Produção estatal de evidências e uso de registros administrativos em políticas públicas

Janine Mello

2022

Introdução

Registros administrativos são dados coletados rotineiramente por órgãos governamentais (ou outras instituições) durante a administração de programas e serviços públicos. Diferentemente de pesquisas estatísticas planejadas como surveys amostrais ou censos demográficos, que são desenhados especificamente para coletar informações de interesse científico ou gerencial, os registros administrativos surgem como subproduto das atividades burocráticas do Estado. Por exemplo, ao registrar um nascimento em cartório, cadastrar uma família em um programa social ou emitir a declaração de imposto de renda, o governo gera informações administrativas. Essas informações, armazenadas em sistemas de cadastro ou sistemas de informação governamentais, compõem grandes bases de dados contínuas sobre pessoas, empresas, benefícios, eventos vitais, entre outros fenômenos sociais.

Um contraste importante em relação às pesquisas amostrais e censos é que os registros administrativos geralmente cobrem exaustivamente a população ou o fenômeno alvo da gestão pública. Em princípio, eles incluem todos os casos registrados de determinado evento ou participação em programa, em vez de uma amostra estatística. Por exemplo, um censo demográfico clássico busca contar todos os habitantes em intervalos definidos (como a cada dez anos), ao passo que um registro administrativo populacional (se existir) manteria uma contagem contínua e atualizada de nascimentos, óbitos e mudanças de residência. Na prática, para que um registro administrativo possa substituir ou complementar um censo, ele precisa ser universal e completo em sua cobertura[1]. Em outras palavras, um cadastro usado com fins estatísticos deve abranger todos os indivíduos do grupo de interesse e conter as informações-chave atualizadas periodicamente, tal como o censo busca fazer em campo. Nos países nórdicos, por exemplo, essa lógica foi levada ao extremo: cadastros nacionais de população, endereços e empresas, desenvolvidos ao longo de décadas, passaram a suprir os censos tradicionais, integrando um sistema unificado de estatísticas baseadas em registros[2].

Outra diferença crucial está no propósito original da coleta. Pesquisas amostrais e censos são planejados com objetivos estatísticos definidos – por exemplo, medir a taxa de

[1]A autora agradece os comentários atentos e generosos feitos por Paulo Jannuzzi e Isabele Bachtold sobre este capítulo. Eventuais erros e omissões são de inteira responsabilidade da autora.

[2]Especialista em políticas públicas e gestão governamental em exercício na Diretoria de Estudos e Políticas do Estado, das Instituições e da Democracia (Diest) do Ipea. E-mail: janine.mello@ipea.gov.br.

desemprego ou estimar a renda média – e seguem protocolos metodológicos padronizados. Já os registros administrativos existem primariamente para fins operacionais e legais, como gerir um benefício, fiscalizar obrigações ou prestar um serviço público (educação, saúde, previdência etc.)[3]. Consequentemente, nem sempre os conceitos e variáveis dos registros correspondem exatamente às categorias analíticas da pesquisa social. Pode haver diferenças de definição: o que conta como "empregado" num sistema administrativo (como o registro trabalhista) pode divergir do conceito de desemprego utilizado em pesquisas de mercado de trabalho. Além disso, os registros administrativos frequentemente refletem regras institucionais – por exemplo, faixas de renda para elegibilidade em programas – e alterações nessas regras ao longo do tempo podem afetar a comparabilidade dos dados.

Quanto ao ciclo de vida de um registro administrativo, podemos pensar nas etapas pelas quais os dados passam: (1) Geração/Coleta – ocorre no ponto de serviço ou interação do cidadão com o Estado (por exemplo, no preenchimento de um formulário, atendimento em unidade de saúde, registro escolar anual etc.); (2) Armazenamento e Gestão – os dados são inseridos em bancos de dados governamentais e atualizados conforme novas ocorrências (um mesmo indivíduo pode ter seu registro alterado ou complementado diversas vezes, por exemplo, quando renova um cadastro anual); (3) Consolidação – muitas vezes os dados locais alimentam uma base nacional centralizada (como o banco de dados unificado de um ministério), onde passam por verificações de consistência; (4) Disponibilização/Compartilhamento – dependendo da sensibilidade, alguns registros são agregados em estatísticas oficiais, outros podem ser disponibilizados sob demanda para pesquisadores mediante acordos, e outros permanecem de uso interno; e finalmente (5) Arquivamento – após cumprir sua função imediata, o registro pode ser mantido historicamente para análises futuras ou auditagem. Esse ciclo de vida implica que a qualidade e forma dos dados podem mudar ao longo do processo – por exemplo, podem ocorrer erros na entrada inicial, mas etapas de consolidação podem corrigir ou padronizar certos campos, ao mesmo tempo em que eventualmente se perdem detalhes ao anonimizar dados para uso público.

