.

Gabarito Ciência de Dados em Saúde Pública

1) Do artigo de Christen e Schnell (Thirty-three myths and misconceptios about population data: from data capture and processing to linkage), descreva três deles e exemplifique.

Selecionar três dos mitos listados abaixo.

(1) Um banco de dados populacional contém todos os indivíduos de uma população

Mesmo bases de dados que deveriam cobrir populações inteiras, como dados de óbitos, nascidos vivos ou censos, muito provavelmente têm subpopulações sub-representadas ou ausentes. Mesmo dados primários podem não conter toda a população, com, por exemplo, coleta via celular, busca em domicílios em horário comercial ou escolar.

(2) A população coberta em uma base de dados é bem definida

Algumas bases podem se basear na inclusão obrigatória de indivíduos (pense em bases de dados de residência ou sistemas de informação de nascidos vivos), enquanto outras se baseiam na inclusão voluntária ou auto-selecionada (pense em bases de dados de notificação, como registros de COVID-19, onde pacientes poderiam optar por usar um aplicativo na cidade do Rio de Janeiro). As definições e regras usadas para extrair registros sobre indivíduos em uma base de dados populacional podem não ser conhecidas por aqueles que estão processando e pareando a base, e menos provável ainda para os pesquisadores que a analisarão.

(3) Bases de dados populacionais contêm informações completas para todos os registros

Muitas bases de dados populacionais contêm diferentes partes de informação para diferentes conjuntos de registros. Essa dispersão ocorre porque grandes bases de dados são comumente geradas pela compilação de diferentes bases individuais, cada uma cobrindo apenas uma parte da população, ou pela coleta de registros ao longo do tempo, onde mudanças nas regulamentações e métodos e processos de captura de dados podem levar à coleta de atributos diferentes. Na pandemia de COVID-19, o eSUS Notifica registrava casos provenientes de unidades da Atenção Primária à Saúde (APS), de farmácias etc. Cada uma dessas organizações registra de forma diferente; nem todas as variáveis são preenchidas, por exemplo.

(4) Todos os registros em uma base de dados populacional estão dentro do escopo de interesse

Indivíduos podem ter saído da população de interesse porque os critérios de inclusão em uma base de dados não são atendidos. Em estudo sobre cura e abandono, casos que não atualizarem o status para óbito, por exemplo, podem entrar como uma das categorias do estudo e não serem eliminados. Casos de tuberculose drogarresistente que encerram erroneamente como mudança de diagnóstico podem ser eliminados da base de tuberculose. Pessoa que se mudou após cadastro no Prontuário Eletrônico constará

como população residente. Notificação no SINAN cujo resultado dos exames laboratoriais tenham resultados após a notificação e tenham sido todos negativos, não tenham continuado o tratamento e porventura tenham sido mantidos na base, embora não sejam um caso da doença.

(5) Cada indivíduo em uma população é representado por um único registro em uma base de dados

Não é incomum que uma base de dados populacional contenha registros duplicados referentes à mesma pessoa devido a erros ou variações nos registros de identificação, como CPF ou Cartão Nacional de Saúde. No mundo real, a mesma pessoa pode, portanto, ser registrada múltiplas vezes em diferentes instituições, e seus registros duplicados não são identificados como referentes ao mesmo indivíduo. Por outro lado, pessoas com características semelhantes, como gêmeos que só têm o primeiro nome ligeiramente diferentes, podem não ser reconhecidas como dois indivíduos, mas sim como duplicatas. Mulheres que mudaram nome ou endereço ao casar (especialmente quando não lembram do número de seu documento de registro). Uma mesma pessoa pode ser cadastrada em prontuários eletrônicos de diferentes unidades de atenção primária. Erros ou problemas no sistema de entrada de dados podem gerar duplicidade no Sistema de Informação sobre Mortalidade.

(6) Registros em uma base de dados populacional sempre se referem a pessoas reais

Bases de dados do mundo real podem conter registros de pessoas que nunca existiram. Estes podem ser registros adicionados para treinamento de pessoal na entrada de dados ou testar software. Esses registros muitas vezes não são removidos de uma base de dados e são difíceis de detectar por algoritmos de limpeza de dados, porque foram projetados para terem as características de pessoas reais. Bases de dados coletadas de plataformas de mídia social podem corresponder a usuários falsos ou a *bots* de IA (inteligência artificial) que geram conteúdo semelhante ao humano.

