	1	3.8
	No information	1	3.8	13	50	26	100	7	27
Profession	Retired	1	3.8	1	3.9	0	0	2	7.7
	Unemployed	2	7.7	3	11.5	0	0	2	7.7
	Housewife	3	11.6	3	11.5	0	0	6	23.1
	Student	0	0	0	0	0	0	2	7.7
	Other	11	42.3	16	61.6	3	11.5	7	26.9
	No information	9	34.6	3	11.5	23	88.5	7	26.9
Previous treatment	Yes	1	3.8	4	15.4	2	7.7	1	3.8
	No	19	73.1	17	65.4	0	0	0	0
	No information	6	23.1	5	19.2	24	92.3	25	96.2

Clinical form	Pulmonary	15	57.8	15	57.7	2	7.7	4	15.4
	Extra pulmonary	5	19.2	6	23.1	1	3.8	4	15.4
	Combined	3	11.5	2	7.7	0	0	2	7.7
	Disseminated	0	0	1	3.8	0	0	0	0
	No information	3	11.5	2	7.7	23	88.5	16	61.5

Ambulatory 2 collected data is described on Table 2. In contrast to Ambulatory 1, we can observe more similarities for the data provided by the ethnicity, scholar degree and profession of patients. However, information for gender and clinical form still has presented disparities due to the number of patient without valid information.

Next, we analyzed the data according to their scores in reliability and completeness previously described. Table 3 shows the completeness scores for each variable in Ambulatory 1. Gender presented the best score for completeness with an average of 91% of completeness over the four data sources, whereas the worst score for completeness was presented by previous treatment, having a average score of 42% over the four data sources. Hygia has the worst completeness scores presenting the worst scores for every variable aside from gender, followed by HC.

Table 2. Collected data for the systems SISTb, Hygia, HC and the local medical records for Ambulatory 2.

Variables		**Ambulatory 2**							
		SISTb Reported Cases		**Local Medical Records Reported Cases**		**Hygia Reported Cases**		**HC Reported Cases**	
		N = 24	%	N = 24	%	N = 24	%	N = 24	%
Gender	Male	9	37.5	14	58.4	18	75	5	20.9
	Female	4	16.7	5	20.8	6	25	2	8.3
	No information	11	45.8	5	20.8	0	0	17	70.8
Ethnicity	White	2	8.3	2	8.3	3	12.5	3	12.5
	White/Black	2	8.3	3	12.5	3	12.5	2	8.3
	Black	3	12.5	2	8.3	3	12.5	1	4.2
	Yellow	0	0	0	0	0	0	0	0
	No information	17	70.9	17	70.9	15	62.5	18	75
Scholar Degree	Never studied	0	0	0	0	0	0	0	0
	Incomplete fundamental	3	12.5	2	8.3	1	4.2	0	0
	Complete fundamental	0	0	0	0	0	0	4	16.7
	Incomplete high school	0	0	0	0	0	0	0	0
	Complete high school	3	12.5	2	8.3	2	8.3	0	0
	Complete higher education	0	0	0	0	0	0	0	0
	Incomplete higher education	0	0	0	0	0	0	1	4.2
	No information	18	75	20	83.4	21	87.5	19	79.1
Profession	Retired	0	0	0	0	0	0	0	0
	Unemployed	1	4.2	1	4.2	1	4.2	0	0
	Housewife	0	0	0	0	0	0	1	4.2
	Student	1	4.2	0	0	0	0	0	0
	Other	8	33.3	10	41.6	8	33.3	3	12.5
	No information	14	58.3	13	54.2	15	62.5	20	83.3
Previous treatment	Yes	0	0	1	4.2	2	8.3	1	4.2
	No	4	16.7	6	25	13	54.2	0	0
	No information	20	83.3	17	70.8	9	37.5	23	95.8
Clinical form	Pulmonary	11	45.8	17	70.8	21	87.4	1	4.2
	Extra pulmonary	1	4.2	1	4.2	1	4.2	0	0
	Combined	1	4.2	1	4.2	1	4.2	0	0
	Disseminated	0	0	0	0	0	0	1	4.2
	No information	11	45.8	5	20.8	1	4.2	22	91.6

Table 4 shows the completeness scores for Ambulatory 2. Clinical form presents the best score for completeness, having a 59% average score over the data. On the other hand, the scholar degree presented the worst completeness score with an average score of 19% over the data sources. Moreover, we can noticed Hygia and HC again with the worst scores for completeness.

Table 3. Completeness for collected data of Ambulatory 1.

	Ambulatory 1					
	Gender	Ethnicity	Scholar Degree	Profession	Previous treatment	Clinical form
SISTb	0.96	0.77	0.96	0.65	0.77	0.92
Local Medical Records	0.92	0.58	0.5	0.88	0.81	0.92
Hygia	1	0.19	0	0.12	0.08	0.12
HC	0.77	0.77	0.73	0.73	0.04	0.38
Mean	0.91	0.58	0.55	0.60	0.42	0.59

Table 4. Completeness for collected data of Ambulatory 2.

	Ambulatory 2					
	Gender	Ethnicity	Scholar Degree	Profession	Previous treatment	Clinical form
SISTb	0.54	0.29	0.25	0.42	0.17	0.54
Local Medical Records	0.79	0.29	0.17	0.46	0.29	0.79
Hygia	0	0.38	0.13	0.38	0.63	0.96
HC	0.29	0.25	0.21	0.17	0.04	0.08
Mean	0.41	0.30	0.19	0.35	0.28	0.59

Table 5 shows the average reliability scores for Ambulatory 1 and Ambulatory 2 between all the data sources and each variable present in the collected data. We stress that the reliability score as described before, does not account instances where a variable has shown no information for a given data source.

Table 5. Reliability for collected data of Ambulatory 1 and Ambulatory 2.

	Reliability					
	Gender	Ethnicity	Scholar Degree	Profession	Previous treatment	Clinical form
Ambulatory 1	0.91	0.62	0.57	0.52	0.60	0.59
Ambulatory 2	0.69	0.78	0.83	0.80	0.76	0.68

4. Discussion

Ambulatory 1 has shown discrepancies in the number of patients for SISTb according to the scholar degree. Nevertheless, in a later analysis of the system it was noticed that these discrepancies are mainly caused due to the storage of this data as an integer representing the number of years that the patient studied. Since its default value was zero, making hard to differentiate cases of incompleteness from cases that the patient had never studied. This characteristic can generate a bias on any research that would rely on these data to develop further analysis on the patient's educational profile. However, the low scores of completeness for Ambulatory 2 are due to the absence of any patient data entry in the system, caused by the late adhesion to the system.

Moreover, the analysis of Hygia revealed a lack in information for both ambulatories, which was later related to the usage of the system as complementary to the information source to the data already present in the local medical records. Hygia users in the ambulatories reported to use the system almost exclusively to schedule appointments, whereas, the local medical records were used to write down other information.

Finally, HC system has presented low scores for completeness due to many patients that were never registered in the hospital information system. Only in exceptional situations, patients attend the Hospital HC and there isn't interoperability between the systems, therefore, most data related to tuberculosis, such as clinical form and previous treatments were incomplete.

5. Conclusion

The current study pointed out inconsistencies for the reliability and completeness of the collected data in two different ambulatories. Among the reasons for the low scores presented in reliability, the leading causes to

inconsistencies between the data sources were the absence of a semantic and structural pattern for the registry and retrieval of data, which made unfeasible further data comparison between the data sources, even though considering the later standardization of the collected data. Furthermore, the lack of care during validation, and registry, of data done by the healthcare professionals in each system was also a problem, since most of them consider too laborious and not need to register the same information in all the systems. In order to avoid these bias, it is needed to raise awareness of healthcare professionals for the importance of this information and its completeness in all systems. Also, we need to provide these professionals with proper tools so they can register the valid information with the proper semantic. Besides, the lack of interoperability between the systems is a problem that causes unnecessary duplicates that rarely are updated according to new input provided in other systems. Therefore, the analysis and decision making using theses information as source becomes difficult and slow.

Usage of an interoperability pattern in healthcare, such as Health Level Seven (HL7) standard, is well known and recommended for these (and other) reasons[12]. HL7 standard define how information should be packaged and communicated from one party to another, setting how language, structure and data types should behave, providing the sufficient resources for integrating different systems. Therefore, healthcare management systems as those seem in this study need to adopt a pattern so they exchange and interoperate their data among themselves. However, the standardization of semantic is not always as trivial and is normally related to usage of one or more terminologies systems and a terminology service[13].

We can conclude from this study the existence of several inconsistencies between different data sources used in the control of tuberculosis in the Brazilian healthcare system. Besides, on this study it is noticeable the important role that health patterns for interoperability and terminologies systems plays on the maintenance and evaluation of data quality. Without a proper data quality any effort put on decision-making processes will be fruitless.

References

1. Ministério da Saúde (2011). Manual de recomendações para o controle da tuberculose no Brasil. Brasil.
2. WHO (2016). Global tuberculosis report 2016. Switzerland.
3. Ministério da Saúde (2017). Plano nacional pelo fim da tuberculose. Brasil.
4. Martinho, R., Rijo, R., & Nunes, A. (2015). Complexity Analysis of a Business Process Automation: Case Study on a Healthcare Organization. Procedia Computer Science, 64, 1226-1231.
5. Ge, X., Rijo, R., Paige, R., Kelly, T., & McDermid, J. (2012). Introducing Goal Structuring Notation to Explain Decisions in Clinical Practice. Procedia Technology, 5, 686-695.
6. Gonçalves, D., Rijo, R., Gonçalves, R., Cruz, J. B., & Varajão, J. (2007). Novos desafios e oportunidades de investigação na área da gestão de projectos de desenvolvimento de sistemas de informação. In Conferência Ibero-Americana WWW/Internet 2007.
7. Lourenço, J., Santos-Pereira, C., Rijo, R., & Cruz-Correia, R. (2014). Service Level Agreement of Information and Communication Technologies in Portuguese Hospitals. Procedia Technology, 16, 1397-1402.
8. Mazieiro, B., Crepaldi, N., Souza, E., Alves, D., Sanches, T., Rijo, R., Lima, I., Rodrigues, L., Bergamini, F., Bollela, V. (2016). Computational tool for organization and management of meetings between health professionals at tuberculosis care. Abstracts.
9. Crepaldi, N., Orfao, N., Yoshiura, V., Villa, T., Netto, A., Alves, D (2014). Desenvolvimento e implantação de um sistema para gestão de pacientes de tuberculose. Medicina. Ribeirão Preto. 47, 13-17.
10. Gauthier, J. Cabral, I., Santos, I., Tavares, C. (1998). Pesquisa em enfermagem: novas metodologias aplicadas. Rio de Janeiro: Guanabara Koogan. 177-208.
11. Bellato, R., Pereira, W., Gaíva, M. (1999). Algumas reflexões sobre o trabalho de campo na pesquisa qualitativa em enfermagem. Revista Gaúcha de Enfermagem, Porto Alegre. 20(2):6-16.
12. Walker, J., Pan, E., Johnston, D., Adler-Milstein, J. (2005). The value of health care information exchange and interoperability. Health affair. W5:10-24.
13. Ingenerf, J., Reiner, J., Seik, B. (2001). Standardized terminological services enabling semantic interoperability between distributed and heterogeneous systems. International journal of medical informatics. 64(2): 223-240.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 138 (2018) 258–263

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2018

Charlson Comorbidities Index importance evaluation as a predictor to tuberculosis treatments outcome in the state of São Paulo, Brazil

Verena Hokino Yamaguti[a,*], Rui Pedro Charters Lopes Rijo[b], Nathalia Yukie Crepaldi[a], Antonio Ruffino-Netto[a], Isabelle Carvalho[c], Domingos Alves[a]

[a]*School of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil*
[b]*School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal*
[c]*Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, Brazil*

Abstract

Tuberculosis is among the world leading causes of death, where over 0.9% of world tuberculosis cases are reported from Brazil. The World Health Organization have suggested the application of the Directly Observed Treatment Short-course, which have proved to reduced the number of deaths. However, over 4,500 deaths due to tuberculosis have been reported in Brazil during 2016. New methods that assist in the management and detection of tuberculosis could still be applied to further re-duce the number of deaths. This study evaluates the use of Charlson Comorbidity Index as predictor to detect deaths caused due to tuberculosis during the patient treatment. We compared the importance Charlson Comorbidity Index as predictor to different attributes that are collected through the patient treatment, and noticed that the Charlson Comorbidity Index have presented one of the highest scores of importance as predictor to detect cases of death due to tuberculosis. Therefore, al-lowing public health care system to focus on patients that have higher chances of death.

Keywords: Tuberculosis; Attribute selection; Bagging; Charlson comorbidities index;

* Corresponding author. Tel.: +55-16-98807-7010; fax: +55-16-3602-1526.
E-mail address: verena.yamaguti@usp.br

10.1016/j.procs.2018.10.037

1. Introduction

Tuberculosis (TB) is among the world leading healthcare problems, the disease affects millions of people worldwide and is one of leading death causes. Brazil reports 0.9% of world tuberculosis cases, 33% of Americas cases and along with Russia, India, China and South Africa (BRICS) sum approximately 50% of the world tuberculosis cases, using over 90% of the required resources from domestic financing sources to control the disease[1].

The World Health Organization (WHO) propose the Directly Observed Treatment Short-course[2] (DOTS) to control the tuberculosis in Brazil. This strategy is composed of five components: 1) Diagnosis by quality ensured sputum-smear microscopy; 2) Standardized short-course anti-TB treatment given under direct and supportive observation (DOT); 3) Regular, uninterrupted supply of high quality anti-TB drugs; 4) Standardized recording and reporting; 5) Sustained political and financial commitment to control the tuberculosis[3].

