Tese de Doutorado

O uso de método de relacionamento de dados (*record linkage*) para integração de informação em sistemas heterogêneos de saúde: estudo de aplicabilidade entre níveis primário e terciário.

Dra. Kátia Suzuki
Orientador: Prof. Dr. Paulo Mazzoncini de Azevedo Marques

Área de Concentração: Clínica Médica

Universidade de São Paulo
Faculdade de Medicina de Ribeirão Preto

Sumário

Contextualização
Considerações Teóricas
- Relacionamento Determinístico
- Relacionamento Probabilístico

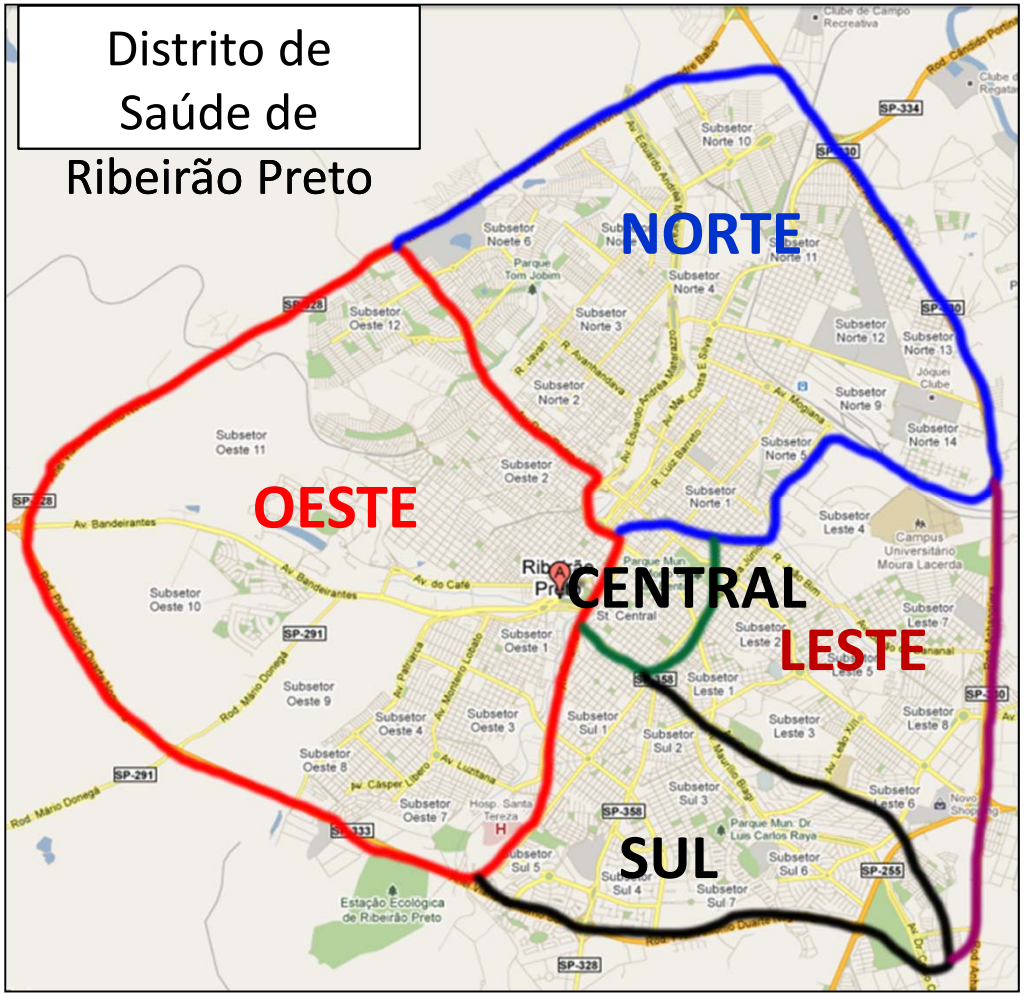
Objetivos
Materiais
Metodologia
Resultados e Discussão
Conclusão

Contextualização

O município de RP está dividido em 5 Distritos de Saúde e a FMRP/USP é corresponsável, em conjunto com a SMS pelo distrito de saúde oeste.

Composto por:

- UBDS
- UBS
- NFS e
- Hospitais.



Contextualização

A SMS-Ribeirão Preto adotou o sistema informatizado ***HygiaWeb*** em suas unidades de saúde.

E o HCFMRP possui um sistema informatizado proprietário.

Contextualização

Considerando os princípios básicos do SUS, é possível conhecer o fluxo dos pacientes dentro da rede de atenção à saúde do município de Ribeirão Preto?

Considerando a falta de identificadores únicos nas bases de dados de saúde e a falta de padronização dos dados nos sistemas informatizados é possível integrar a base de dados utilizada na atenção básica (***HygiaWeb***) com a base de dados da atenção terciária (**HCFMRP**)?

Contextualização

Sim.

Através de técnicas computacionais automatizadas.
Relacionamento de Dados (*record linkage*)

A integração dessas bases heterogêneas poderá contribuir para a recuperação de informações dos usuários, dentro da cadeia de atenção à saúde que está organizada de forma descentralizada e hierarquizada.

Considerações Teóricas

Relacionamento de Dados – *Record Linkage*

Originou-se na área da saúde pública,
e foi encontrado pela primeira vez no trabalho do Dr. Halbert Dunn, chefe do *The U.S. National Office of Vital Statistics*, no Canadá.

(DUNN, 1946).

Considerações Teóricas

Relacionamento de Dados – *Record Linkage*

Definição: processo de comparação entre dois ou mais registros em diferentes bases de dados, que contêm informações de identificação suficientes para determinar se estes registros referem-se à mesma pessoa, ou mais genericamente, a uma entidade.

(HOWE, 1988).

Considerações Teóricas

Relacionamento de Dados – *Record Linkage*

Classificado em:

- Manual
- Determinístico e
- Probabilístico.

Alguns autores definem que o relacionamento de dados está dividido em dois grupos:

- Determinístico ou baseado em regras e
- Probabilístico.

(CHURCHES et al., 2002).

Considerações Teóricas

Relacionamento de Dados – *Record Linkage*

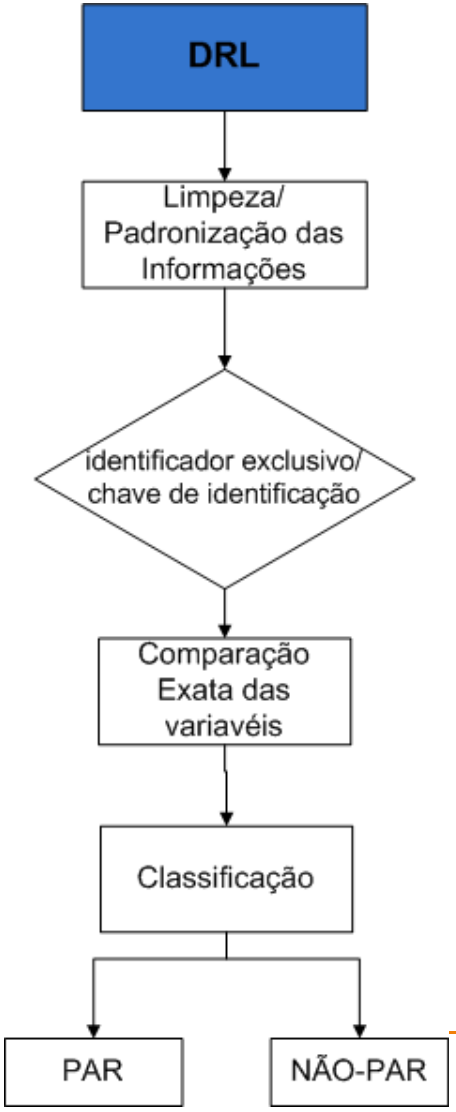
Manual

Comparação manual dos registros entre duas bases de dados para se decidir se são pares ou não.

Desvantagem: processo muito trabalhoso e às vezes pode não ser viável, em virtude da quantidade de dados envolvida no relacionamento.

Considerações Teóricas

Relacionamento de Dados – DRL



Determinístico

Deterministic Record Linkage (DRL)

Comparação exata de um identificador exclusivo ou de um conjunto de identificadores, denominada chave identificadora, comum em ambas as bases de dados e que permite a discriminação, classificando-os como pares ou não-pares.

(LI et al., 2006; GOMATAM et al, 2002).

