A construção da base de dados nacional em Terapia Renal Substitutiva (TRS) centrada no indivíduo: aplicação do método de linkage determinístico-probabilístico

Mariângela Leal Cherchiglia*
Augusto Afonso Guerra Júnior**
Eli Iola Gurgel Andrade***
Carla Jorge Machado****
Francisco de Assis Acúrcio*****
Wagner Meira Júnior******
Bruno Diniz de Paula*******
Odilon Vanni de Queiroz********

Introdução

Os sistemas de informação em saúde são ferramentas fundamentais para subsidiar a tomada de decisões e auxiliar a organização dos serviços, por meio do planejamento das ações e do acompanhamento e avaliação dos objetivos propostos. O sistema de saúde brasileiro tem larga experiência com a captura e o uso de informações. No entanto, Morais e Gómez (2007) fazem uma reflexão de que os atuais pressupostos, práticas e saberes de informação e informática em saúde não mais dão conta da complexidade do processo saúde/doença/cuidado, apontando algumas questões: a fragmentação dos indivíduos entre diversas bases de dados em saúde, o que acarreta em perda da historicidade do indivíduo pleno; e o tratamento de conceito de coletivo como somatório de indivíduos (desintegrados) ou de determinados agravos (piores procedimentos realizados) referidos no tempo e lugares de forma estática, perdendo-se a dinâmica temporal e espacial como dimensão essencial no caminhar da vida ou mesmo da trajetória do paciente pelo sistema de saúde (linha de cuidado).

O monitoramento das linhas de cuidado da alta complexidade no SUS pode ser desenvolvido a partir do Sistema de Informação de Procedimentos de Alta Complexidade (Apac), integrante do Sistema de Informação Ambulatorial (SIA). O sistema Apac/SUS, criado em 1996, tem como principal finalidade registrar a produção, a cobrança e o pagamento desses procedimentos no âmbito do SUS, tais como os de terapia de substituição renal, quimioterapia, radioterapia e medicamentos excepcionais. O Apac/SIA diferencia-se dos demais sistemas de informação em saúde pelo grau de detalhamento dos registros – dados de interesse clínico-epidemiológico e demográficos –, bem como pela forma de identificação do paciente, que inclui a obrigatoriedade do número de inscrição no CPF – Cadastro de Pessoas Físicas (MS/GM, 1996). A princípio, o registro do CPF facilitaria a correta identificação dos indivíduos, eliminando o problema de homônimos (GOMES Jr. et al, 2003). Adicionalmente, a utilização conjunta do Apac/SUS e outros sistemas de informação, tais como o Sistema de Informação de Mortalidade (SIM) e o Sistema de Informação Hospitalar (SIH), torna-se um importante desafio na avaliação da assistência de alta complexidade prestada à população pelo SUS (GOMES Jr.; ALMEIDA, 2004; DE MARTINO; ALMEIDA, 2003).

Entretanto, os aspectos estruturais das bases de dados com finalidade admi-

* Professora do Departamento de Medicina Preventiva e Social da Faculdade de Medicina, UFMG.
** Assessor técnico do Ministério da Saúde.
*** Professora do Departamento de Medicina Preventiva e Social da Faculdade de Medicina, UFMG.
**** Professora do Departamento de Demografia, UFMG.
***** Professor da Faculdade de Farmácia, UFMG.
****** Professor do Departamento de Ciência da Computação, Instituto de Ciências Exatas, UFMG.
******* Doutorando do Departamento de Ciência da Computação, Instituto de Ciências Exatas, UFMG.
******** Mestrando do Departamento de Medicina Preventiva e Social, da Faculdade de Medicina, UFMG.

nistrativa – como as lacunas de informação clínica, as dificuldades na codificação dos procedimentos e o caráter de faturamento – restringem a possibilidade de se realizarem avaliações a partir dessas informações (IEZZONI, 1997). Apesar de reconhecer essas limitações, a autora destaca a grande potencialidade dos dados administrativos em traçar a trajetória do usuário nos serviços de saúde, o que requer que os indivíduos sejam identificados.