The use of DOTS in Brazil has made meaningful improvements ranging different epidemiological indicators related to tuberculosis. In 2005 were detect 26.5% less cases than the estimated and the healing rate reached to 69.3% for all tuberculosis types and 71.3% for the pulmonary TB. Furthermore, the incidence rate dropped by 5.7% from 1999 to 2005, the abandonment rates are dropping and the supply of anti-TB drugs is being done even for multi resistant TB (MRTB)[4]. The report published in 2016 by Brazilian Ministry of Health shows that in the past 10 years the number of TB cases have been reduced by 20%. However, there are over 70,000 new TB cases and 4,500 deaths due to TB every year in Brazil. Therefore, Brazil must implement new actions to control TB and reach the WHO's goal to reduce the number of deaths and infections caused by TB that happened in 2015 by 35% and 20%, respectively, until 2020.

Comorbidities are coexisting medical conditions that differ from the primary medical condition under study and can affect the patient treatment through detection, therapy and result[5]. They can share with morbidity the same risk factor, and be cause or consequence for a given disease. Moreover, they affect prognosis and the follow up for a randomized clinical trial[6]. Therefore, they can act as a predictor to death, prognosis and allow the use of statistical methods to raise hypotheses[6] about the primary medical condition.

The Charlson Comorbidities Index[7] (CCI) is an indicator that uses comorbidities for estimating the risk of death and to classify the severity of a patient clinical condition. It can be used in any database that has data for the used comorbidities[8]. Therefore, the CCI can be used as a predictor for death and a score for patient morbidity, assisting in the detection of patients with higher chances of death and in the control of TB. The CCI uses 19 comorbidities categories and their weights to provide a single score for each patient, each comorbidity weight is adjusted accordingly to the relative chance of death for one year. Devo et al.[9] mapped the CID-9 CM diagnosis codes and procedures for each of the CCI categories, consolidating "Leukemia" and "Lymphomas" in the category "Any Malignancy", resulting in 17 categories. Later, Romano et al.[10] showed that the use of CID-9 CM codes may vary and it depends on the data used and the its availability. Therefore, in some situations, the differences may not affect the final results provided by the CCI.

Seeing that, this study evaluates the use of CCI as predictor for death in the context of TB.

Section 2 we explain the materials and methods used for this study, specifying the methods to analyze the use of CCI as predictor. Section 3 shows the quantitative results obtained through the proposed methods. Section 4 discuss and analyze the results that evaluates our main hypotheses. Lastly, Section 5 shows our finals considerations and future goals on this matter.

2. Materials and Methods

2.1. Data

The initial dataset had 208.620 tuberculosis cases from the state of São Paulo. Data was collected through the TBWEB system between 2006 and 2016. TBWEB is the official system for processing TB data of the State Health Secretariat with the approval of the Health Ministry. TBWEB goal is epidemiological vigilance and monitoring of TB cases in the state of São Paulo. The system was built online, enabling TB notification register through the whole state, and in real time, where new cases can be submitted and their data can retrieved through the internet during the whole TB treatment course. The system provides a single entry for each patient and a history of past treatments attached to this entry. Besides acting as centralized database for TB cases, the TBWEB provides tools for information

management as: patient data reports, data cohort analysis, treatment monitoring and a better communication among the several epidemiological monitoring levels and their care facilities[11].

2.2. Model of study

This study is defined as case-control type study where we define CASE as patient that have died through the treatment due to TB, and CONTROL as those who have successfully healed from TB after the treatment. The initial dataset was collected from TBWEB, where we kept a 1-1 ratio for CASE and CONTROL to avoid any bias during the attribute selection and analysis. This reduced the initial dataset to a total 10,562 entries, where we had 5,281 CONTROL entries and 5,281 CASE entries. In addition, each CASE and CONTROL were described throughout 63 attributes. All patients in the prison system or with death causes not related to TB were excluded from this study.

Afterwards, we pre-processed all nominal attributes of our dataset to remove any attribute with a high correlation value, and attributes with near zero variance, therefore, eliminating attributes that would not provide any information gain to predict the treatment outcome. This reduced the number of attributes to 43. The correlation between nominal attributes was tested through Pearson Chi-Square Test[12]. The p-value for each test was computed through the Chi-Square asymptotic, where the correlation value between the attributes was give thought the Crammer value, describe as

$$v = \sqrt{\frac{X^2}{N\ x\ (min(N,P)-1)}} \quad (1),$$

where N is the number of entries, P is number of attributes.

2.3. Charlson Comorbidities Index (CCI) and Patient Estimated 10-year Survival

The CCI was calculated accordingly the following comorbidities: diabetes, AIDS and dementia. Additionally, in this study the CCI was adjusted by the patient age following the combined index proposed by Charlson et al[13]. The CCI was calculated using only these attributes because there was no further information about other comorbidities in on our dataset. Besides, the system information could not be related to any CID-9 CM codes, what made difficult the identification of the comorbidities and their respective weights[7]. Additionally, the findings, updates and validations done by Quant et al.[14], Charlson et. al.[13], and Zavascki & Funchs[15] were considered when calculating the CCI and the patient 10-year survival rate. The former was calculated as the sum of all used comorbidities weights, where the weights were: +1 for dementia and diabetes; +6 for AIDS; +1 for ages between 50-59 years; +2 for ages between 60-69 years; +3 for ages between 70-79 years; and +4 for ages above 80 years. The later was calculated accordingly to

$$10 - year\ survival = 0.983^{e^{CCIx0.9}} \quad (2),$$

where CCI is the calculated Charlson Comorbidities Index for the patient.

Both the CCI and the estimated 10-year survival were added as predictors to our dataset. Finally, each attribute was ranked as a predictor to the possible outcomes. The rank was created using the Bagging tree[16]. This method bootstraps and generates different tree models, the final importance value for an attribute is total Entropy of all bootstrapped trees[17]. Therefore, it is possible to detect the attributes that have a higher Entropy across all bootstrapped trees, and rank all attributes according to its importance as a predictor to the TB treatment outcome.

3. Results

The CCI distribution among TB treatment death and cure cases is shown in Figure 1. The figure shows that the amount of treatments in which the patient is cured from TB usually have a lower CCI value, having a rapid decay as the CCI value increases. This behavior is also shown in histogram for treatments which resulted in death due to TB. However, the mean value of CCI is higher for treatments that have this outcome.

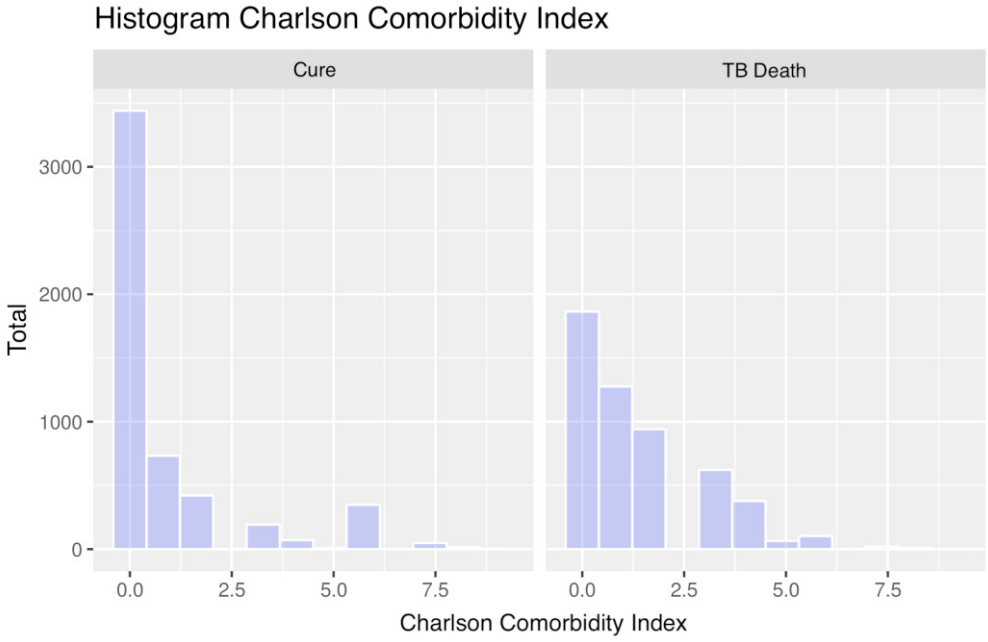


Figure 1. Histogram for the total number of treatments in both outcomes vs. Charlson Comorbidities Index (CCI).

Figure 2 shows the relationship between the number of treatments for both outcomes, death and cure, and the estimated 10-year survival rate of the patients, derived from the CCI. Also, we can notice an inverse distribution to the one in Figure 1, where most cured patients are estimated to have a better 10-year survival rate. Meanwhile, death outcomes show a lower estimated 10-year survival chance. Thus, it is possible to observe a certain degree of correlation between the CCI and the estimated 10-year survival rate of the patients.

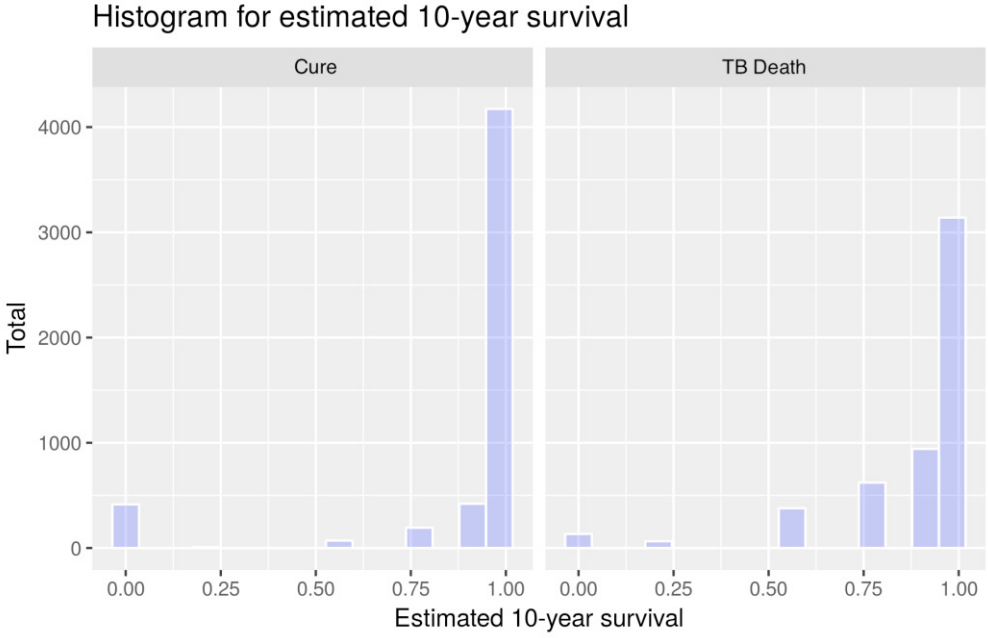


Figure 2. Histogram for the total number of treatments in both outcomes vs. Estimated 10-year survival rate.

The classification process results for the tree-Bagging method are summarized in Figure 3, which shows the 10-most important attributes as predictors for the treatment outcome, among which the attributes age, CCI, the estimated 10-year survival rate, and the number of times the patient took medication under the supervision of a health agent are present. Also, we can observe, respectively, the variables that represent the total number of people who had contact with the patient (communicants total), the number of these people examined (communicants examined), the presence of alcoholism, the initial drug scheme, and the type of treatment performed. We stress, that the correlation of the communicants total and communicants examined to the patients death during treatment do not indicate any causation. The causation of these attributes could only be indicated through further statistical study and clinical analysis, which is not performed here.

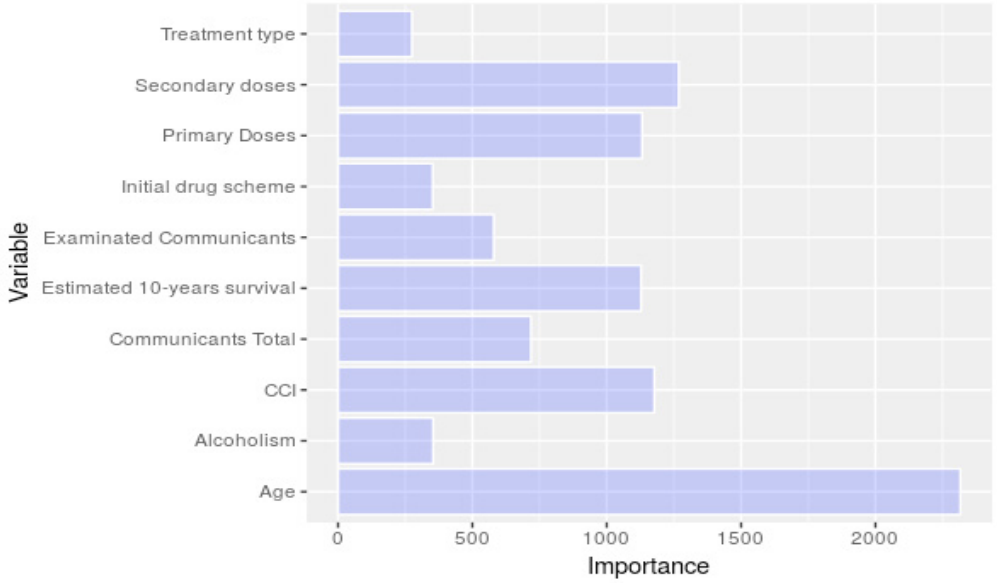


Figure 3. Rank with the 10-most important attributes and their importance value.

4. Discussion

As seeing in the results, the CCI and the estimated 10-year survival rate could be used as predictors to classify between both outcomes of TB treatment, death and cure. It is possible to characterize the correlation of both attributes with the outcomes through the Figure 1 and Figure 2. These correlations are also validated by Kuo et al[18] in study on incidence and outcome of newly-diagnosed TB. In addition, the analysis of the other attributes, through the use of the tree-Bagging technique, serve as means to compare the CCI importance as predictor to the TB treatment outcome. These findings suggest that a new comorbidity index and predictor to TB treatment outcome could be developed through CCI extension, taking into account the attributes that have displayed higher importance as predictors. Several authors given a new cause-of-death structure suggest the development and validation of existing comorbidity indexes[19,20,21].