Considerações Teóricas

Relacionamento de Dados – DRL

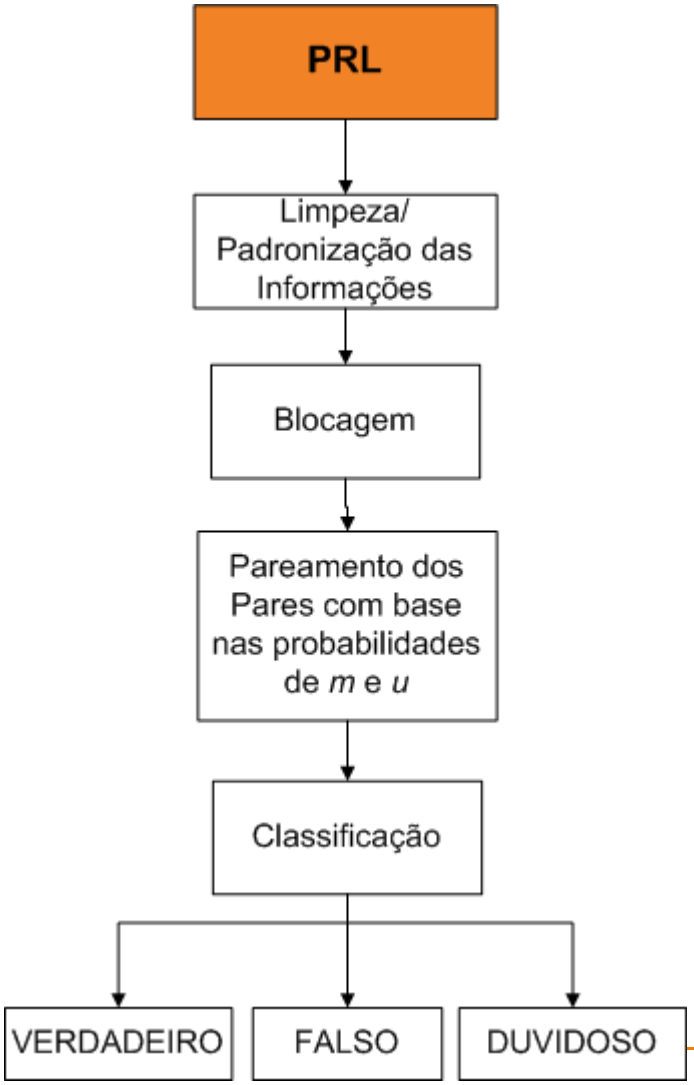
DRL

Vantagem: simples entendimento e implementação, principalmente em virtude da inexistência de conceitos estatísticos.

Desvantagem: Já em situações em que há a necessidade de solucionar questões de subjetividade, a simplicidade do método pode ser comprometida tornando-se laboriosa e consumindo muito tempo.

Considerações Teóricas

Relacionamento de Dados – PRL



Probabilístico

Probabilistic Record Linkage (PRL)

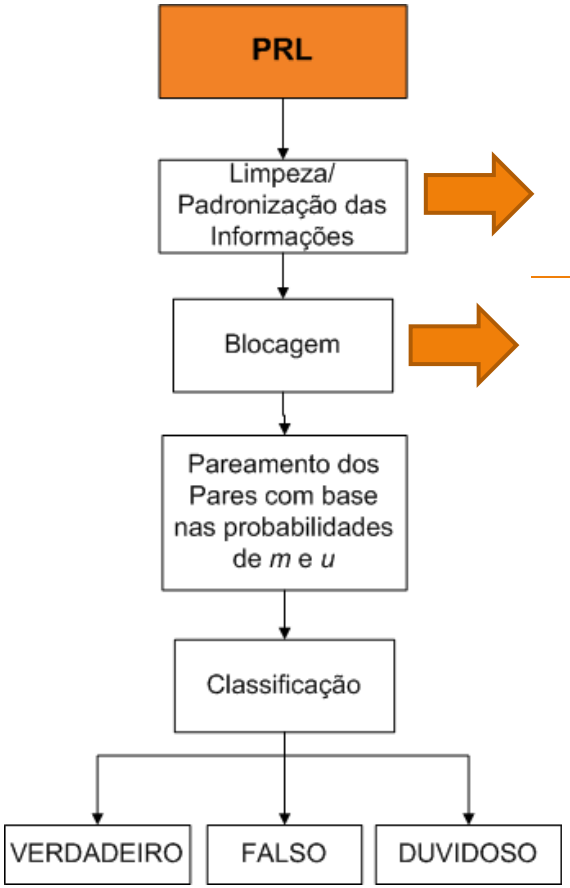
Uso de variáveis comuns existentes para relacionar bases de dados utilizando a estimativa de dois parâmetros de probabilidade *m* e *u* para cada variável comum.

A sua teoria estatística foi fundamentada por Fellegi e Sunter.

(CHRISTEN; CHURCHES, 2006).

Considerações Teóricas

Relacionamento de Dados – PRL

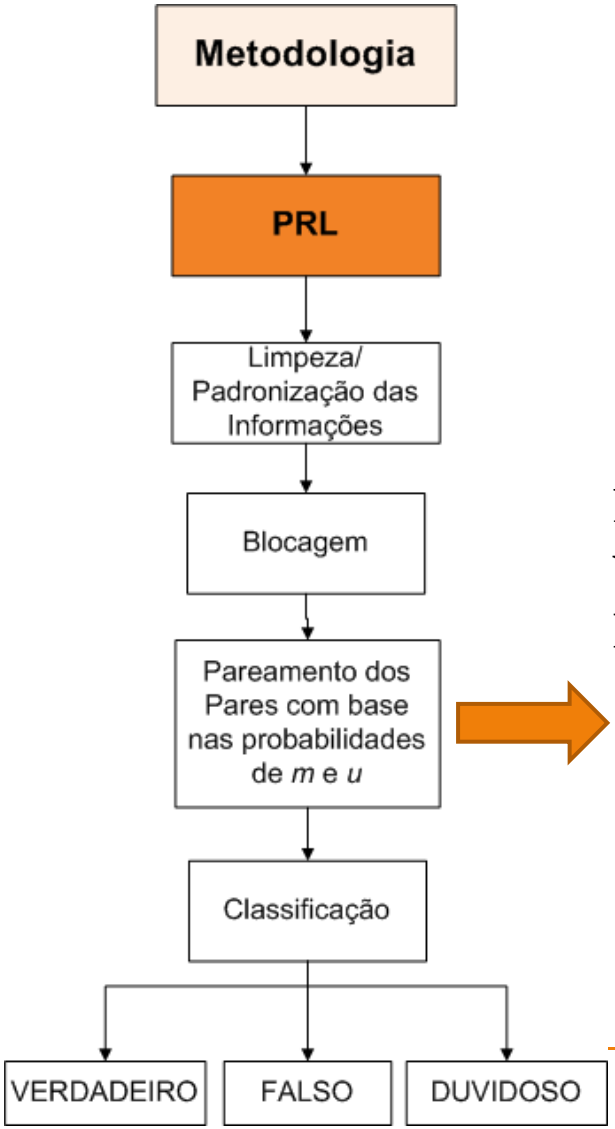


Limpeza e Padronização dos dados para potencializar a qualidade das informações.

Realizar a criação de blocos lógicos para minimizar o tempo de comparação dos registros.

As comparações são realizadas somente entre os registros correspondentes aos blocos.

Aplicando PRL



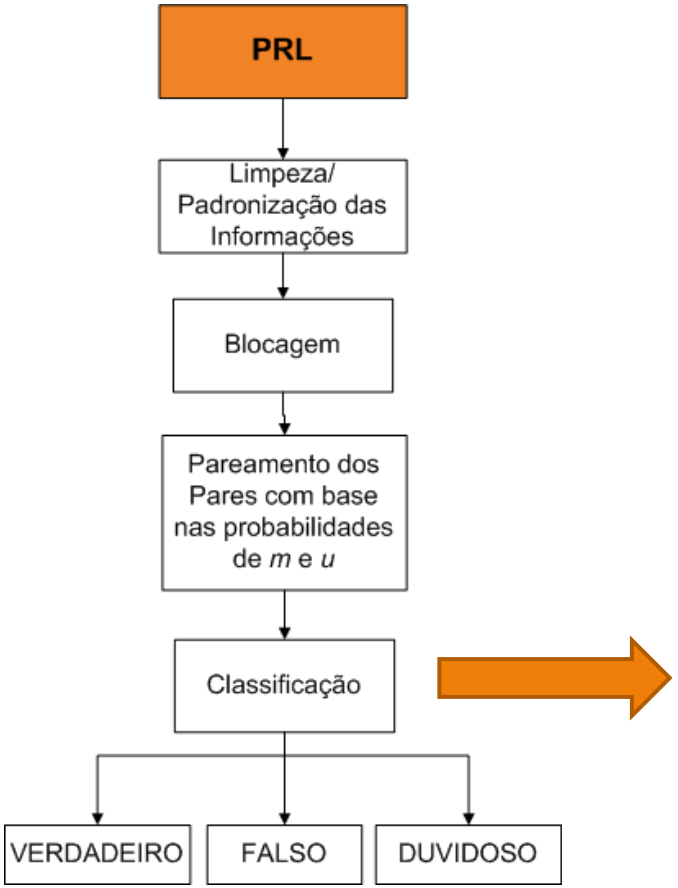
Pareamento de registros

1. Estimar a probabilidade de *m* e *u;*
2. Calcular os pesos de concordância e discordância;
3. Calcular o escore de cada par;
4. Definir a função de comparação.

No contexto deste trabalho foram adotados os valores estimados pelo algoritmo EM para as probabilidades *m* e *u*.

Considerações Teóricas

Relacionamento de Dados – PRL



Classificação

A classificação ocorrerá de maneira automática dos pares de acordo com os valores de limiar superior e inferior, onde:

- **Verdadeiro** = Pares com scores superiores ou iguais ao limiar superior;
- **Falso** = Pares com scores inferiores ou iguais ao limiar inferior;
- **Duvidoso** = os pares com scores entre o limiar superior e inferior. Estes poderão ser revisados manualmente.

Considerações Teóricas

Relacionamento de Dados – PRL

PRL

Desvantagem: Complexidade do método, principalmente para estimar os parâmetros de probabilidade *m* e *u*.