O *linkage* de registros pode ser realizado de forma determinística probabilística ou combinada. É possível conduzir, com bastante sucesso, um *linkage* determinístico quando existem identificadores únicos e confiáveis. Se o identificador único não for confiável, ainda assim é possível realizar um *linkage* determinístico, baseando-se na comparação de múltiplos identificadores. Já o *linkage* probabilístico é empregado quando os registros em bancos de dados apresentam problemas de consistência, erros ou informações não declaradas. Utilizam-se métodos para ponderar os identificadores, baseando-se no grau de certeza e precisão do pareamento (JARO, 1995).

Este trabalho descreve as etapas do processo de adaptação do banco de dados da Apac, utilizando o *linkage* determinístico-probabilístico. Enfocando os registros referentes aos pacientes em terapia renal substitutiva (TRS), entre 2000 e 2004, elaborou-se uma base nacional de dados, incluindo-se indicadores econômicos e epidemiológicos úteis ao acompanhamento e avaliação da atenção prestada, com o objetivo de traçar a trajetória do cuidado a esses pacientes no SUS.

Métodos

Utilizaram-se as informações do subsistema Apac do SIA/SUS, no período de 2000 a 2004. O banco de dados completo foi disponibilizado em abril de 2005, pelo Datasus, através do DES/SCTIE do Ministério da Saúde, conforme termo de compromisso e responsabilidade firmado entre o DES e o GPES/UFMG. Selecionou-se este período pela maior consistência e homogeneidade dos dados, após mudanças operacionais ocorridas no subsistema Apac/SIA, em 1999.

Para cumprir os requisitos éticos, a única variável identificadora do paciente após o processo de *linkage* passou a ser um código numérico. O projeto foi aprovado pela Comissão de Ética em Pesquisa da UFMG (Coep/UFMG), parecer ETIC nº 397/2004. A seguir, apresentam-se as etapas da construção da base.

Preparação dos dados originais da Apac para linkage

Constituiu-se um banco de dados único, a partir da importação dos arquivos em formato dBase da Apac (Quadro 1) para o sistema gerenciador de banco de dados MySQL.

Cada um dos tipos de arquivos, agrupados por mês de competência e por Unidade da Federação (UF), foi unificado, com a inserção destas informações em cada registro. Foi incluído um campo para identificação (ID seqüencial) de cada registro, possibilitando retornar ao arquivo original.

Gerou-se um arquivo em formato texto com os dados necessários à identificação dos indivíduos, contendo CPF, nome do paciente, UF de nascimento, nome da mãe, logradouro, número da residência, município de residência, data de nascimento, sexo, ID seqüencial e código do arquivo de origem.

"Limpeza" e padronização dos dados

Na limpeza dos dados eliminaram-se valores inválidos (por exemplo, "desconhecido") e foram removidos acentos, cedilha e espaços extras nos nomes. A padronização foi feita para datas e nomes. Os nomes foram fragmentados em: nome, sobrenome1 e sobrenome2. Para esse passo, utilizaram-se programas desenvolvidos internamente e o *software* Freely Extensible Biomedical Record Linkage – Febrl, versão 0.3 (CHRISTEN et al, 2005), que possibilita o *linkage* probabilístico e possui ferramentas para padronização, blocagem e comparação de registros, permitindo a execução

do *linkage* em paralelo (a paralelização permite a decomposição do processo em pequenas tarefas que podem ser executadas de forma independente). Ainda na padronização, procurou-se um mesmo indivíduo (pares duplicados ou repetidos) em um mesmo arquivo, entre os descritos no Quadro 1.

Linkage dos registros Apac/SIA

O *linkage* teve como objetivo encontrar os registros para um mesmo paciente nos arquivos e unificá-los em um único registro. Para otimizar e viabilizar o processo de comparação, realizou-se a blocagem do arquivo unificado: os registros só poderiam ser comparados se estivessem nesta blocagem. Foram utilizadas três estratégias (Quadro 2), sendo a primeira mais restritiva e as seguintes mais flexíveis.