The number of doses taken during treatment, acts as one of the most important predictors for the classification between treatment outcome, along with the type of treatment, reinforce the role that DOTS is performing in the TB control and monitoring. Alcoholism indicate a social context very related to tuberculosis[22]. On the other hand, other attributes such as number of people who had contact with the patient (communicants total), and the number of these people examined (examined communicants), are correlated to TB death during treatment but do not bear enough meaning to be presented as possible cause to TB treatment outcomes. The use of this information with the CCI can create a relevant feature to predict patient death due to TB during treatment. This would aid treatment monitoring, enabling fast and precise decision-making, and resource reallocation accordingly to the patient morbidity.

5. Conclusion

Finally, in this study we could observe that the CCI has a high potential as a pre-dictor for the classification of TB treatment outcomes. In addition, we could also observe different attributes that can act as predictor for TB treatment outcomes, which can be used to develop a new predictor that extends the CCI.

Acknowledgements

This project was financed by CAPES – Coordination for the Improvement of Higher Education Personnel.

References

[1] Ministério da Saúde. (2017) "Plano nacional pelo fim da tuberculose como problema de saúde pública". Brasil.

[2] World Health Organization (WHO). (2005) "Global tuberculosis control: surveillance, planning, financing". Geneva.

[3] Organización Panamericana De La Salud (OPAS) and Organización Mundial de la Salud (OMS). (1997) "Reunión regional de directores nacionales de programas de control de la tuberculosis: informe final". Ecuador.

[4] Santos, Joseney. (2007) "Resposta brasileira ao controle da tuberculose". *Rev Saude Publica* **41**(**1**):89-93.

[5] Farley, Joel. F., Carolyn R. Harley and Joshua. W. Devine. (2006) "A comparison of comorbidity measurements to predict healthcare expenditures". *The American journal of managed care* **12**(**2**):110–9.

[6] De Groot, Vincent, Heleen Beckerman, Gustaaf J. Lankhorst and Lex M. Bouter. (2003) "How to measure comorbidity. a critical review of available methods". *Journal of clinical epidemiology* **56**(**3**):221–9.

[7] Charlson, Mary E., Peter Pompei, Kathy L. Ales and Charles Ronald Mackenzie. (1987) "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation". *J Chronic Dis* **40**(**5**):373-383.

[8] Carvalho, Isabelle. (2014) "Estudo do risco de óbito por meio da análise de comorbidade nos pacientes internados nos hospitais gerais do DRS XIII em 2011". *Dissertação (Mestrado em Ciências)* – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto.

[9] Deyo, Richard A., Daniel C. Cherkin and Márcia A. Ciol. (1992) "Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases". *Journal of Clinical Epidemiology* **45**(**6**):613-619.

[10] Romano, Patrick S., Leslie L.Roost and James G. Jollis. (1993) "Response: Further evidence concerning the use of a clinical comorbidity index with ICD-9-CM administrative data". *J Clin Epidemiol* **46**(**10**):1085-1090.

[11] Secretaria Estadual de Saúde. Centro de Vigilância Epidemiológica. Prof. Alexandre Vranjac. Divisão de Controle da Tuberculose. "TB em números". Available online: < http://www.cve.saude.sp.gov.br/htm/cve_tb.html>. Access in: 16th march, 2018.

[12] McHugh, Mary L. (2013). "The Chi-square test of independence". *Biochemia Medica* **23**(**2**):143–149.

[13] Charlson, Mary E., Ted P. Szatrowski, Janey Peterson and Jeffrey Gold. (1994) "Validation of a combined comorbidity index". *Journal of clinical epidemiology* **47**(**11**):1245-1251.

[14] Quan, Hude, Bing Li, Chantal M. Couris, Kiyohide Fushimi, Patrick Graham, Phil Hider, Jean-Marie Januel and Vijaya Sundararajan. (2011) "Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries". *American journal of epidemiology* **173**(**6**):676-682.

[15] Zavascki, Alexandre P. and Sandra C. Fuchs. (2007) "The need for reappraisal of AIDS score weight of Charlson comorbidity index". *Journal of clinical epidemiology* **60**(**9**):867-868.

[16] Breiman, Leo. (1996) "Bagging predictors". *Machine learning* **24**(**2**):123-140.

[17] Breiman, Leo. (1996) "Technical Note: Some Properties of Splitting Criteria". *Machine learning* **24**(**1**):41-47.

[18] Kuo, Shu-Chen, Yung-Tai Chen, Szu-Yuan Li, Yi-Tzu Lee, Albert C. Yang, Te-Li Chen, Chia-Jen Liu, Tzeng-Ji Chen, Ih-Jen Su, Chang-Phone Fung. (2013) "Incidence and outcome of newly-diagnosed tuberculosis in schizophrenics: a 12-year, nationwide, retrospective longitudinal study". *BMC Infectious Diseases* **13**(**1**):351.

[19] Kil, Seol-Ryoung, Sang-Il Lee, Young-Ho Khang, Moo-Song Lee, Hwa-Jung Kim, Seon-Ok Kim and Min-Woo Jo. (2012). "Development and validation of comorbidity index in South Korea". *International Journal for Quality in Health Care* **24**(**4**):391-402.

[20] Park, Jae Y., Myoung-Hee Kim, Seung S. Han, Hyunjeong Cho, Ho Kim, Dong-Ryeol Ryu, Hyunwook Kim, Hajeong Lee, Jung P. Lee, Chun-Soo Lim, Kyoung H. Kim, Kwon W. Joo, Yon S. Kim and Dong K. Kim. (2015). "Recalibration and validation of the Charlson comorbidity index in Korean incident hemodialysis patients". *PloS ONE* **10**(**5**):e0127240.

[21] Takenaka, Yukinori, Norihiko Takemoto, Ryohei Oya, Naoki Ashida, Takahiro Kitamura, Kotaro Shimizu, Kazuya Takemura, Takahiro Michiba, Atsushi Hanamoto, Motoyuki Suzuki, Yoshifumi Yamamoto, Atsuhiko Uno and Hidenori Inohara. (2017). "Development and validation of a new comorbidity index for patients with head and neck squamous cell carcinoma in Japan". *Scientific reports* **7**(**1**):7297.

[22] Centers for Disease Control and Prevention (CDC). (2012) "Integrated prevention services for HIV infection, viral hepatitis, sexually transmitted diseases, and tuberculosis for persons who use drugs illicitly: summary guidance from CDC and the US Department of Health and Human Services". *MMWR Recomm Rep* **61**(**RR-5**):1-40.

International Journal of Medical Informatics 141 (2020) 104198

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Development of CART model for prediction of tuberculosis treatment loss to follow up in the state of São Paulo, Brazil: A case–control study

Verena Hokino Yamaguti[a,*], Domingos Alves[a,d], Rui Pedro Charters Lopes Rijo[a,b,c,d], Newton Shydeo Brandão Miyoshi[a], Antônio Ruffino-Netto[a]

[a] Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, State of São Paulo, Brazil
[b] School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal
[c] Institute for Systems and Computers Engineering (INESCC), Coimbra, Portugal
[d] Center for Research in Health Technologies and Services (CINTESIS), University of Porto, Porto, Portugal

ARTICLE INFO

Keywords:
Tuberculosis
Treatment loss to follow up
Prediction model
Feature selection

ABSTRACT

Background: Tuberculosis is the leading cause of infectious disease-related death, surpassing even the immunodeficiency virus. Treatment loss to follow up and irregular medication use contribute to persistent morbidity and mortality. This increases bacillus drug resistance and has a negative impact on disease control.
Objective: This study aims to develop a computational model that predicts the loss to follow up treatment in tuberculosis patients, thereby increasing treatment adherence and cure, reducing efforts regarding treatment relapses and decreasing disease spread.
Methods: This is a case-controlled study. Included in the data set were 103,846 tuberculosis cases from the state of São Paulo. They were collected using the TBWEB, an information system used as a tuberculosis treatment monitor, containing samples from 2006 to 2016. This set was later resampled into 6 segments with a 1-1 ratio. This ratio was used to avoid any bias during the model construction.
Results: The Classification and Regression Trees were used as the prediction model. Training and test sets accounted for 70% in the former and 30% in the latter of the tuberculosis cases. The model displayed an accuracy of 0.76, *F*-measure of 0.77, sensitivity of 0.80 and specificity of 0.71. The model emphasizes the relationship between several variables that had been identified in previous studies as related to patient cure or loss to follow up treatment in tuberculosis patients.
Conclusion: It was possible to construct a predictive model for loss to follow up treatment in tuberculosis patients using Classification and Regression Trees. Although the fact that the ideal predictive ability was not achieved, it seems reasonable to propose the use of Classification and Regression Trees models to predict likelihood of treatment follow up to support healthcare professionals in minimising the loss to follow up.

1. Introduction

Tuberculosis (TB) is the leading cause of infectious disease-related death, surpassing the immunodeficiency virus (HIV) [1]. TB control is based on early and appropriate diagnosis and treatment, interrupting the transmission chain [2]. The National Tuberculosis Control Plan (PNCT) is a strategy implemented by the Brazilian government. The Directly Observed Treatment Short-Course (DOTS), is a part of this strategy and has assisted in the country's reach to an 85% cure rate and has reduced the loss to follow up rate in TB treatments by 5% [2]. However, loss to follow up treatment and irregular medication use contribute to the persistence of morbidity and mortality, increasing drug resistance cases [3,4].

TB is related to poverty and different social causes. Therefore, patients are difficult to monitor. Factors such as initial improvement, lack of patient follow-up and monitoring are all associated with poor treatment adherence [5]. After the first few months of using the correct medications during treatments, patients present improvements in their overall physical well-being and most of their symptoms are gone, which then leads to patients interrupting the treatment themselves, even though they are not cured. Social factors, treatment methodology and provided health services are factors associated with treatment loss to follow up [6–8].

In the last 10 years, computational techniques for predicting the

* Corresponding author.
E-mail addresses: verena.yamaguti@usp.br (V. Hokino Yamaguti), quiron@fmrp.usp.br (D. Alves), rui.rijo@ipleiria.pt (R.P. Charters Lopes Rijo), newton.sbm@gmail.com (N.S. Brandão Miyoshi), aruffino@fmrp.usp.br (A. Ruffino-Netto).

https://doi.org/10.1016/j.ijmedinf.2020.104198
Received 19 December 2019; Received in revised form 3 April 2020; Accepted 23 May 2020
Available online 15 June 2020

outcomes of TB treatment have been explored. Neural networks have been applied to predict the TB treatment loss to follow up in some cities of Espírito Santo in Brazil [9]. A logistic regression model was used to select features [10] and predict loss to follow up TB treatment [5]. Another study compared the application of different machine learning methods to classify treatment outcomes, among which the C4.5 decision tree showed better performance [11].

With a computational model that is able to identify patients that are prone to not follow-up treatment, teams would then be able to focus their efforts in reducing the rate of relapse, dissemination and loss to follow-up treatment. Although there are studies about the subject [9,10,5,11], no study has developed an effective human-readable model that can be applied by health professionals in TB treatment control. Decision trees can be represented as a set of rules that would assist healthcare professionals, the ones with and the ones without access to an information system, to control TB. Thus, this study aims to develop a CART predictive model to predict TB treatment loss to follow up cases.

2. Materials and methods

For the development of this case-control study, 208,620 TB treatment cases were used, all between the years of 2006 and 2016 and in the state of São Paulo. The sample had sensitive and drug-resistant tuberculosis cases. The state of São Paulo is one of 27 federal states of Brazil and is located in the southeast region of the country. It has 645 cities and an area of 248,219,491 km^2 [12]. It has the greatest population in the country, with around 45.9 million inhabitants [12].

The inclusion criteria considered treatment only when the outcome was either cure or loss to follow up, totaling in 180,100 cases. Data was collected through TBWEB, a system used by the public healthcare system for reporting and monitoring TB treatments in the state of São Paulo. The Health Secretary of the State of São Paulo granted access to the data and all information was previously anonymized.

Inmates and patients with invalid information for the variables in 3 were excluded. The resulting data set had 103,846 TB treatment cases, and it was split into 6 sub-samples with a 1-1 ratio of cases and controls. The ratio was chosen to reduce any class imbalance bias. The sample numbers were empirically chosen to maximize the model precision, and sampling was done following the random under-sampling technique [13]. There are 46 attributes that describe each instance in the dataset S1 File. We defined it as case when there was a treatment loss to follow up and as control when the patient was cured.

Cured patients are those with two consecutive sputum culture-negative tests within a minimum interval of 30 days after the 12th month of treatment, without clinical and radiological signs of active disease. In the case of a sputum culture-positive test in the 12th month, the treatment is prolonged for up to 24 months. The patient is then only considered cured after three consecutive negative tests with a minimum interval of 30 days and without any clinical and radiological signs of active disease. Treatment loss to follow up refers to patients who did not attend the healthcare establishments for more than 30 consecutive days after the expected date of their return, or, in the cases of supervised treatments, after the date of the last taken medication. Patients that completed the treatment but were not cured were classified as treatment failures and not addressed in this study.

The cross-validation applied to the 6 balanced folds allows us to perform a non-skewed analysis of the discrepancies between the number of cases and controls in the initial dataset. It also allows us to use the full set of available information. Since the attributes presented in the database are mostly categorical, they were transformed into indicative variables and totaled 194 binary attributes, each of these new attributes representing the presence or not of its possible categories. This transformation, called one-hot encoding, prevents any missing data from affecting the prediction process and prevents us from generating any bias due to the transformation of categories into numbers.

Table 1
Number of doses classifiers correlation to the treatment loss to follow up.

Attribute	Loss to follow up
Number of doses from 3rd to 6th month	−0.239080
Number of doses up to the 2nd month	−0.164101

On the other hand, quantitative attributes were standardized and missing values were not considered.