Objetivos Gerais/Específicos

Gerais

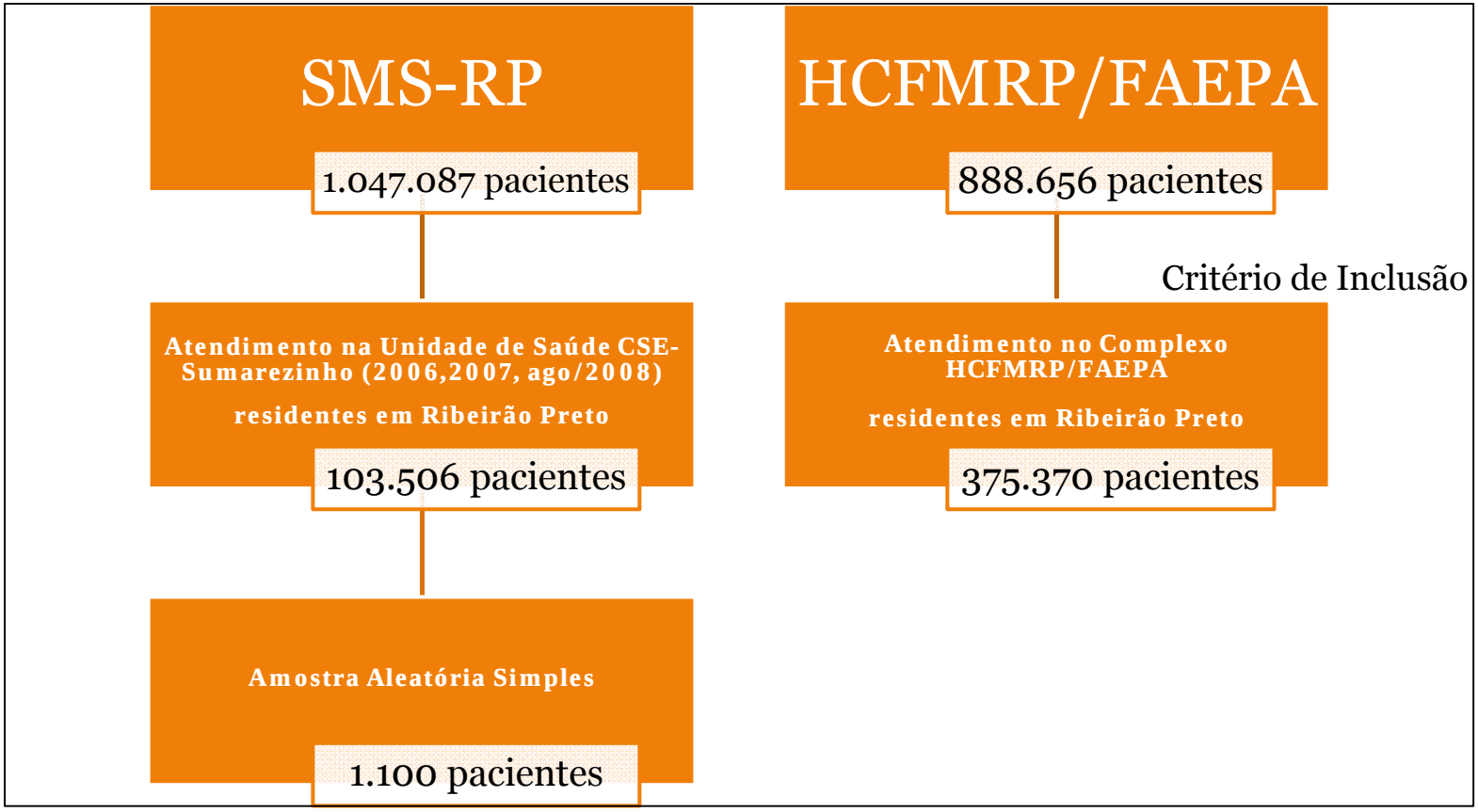
Aplicar o ***Record Linkage*** nas bases de dados Hygiaweb e HCFMRP/USP para **identificar a estratégia apropriada a ser adotada em bases de dados nacionais da área da saúde**.

Específicos

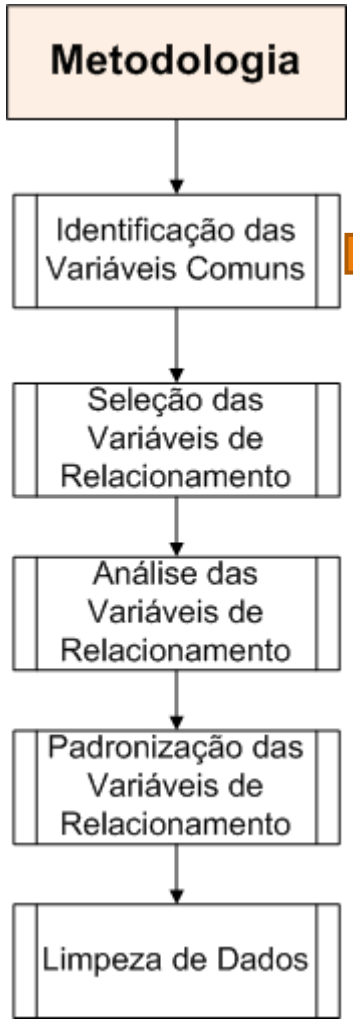
- Avaliar a qualidade das informações;
- Avaliar os resultados da estratégia DRL exato, DRL com discordância em uma variável e baseado em métricas de similaridade (*Dice*, *Jaro*, *Jaro-Winkler e Levenshtein*);
- Avaliar os resultados PRL;
- Comparar o desempenho dos métodos abordados.

Materiais

População do Estudo/Amostragem

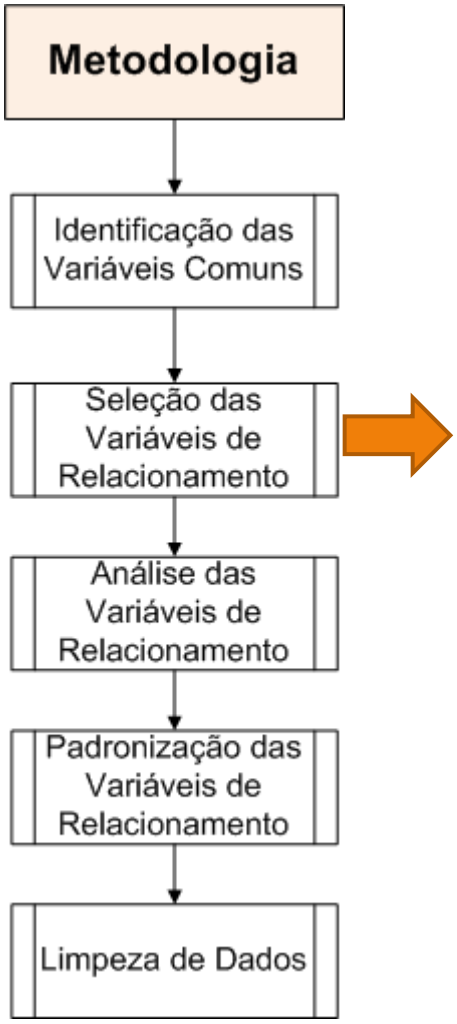


Metodologia – Variáveis comuns existente nas bases de dados



SMS-RP/Hygiaweb		HCFMRP	
Variável	Tipo de Dados	Variável	Tipo de Dados
CPF	Number(11)	CPF	Varchar(11)
Nome do paciente	Varchar (70)	Nome paciente	Varchar (60)
		Sobrenome paciente	Varchar (30)
Sexo	Varchar(1)	Sexo	Varchar (1)
Data de Nascimento	Date	Data de Nascimento	Date
Raça	Varchar(15)	Cor	Number (1)
Nome da Mãe	Varchar(70)	Nome da Mãe	Varchar(45)
Nome do Pai	Varchar(70)	Nome do Pai	Varchar(45)
Estado Civil	Varchar(100)	Estado Civil	Number(1)
Cidade	Varchar(50)	Cidade	Varchar(60)
Estado	Varchar(2)	Estado	Varchar(2)
Tipo de Endereço	Varchar(5)	Tipo de Logradouro	Varchar (10)
Logradouro	Varchar(100)	Endereço	Varchar(60)
Número	Varchar(5)	Número	Varchar(10)
Complemento	Varchar(50)	Complemento	Varchar(20)
CEP	Varchar(8)	CEP	Char(8)

Metodologia – Variáveis comuns existente nas bases de dados

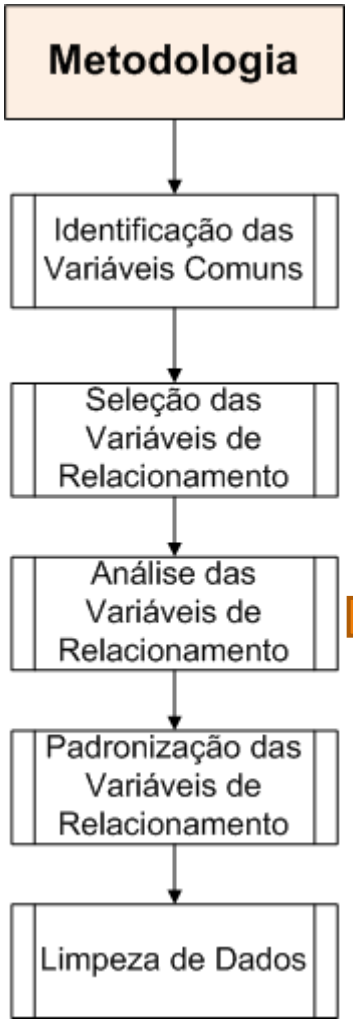


Foram identificadas 29 variáveis comuns entre as bases de dados CSE-Sumarezinho e HCFMRP.