Adicionalmente à blocagem, compararam-se os registros com base em campos específicos através de escores, utilizando metodologia probabilística. Os campos usados foram município de residência, logradouro, UF de nascimento, dia, ano e mês de nascimento, sexo, CPF, nome, sobrenome(s) do meio, último sobrenome, nome da mãe, sobrenome(s) do meio da mãe e último sobrenome da mãe. Os pesos variaram de -57,14% a 47,94%.

O processo de obtenção de pares ocorreu seqüencialmente. Após cada blocagem e definido o escore limiar, eram aceitos os pares preliminares acima deste. A seguir, modificava-se a blocagem e o processo era repetido, sendo que os novos resultados eram mesclados com os anteriores.

O arquivo foi posteriormente dividido em 20 subarquivos, e a blocagem e o linkage realizados em cada subarquivo separadamente. De fato, mesmo com a blocagem, o número de comparações era imenso e traria um problema técnico devido às limitações impostas pela memória RAM disponível. Esta partição foi feita aleatoriamente, para que cada subarquivo fosse representativo do arquivo completo. O potencial problema foi a impossibilidade de se compararem registros de um subarquivo com outro, em um primeiro momento. Visando solucionar a perda de pares, após terminado o *linkage*, verificou-se em todos os outros subarquivos se havia registros com todos os campos iguais ou CPFs iguais. Caso houvesse, repetia-se o *linkage* destes registros.

Resultados

Resultados quantitativos e validação

Foram pareados 34.645.811 registros, formados a partir dos arquivos originais, chegando-se, ao final, a 8.569.949 *clusters* (média de 4,04 registros por *cluster*).

Após o *linkage* calculou-se a média dos escores dos pares por *cluster* e obtiveram-se decis de escores médios por *cluster*. O menor escore médio correspondeu a 15,92. Foi selecionada, por decil, uma amostra aleatória de 30 *clusters,* utilizada para verificar a qualidade do *linkage*, encontrando-se um resultado satisfatório (de 300, apenas um par foi considerado falso).

A construção da base centrada no indivíduo

Para cada indivíduo, foram gerados um determinado número de pares e um ID-único (*cluster*). Cada *cluster* continha, no mínimo, um registro por indivíduo.

Em seguida, relacionou-se cada *cluster* aos arquivos PC, PF, PQ, PR e OP através do sistema MySQL, por meio do ID seqüencial e do código do arquivo de origem.

Em alguns *clusters* havia pares com diferentes CPFs. Nestes casos, pesquisou-se se um cluster com registros formados por CPFs diferentes podia ser considerado o mesmo indivíduo. Assim, após identificação e listagem do número de CPFs por *cluster*, os registros com mais de dois CPFs foram novamente pareados. Uma provável causa da formação de falsos pares foi a existência dos valores "não informados" e "a mesma", no campo nome da mãe, o que artificialmente aumentava o escore destes pares. Esses valores foram, então, substituídos por um número aleatório, dado que sempre havia uma discordância quando este tipo de *linkage* ocorria. Isto permitiu novo

linkage de registros, que receberam um novo ID-único e formaram novo *cluster*. Os restantes mantiveram-se em seu *cluster* original.

Formados os *clusters*, era preciso determinar quais CPF, sexo, nome, UF de nascimento e data de nascimento deveriam ser alocados para cada *cluster*. Utilizou-se o critério de campo válido e mais freqüente. Em caso de empate, escolheu-se um dos CPFs (ou data de nascimento) aleatoriamente. Elaborou-se, assim, um cadastro único dos pacientes, com os campos válidos e mais freqüentes, denominado "Cadunico".

Finalmente, elaborou-se uma tabela única denominada "Base Nacional em TRS", com 176.773 pacientes em TRS, no período de 2000 a 2004. Essa base contém registros de identificação de cada paciente (sexo, data de nascimento, cidade de residência), modalidade de entrada e saída da TRS, diagnóstico na entrada, resultado final (óbito, continuidade do tratamento, perda do seguimento) e gastos.