In the attribute analysis step, each attribute was evaluated as a predictor, either for the loss to follow-up treatment or for the cure outcome. The ranking was created using the Bagging Tree technique [14]. This method initializes and generates different decision tree models. Lastly, the importance of an attribute was given by the total entropy of the initialized trees [15] and therefore, it is possible to detect which of the attributes have the most entropy among all trees. With this, we classify all the attributes in accordance to their importance as predictors for the outcomes of TB treatment. Additionally, to avoid redundancy between some classifiers and treatment outcomes, we have inspected the correlation between them. Classifier correlation was tested using the Point-biserial correlation coefficient.

By consulting with a specialist, we were able to identify some attributes as possible tautology, such as the number of doses confirmed from the 3rd month to the 6th month and the number of doses up to the 2nd month. We analyzed the correlation of these classifiers and the treatment outcome in Table 1 to exclude any classifier that could indicate a tautology. None of them showed a correlation higher than 0.5 or lower than −0.5, and as a result, we considered these classifiers in the final model, since they could not be considered tautologies in our study dataset.

Finally, by using the most important attributes identified through the aforementioned ranking, the prediction model was created. The CART algorithm was used to develop the prediction model. This model was selected because of its robustness against outliers and for being a decision tree. As a decision tree, the model is human-readable, and can be used either as a paper printable guideline or as an algorithm that can be implemented into a computational system, therefore providing a model that can be used in either the absence or presence of a proper computational system infrastructure. Having it be available to those without access to an information system was taken into account because the model aims to reduce treatment loss to follow up in a country where the infrastructure for a computation system is uncertain in some locations.

The model was trained and tested with 70% and 30% of the population, from each of the segments, respectively. The model's predictive capacity evaluation was performed by using the metrics extracted for the test population. The metrics used were: Receiver Operating Characteristic (ROC) curve, Sensitivity, Specificity, Confusion Matrix and F-Measure. Decision trees have already been used by other authors as predictors of TB treatment. The current study is differentiated by the use of the CART algorithm, which is characterized by the construction of binary trees, in which each node has exactly two leaves and is more robust against outliers.

This study was submitted to Plataforma Brasil and has the number CAAE: 70947317.3.0000.5440. This study was approved by CEP (Comitê de Ética em Pesquisa, i.e., Committee of Research Ethics)/ CONEP (Comissão Nacional de Ética em Pesquisa i.e., National Research Ethics Commission) of the Brazilian Government.

3. Results

The final population of the study contained 91,823 controls and 12,023 cases (Fig. Figure 1). Table Table 2 shows the distribution for the population demographic data.

Most of the patients are male adults between the ages of 20 and 39

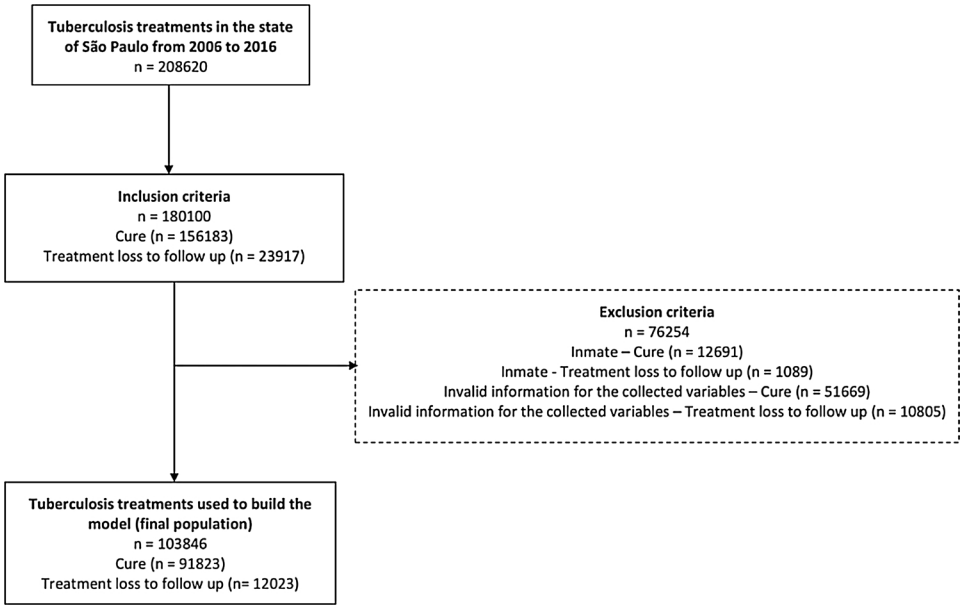


Fig. 1. Patient inclusion flowchart.

(24.67%). Less than 2% of the population represents the youth and elders.

The sampled population is mainly composed of white (39.84%) and brown (26.69%) ethnic groups. Only 0.52% of them were from an indigenous ethnic group. Most patients had basic education (less than 11 years of study) or had never studied (none).

The ranking of the 20 most important attributes that classify between cure and loss to follow up treatment can be seen in Table Table 3. It is possible to observe that the number of doses confirmed from the 3rd month to the 6th month is the attribute that stands out the most for identifying loss to follow up treatment cases, representing a much greater entropy when compared to other attributes. Other attributes with similar importance for classifying treatment loss to follow up are age, total number of people residing with the patient, and the number of people who attend the exam. As follows, other variables related to social vulnerabilities and comorbidities.

Only the first 20 analyzed variables were used to generate the prediction model. The 3rd to 6th month doses were excluded as well as the doses from the 1st and 2nd month. The number of variables used were determined empirically, in order to maximize the defined quantitative metrics. Each of the segments that were used for training contained 12,023 loss to follow-up treatment cases and 12,023 cure cases. To train the model, 16,842 (70%) cases from each of the 6 segments of the population were used. Subsequently, the model test used 7218 (30%) individuals from each segment.

The generated model is visually represented by a tree, which can be

Table 2
Distribution of patients treated by age group, sex, ethnic groups and scholarity.

Variables	Value	Control (N = 91, 823)	Case (N = 12, 023)	Total (N = 103, 846)
Age group	< 1 year	385 (0.37%)	33 (0.03%)	418 (0.40%)
	1–4 years	1085 (1.04%)	51 (0.05%)	1136 (1.09%)
	5–9 years	882 (0.85%)	29 (0.03%)	911 (0.88%)
	10–14 years	1550 (1.49%)	49 (0.05%)	1599 (1.54%)
	15–19 years	6109 (5.88%)	733 (0.71%)	6842 (6.59%)
	20–29 years	21,735 (20.93%)	3882 (3.74%)	25,617 (24.67%)
	30–39 years	19,115 (18.41%)	3451 (3.32%)	22,566 (21.73%)
	40–49 years	17,086 (16.45%)	2253 (2.17%)	19,339 (18.62%)
	50–59 years	13,440 (12.94%)	1009 (0.97%)	14,449 (13.91%)
	60–69 years	6691 (6.44%)	344 (0.33%)	7035 (6.77%)
	70–79 years	2863 (2.76%)	143 (0.14%)	3006 (2.89%)
	> =80 years	882 (0.85%)	46 (0.04%)	928 (0.89%)
Sex	Female	32,107 (30.92%)	2984 (2.87%)	35,091 (33.79%)
	Male	59,716 (57.50%)	9039 (8.70%)	68,755 (66.21%)
Ethnic	Yellow	880 (0.85%)	74 (0.07%)	954 (0.92%)
	White	41,372 (39.84%)	4445 (4.28%)	45,817 (44.12%)
	Indigenous	541 (0.52%)	77 (0.07%)	618 (0.60%)
	Brown	27,715 (26.69%)	4154 (4.00%)	31,869 (30.69%)
	Black	9023 (8.69%)	1530 (1.47%)	10,553 (10.16%)
	No information	12,292 (11.84%)	1743 (1.68%)	14,035 (13.52%)
Scholarity	None	3681 (3.54%)	301 (0.29%)	3982 (3.83%)
	1–3 years	8355 (8.05%)	1010 (0.97%)	9365 (9.02%)
	4–7 years	25,055 (24.13%)	3838 (3.70%)	28,893 (27.82%)
	8–11 years	28,437 (27.38%)	3488 (3.36%)	31,925 (30.74%)
	12–14 years	6037 (5.81%)	429 (0.41%)	6466 (6.23%)
	> =15 years	2887 (2.78%)	168 (0.16%)	3055 (2.94%)
	No information	17,371 (16.73%)	2789 (2.69%)	20,160 (19.41%)

Table 3
Ranking of the most important attributes.

Ranking	Sum of entropy of trees	Attribute
1°	0.077879	Number of doses from 3rd to 6th month
2°	0.039481	Age in completed years
3°	0.035257	Number of doses up to the 2nd month
4°	0.032344	Total number of people residing with the patient
5°	0.031436	Number of people residing with the patient who attended the examination
6°	0.024282	Non-drug addicts
7°	0.020861	Negative HIV test
8°	0.017036	Drug addicts
9°	0.016322	Total 8–11 years of study completed
10°	0.016327	Chest XR as suspected TB
11°	0.014258	Type of occupation as unemployed
12°	0.015270	Discovery of the case was in outpatient demand
13°	0.016537	Total 4–7 years of study completed
14°	0.014184	Self-reported race or color as brown
15°	0.013968	Type of treatment as supervised
16°	0.014153	Type of occupation as other
17°	0.014522	Sputum culture not performed
18°	0.013160	Necropsy examination not performed
19°	0.013338	Discovery of the case in Urgency and Emergency Unit
20°	0.012944	Race or self color reported by the patient as white

seen in Fig. Figure 2 of this article. Table Table 4 features the averages for accuracy, *F*-Measure, sensitivity and specificity metrics for the test folders that were used to quantitatively evaluate the developed predictive model. From which it is possible to highlight the high sensitivity and *F*-Measure presented by the model.

Table Table 5 shows the total count number of false positives, false negatives, true positives and true negatives obtained from the test folders. In it we highlight the true positive and true negative numbers.

Fig. Figure 3 shows the obtained ROC curve for each of the test folds of the model contrasted by the dashed line that represents a model that randomly classifies the cases. We can observe that the curve and its area under it is similar for all the analyzed segments, the average Area Under the Curve (AUC) can be viewed between segments 0.73.

4. Discussion

The decision tree created can be used as a guideline among health professionals to detect treatment loss to follow up. The accuracy of the model is similar to other models already presented in the literature [11]. For health professionals, the decision tree is a much more intuitive and understandable approach than a statistical model, since it can be visualized and explored by anyone, as well as be applied without a computational system. To exemplify, the decision tree for this study can be visualized in Fig. Figure 2 .

The model emphasizes the relationship between several variables already identified in the literature as patient loss to follow-up treatment

Table 4
Quantitative metrics used to evaluate the predictive model.

Metrics	Value
Accuracy	0.7601
F-measure	0.7703
Sensitivity	0.8048
Specificity	0.7156

Table 5
Matrix of confusion for the predictive model.

	Positive	Negative
True	17,413	15,509
False	6163	4223

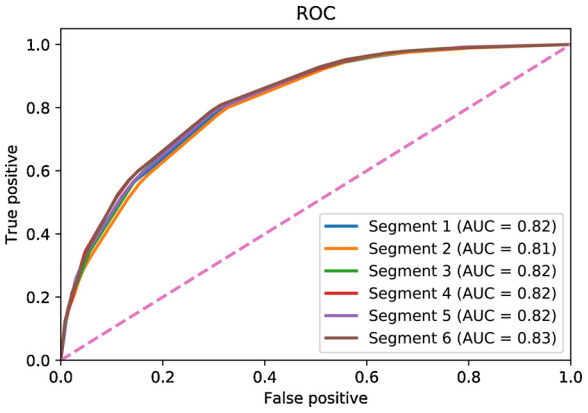


Fig. 3. Folder ROC curve.

or patient cure. Among which are patient ethnicity, unemployment, schooling, drug addiction and age [10,5,11]. However, other factors not raised by the authors appear as factors of great importance to predict the treatment loss to follow up that are cited in this study.

Among these variables, we have highlighted the variables that represent the number of doses received by the patient up until the 2nd month as well as up until the 3rd to the 6th month. Although some specialists believe that it is possible that these variables represent a tautology, they do not have a high enough correlation to be considered one. This further emphasizes that what determines a treatment's success is the frequency of doses and correct medication taken throughout it and not the number of doses taken throughout the treatment. The monitoring of dose numbers taken and its frequency can act as an alarm trigger, notifying health professionals automatically.

In general, our model presented a lower number of false positives than other models presented in the literature [5]. This is mainly due to the balance between cure and treatment loss to follow up cases. This

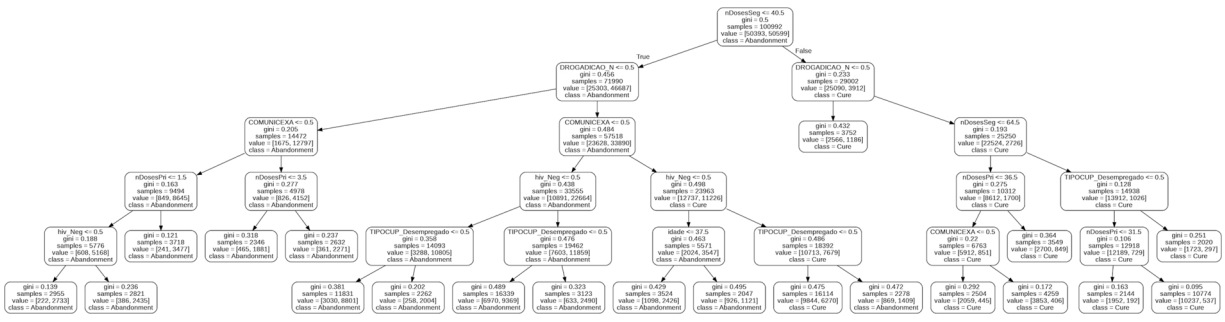


Fig. 2. Decision tree.

approach reduces the bias generated by the imbalance classes. On the other hand, our model failed to adequately optimize the accuracy indexes already presented by similar models in the literature [11]. Thus, we believe experimentation with other models may be appropriate.