Somente 11 variáveis possuíam mais de 50% dos dados preenchidos. São elas:

- Dados Pessoais
  - Nome do paciente, **Data de Nascimento, Sexo, Nome da Mãe**, Nome do Pai, Raça e Naturalidade;
- Identificação de Endereço
  - Logradouro, CEP, Cidade e Estado.

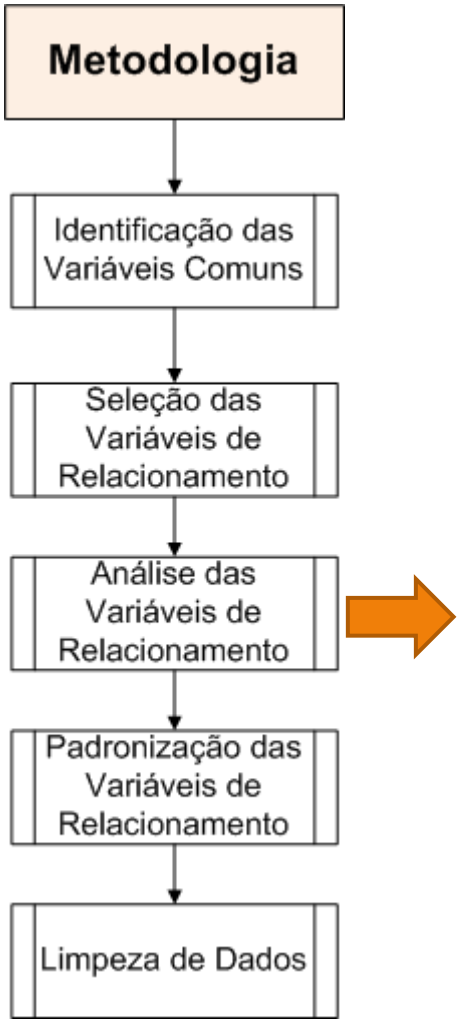
Metodologia – Análise das variáveis de relacionamento



Quanto à completude

Variável	CSE-Sumarezinho	HCFMRP/USP
Nome	1	1
Nome da Mãe	1	0,99668
Sexo	1	0,99998
Ano de Nascimento	1	0,99296
Data de Nascimento	1	0.99296

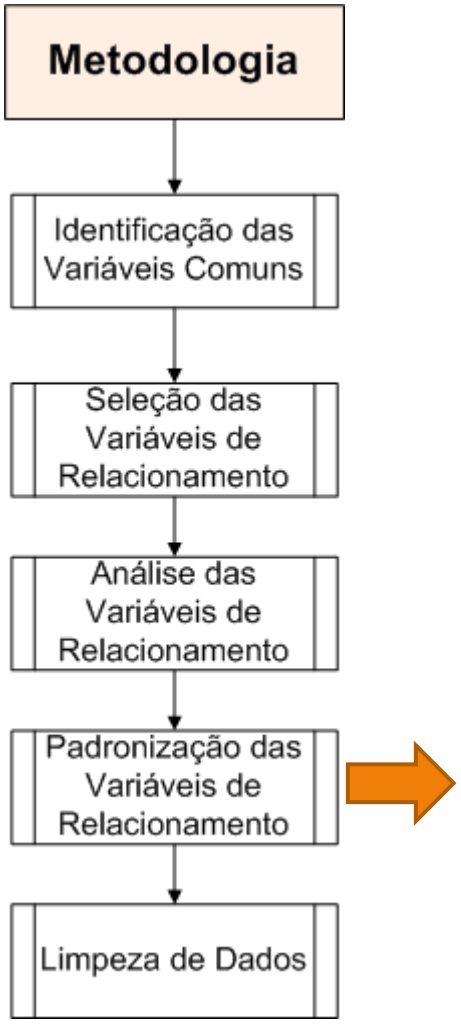
Metodologia – Análise das variáveis de relacionamento



Quanto à consistência dos dados

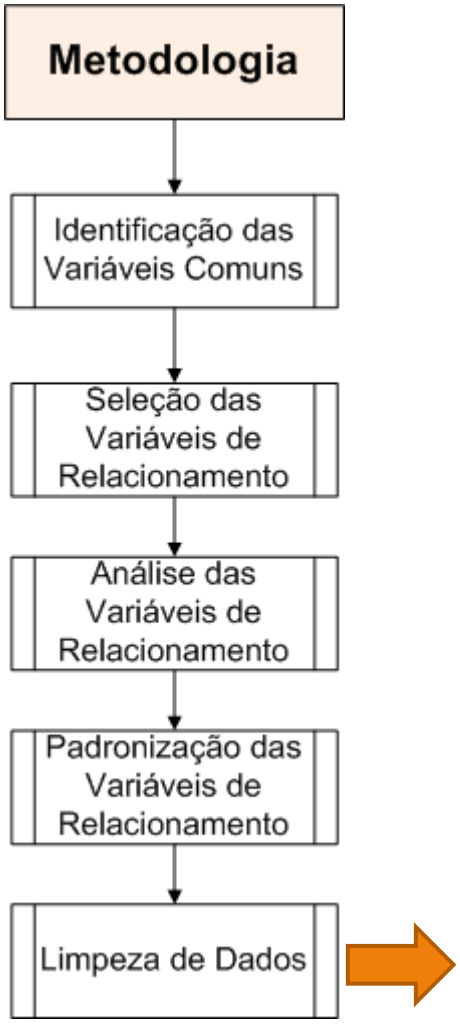
	Nome		Nome da Mãe		Data Nasc.	
Valor/Variável	**CSE-S**	**HCFMRP**	**CSE-S**	**HCFMRP**	**CSE-S**	**HCFMRP**
RN	0,27	7,20	0,00	0,00	-	-
IGN, IGNORADO, FALECIDO, AUS DESCONHECIDO	0,00	0,00	0,73	0,32	-	-
Caracteres Especiais (*, ?, #, -, /)	0,00	0,00	0,00	19,61	-	-
Ano > 2008	-	-	-	-	0,00	0,00

Metodologia – Padronização das variáveis de relacionamento



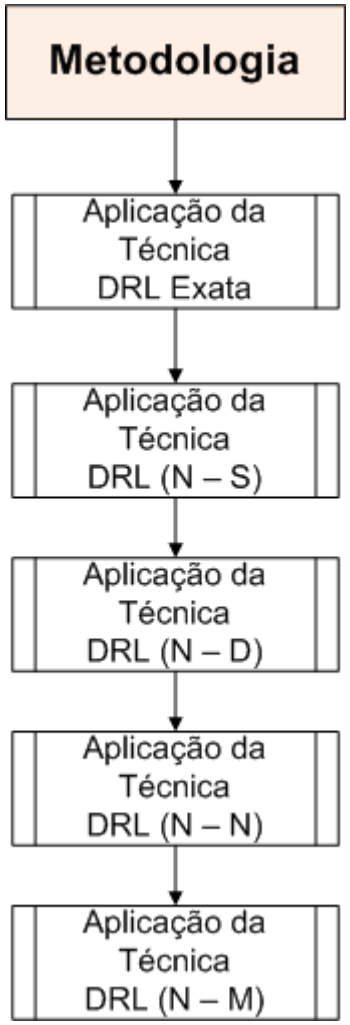
Variável	HCFMRP/USP	CSE-Sumarezinho	Código Padronizado
Sexo	Varchar(1) F = Feminino M = Masculino D = Desconhecido	Char(1) F = Feminino M = Masculino	Varchar(1) F = Feminino M = Masculino
Data Nascimento	Date ano-mês-dia (aaaa-mm-dd)	Date dia/ mês/ano (dd/mm/aaaa)	Date dia/ mês/ano (dd/mm/aaaa)
Nome	Nome - Varchar(60) Sobrenome – Varchar (30)	Nome – Varchar(70)	Varchar(70)
Nome da Mãe	Varchar(45)	Varchar(70)	Varchar(70)

Metodologia



Foram submetidas ao processo de limpeza de dados as variáveis “nome” e “nome da mãe”, totalizando 74.829 registros, ou seja, aproximadamente 20% dos dados.