(7) Erros em dados pessoais não são intencionais

Existem razões sociais, culturais, bem como pessoais, pelas quais os indivíduos podem decidir fornecer detalhes pessoais incorretos. Isso inclui medo de vigilância por governos, tentar evitar anúncios não solicitados de empresas ou simplesmente o desejo de manter dados pessoais sensíveis privados. Ausência ou falsas respostas sobre renda, religião ou gênero.

(8) Certos detalhes pessoais não mudam ao longo do tempo

Enquanto alguns detalhes pessoais, como nomes e endereços, são conhecidos por mudar ao longo do tempo para muitos indivíduos, muitas vezes se presume que outros são fixados no nascimento. Isso inclui identificação étnica e de gênero, bem como local e país de nascimento. Em muitas bases de dados populacionais, a identificação étnica é auto-relatada, onde as categorias disponíveis dependem de como uma sociedade valoriza diferentes subpopulações. Nome por ocasião do casamento, endereço, escolaridade podem mudar ao longo do tempo.

(9) Variações de nomes pessoais são incorretas

Os nomes das pessoas são um componente chave em muitas bases de dados. Muitos nomes pessoais têm múltiplas variações de grafia (Stefany, Estéfani, Stephanie, Estefânia, Sthefany, Stefhany, Stephany, Stefaany, Stefani, Stefanie etc). Quando os dados são inseridos, por exemplo, por telefone, variações de nomes com sons diferentes podem ser registradas para o mesmo indivíduo devido à pronúncia incorreta ou mal-entendido. Além disso, existem muitos aspectos culturais dos nomes, incluindo diferentes ordens dos nom e sobrenome, estruturas de nomes, transliterações ambíguas de alfabetos não-romanos para o alfabeto romano, ou mudanças de nome ao longo do tempo por razões religiosas, para citar alguns. Variações de nomes são um problema conhecido quando os nomes são comparados entre registros ao vincular bases de dados.

(10) Sistemas de codificação não mudam ao longo do tempo

Campos de bases de dados que são comumente codificados usando sistemas de codificação podem mudar ao longo do tempo. A Classificação Internacional de Doenças (CID) possui versões (estamos na 11ª), Código Brasileiro de Ocupações, Tabela de procedimentos do SUS (alterou substancialmente em 2008), novos procedimentos podem ser incluídos e outros, eliminados.

(11) Definições de dados são inequívocas

Assim como os sistemas de codificação, as definições de dados podem mudar ao longo do tempo e também podem ser interpretadas de maneira diferente. No início da COVID-19, o agravo foi registrado de maneiras diferentes, até ter um código específico. Ações de minimização de registro de causa mal definida de morte podem produzir registros inverídicos.

(12) Aspectos temporais dos dados não importam

O tempo e a data em que os dados populacionais são capturados e armazenados em uma base de dados podem ser cruciais, pois diferenças no atraso dos dados (*data lag*) podem levar a dados inconsistentes. Notificações de casos de COVID-19 variavam dependendo da agilidade de cada ponto de coleta e registro nas bases de dados, gerando indicadores diários não verdadeiros. Atendimentos e exames para detecção de agravos podem ser mais frequentes durante a semana, menos intensos durante o final de semana ou final de ano.

(13) O significado dos dados é sempre conhecido

Não é incomum que bases de dados populacionais contenham atributos que não estão (bem) documentados. Estes podem incluir códigos sem significado conhecido, números de sequência irrelevantes ou valores temporários que foram adicionados em algum momento para algum propósito específico. Como resultado, padrões espúrios podem ser detectados. Ausência ou dicionários de dados incompletos, como o do SIH-SUS (o mais facilmente disponível está incompleto), do Gerenciador de Ambiente Laboratorial (GAL), da Comunicação de Informação Hospitalar e Ambulatorial (CIHA, que registra internações e atendimento ambulatorial não SUS).