The final model has an accuracy higher than the logistic regression model and the C4.5 presented in the literature [5,11]. We also showed a better value for the *F*-measure than the C4.5 model [11]. However, the C4.5 model [11] presents a better result for the ROC curve.

5. Conclusion

It was possible to build the predictive model of treatment loss to follow up of tuberculosis through CART. During model development, new features for TB treatment loss to follow up predictors were identified. Although the fact that the ideal predictive ability was not achieved, it seems reasonable to propose the use of CART models. We aim in future studies to improve the model accuracy and implement the model in an informational system for tuberculosis treatment and management.

Summary Points

What was already known on the topic:

- Treatment loss to follow up and irregular medication use contribute to increased bacillus drug resistance and has a negative impact on disease control.
- Other studies apply different models to predict the TB treatment outcome.
- There is evidence that, among different models, decision trees present better performance.

What this study added to our knowledge:

- The CART model was able to perform equally or better than other models in the literature, additionally providing a human-readable model that can be used without an informational system.
- Features such as age, the number of doses taken by the patient between the 3rd and 6th month, the number of people resident with the patient were ranked among the feature with greater entropy to predict treatment loss to follow up.
- The number of doses taken by the patient and the treatment outcome presented a correlation lower than expected, discarding the possibility of a possible tautology, as suggested by specialists.

Acknowledgements

We would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and São Paulo Research Foundation (FAPESP) (Grant Nos. 2018/23963-2 and 2018/00307-2).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ijmedinf.2020.104198.

References

[1] WHO, Global Tuberculosis Report 2019, Geneva, (2019).
[2] Ministério da Saúde, Plano Nacional pelo Fim da Tuberculose como Problema de Saúde Pública, (2017) http://bvsms.saude.gov.br/bvs/publicacoes/brasil_livre_tuberculose_plano_nacional.pdf.
[3] L.M.M. Paixão, E.D. Gontijo, Perfil de casos de tuberculose notificados e fatores associados ao abandono, Belo Horizonte, MG, Rev. Saúde Pública 41 (2) (2007) 205–213, https://doi.org/10.1590/S0034-89102007000200006.
[4] K.M.J. de Souza, Abandono do tratamento da tuberculose na Atenção Primária à Saúde: uma análise segundo o enfoque familiar do cuidado, (2008), pp. 1–112.
[5] E.d.A. Silva, U.U. dos Anjos, J.d.A. Nogueira, Modelo preditivo ao abandono do tratamento da tuberculose, Saúde em Debate 38 (101) (2014) 200–209, https://doi.org/10.5935/0103-1104.20140018.
[6] H.B. De Oliveira, D.C. De Moreira Filho, Abandono de tratamento e recidiva da tuberculose: Aspectos de episódios prévios, Campinas, SP, Brasil, 1993–1994, Rev. Saude Publica 34 (5) (2000) 437–443, https://doi.org/10.1590/S0034-89102000000500002.
[7] M.F. Rabahi, A.B. Rodrigues, F. Queiroz de Mello, J.C. de Almeida Netto, A.L. Kritski, Noncompliance with tuberculosis treatment by patients at a tuberculosis and AIDS reference hospital in midwestern Brazil, Braz. J. Infect. Dis. 6 (2) (2002) 63–73, https://doi.org/10.1590/S1413-86702002000200002.
[8] N. Schluger, C. Ciotoli, D. Cohen, H. Johnson, W.N. Rom, Comprehensive tuberculosis control for patients at high risk for noncompliance, Am. J. Respir. Crit. Care Med. 151 (5) (1995) 1486–1490, https://doi.org/10.1164/ajrccm.151.5.7735604.
[9] T.N. Do Prado, R. Dietze, A.R. Netto, E. Zandonade, E.L.N. Maciel, prediction of dropout tuberculosis treatment on prioritary cities to control in Espi'rito Santo State, Brazil, J. Epidemiol. Commun. Health 65 (Suppl 1) (2011) A138, https://doi.org/10.1136/jech.2011.142976e.53.
[10] G. Harling, A.S. Lima Neto, G.S. Sousa, M.M. Machado, M.C. Castro, Determinants of tuberculosis transmission and treatment abandonment in Fortaleza, Brazil, BMC Public Health 17 (1) (2017) 1–10, https://doi.org/10.1186/s12889-017-4435-0.
[11] S.R.N. Kalhori, X.-j. Zeng, Evaluation and comparison of different machine learning methods to predict outcome of tuberculosis treatment course, J. Intell. Learn. Syst. Appl. 05 (03) (2013) 184–193, https://doi.org/10.4236/jilsa.2013.53020.
[12] I.B. de Geografia e Estatística, Estatísticas do estado de são paulo, (2020) (accessed 15.03.20), https://www.ibge.gov.br/cidades-e-estados/sp.html.
[13] N. Japkowicz, The class imbalance problem: significance and strategies, Proc. of the Int'l Conf. on Artificial Intelligence (2000).
[14] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140, https://doi.org/10.1007/BF00058655.
[15] L. Breiman, Technical note: some properties of splitting criteria, Mach. Learn. 24 (1996) 41–47, https://doi.org/10.1007/BF00117831.

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2022

Clinical Pathways and Hierarchical Clustering for Tuberculosis Treatment Outcome Prediction

Verena Hokino Yamaguti[a,*]; Alberto Freitas[c]; Anderson Chidi Apunike[a]; Rui Pedro Charters Lopes Rijo[b]; Domingos Alves[a]; Antonio Ruffino Netto[a]

[a]*University of São Paulo, Ribeirão Preto, Brazil*
[b]*Polytechnic Institute of Leiria, Leiria, Portugal*
[c]*CINTESIS@RISE, FMUP, University of Porto, Porto, Portugal*

Abstract

Clinical pathways are chronological event series that happen throughout a patient's treatment. They can be extracted from the Electronic Health Record medical information and this can be used to correlate the pathway to possible healthcare outcomes. This can be applied to a wide variety of diseases to point pathways related to bad outcomes. These pathways can be audited and patients that start to follow such patterns can be put in special observation and care. Tuberculosis (TB) is one of the leading causes of death through infectious disease and its control is based on search for cases, accurate and premature identification, and treatment. The use of the aforementioned method can help in disease control and premature identifications of bad outcomes for ongoing treatments. Therefore, the current study goals are: 1) identify the existing clinical pathways; 2) group these pathways using hierarchical clustering; 3) create a classification model based on the generated clusters to predict bad outcomes. The dataset used consisted of 277,870 TB treatment cases from the state of São Paulo collected through TBWEB, a information system for monitoring and follow-up of TB cases. All cases with ongoing treatment were excluded from the study and the resulting dataset was splitted in training and test samples. To reduce bias due to imbalance the undersampling technique was applied to the training dataset resulting in a final sample size of 90,184. The test dataset had a size of 52,639 cases. Both datasets had 16 attributes describing the patient diagnosis and drug scheme evolution through the treatment. All attributes unique values were mapped and a representation character was assigned to each one. Later, these representation characters were concatenated in the chronological order of the events and diagnosis creating a representational string for the clinical pathway. The resulting pathways of the training dataset were used to build the clusters which were later used to build the classifier to predict the treatment outcome based on the test dataset clinical pathways. The final model overall accuracy is at 0.829. The model showed a significant improvement of accuracy from previous studies and had similar or better performance than others in the literature. We believe this model can be implemented to a informational system to further improve treatments management and tuberculosis control.

Peer-review under responsibility of the scientific committee of the CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2022

Keywords: Tuberculosis; Clinical pathways; Process mining; Public health; Clustering; Machine Learning.

1. Introduction

Clinical pathways describe a series of clinical events that occur throughout a patient's treatment in a timely manner. These clinical interventions are decided by a medical staff and have specific goals. They are initialized at patient admission and end at patient discharge [1]. Clinical pathways can be designed to deliver a straight and structured healthcare to patients that reduces clinical variation and enhances operational execution and healthcare results [2]. Also, they serve as documentation which can improve overall teamwork and communication [3].

However, to ascertain which clinical pathways relate to a specific outcome, different permutations must be evaluated by a clinical team. This imposes a challenge due to the number of possible permutations in a clinical pathway. Therefore, the clinical pathway design would benefit from a more explicit design based on informational systems [4] using data mining and pattern recognition methods to speed up the process and clinical staff to determine possible pathways.

Clinical pathways can be extracted from the patient's Electronic Health Record (EHR) through the use of data mining methods [5]. EHRs have miscellaneous medical information about the patient profile and clinical events that happened throughout the patient's care. This clinical information can be used to correlate these events to possible healthcare outcomes [6].

The clinical pathway extraction process relies on determining the chronological events and attributing them to a single character which results in a representational string with all the events. These pathways can be visualized as flowchart or Petri's networks [7]. Their visualization provides an overview of most common pathways and enables the risk assessment and prediction through machine learning methods. Pathways that are mostly related to negative outcomes can be audited to verify if the clinical events follow the recommended guidelines and healthcare protocols.

Tuberculosis (TB) is one of the main leading causes of infectious disease death worldwide [8]. This disease control is based on search for cases, accurate and premature identification, and treatment [9]. Therefore, the use of clinical pathways for TB treatment can point pathways related to bad outcomes. These pathways can be audited and patients that start to follow such patterns can be put in special observation and care [10].

The current study goal is to develop a model for multi label classification using clinical pathways to predict the TB outcome. This model will use all clinical events recorded in the patient's EHR throughout the TB treatment. Also, the proposed model uses unsupervised hierarchical clustering for predicting a treatment outcome. The records have all health related events and information about the patient's diagnosis and medication through the treatment.

2. Materials and methods

2.1 Dataset and Software

The initial dataset had 277,870 TB treatment cases from the state of São Paulo. Data was collected through the TBWEB system between 2006 and 2019. TBWEB [11] is a system for notification and tuberculosis treatment follow-up in the state of São Paulo which belongs to the State Health Secretariat of São Paulo. It performs the role of a centralized database for all TB treatments. All data and information used in this study was previously anonymized. The state of São Paulo is one of 27 federal states of Brazil and is located in the southeast region of the country. It has 645 cities and an area of 248,219,491 km^2 [12]. It has the greatest population in the country, with around 45.9 million inhabitants [12]. All code implemented and used in this study was developed in *Python* 3.9.10.

2.2 Study and Preparation of the Data

The first step in this study was to analyze the dataset to understand its structure and variables. The dataset was prepared by selecting the variables of interest and defining the exclusion and inclusion criteria. The inclusion criteria considered treatment only with either a good or bad outcome. Therefore, excluding any ongoing treatment (14,677), which totaled in 263,193 cases. Later, the dataset was split in test (52,639) and training (210,554) sets. To avoid bias from class imbalance we applied the undersampling method to balance good and bad outcomes in the training dataset to 1-1 ratio. The original training dataset had only 21.4% of bad outcomes. The final training set had 90,184 instances.

The initial dataset consisted of 115 attributes which described the patient’s demographic information, medication, interventions and clinical diagnosis throughout the treatment. The final dataset used only 16 attributes that described the diagnosis and drug scheme evolution through the patient's treatment. These attributes had been used in a previous study [10]. The order of these attributes is relevant for evaluating the patient treatment evolution in chronological order and is later used to assemble the string that will represent the clinical pathway.

2.3 Setup and Filtering of the Clinical Pathways

In order to determine the clinical pathway for a patient all attributes unique values were mapped and a representation character was assigned to each one. Later, these representation characters were concatenated in the chronological order of the events and diagnosis creating a representational string for the clinical pathway. For this study we considered all clinical pathways that had a good (Cure) or bad outcome (Death, Loss to follow up or Drug resistance). Pathways for ongoing treatments or with a different outcome were filtered and removed.

2.4 Hierarchical clustering

After pathways were identified for the training set, the distance matrix was calculated using the Levenshtein distance [13]. The Levenshtein distance is a relevant string metric for measuring the difference between two-character sequences and is used in various domains such as information retrieval, pattern recognition, error correction, and molecular genetics [14].

The generated distance matrix was used to apply the Weighted Pair Group Method with Arithmetic Mean (WPGMA) [15] clustering hierarchical algorithm and generating a dendrogram from the training set. The cutoff point of the dendrogram was determined by measuring the clusters mean sum of squared distances to their centers using the Elbow-curve method. Fig. 1 shows the Elbow curve on which the optimal k is 7.

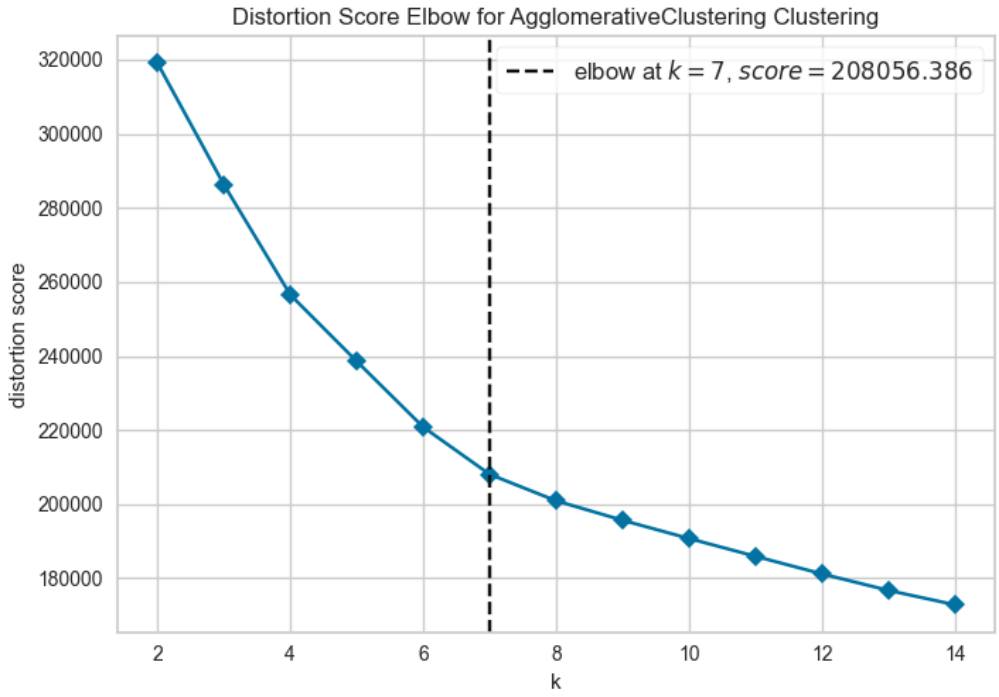


Fig. 1. Elbow curve for clusters between 2 to 14.