Aplicando DRL Exato e DRL (N-1)



Variáveis de Relacionamento:

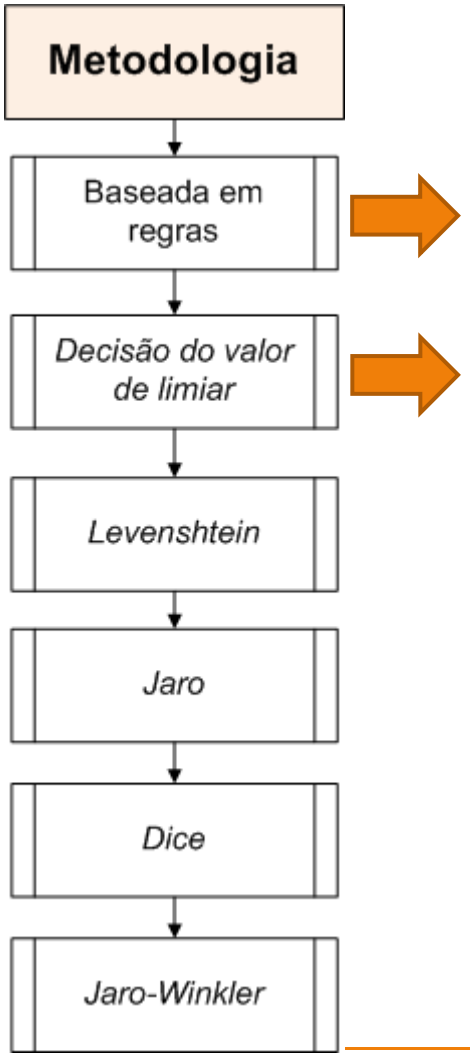
Nome

Nome da Mãe

Sexo

Data de Nascimento

Baseados em Funções de Similaridade

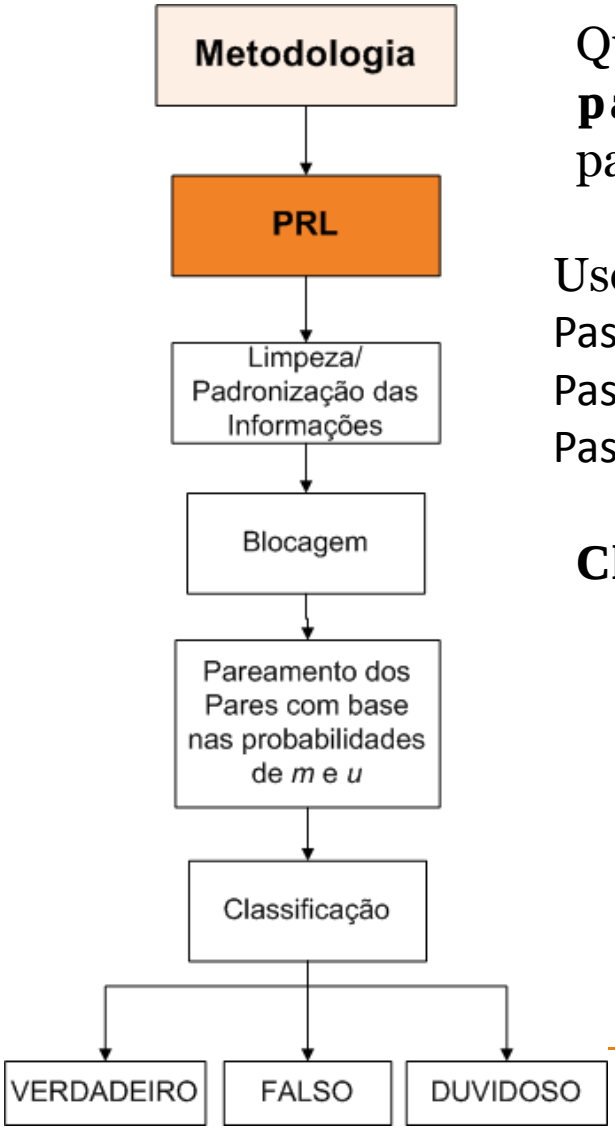


Foram utilizadas as funções de similaridade mais adequadas para comparação de caracteres que apresentam bons resultados quanto ao tamanho da cadeia de caracteres e voltadas para minimizar erros de grafia, de acordo com o encontrado na literatura.

Os valores de limiares foram testados com a variação entre 1 a 0,7.

Os resultados apresentados serão dos limiares de 0,9 e 0,8.

Aplicando PRL



Quanto ao **processo de limpeza e padronização** manteve-se o mesmo realizado para a técnica de relacionamento determinística.

Uso da **Blocagem** em 3 passos:

Passo 1 : PBloco+UBloco+Sexo+Anonasc = 711 blocos

Passo 2: Pbloco+Sexo+Anonasc = 479 blocos

Passo 3: PBloco+Ubloco = 373 blocos

Classificação

Par Verdadeiro ➔ escore >= 20.9576

Par Falso ➔ escore <= -14.7356

Par Duvidoso ➔ escore entre 20.9576 e -14.7356

Avaliação dos Métodos

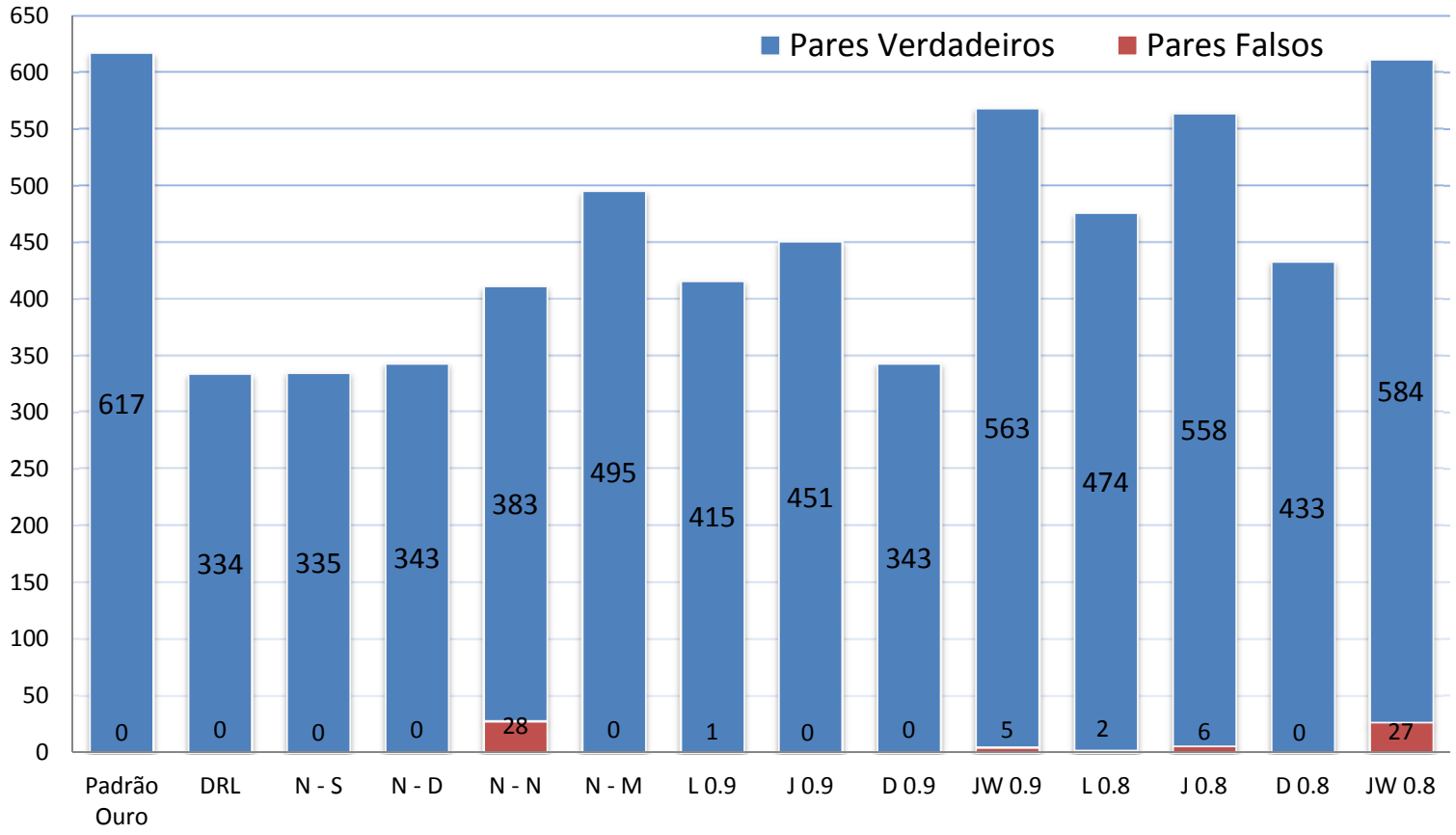
As medidas de acurácia foram realizadas com base no Padrão-Ouro.

O Padrão-Ouro foi composto pelos pares verdadeiros relacionados pela técnica DRL Exata e a verificação manual dos demais pares formados com o uso das funções de similaridade.

A revisão manual foi realizada por dois revisores.