(14) Dados faltantes não têm significado

Dados faltantes podem levar a problemas com o processamento, vinculação e análise de dados. Dados faltantes podem ocorrer no nível de registros ausentes (nenhuma informação está disponível sobre certos indivíduos em uma população), valores de atributos ausentes (ausência de valores para atributos específicos). Em alguns casos, um valor faltante não contém nenhuma informação valiosa, em outros pode ser o único valor correto (crianças com menos de certa idade não devem ter uma ocupação) ou pode ter múltiplas interpretações. Um valor faltante para uma pergunta sobre religião em um censo pode significar que um indivíduo não tem religião ou optou por não divulgá-la. Dados faltantes também podem ocorrer em contextos onde os recursos são limitados e, portanto, a entrada de dados teve que ser priorizada, como em prontos-socorros movimentados. Remover atributos ou até mesmo registros com valores faltantes, ou imputar valores faltantes, pode resultar na introdução de erros e viés estrutural em uma base de dados, o que pode levar a resultados incorretos em um estudo.

(15) Todos os registros em uma base de dados populacional foram capturados usando o mesmo processo

Como as bases de dados populacionais são frequentemente coletadas por longos períodos de tempo e em amplas áreas geográficas, os registros são comumente inseridos por um grande número de funcionários, o que dá origem a diferentes interpretações das regras de entrada de dados.

Por exemplo, se um campo de entrada requer um valor obrigatório, podem ser inseridos todos os tipos de conteúdos não padronizados para dados faltantes, variando de símbolos únicos (como '–' ou '.'), acrônimos ('NA' ou 'MD'), a textos que explicam os dados faltantes (como 'desconhecido').

Os dados podem ser capturados em diferentes resoluções temporal e espacial, como apenas códigos postais ou nomes de cidades *versus* endereços de rua detalhados, tornando sua comparação e análise desafiadoras.

(16) Valores de atributos são corretos e válidos

Quaisquer valores de dados capturados podem estar sujeitos a erros provenientes de mau funcionamento de equipamentos, entrada de dados humana (erro de digitação) ou erros cognitivos (como confusão sobre os dados exigidos ou dificuldades em lembrar a informação correta), ou até mesmo intenção maliciosa.

No domínio médico, erros de digitação manual, interpretações erradas de formulários (pense em receitas manuscritas por médicos), inserção de valores nos campos de entrada errados ou erros na interpretação de instruções (ao prescrever medicamentos) são erros que ocorrem frequentemente.

(17) Valores de dados estão em seus atributos corretos

O pessoal de entrada de dados nem sempre insere os valores no atributo correto. Muitos nomes asiáticos e alguns ocidentais, por exemplo, podem ser usados indistintamente como primeiro e último nome. A ordem de como o primeiro e o último nome são escritos também pode depender da cultura e origem de um indivíduo. Ordem da data (ano, mês, dia ou dia, mês, ano).

(18) Regras de validação de dados produzem dados corretos

Para garantir dados de alta qualidade, muitos sistemas de gerenciamento de dados contêm regras que precisam ser cumpridas quando os dados estão sendo capturados. Por exemplo, registrar um novo paciente em um hospital requer um endereço válido e uma data de nascimento válida. Em alguns casos, como em admissões de emergência, nem todas essas informações serão conhecidas, e valores padrão são frequentemente usados (1º de janeiro).

(19) Todos os dados relevantes foram capturados

Como o objetivo principal da maioria das bases de dados populacionais não é o seu uso para estudos de pesquisa, nem todas as informações relevantes que são importantes para um determinado estudo podem estar disponíveis para todos os registros em uma base de dados. Isso pode ser devido, por exemplo, a mudanças nos requisitos de entrada de dados ao longo do tempo, ou porque os dados podem ter sido retidos pelo proprietário devido a preocupações de confidencialidade, ou os dados podem ser fornecidos apenas em forma agregada ou anonimizada.

(20) Dados populacionais fornecem as mesmas respostas que dados de inquéritos

Dados populacionais referem-se ao que as pessoas são e o que fazem. Isso é diferente dos dados de inquérito, onde geralmente são feitas perguntas sobre atitudes, crenças, expectativas ou intenções, com o objetivo de entender o comportamento das pessoas.