2.5 Model classification and performance evaluation

Once all clusters were built, we used the test instance to measure the model performance. Each cluster is evaluated and related to the outcome with the highest probability of the instances that belong to it. The test instances are compared to the clusters by calculating distance to cluster centroid. To simulate a real-world scenario all instances from the test dataset had the representational letter for the treatment outcome removed. Then, for each test instance the closest cluster gives a predicted outcome. To evaluate the model performance, we considered the average Precision, Recall and F1-score.

3. Results

The generated model produced k=7 hierarchical clusters. The uncut dendrogram for the hierarchical clustering is displayed in Fig. 2 in which the different colors denote the similarity between the nodes. Table 1 shows the clusters most related outcomes where it is observable their even distribution among good and bad outcomes. We stress that this even distribution among clusters is due to the undersampling applied to the dataset. Otherwise, we expected to see most clusters representing good outcomes.

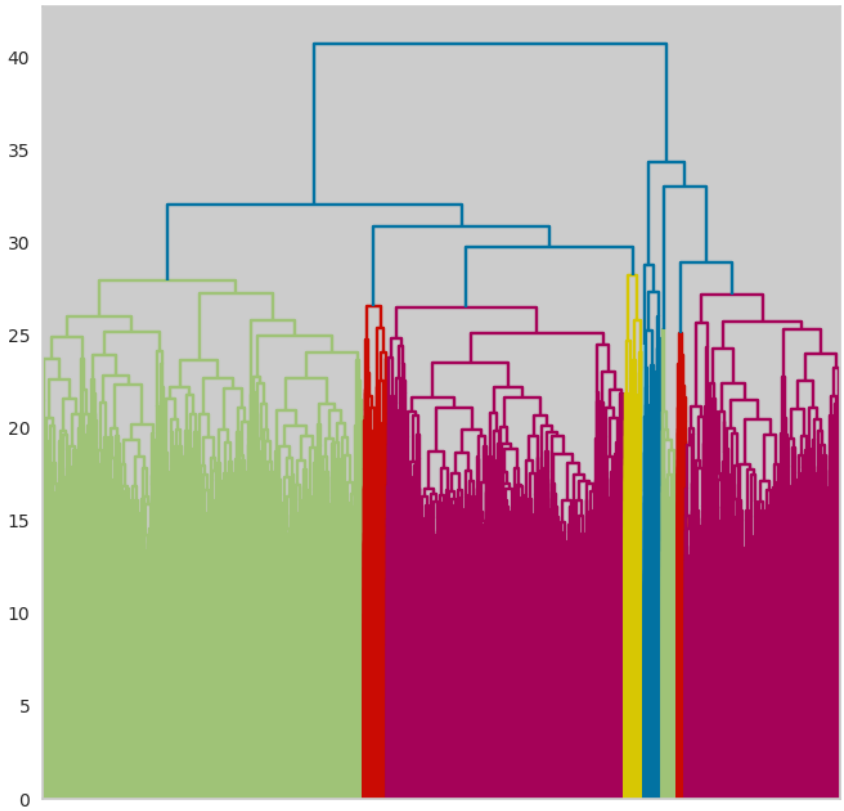


Fig. 2. Uncut dendrogram generated from the hierarchical clustering applied to the training dataset.

After the clusters were generated, we used them as a model for predicting each instance outcome in the test dataset. To predict an outcome for an instance we calculated the closest cluster to that instance pathway and defined the predicted outcome as the cluster most related outcome. Table 2 shows the Precision, Recall and F1-score for good and bad outcomes. We can observe that the model is precise to describe good outcomes rather than bad outcomes showing a precision of 0.883, recall of 0.866 and f1-score of 0.875. The model displays a weighted average precision of 0.832, recall of 0.748, and f1-score of 0.830. The overall model accuracy is at 0.829.

Table 1. Most representative outcome per cluster.

Clusters	Outcome
Cluster 1	Bad
Cluster 2	Good
Cluster 3	Good

Cluster 4	Bad
Cluster 5	Good
Cluster 6	Bad
Cluster 7	Bad

Table 2. Model accuracy per outcome.

Outcome	Precision	Recall	F1-Score
Bad	0.717	0.748	0.732
Good	0.883	0.866	0.875

4. Discussion

The produced clusters allow the prediction of an on-going treatment outcome. If incorporated into a health system this can provide means to alert health care professionals to perform additional actions and efforts to prevent bad outcomes. Additionally, this study enables us to evaluate the most common pathways for good and bad outcomes by analyzing the most representative pathway for each cluster.

Through this study the main problem faced for the prediction of TB treatments was the class imbalance in favor of good outcomes present in the dataset. This imbalance is mostly due to the National Tuberculosis Control Plan enforced by the Brazilian Health Ministry which among other policies implements the Directly Observed Therapy (DOT) which assists the country reducing the number of deaths due to TB and treatment loss to follow-up. In order to reduce this bias, we opted for the undersampling method instead of others because we would still have a significant sample even after applying the undersampling technique.

In general, the present model has improved the accuracy of previous studies [16,17,18] done using a similar dataset. Also, it shows better accuracy when compared to a logistic regression model for TB prediction [19]. We believe this improvement has come due to the temporal analysis provided by the clinical pathways. This allows the model to identify temporal patterns that affect the treatment outcome, something the previous model disregards. Additionally, it provides a more intuitive approach than other predicting models in the literature [20] due to the fact that the generated clusters can be defined according to their more representative clinical pathways. This gives to the involved healthcare professionals a guide to better understand problematic clinical pathways and identifies bad patterns through the patient care.

5. Conclusion

It was possible to build a predictive model of TB treatment outcome through an unsupervised learning model. The generated clusters of the unsupervised model can be used for prediction by calculating the closest cluster for an on-going treatment clinical pathway. Also, the clusters can be used to determine their most common pathway and this can be used to establish a guideline for healthcare professionals [10]. The model presented a significant improvement of accuracy from previous studies and had similar or better performance than others in the literature [19,20]. We aim in future studies to implement the model in an informational system for tuberculosis [21,22,23] treatment and management to help healthcare professionals to manage and monitor TB treatment.

Acknowledgements

We would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and São Paulo Research Foundation (FAPESP) (Grant Nos. 2018/23963-2, 2021/08341-8 and 2020/01975-9).

References

[1] Hunter, Billie and Jeremy Segrott. (2010) “Using a clinical pathway to support normal birth: Impact on practitioner roles and working practices.” *Birth* **37** (**3**): 227–236.

[2] Lin, Fu-ren, Shien-chao Chou, Shung-Mei Pan and Yao-mei Chen. (2001) “Mining time dependency patterns in clinical pathways.” *International Journal of Medical Informatics* **62** (**1**):11-25.

[3] Deneckere, Svin, Martin Euwema, Pieter Van Herck, Cathy Lodewijckx, Massimiliano Panella, Walter Sermeus and Kris Vanhaecht. (2012) “Care pathways lead to better teamwork: Results of a systematic review.” *Social Science and Medicine* **75** (**2**):264–268.

[4] Kempa-Liehr, Andreas, Christina Lin, Randall Britten, Delwyn Armstrong, Jonathan Wallace, Dylan A Mordaunt and Michael O'Sullivan. (2020) “Healthcare pathway discovery and probabilistic machine learning.” *International Journal of Medical Informatics* **137**:104087.

[5] Caron, Filip, Jan Vanthienen, Kris Vanhaecht, Erik Van Limbergen, Jochen Deweerdt and Bart Baesens. (2014) “A process mining-based investigation of adverse events in care processes.” *Health Information Management Journal* **43** (**1**):16–25.

[6] Huang, Zhengxing, Wei Dong, Lei Ji and Huilong Duan. (2016) “Outcome Prediction in Clinical Treatment Processes.” *Journal of Medical Systems* **40** (**1**):8.

[7] Van Der Aalst, Wil. (2011) “Process Mining: Discovery, Conformance and Enhancement of Business Processes.” *Springer*, London.

[8] World Health Organization (WHO). (2021) “Global Tuberculosis Report 2021.” Geneva.

[9] Ministério da Saúde. (2017) “Plano nacional pelo fim da tuberculose como problema de saúde pública.” Brasil.

[10] Apunike, Anderson Chidi., Lívia Maria de Oliveira-Ciabati, Tiago L. M. Sanches, Lariza Laura de Oliveira, Mauro N. Sanchez, Rafael Mello Galliez, and Domingos Alves. (2020) “Analyses of Public Health Databases via Clinical Pathway Modelling: TBWEB.” *International Conference on Computational Science: Computational Science – ICCS 2020* **12140**:550-552.

[11] Galesi, Vera Maria Neder. (2007) “Data on tuberculosis in the state of São Paulo, Brazil.” *Revista de Saúde Pública* **41** (**1**):121.

[12] Instituto Brasileiro de Geografia e Estatística (IBGE). (2020) “Estatísticas do estado de São Paulo.” Available online: <https://www.ibge.gov.br/cidades-e-estados/sp.html>. Access in: 16th may, 2022.

[13] Levenshtein, Vladimir Iossifowitsch. (1966) “Binary codes capable of correcting deletions, insertions, and reversals.” *Soviet Physics-Doklady* **10**(**8**):707–710.

[14] Will, Sebastian, Kristin Reiche, Ivo L Hofacker, Peter F. Stadler and Rolf Backofen. (2007) “Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering.” *PLoS Computational Biology* **3** (**4**): e65.

[15] Sokal, Robert Reuven. (1958) "A statistical method for evaluating systematic relationships." *University of Kansas Science Bulletin* **38**: 1409–1438.

[16] Yamaguti, Verena Hokino, Domingos Alves, Rui Pedro Charters Lopes Rijo, Newton Shydeo Brandão Miyoshi and Antônio Ruffino-Netto. (2020) “Development of CART model for prediction of tuberculosis treatment loss to follow up in the state of São Paulo, Brazil: A case–control study.” *International Journal of Medical Informatics* **141**:104198.

[17] Carvalho, Isabelle, Mariane Barros Neiva, Newton Shydeo Brandão Miyoshi, Nathalia Yukie Crepaldi, Filipe Andrade Bernardi, Vinícius Costa Lima, Ketlin Fabri dos Santos, Ana Clara de Andrade Mioto, Mariana Tavares Mozini, Rafael Mello Galliez, Mauro Niskier Sanchez, Afrânio Lineu Kritski, and Domingos Alves. (2022) “Knowledge Discovery in Databases: Comorbidities in Tuberculosis Cases.” *International Conference on Computational Science:Computational Science – ICCS 2022* **13352**:3–13.

[18] Da Costa, Luana M.A., Filipe Andrade Bernardi, Tiago Lara Michelin Sanches, Afranio Lineu Kritski, Rafael Mello Galliez, and Domingos Alves. (2021) “Operational modeling for testing diagnostic tools impact on tuberculosis diagnostic cascade: A model design.” *Procedia Computer Science* **181**:650-657.

[19] Silva, Eveline de Almeida, Ulisses Umbelino dos Anjos and Jordana de Almeida Nogueira. (2014) “Modelo preditivo ao abandono do tratamento da tuberculose.” *Saúde em Debate* **38** (**101**):200–209.

[20] Kalhori, Sharareh Rostam Niakan and Xiao-Jun Zeng. (2013) “Evaluation and comparison of different machine learning methods to predict outcome of tuberculosis treatment course.” *Journal of Intelligent Learning Systems and Applications* **05** (**03**):184–193.

[21] Pellison, Felipe Carvalho, Rui Pedro Charters Lopes Rijo, Vinícius Costa Lima, Nathalia Yukie Crepaldi, Filipe Andrade Bernardi, Rafael Mello Galliez, Afranio Lineu Kritski, Kumar Abhishek, and Domingos Alves. (2020) “Data Integration in the Brazilian Public Health System for Tuberculosis: Use of the Semantic Web to Establish Interoperability.” *JMIR Medical Informatics* **8** (**7**):e17176

[22] Lima, Vinícius Costa, Filipe Andrade Bernardi, Michael Domingues, Afranio Lineu Kritski, Rui Pedro Charters Lopes Rijo, and Domingos Alves. (2022) “A computational infrastructure for semantic data integration towards a patient-centered database for Tuberculosis care.” *Procedia Computer Science* **196**:434–438.

[23] Crepaldi, Nathalia Yukie, Vinícius Costa Lima, Filipe Andrade Bernardi, Luiz Ricardo Albano dos Santos, Verena Hokino Yamaguti, Felipe Carvalho Pellison, Tiago Lara Michelin Sanches, Newton Shydeo Brandão Miyoshi, Antonio Ruffino-Netto, Rui Pedro Charters Lopes Rijo, and Domingos Alves. (2019) “SISTB: an ecosystem for monitoring TB.” *Procedia Computer Science* **164**:587–94.

5. DISCUSSÃO GERAL

Durante o desenvolvimento do Artigo 1, foi possível perceber que o levantamento de informações depende de uma série de fatores como, por exemplo, de quem está coletando, dos sujeitos participantes da pesquisa e das situações de vulnerabilidade em que podem se encontrar no momento do trabalho e dos instrumentos utilizados para a coleta (e.g.: registros médicos do paciente, questionários utilizados, entre outros) (BELLATO; PEREIRA, 1999).