Padrão-Ouro = 617 pares verdadeiros e 483 não-pares

Resultados das Técnicas DRL Exato, DRL (N – 1) e baseada em regras



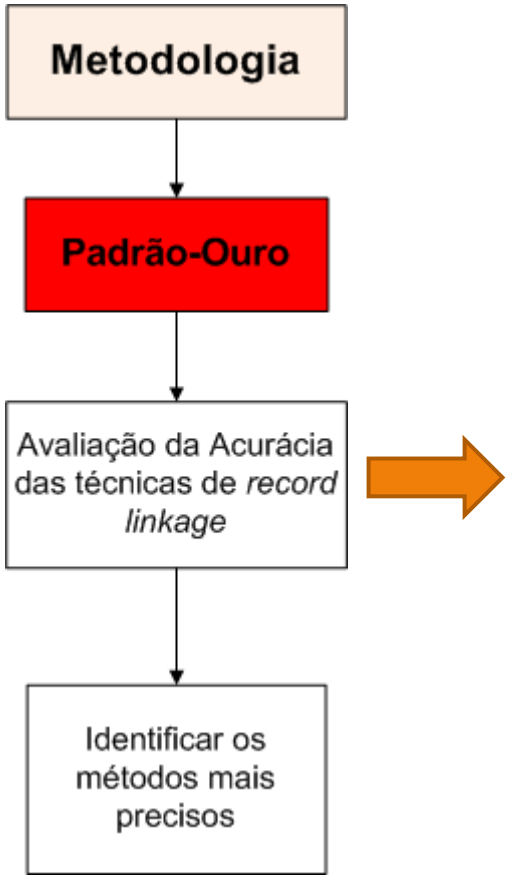
Resultados da Técnica PRL

	Passo1	Passo2	Passo 3	Total
Base de Dados - CSE	1.100	539	512	
Base de Dados - HCFMRP/USP	375.370	374.803	374.776	
Número de Bloco	711	479	373	
Registros Blocos - CSE	714	500	397	
Registros Blocos - HCFMRP/USP	1.706	25.377	23.844	
Pares Formados Possíveis	1.720	29.423	36.585	
Pares Verdadeiros	539	3	6	548
Pares Falsos	755	20.063	35.067	55.885
Pares Duvidosos	426	9.357	1.512	11.295
Revisão Manual (pares encontrados)	28	24	1	55
Tempo Processamento (segundos)	1.463	1.146	636	4.181

Pares Pareados = 601 com a revisão manual*
Falso Positivos = 7

* sem os registros duplicados

Avaliação das Técnicas DRL Exato



	DRL	
	%	95% CI
Sensibilidade	54,13	50,1 - 58,1
Especificidade	100	99,2 - 100
VPN	63,1	59,5 - 66,5

Abreviações: CI – Intervalo de Confiança; VPN – Valor Preditivo Negativo; DRL – Relacionamento Determinístico.

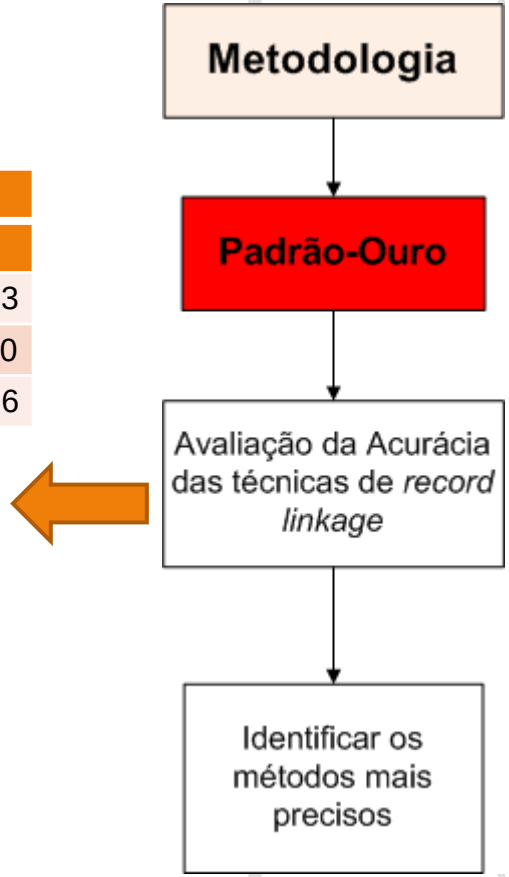
BING LI et al., 2006 (utilizou 3 bases de dados) obteve-se sensibilidade entre 74 % a 95%

BRONHARA et al., 2008 obteve sensibilidade entre 91,6% a 97,1%

Avaliação das Técnicas DRL (N – 1)

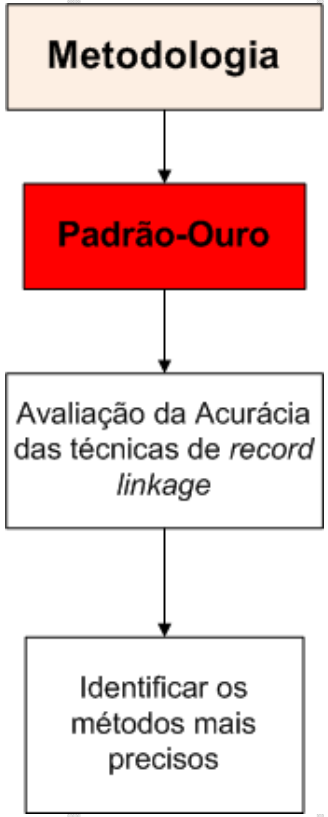
	DRL (N-D)		DRL (N-M)		DRL (N-N)		DRL (N-S)	
	%	95% CI	%	95% CI	%	95% CI	%	95% CI
Sensibilidade	55,59	51,6 - 59,6	80,06	76,7 - 83,1	62,07	58,1 - 65,9	54,29	50,3 - 58,3
Especificidade	100	99,2 - 100	100	99,2 – 100	94,2	91,7 - 96,1	100	99,2 – 100
VPN	63,8	60,3 - 67,2	79,7	76,3 - 82,8	66,0	62,4 - 69,6	63,1	59,6 - 66,6

Abreviações: DRL – Relacionamento Determinístico; (N-D) variável – Data de Nascimento; (N-M)- variável – nome da mãe; (N-N)- variável – nome; (N – S) – variável – Sexo.

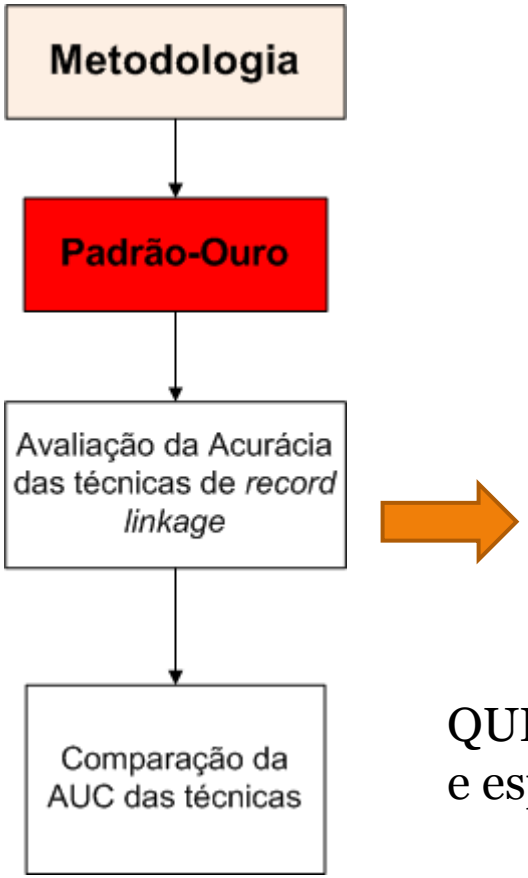


Avaliação das Técnicas baseada nas funções de similaridade

	DICE		LEVENSHTEIN		JARO		JARO WINKLER	
	%	95% CI	%	95% CI	%	95% CI	%	95% CI
Valor de limiar 0,9								
Sensibilidade	55,6	51,6 - 59,6	67,3	63,4 - 71,0	73,1	69,4 - 76,6	91,3	88,7 - 93,4
Especificidade	100,0	99,2 - 100,0	99,8	98,9 - 100,0	99,6	98,5 - 99,9	99,0	97,6 - 99,7
VPN	63,8	60,3 - 67,2	70,5	66,9 - 73,9	74,3	70,8 - 77,7	89,8	87,0 - 92,3
Valor de limiar 0,8								
Sensibilidade	70,0	66,2 - 73,6	76,8	73,3 - 80,1	90,4	87,8 - 92,6	94,7	92,6 - 96,3
Especificidade	100,0	99,2 - 100,0	99,4	98,2 - 99,9	98,8	97,3 - 99,5	93,4	90,8 - 95,4
VPN	72,3	68,7 - 75,7	77,0	73,5 - 80,3	89,0	86,0 - 91,5	93,2	90,6 - 95,3



Avaliação das Técnicas PRL



	PRL	
	%	95% CI
Sensibilidade	96,27	94,5 – 97,6
Especificidade	98,55	97,0 – 99,4
VPN	95,4	93,2 – 97,1