Conclusões

Os resultados deste estudo indicam a viabilidade de unificar os dados da Apac. Como dificuldades, destaca-se a impossibilidade de utilizar o identificador único CPF, o que gerou a necessidade de utilização da técnica probabilística. Não foi possível utilizar apenas a chave CPF para o *linkage* por dois motivos: crianças não possuem CPF e havia CPFs inválidos na base.

O *linkage* possibilitou a (re)construção da trajetória dos pacientes em TRS em larga escala. Isto permite a construção de indicadores econômico-epidemiológicos que podem contribuir para o aperfeiçoamento das políticas de TRS no Brasil.

QUADRO 1
Arquivos Apac – 2000-2004

Tipo	Informações sobre os pacientes
PC	Nome, CPF, sexo, data de nascimento, início do tratamento, endereço completo, diagnósticos principal e secundário (CID10).
PF	Identificação dos inscritos no Programa de Medicamentos Excepcionais e dos medicamentos recebidos
PQ	Identificação dos pacientes em quimioterapia.
PR	Identificação dos pacientes em radioterapia.
OP	Identificação daqueles que realizaram procedimentos de alta complexidade e sem arquivo específico (e.g. pacientes transplantados). Permite relacionar os dados do paciente com o sistema SIH/SUS.

QUADRO 2
Estratégias utilizadas para *linkage*

Blocagem	Descrição	Escores limiares
1	Concordância exata em dia, mês e ano do nascimento	14; pares corretos foram os registros que concordavam em UF de residência, CPF, sexo, nome e todos sobrenomes (escore 14,25).
2	Concordância no primeiro nome do indivíduo e da mãe, último sobrenome do indivíduo e da mãe e em dia, mês e ano do nascimento	12; pares corretos foram os registros que concordavam em UF de residência, sexo, nome do indivíduo, primeiro nome da mãe, último sobrenome do indivíduo, último sobrenome da mãe fossem referentes ao mesmo indivíduo (escore 12,25)
3	Concordância exata em CPF OU no primeiro nome do indivíduo e da mãe OU em dia, mês, ano, UF de nascimento e sobrenomes da mãe e do indivíduo	- 7; pares corretos foram os registros que concordavam em UF de nascimento, sexo, CPF, nome do indivíduo, segundo sobrenome do indivíduo e nome da mãe (escore de -6,23).

Referências bibliográficas

GOMES Jr., S.; ALMEIDA, R. Identificação de um caso novo de câncer no Sistema de Informação Ambulatorial do Sistema Único de Saúde. **Cadernos Saúde Coletiva**, Rio de Janeiro, volume 12, número 1: 57-68, 2004.

GOMES Jr., S.; DE MARTINO, R.; ALMEIDA, R. Rotinas de integração das tabelas do Sistema de Autorização de Procedimentos de Alta Complexidade em Oncologia do Sistema Único de Saúde. **Cadernos Saúde Coletiva**, v. 11, n. 2, p. 231-254, 2003.

CHRISTEN P.; CHURCHES T.; HEGLAND, M. A parallel open source data linkage system. Disponível em: <http://citeseer.ist.psu.edu/christen04parallel.html>. Acesso em: 06/mar./2005.

IEZZONI, L. Assessing quality using administrative data. **Annals of Internal Medicine,** volume 127, issue 8 (Suppl): 666-74, 1997.

JARO, M. Probabilistic linkage of large public health data files. **Statistics in Medicine**, volume 15, issue 14: 491-498, Mar./Apr. 1995.

GOMES Jr., S.; ALMEIDA, R. Identificação de um caso novo de câncer no Sistema de Informação Ambulatorial do Sistema Único de Saúde. **Cadernos Saúde Coletiva**, Rio de Janeiro, volume 12, número 1: 57-68, 2004.

MORAES, I.; GÓMEZ, M. Informação e informática em saúde: caleidoscópio contemporâneo da saúde. **Ciência e Saúde Coletiva**, Rio de Janeiro, volume 12, número 13: 553-564, maio/jun. 2007.

Recebido para publicação em 22/02/2007.
Aceito para publicação em 31/05/2007.