O processo de coleta de dados se mostrou demorado devido a diversas limitações: 1) a indisponibilidade de uma conexão de internet acessível em todos os locais nos ambulatórios; 2) a existência de inúmeros protocolos burocráticos demorados para o acesso às informações dos pacientes; 3) a má organização física dessas informações para os casos na qual essas informações se referiam a prontuários físicos; 4) a divergência nas informações e semânticas encontradas entre diferentes fontes de dados.

Durante a análise dos dados coletados foram identificadas diversas inconsistências que podem ter sido causadas por erros de entrada do usuário, como por exemplo o preenchimento incorreto de informações no sistema ou prontuário do paciente pelo profissional de saúde, ou devido pela fadiga dos coordenadores de campo envolvidos na coleta de dados secundários.

Finalmente, o uso do TBWEB entre diferentes unidades de saúde para o controle e acompanhamento da TB no estado de São Paulo se mostrou mais disseminado. Isso é observável pelos maiores índices de completude das informações do sistema. Além disso, o aumento da confiabilidade das informações de uma forma geral garante que as informações coletadas no TBWEB são confirmadas pelos dados presentes nos outros sistemas. Isso garante a confiabilidade dos dados presentes no sistema.

A maior completude, confiabilidade, quantidade e abrangência estadual e temporal das informações presentes no sistema TBWEB nos levou a considerá-lo como nossa base principal para o desenvolvimento dos Artigos 2, 3 e 4.

No Artigo 2 foi possível observar pelos resultados que o Índice de Comorbidade de Charlson (ICC) e a taxa de sobrevivência estimada em 10 anos poderiam ser usados como preditores para classificar entre os desfechos (morte e curta) do tratamento da TB. Essas correlações também são validadas por KUO et al. (2013) no estudo sobre incidência e desfecho da TB recém-diagnosticada. Além disso, a análise dos demais atributos, por meio do uso da técnica *Tree Bagging*, serve como meio de comparação da importância do ICC como preditor para o desfecho do tratamento da TB. Esses achados sugerem que um novo índice de comorbidade e preditor para o desfecho do tratamento da TB poderia ser desenvolvido através da extensão do ICC, levando em consideração os atributos que apresentaram maior importância como preditores. Vários autores que apresentam uma nova estrutura de causa de morte sugerem o desenvolvimento e validação de índices de comorbidade existentes (KIL et al., 2012; PARK et al., 2015; TAKENAKA et al., 2017). Tal índice ou score pode ser desenvolvido utilizando uma técnica semelhante à técnica aplicada por ARROYO et. al (2019) que transforma diversos preditores em um score de probabilidade para os diferentes desfechos do tratamento de tuberculose por meio dos nomogramas.

Além disso, o estudo permitiu que desenvolvêssemos um ranqueamento de importância dos atributos para prever se um tratamento terá como desfecho o abandono.

No Artigo 3, desenvolvemos um modelo CART (*Classification And Regression Trees*) para predição do abandono do tratamento de TB. O modelo enfatiza a utilização dos atributos identificados durante o estudo desenvolvido no Artigo 2. O modelo final possui uma capacidade preditiva que se equipara a outros já apresentados na literatura (PEETLUK, et al. 2021) e superior a outros modelos já desenvolvidos para estudos utilizando a mesma base (ARROYO, et. al. 2019). Além disso, pelo modelo se derivar de uma árvore preditiva apresenta uma abordagem muito mais intuitiva e compreensível para profissionais de saúde quando comparado a outros modelos como RNAs e modelos estatísticos.

Já que a árvore de decisão proporciona ao profissional de saúde um desfecho do tratamento baseado em uma sequência de decisões a serem tomadas baseadas em características apresentadas pelo paciente ou durante o tratamento.

Por fim, ressaltamos que o modelo enfatiza variáveis que já eram conhecidas e identificadas como preditoras para o abandono do tratamento ou cura do paciente em outros estudos.

No Artigo 4, aplicamos uma nova metodologia baseada na definição de caminhos clínicos para o desenvolvimento de outro classificador para o desfecho do tratamento de TB. Este, diferentemente do modelo desenvolvido no Artigo 3, utiliza a ordem e relação cronológica dos eventos clínicos que descrevem o diagnóstico clínico e esquema medicamentoso ao longo do tratamento de TB do paciente. Preditores até então pouco explorados na literatura (PEETLUK et. al 2021), mas que podem apresentar uma alta capacidade preditiva por serem capazes de segmentar e agrupar tratamentos que possuem alta correlação com determinados desfechos (APUNIKE, et. al 2020).

Isso possibilitou o desenvolvimento de um modelo de agrupamento não hierárquico que agrupou os caminhos clínicos de todos os tratamentos em 7 grupos. Estes, após serem relacionados aos seus principais desfechos, foram utilizados como preditores para o desfecho do tratamento de TB. Possibilitando também uma interpretação intuitiva, pois cada grupo pode ser representado por seu caminho clínico mais comum.

6. CONCLUSÕES

Existem várias inconsistências entre as diferentes fontes de dados utilizadas no controle da TB no sistema de saúde brasileiro (Artigo 1). Além disso, é perceptível a importância do papel que os padrões de saúde desempenham nos sistemas de interoperabilidade e terminologias para a manutenção e avaliação da qualidade dos dados. Sem a qualidade de dados adequada qualquer esforço colocado nos processos de tomada de decisão será infrutífero.

A escolha do TBWEB como base de dados principal se deu ao seu melhor grau de completude de informações e a confiabilidade apresentada quando comparado com outros sistemas.

No Artigo 2, é possível observar quais variáveis apresentam maior relação de importância para o desfecho da TB. Também, observamos que o ICC tem um alto potencial como preditor para a classificação dos desfechos do tratamento da TB e observarmos diferentes atributos que podem atuar como preditor para os resultados do tratamento da TB. Estes podem ser usados como preditores para descrever determinados desfechos da TB como o óbito, o abandono, e até mesmo a cura.

No Artigo 3 é desenvolvido o modelo preditivo de abandono ao tratamento da TB através do CART utilizando como base os resultados obtidos nos Artigos 1 e 2. Embora o fato de a capacidade preditiva ideal não tenha sido alcançado, parece razoável propor o uso de modelos CART devido a sua fácil interpretação. Além disso, a capacidade preditiva do modelo desenvolvido se equipara a outros já apresentados na literatura (PEETLUK, et al. 2021) e superior a outros modelos já desenvolvidos para estudos utilizando a mesma base (ARROYO, et. al. 2019).

Por fim, no Artigo 4 construímos outro modelo preditivo de abandono utilizando como base os agrupamentos gerados a partir de um modelo de aprendizado não supervisionado hierárquico. Para isso foi necessário realizar o agrupamento hierárquico de todos os caminhos clínicos existentes na base do TBWEB.

O modelo gerado a partir dos agrupamentos apresentou um desempenho significativamente melhor que o desenvolvido no Artigo 3. Dessa forma, sugerindo

que a relação cronológica entre os eventos clínicos transcorridos ao longo do tratamento tenha grande importância na determinação do desfecho do mesmo. Além disso, como visto anteriormente, o modelo explora um conjunto de preditores pouco utilizados na literatura (PEETLUK, et al. 2021) obtendo uma capacidade preditiva melhor que outros modelos que fazem uso da mesma base de dados (ARROYO, et. al. 2019).

As possíveis limitações do projeto foram já discutidas no transcorrer dos textos respectivos a cada trabalho publicado.

REFERÊNCIAS

APUNIKE, A.; CIABATI, L.; SANCHES, T.; OLIVEIRA, L.; SANCHEZ, M.; GALLIEZ, R.; ALVES, D. Analyses of public health databases via clinical pathway modelling: TBWEB. **Computational Science – ICCS 2020: 20th International Conference, Amsterdam, The Netherlands**, v. 12140, p. 550-562, 2020. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7303689>.

ARROYO, L. H.; RAMOS, A. C. V.; YAMAMURA, M.; BERRA, T. Z.; ALVES, L. S.; BELCHIOR, A. S.; SANTOS, D. T.; ALVES, J. D.; CAMPOY, L. T.; ARCOVERDE, M. A. M.; BOLLELA, V. R.; BOMBARDA, S.; NUNES, C.; ARCÊNCIO, R. A. Predictive model of unfavorable outcomes for multidrug-resistant tuberculosis. **Revista de saude publica**, 53, 77, 2019. Disponível em: <https://doi.org/10.11606/s1518-8787.2019053001151>.

BELLATO, R.; PEREIRA, W. R. ALGUMAS REFLEXÕES SOBRE O TRABALHO DE CAMPO NA PESQUISA QUALITATIVA EM ENFERMAGEM Some througts about field work in nursing qualitative research. n. May 2008, p. 6–16, 1999.

CHRISTENSEN, A. J.; SMITH, T. W. Personality and patient adherence: Correlates of the five-factor model in renal dialysis. **Journal of Behavioral Medicine**, v. 18, n. 3, p. 305–313, 1995.

CREPALDI, N.; ORFÃO, N.; YOSHIURA, V.; VILLA, T.; RUFFINO-NETTO, A.; ALVES, D. DESENVOLVIMENTO E IMPLANTAÇÃO DE UM SISTEMA PARA GESTÃO DE PACIENTES DE TUBERCULOSE. **Revista da Faculdade de Medicina de Ribeirão Preto e do Hospital das Clínicas da FMRP**, v. 47, n. 1, p. 13–17, 2014.

DO PRADO, T. N.; DIETZE, R.; NETTO, A. R.; ZANDONADE, E.; MACIEL, E. L. N. PREDICTION OF DROPOUT TUBERCULOSIS TREATMENT ON PRIORITARY CITIES TO CONTROL IN ESPI ´RITO SANTO STATE, BRAZIL. **Journal of Epidemiology & Community Health**, v. 65, n. Suppl 1, p. A138–A138, 2011. Disponível em: <http://jech.bmj.com/cgi/doi/10.1136/jech.2011.142976e.51>.

GALDÓS TANGÜIS, H.; CAYLÀ, J. A.; GARCÍA DE OLALLA, P.; JANSÀ, J. M.; BRUGAL, M. T. Factors predicting non-completion of tuberculosis treatment among HIV- infected patients in Barcelona (1987-1996). **International Journal of Tuberculosis and Lung Disease**, v. 4, n. 1, p. 55–60, 2000.

GE, X.; RIJO, R.; PAIGE, R. F.; KELLY, T. P.; MCDERMID, J. A. Introducing Goal Structuring Notation to Explain Decisions in Clinical Practice. **Procedia Technology**, v. 5, p. 686–695, 1 jan. 2012. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2212017312005075>.

GONÇALVES, D.; RIJO, R.; GONÇALVES, R.; CRUZ, J. B.; VARAJÃO, J. Novos Desafios e Oportunidades de Investigação na Área da Gestão de Projectos de Desenvolvimento de Sistemas de Informação. **Conferência IADIS Ibero-Americana WWW/Internet**, n. October 2014, p. 324–29, 2007. Disponível em: <https://www.researchgate.net/profile/Ramiro_Goncalves/publication/266881838_NOVOS_DESAFIOS_E_OPORTUNIDADES_DE_INVESTIGACAO_NA_AREA_DA_GESTAO_DE_PROJECTOS_DE_DESENVOLVIMENTO_DE_SISTEMAS_DE_INFORMACAO/links/54427fd10cf2a6a049a89927/NOVOS-DESAFIOS-E-OPORTUNI>.

HARLING, G.; LIMA NETO, A. S.; SOUSA, G. S.; MACHADO, M. M. T.; CASTRO, M. C. Determinants of tuberculosis transmission and treatment abandonment in Fortaleza, Brazil. **BMC Public Health**, v. 17, n. 1, p. 1–10, 2017.

KALHORI, S. R. N.; ZENG, X. Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course. **Journal of Intelligent Learning Systems and Applications**, v. 05, n. 03, p. 184–193, 2013. Disponível em: <http://www.scirp.org/journal/PaperDownload.aspx?DOI=10.4236/jilsa.2013.53020>.

KIL, S.-R.; LEE, S.-I.; KHANG, Y.-H.; LEE, M.-S.; KIM, H.-J.; KIM, S.-O.; JO, M.-W. Development and validation of comorbidity index in South Korea. **International Journal for Quality in Health Care**, v. 24, n. 4, p. 391–402, 1 ago. 2012. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/22718515>.

KUO, S. C.; CHEN, Y. T.; LI, S. Y.; LEE, Y. T.; YANG, A. C.; CHEN, T. L.; LIU, C. J.; CHEN, T. J.; SU, I. J.; FUNG, C. P. Incidence and outcome of newly-diagnosed tuberculosis in schizophrenics: A 12-year, nationwide, retrospective longitudinal study. **BMC Infectious Diseases**, v. 13, n. 1, 2013.

LIPPEVELD, T.; SAUERBORN, R.; BODART, C. Design and implementation of health information systems Edited by. **World Health Organization**, p. 280, 2000.

LOURENÇO, J.; SANTOS-PEREIRA, C.; RIJO, R.; CRUZ-CORREIA, R. Service Level Agreement of Information and Communication Technologies in Portuguese Hospitals. **Procedia Technology**, v. 16, p. 1397–1402, 1 jan. 2014.

Disponível em: <https://www.researchgate.net/profile/Ramiro_Goncalves/publication/266881838_NOVOS_DESAFIOS_E_OPORTUNIDADES_DE_INVESTIGACAO_NA_AREA_DA_GESTAO_DE_PROJECTOS_DE_DESENVOLVIMENTO_DE_SISTEMAS_DE_INFORMACAO/links/54427fd10cf2a6a049a89927/NOVOS-DESAFIOS-E-OPORTUNI>.

MARTINHO, R.; RIJO, R.; NUNES, A. Complexity Analysis of a Business Process Automation: Case Study on a Healthcare Organization. **Procedia Computer Science**, v. 64, p. 1226–1231, 1 jan. 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050915026459>.