QUEIROZ et al., 2010 obteve-se sensibilidade = 95,7 % e especificidade 99%

MIGOWSKI et al., 2008 obteve sensibilidade = 90,6 e especificidade de 100%

Curva ROC - DRL Exato, DRL (N – 1) e PRL

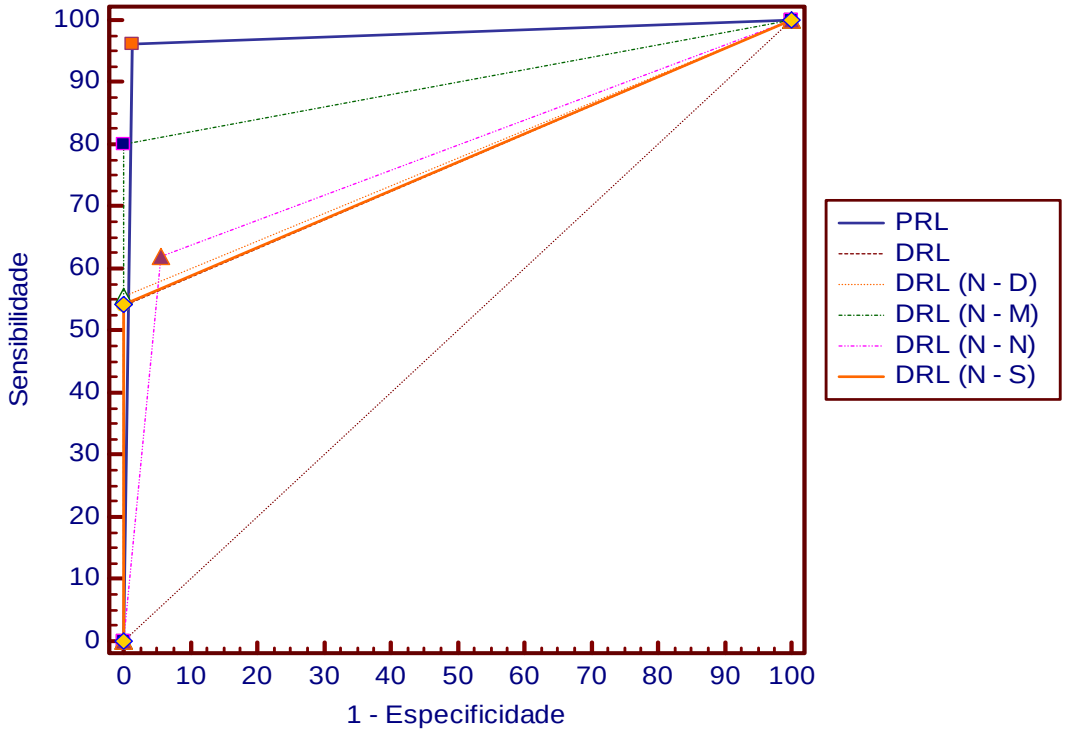
Metodologia

Padrão-Ouro

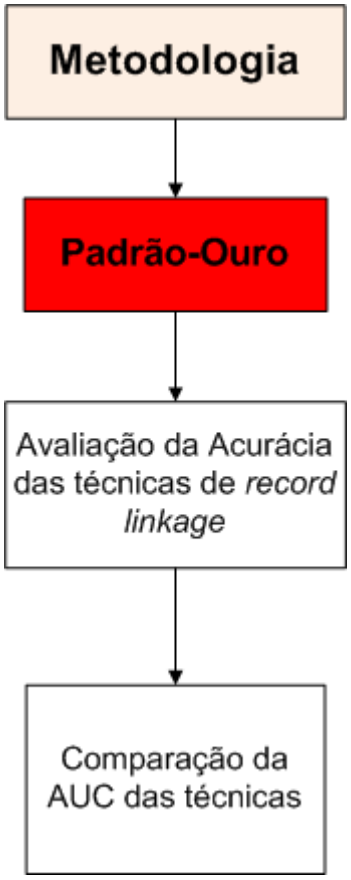
Avaliação da Acurácia das técnicas de *record linkage*

Comparação da AUC das técnicas

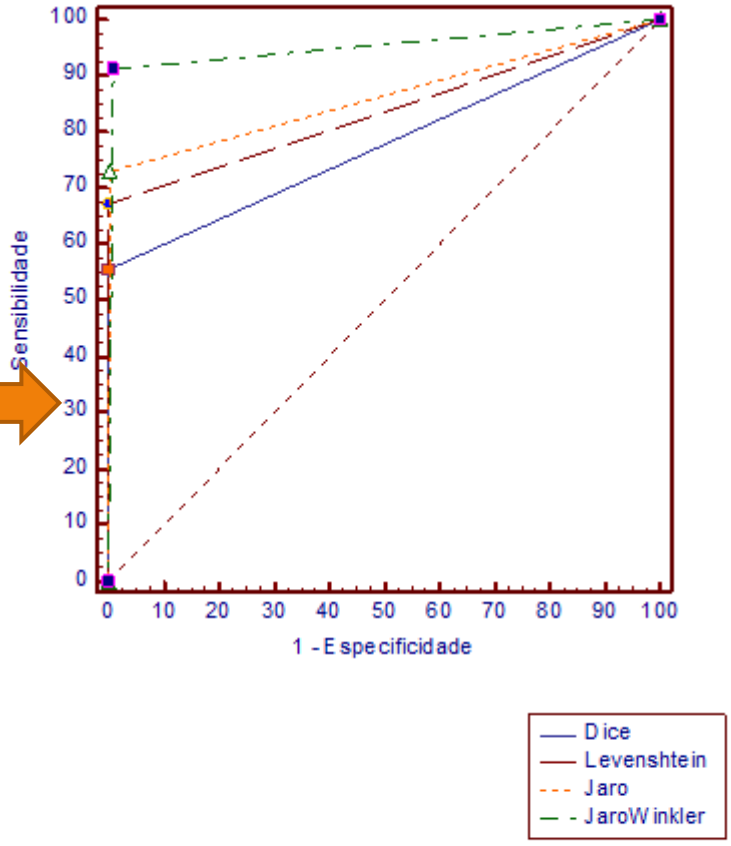
Curva ROC - área sob a curva (AUC) utilizada para avaliar o desempenho do método e Reconhecimento de Pares. Quanto maior a área sob a curva, melhor o desempenho



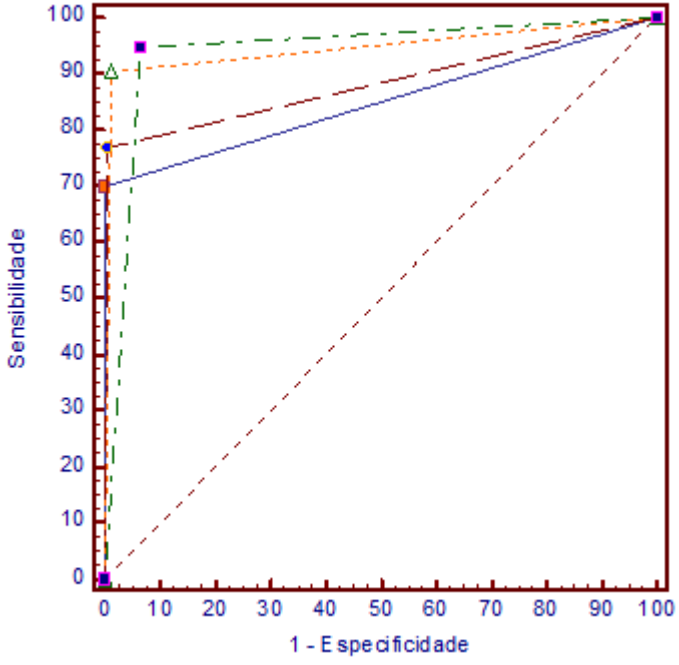
Curva ROC – Funções de Similaridade



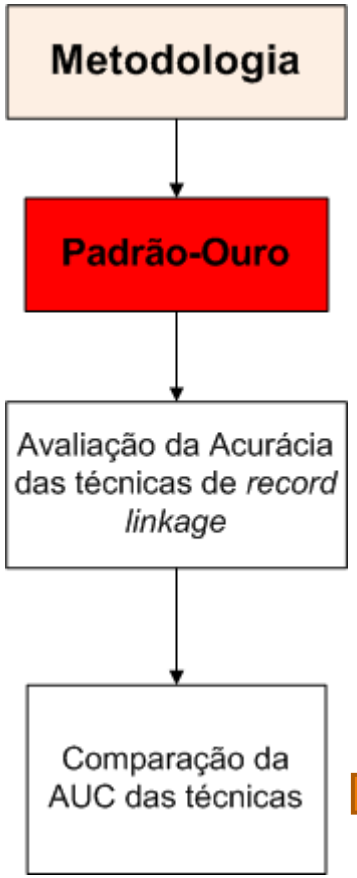
limiar de 0.9



limiar de 0.8



Comparação da AUC ROC



Método	AUC	Erro Padrão	95% CI
PRL	0,97400	0,00543	96,3 a 98,3
DRL	0,77100	0,01000	74,5 a 79,5
DRL (N - D)	0,77800	0,01000	75,2 a 80,2
DRL (N - M)	0,90000	0,00805	88,1 a 91,7
DRL (N – N)	0,78100	0,01110	75,6 a 80,5
DRL (N – S)	0,77100	0,01000	74,5 a 79,6
Valor de Limiar 0.9			
Levenshtein	0,83500	0,00951	81,2 a 85,7
Jaro	0,86300	0,00905	84,2 a 88,3
Dice	0,77800	0,01000	75,2 a 80,2
JaroWinkler	0,95100	0,00614	93,7 a 96,3
Valor de Limiar 0.8			
Levenshtein	0,88100	0,00869	86,0 a 90,0
Jaro	0,94600	0,00644	93,1 a 95,9
Dice	0,85000	0,00923	82,8 a 87,1
JaroWinkler	0,94000	0,00725	92,4 a 95,3