(21) Dados populacionais são sempre valiosos

Muitas das bases de dados carecem de metadados ou contexto para serem úteis, ou são agregadas ou anonimizadas devido a preocupações com privacidade e confidencialidade. Ter o nível educacional dos indivíduos em uma base de dados só se torna útil se essa base de dados for atualizada regularmente, devido à natureza dinâmica da vida das pessoas. Sem metadados adequados, contexto, microdados detalhados úteis e atualizações regulares, muitas bases de dados disponíveis publicamente têm pouco valor para a pesquisa.

(22) O processamento de dados pode ser totalmente automatizado

Grande parte do pré-processamento de dados populacionais tem que ser conduzida de forma iterativa, onde a exploração de dados levam a uma melhor compreensão de uma base de dados, o que, por sua vez, ajuda a aplicar técnicas de processamento de dados apropriadas. Esse processo requer exploração manual, experiência no domínio com relação à proveniência e conteúdo de uma base de dados, bem como compreensão do uso final de uma base de dados. O pré-processamento de dados é frequentemente a etapa mais demorada e intensiva em recursos de todo o pipeline de análise de dados, exigindo comumente substancial experiência no domínio, bem como em dados. Limpeza de campos, como exemplo nomes e endereços, pode gerar conteúdos errados, caso seja realizada de forma automatizada.

(23) O processamento de dados está sempre correto

Muitas vezes, vários métodos estão disponíveis para processar dados, por exemplo, para normalizar valores numéricos, imputar dados faltantes ou padronizar texto de formato

livre. Converter dados 'sujos' em dados 'limpos' pode, portanto, resultar em dados incorretamente limpos. Às vezes, não há um único valor correto para um determinado valor de entrada ambíguo. Por exemplo, em um endereço de rua, a abreviação 'St' pode significar 'Street' (rua) ou 'Saint' (santo, como usado em um nome de cidade como 'Saint Marys'). Foi relatado que em 2 de outubro de 2020, um total de 15.841 casos positivos de COVID-19 (cerca de 20%) na Inglaterra foram perdidos porque, ao registrar casos diários, foi usado um formato de arquivo antigo do software de planilha Microsoft Excel, que permitia um máximo de 65.536 linhas.

(24) Dados agregados são suficientes para pesquisa

Dados altamente agregados, por exemplo, no nível de estados ou grandes unidades geográficas, dificilmente são úteis para pesquisas científicas que visam causalidade. Um grande problema é a falácia ecológica, que descreve o erro de que uma relação agregada implica a mesma relação para indivíduos. Se taxas de mortalidade aumentadas forem observadas em regiões onde as taxas de vacinação são altas, a conclusão falsa seria que as pessoas vacinadas têm uma probabilidade maior de morrer. Mas, na verdade, o inverso pode ser verdadeiro: pessoas observando outras pessoas morrendo podem estar mais dispostas a se vacinar.

(25) Metadados estão corretos, completos e atualizados

Metadados (também conhecidos como dicionários de dados) descrevem uma base de dados, como ela foi criada, preenchida e seu conteúdo capturado e processado. A falta de metadados pode levar a mal-entendidos durante o processamento, vinculação e análise de dados, relatórios incorretos, ou pode tornar uma base de dados totalmente inútil.

(26) Um conjunto de dados vinculados corresponde a uma população real

Devido a problemas de qualidade de dados e à(s) técnica(s) de vinculação de registros empregada(s), um conjunto de dados vinculados provavelmente contém ligações erradas (erros Tipo I, dois registros referentes a dois indivíduos diferentes foram ligados erroneamente), enquanto algumas ligações verdadeiras foram perdidas (erros Tipo II, dois registros referentes à mesma pessoa não foram ligados). O aumento da sensibilidade no processo de vinculação, com mais registros pareados, tende a gerar mais falsos positivos.

(27) Bases de dados populacionais representam as condições das pessoas ao mesmo tempo

As atualizações de dados sobre indivíduos geralmente ocorrem em diferentes momentos, geralmente quando um evento como uma condição médica ocorre. Escolaridade e estado civil registrados em uma base podem não ser atuais

(28) Um conjunto de dados vinculados não contém duplicatas

Ao parear bases de dados, pares ou grupos de registros que se referem ao mesmo indivíduo podem não ser vinculados corretamente. Uma razão para isso ocorrer é se um identificador de entidade errado tiver sido atribuído a um indivíduo (por acidente ou de

propósito). Portanto, muitos conjuntos de dados pareados contêm mais de um registro para alguns indivíduos. Agravos de notificação compulsória podem ser notificados em mais de uma oportunidade e erros de *linkage* podem gerar duplicidade de pares.