MINISTÉRIO DA SAÚDE. Manual de Administração/ Programa Nacional de Controle da Tuberculose. **Boletim de Pneumologia Sanitária**, v. 4, p. 7–56, 1996.

MINISTÉRIO DA SAÚDE. **Plano Nacional pelo Fim da Tuberculose como Problema de Saúde Pública**, 2017.

MINISTÉRIO DA SAÚDE. Secretaria de Vigilância em Saúde. Departamento de Vigilância das Doenças Transmissíveis. Manual de Recomendações para o Controle da Tuberculose no Brasil / Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância das Doenças Transmissíveis. – Brasília: Ministério da Saúde, 2019.

OPAS, O. P. D. L. S.; OMS, O. M. de la S. Reunión regional de directores nacionales de programas de control de la tuberculosis: informe final. 1997.

PARK, J. Y.; KIM, M. H.; HAN, S. S.; CHO, H.; KIM, H.; RYU, D. R.; KIM, H.; LEE, H.; LEE, J. P.; LIM, C. S.; KIM, K. H.; JOO, K. W.; KIM, Y. S.; KIM, D. K.; DO, J. Y.; SONG, S. H.; KIM, S. E.; KIM, S. H.; KIM, Y. H.; LEE, J. S.; JIN, H. J.; LIM, C. S.; LEE, J. P.; CHANG, J. H.; YOO, T. H.; PARK, J. T.; OH, H. J.; PARK, H. C.; CHANG, T. I.; RYU, D. R.; OH, D. J.; CHANG, Y. S.; KIM, Y. O.; KIM, S. H.; JIN, D. C.; KIM, Y. K.; KIM, H. Y.; KIM, W.; LEE, K. W.; LEE, C. S. Recalibration and validation of the Charlson comorbidity index in Korean incident hemodialysis patients. **PLoS ONE**, v. 10, n. 5, p. 1–14, 2015.

PEETLUK, L. S.; RIDOLFI, F. M.; REBEIRO, P. F.; LIU, D.; ROLLA, V. C.; STERLING, T. R. Systematic review of prediction models for pulmonary tuberculosis treatment outcomes in adults. **BMJ open**, v.11, n.3, p. e044687, 2021. Disponível em: <https://doi.org/10.1136/bmjopen-2020-044687>

RUFFINO-NETTO, A. Tuberculose: a calamidade negligenciada. **Revista da**

Sociedade Brasileira de Medicina Tropical, v. 35, n. 1, p. 51–58, 2002.

SHORTLIFFE, E. H.; PERREAULT, L. E. **Medical Informatics**. New York: Springer New York, 2001.

SILVA, E. de A.; ANJOS, U. U. dos; NOGUEIRA, J. de A. Modelo preditivo ao abandono do tratamento da tuberculose. **Saúde em Debate**, v. 38, n. 101, p. 200–209, 2014. Disponível em: <http://www.gnresearch.org/doi/10.5935/0103-1104.20140018>.

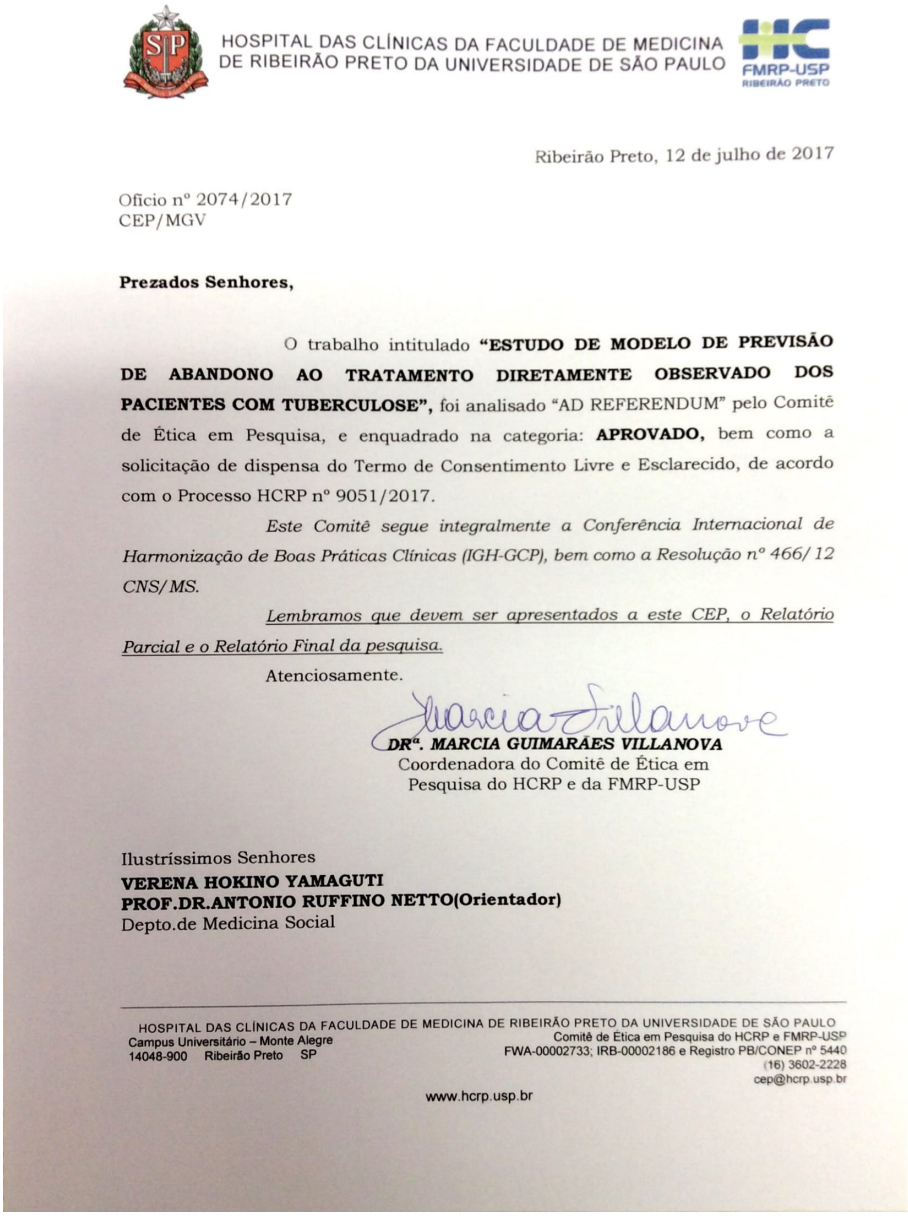
TAKENAKA, Y.; TAKEMOTO, N.; OYA, R.; ASHIDA, N.; KITAMURA, T.; SHIMIZU, K.; TAKEMURA, K.; MICHIBA, T.; HANAMOTO, A.; SUZUKI, M.; YAMAMOTO, Y.; UNO, A.; INOHARA, H. Development and validation of a new comorbidity index for patients with head and neck squamous cell carcinoma in Japan. **Scientific Reports**, v. 7, n. 1, p. 1–11, 2017. Disponível em: <http://dx.doi.org/10.1038/s41598-017-07752-1>.

WHO. **Global Tuberculosis Report 2022**.

ZAVASCKI, A. P.; FUCHS, S. C. The need for reappraisal of AIDS score weight of Charlson comorbidity index. **Journal of Clinical Epidemiology**, v. 60, n. 9, p. 867–868, 2007.

ANEXOS

Anexo 1: Carta de aprovação do Comitê de Ética em Pesquisa (CEP)



HOSPITAL DAS CLÍNICAS DA FACULDADE DE MEDICINA
DE RIBEIRÃO PRETO DA UNIVERSIDADE DE SÃO PAULO

FMRP-USP
RIBEIRÃO PRETO

Ribeirão Preto, 12 de julho de 2017

Ofício nº 2074/2017
CEP/MGV

Prezados Senhores,

O trabalho intitulado **"ESTUDO DE MODELO DE PREVISÃO DE ABANDONO AO TRATAMENTO DIRETAMENTE OBSERVADO DOS PACIENTES COM TUBERCULOSE",** foi analisado "AD REFERENDUM" pelo Comitê de Ética em Pesquisa, e enquadrado na categoria: **APROVADO,** bem como a solicitação de dispensa do Termo de Consentimento Livre e Esclarecido, de acordo com o Processo HCRP nº 9051/2017.

Este Comitê segue integralmente a Conferência Internacional de Harmonização de Boas Práticas Clínicas (IGH-GCP), bem como a Resolução nº 466/12 CNS/MS.

Lembramos que devem ser apresentados a este CEP, o Relatório Parcial e o Relatório Final da pesquisa.

Atenciosamente.

DRª. MARCIA GUIMARÃES VILLANOVA
Coordenadora do Comitê de Ética em
Pesquisa do HCRP e da FMRP-USP

Ilustríssimos Senhores
VERENA HOKINO YAMAGUTI
PROF.DR.ANTONIO RUFFINO NETTO(Orientador)
Depto.de Medicina Social

HOSPITAL DAS CLÍNICAS DA FACULDADE DE MEDICINA DE RIBEIRÃO PRETO DA UNIVERSIDADE DE SÃO PAULO
Campus Universitário – Monte Alegre
14048-900 Ribeirão Preto SP
Comitê de Ética em Pesquisa do HCRP e FMRP-USP
FWA-00002733; IRB-00002186 e Registro PB/CONEP nº 5440
(16) 3602-2228
cep@hcrp.usp.br
www.hcrp.usp.br

Anexo 2: Aceite de Publicação no HCist 2022

Verena Hokino Yamaguti <verena.yamaguti@usp.br>

[HCist 2022] Your submission 552 has been accepted as a FULL PAPER

1 message

HCist 2022 <francesca.m.bianchi@gmail.com> Fri, Jul 22, 2022 at 2:01 PM
Reply-To: francesca.m.bianchi@gmail.com
To: verena.yamaguti@usp.br

***
Please excuse us for repeating this notification, but many authors are reporting that they are not receiving our emails or found the emails in the spam box.
Hence we are sending this notification from another email account.
***

Dear Author,

On behalf of the HCist 2022 - International Conference on Health and Social Care Information Systems and Technologies Organising Committee, we are pleased to inform that your submission titled

552 - Clinical Pathways and Hierarchical Clustering for Tuberculosis Treatment Outcome Prediction

has been accepted for presentation at the conference and for inclusion in the Conference Proceedings as a FULL PAPER.

Congratulations!

Please note that the publication of the paper is conditioned to the corrections indicated by the reviewers, and to the submission of the final/camera ready version of your paper strictly following the formatting guidelines.

We have included the reviewers' comments at the end of this message, which must be taken into account when preparing the final/camera ready version of the paper.

Please have in mind that a FULL paper should have between six to eight pages in length, considering the template available at http://hcist.scika.org/CONTENTS/downloads/template.docm, and the guidelines provided at http://hcist.scika.org/?page=submissionguidelines

Submissions will open next July 22.

Please CAREFULLY follow the steps below to proceed with your final/camera ready submission (hard deadline: August 31, 2022):

1. Access the final submissions platform at http://scms.sciencesphere.org. If you have already created an account for previous conference editions, you can jump directly to step 4;

2. Create an account by clicking the "Sign up" (green) button, and filling in the required info. A confirmation link will be sent to your email;

3. Go to your email inbox, and look for an email from Conferences ScMS <scms@sciencesphere.org>, with the subject "Conferences ScMS - New account created". Follow the link provided, and you'll see (again) the website http://scms.sciencesphere.org, with a confirmation message at the top "Your account was activated. You can now sign in."

4. Click the "Sign in" (blue) button, fill in your account's username and password and click the "Login" (purple) button. You'll be redirected to your home/dashboard, where you'll be able see your profile info, and to submit/edit all your papers, invoice requests, registrations and payment proofs;

5. Click the "My Profile" (red) button, and fill it in as much as you can. It will be useful later on;

6. To submit the final/camera ready version of your paper(s), go back to your home (dashboard) web page, and click the "Papers" (green) button. On the papers list web page, click "Add new", and fill in all the required info. A link for downloading the copyright form template will be available once you select the according Conference of your paper;

7. Then, go back to your dashboard, click the "Registrations" (black) button, to register yourself and/or any other authors to participate in the Conference. Click "Add new" at the list of registrations web page, select the according

Conference, and fill in your personal information. Select the paper(s) under "Publications" which this registration refers to, and select the registration fees that apply;

8. You can then go back to your dashboard, and submit a payment proof (pdf) file similarly, by clicking the "Payment proofs" (purple) button, and the "Add new" one next. Select the according Conference, and upload a pdf file with the payment proof. You can also select the corresponding registration, and fill in the associated amount;

9. Finally, you can file for one or more "Invoice requests" (blue button at your home/dashboard web page). Once you click the "Add new" button, you can fill in the invoice request for an already created registration, or simply fill in the info manually. Please be sure to add a custom description if you need one!

Note 1: Your registration will only be considered as "valid" after all information is submitted;
Note 2: Warnings will be shown in your home/dashboard page in case of any inconsistency between existing registration values, invoice requests and payment proofs;
Note 3: You can edit/complement all this information as many times as you want until the final/camera ready deadline (August 31, 2022).

For every accepted paper, at least one registration fee must be paid by August 31, so that the paper can be included in the conference program and published.

Thank you very much for your work, which will greatly contribute to the high quality standards of this conference!

See you next November, on-line or in-person in Lisbon, for another great HCist conference!

Sincerely,
The Program Committee
****

Reviewer's comments:

****

The subject of the work is of significant importance to the scientific community.
The article superficially (perhaps intentionally) describes the problem and a more efficient way of dealing with it.
However, I believe that the abstract could be better written, providing better context to the subject and describing the proposed solution more explicitly.

****

This is a very interesting and relevant contribution. Nice work! Good, methodology, relevant literature. I would suggest to improve the discussion of the achieved results.

****

The authors present an interesting study about a predictive model of TB treatment. The methodoly is well introduced, and the discussion of results is rich enough to assure a good contribution to Hcist.

****