Comparação da AUC ROC

Metodologia

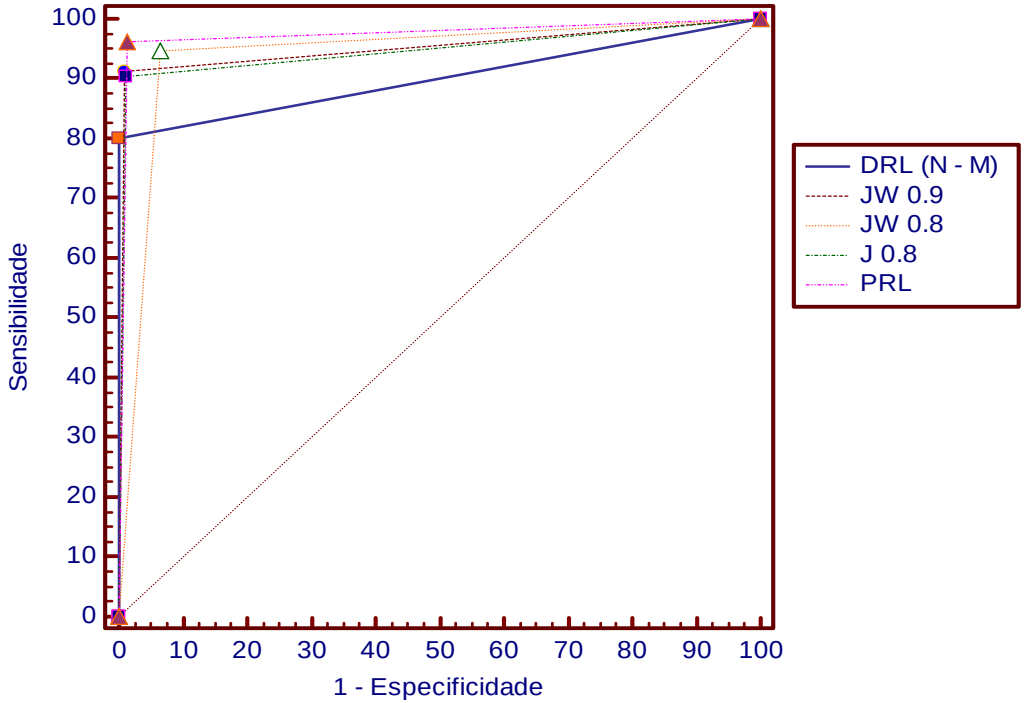
Padrão-Ouro

Avaliação da Acurácia das técnicas de *record linkage*

Identificar os métodos mais precisos

	DRL (N - M)	JW (0.9)	JW (0.8)	Jaro (0.8)	PRL
DRL (N - M)	-	0,0001	0,0008	0,0001	0,0001
DRL JaroWinkler (0.9)	-	-	0,1847	0,1411	0,0001
DRL JaroWinkler (0.8)	-	-	-	0,4855	0,0001
DRL Jaro (0.8)	-	-	-	-	0,0001

O PRL com as demais técnicas apresentaram diferença em suas AUC, sendo mais preciso.



Considerações

- O uso de padrão-ouro auxilia na verificação da acurácia dos métodos;
- A qualidade das informações influencia diretamente no sucesso do *record linkage*, proporcionando agilidade na aplicação da técnica e com baixo custo.
- Erros comuns na variável “nome da mãe” e “nome do paciente”, em virtude da mudança de sobrenome para o sobrenome de casada, erros de grafia estabelecendo-se uma forte relação com o poder de discriminação das variáveis.

Sensibilidade => Aumenta e

Especificidade => tende a diminui.

Considerações

- No uso PRL deve-se destacar a complexidade do método, principalmente para a estimativa dos parâmetros das probabilidade *m* e *u*
- A escolha inadequada de uma chave de blocagem pode resultar em perda desses possíveis pares
- Chave de blocagem menos restritiva pode inviabilizar a revisão manual
- Ao definir chaves de blocagem com base no código fonético para as bases de dados nacionais deve-se observar o uso do *Soundex*, pois a sua origem é inglesa

Considerações

- *Jaro-Winkler* deve ser aplicado com ressalva em virtude de sua característica que atribui maior peso ao início das sentenças e na língua portuguesa é comum o uso de nomes compostos

- O projeto de *Record Linkage* deve ser cuidadosamente avaliado, para se definir a melhor escolha dentre as opções existentes

Conclusão

Hipótese

A técnica PRL poderá ser uma opção viável para integrar as bases de dados nacionais de acordo com a realidade e peculiaridade das mesmas.

- DRL e DRL (N – 1) apresenta maior especificidade
- As métricas de similaridade que mais se destacaram quanto à medida de sensibilidade foram: *Jaro* e *Jaro-Winkler*. Essas métricas oferecem bons resultados para detectar erros de grafia e apresentaram valores de sensibilidade/especificidade próximo à técnica PRL
- Comparando a AUC a técnica PRL apresenta melhor desempenho.

Conclusão

Hipótese

A técnica PRL poderá ser uma opção viável para integrar as bases de dados nacionais de acordo com a realidade e peculiaridade das mesmas.

De acordo com os resultados apresentados a técnica **PRL**, em termos da medida de sensibilidade/especificidade é viável, mas deve-se levar em consideração a sua complexidade , a qualidade da base de dados e a definição das chaves de blocagem.

Em conclusão, recomenda-se fortemente o planejamento do método de *record linkage* segundo o desenho do estudo a que ele se destina.

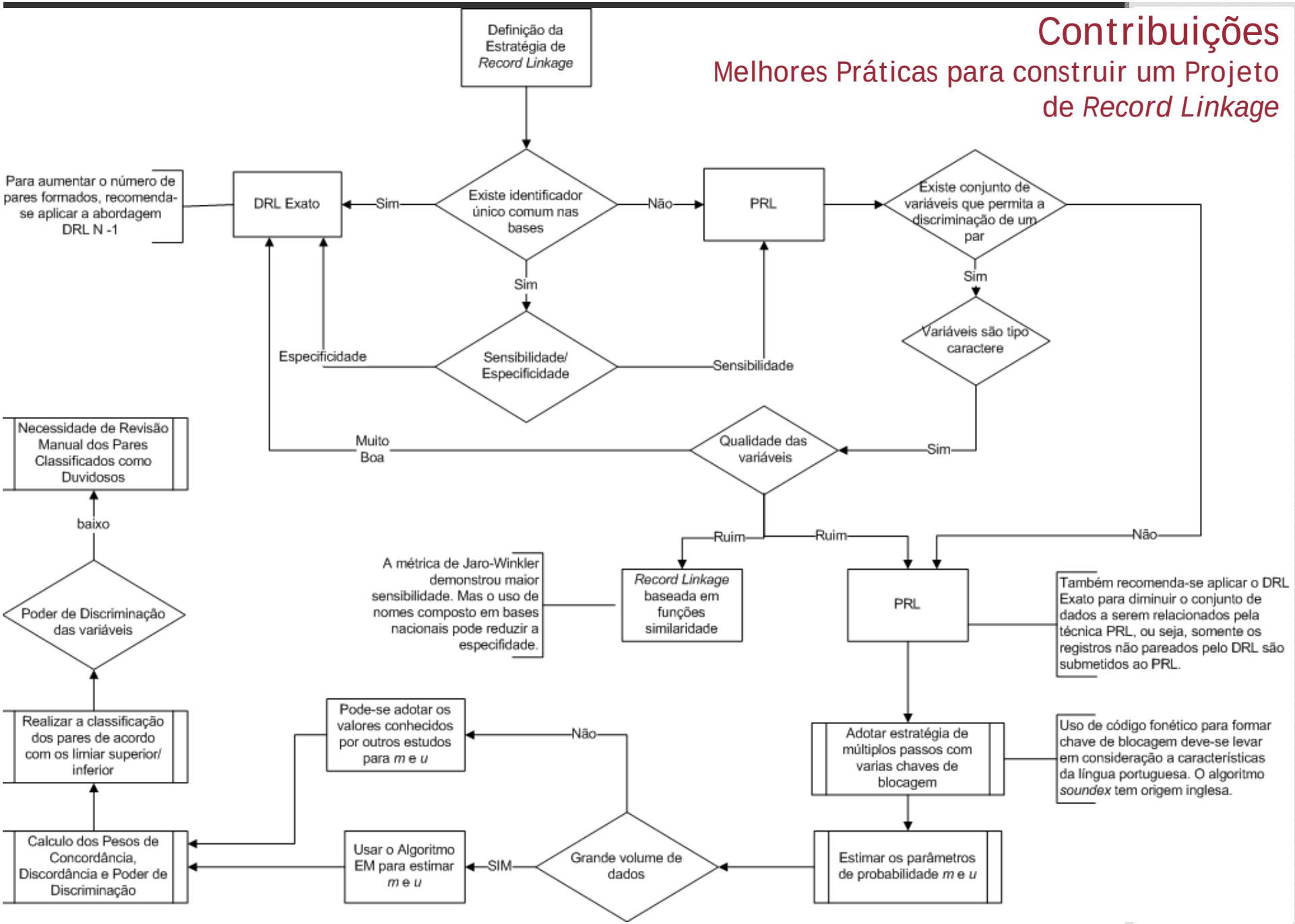
Contribuições

Melhores Práticas para construir um Projeto de *Record Linkage*

Seleção das Base de Dados	Análise dos Dados	Limpeza de Dados e Padronização
Selecionar a base de dados a serem integradas	Consistência; Preenchimento; Frequência de distribuição; Identificadores Únicos; Campos Comuns entre as bases;	Remover caracteres especiais (*, ?); Padronizar formato de dados; Remover abreviações (º,ª), Acentuação; Padronizar maiúsculo e minúsculo;

Contribuições

Melhores Práticas para construir um Projeto de *Record Linkage*



Questões.

Contatos

Kátia Suzuki
kmsuzuki@fmrp.usp.br

Prof. Dr. Paulo Mazzoncini de Azevedo Marques
pmarques@fmrp.usp.br