(29) Um conjunto de dados vinculados é imparcial (*unbiased*)

Erros de pareamento geralmente não ocorrem aleatoriamente. Exemplos incluem estruturas de nomes de estrangeiros podem ser diferentes do padrão nome, sobrenome da mãe e sobrenome do pai. A mobilidade (mudanças de endereço) é menor para pessoas mais velhas. Como resultado, pode haver viés estrutural em um conjunto de dados vinculados em subpopulações definidas por categorias étnicas ou sociais, idade ou gênero (por exemplo, se as mulheres são mais propensas a mudar seus nomes em comparação com os homens quando se casam).

(30) Valores de atributos em registros pareados estão corretos

A fusão de dados é o processo de resolver tais inconsistências, onde muitas vezes uma decisão precisa ser tomada sobre qual aplicar dentre as muitas operações de fusão disponíveis. Diferentes registros que se referem à mesma pessoa foram ligados corretamente, porém cada registro contém um valor de salário diferente.

(31) As taxas de erro de pareamento são independentes do tamanho da base de dados

Uma variável identificadora usada para vincular registros pode não ser unívoca e ser compartilhada por múltiplos indivíduos. Potencialmente, usar nomes de cidades ou nomes e sobrenomes populares para parear registros pode gerar falsos positivos. Portanto, quando bases de dados maiores estão sendo pareadas, o número de pares de registros com os mesmos valores de variável de pareamento provavelmente aumenta, resultando em pares mais altamente semelhantes.

(32) Técnicas modernas de pareamento de registros podem lidar com bases de dados de qualquer tamanho

> Muitos pesquisadores, especialmente nos domínios da ciência da computação e estatística, que desenvolvem técnicas de pareamento de registros, não têm acesso a grandes bases de dados do mundo real devido à natureza sensível dos dados populacionais. Como resultado, novas técnicas de pareamento são frequentemente avaliadas em pequenos conjuntos de dados públicos ou em dados gerados sinteticamente.
>
> (33) Técnicas de pareamento e suas configurações são facilmente transferíveis
>
> Se um método de pareamento juntamente com suas configurações de parâmetro (por exemplo, como a blocagem é conduzida, como os valores são comparados e como um limiar de classificação é definido) foi implantado com sucesso em um determinado projeto de pareamento, isso não significa que o mesmo método e configurações fornecerão resultados de qualidade quando utilizados para bases de dados diferentes.

2) Discorra sobre duas dimensões de qualidade de dados discutidas no artigo de Lima et al. (2009), intitulado "Revisão das dimensões de qualidade dos dados e métodos aplicados na avaliação dos sistemas de informação em saúde"

Resposta: Escolher duas das dimensões a seguir.

Acessibilidade: grau de facilidade e rapidez na obtenção dos dados ou informações (regras claras definindo preço, permissões e onde obtê-los), no trato (instrumentos para manuseio e for mato) e na compreensão da informação.

Clareza metodológica: grau no qual a documentação que acompanha o SIS (instruções de coleta, manuais de preenchimento, tabelas de domínios de valores de variáveis, modelos de da dos etc.) descreve os dados sem ambiguidades, de forma sucinta, didática, completa e numa linguagem de fácil compreensão.

Cobertura: grau em que estão registrados no Sistema de Informação em Saúde os eventos do universo (escopo) para o qual foi desenvolvido.

Completitude: grau em que os registros de um Sistema de Informação em Saúde possuem valores não nulos.

Confiabilidade: grau de concordância entre aferições distintas realizadas em condições similares.

Consistência: grau em que variáveis relacionadas possuem valores coerentes e não contraditórios.

Não-duplicidade: grau em que, no conjunto de registros, cada evento do universo de abrangência do SIS é representado uma única vez.

Oportunidade: grau em que os dados ou in formações estão disponíveis no local e a tempo para utilização de quem deles necessita.

Validade: grau em que o dado ou informação mede o que se pretende medir.