Os principais tipos de registros administrativos podem ser categorizados conforme a área ou finalidade original. No Brasil, há uma variedade extensa de bases administratives federais e estaduais. Exemplos importantes incluem: registros demográficos e vitais, como o Registro Civil (nascimentos, óbitos, casamentos, etc., compilados em estatísticas pelo IBGE); registros sociais, como o Cadastro Único (CadÚnico) de programas sociais, que reúne informações de famílias de baixa renda para políticas como Bolsa Família; registros trabalhistas, a exemplo da RAIS (Relação Anual de Informações Sociais) e do CAGED, que registravam vínculos de emprego formal reportados por empregadores; registros educacionais, como o Censo Escolar e o Censo da Educação Superior, coletados anualmente pelo Inep, que cobrem respectivamente toda a educação básica e o ensino superior do país; sistemas de saúde pública, como os vários sistemas do DATASUS (por exemplo, o SIM – Sistema de Informações sobre Mortalidade; o SINASC – Sistema de Nascidos Vivos; o SIH – Sistema de Informações Hospitalares; o SINAN – agravos de notificação compulsória; entre outros); registros previdenciários, como cadastros de contribuintes e beneficiários do INSS; registros fiscais, como declarações de imposto de renda e cadastros de empresas (CNPJ) ou indivíduos (CPF); entre outros como registros agrários, de segurança pública, meio ambiente etc. Em geral, cada órgão gestor mantém os

registros vinculados à sua missão, embora haja esforços de integração. Uma característica marcante é a escala populacional e temporal desses registros: eles costumam cobrir milhões de unidades (pessoas, empresas, eventos) em todo o território nacional e ao longo de vários anos, com atualização frequente. Isso contrasta com pesquisas amostrais, que cobrem algumas milhares de pessoas e geralmente oferecem "fotografias" em momentos isolados. De fato, registros administrativos em geral oferecem séries históricas longas e abrangência territorial ampla, com rotinas de atualização contínuas ou periódicas mais frequentes do que pesquisas pontuais[4].

Em resumo, um registro administrativo pode ser visto como um cadastro organizado que descreve exaustivamente certo fenômeno ou população de interesse público (por exemplo, todas as escolas e alunos matriculados no país, no caso do Censo Escolar, ou todas as empresas formais ativas, no caso do cadastro de CNPJ)[1]. Enquanto as pesquisas amostrais e censos nos fornecem informações planejadas porém esporádicas (anuais, decenais etc.), os registros administrativos oferecem um retrato contínuo, ainda que não originalmente concebido para análise estatística. A seguir, veremos como esses registros podem ser aproveitados na pesquisa em ciências sociais e na avaliação de políticas públicas.

Notas breves sobre o conceito de evidência

Os dados de registros administrativos vêm ganhando importância como fonte de evidências empíricas para subsidiar políticas públicas e pesquisas sociais. Eles podem ser utilizados em múltiplas etapas do ciclo de políticas – desde o diagnóstico de problemas até o monitoramento e avaliação de programas implementados[5][6]. Em termos de diagnóstico, registros administrativos permitem traçar um quadro detalhado da realidade social em determinada área. Por exemplo, dados administrativos educacionais podem identificar com precisão quantas crianças estão matriculadas em determinada série e quantas abandonam a escola durante o ano, fornecendo insumos para diagnosticar problemas de evasão ou atraso escolar. Registros de saúde pública, por sua vez, ajudam a caracterizar incidência de doenças e demandas por serviços médicos em diferentes regiões, servindo para apontar carências e grupos de risco. Uma vantagem nesse contexto é que, por serem exaustivos (no âmbito da política em questão), os RA fornecem contagens precisas e desagregadas – até níveis municipais ou menores – sem erros amostrais. Isso é valioso para localizar populações vulneráveis ou disparidades regionais. Por exemplo, o Cadastro Único permite mapear as famílias em pobreza extrema por bairro, informação essencial para o desenho focalizado de programas sociais.

Quando a política entra em fase de implementação, os registros administrativos continuam a ter uso fundamental no acompanhamento e monitoramento. Gestores podem utilizar sistemas de informação gerados pela própria execução da política para acompanhar metas e indicadores em tempo quase real. Um caso simples: o sistema de condicionalidades do Bolsa Família (Sicon) registra a frequência escolar e vacinação das crianças beneficiárias; essas informações administrativas servem para monitorar se os objetivos imediatos do programa (manter crianças na escola e vacinadas) estão sendo cumpridos e onde existem falhas operacionais. Outro exemplo é o orçamento público: os dados de execução orçamentária (despesas e repasses) são registros administrativos financeiros que podem ser monitorados

para verificar se recursos estão chegando onde previsto, permitindo correções de rumo durante a implementação. Muitas vezes, os RA são usados internamente como instrumentos de gestão: painéis e relatórios gerenciais são gerados a partir deles, facilitando decisões táticas. Essa prática de usar dados administrativos para monitorar ações também reforça a transparência e a prestação de contas, já que indicadores de desempenho podem ser divulgados publicamente para controle social.

Na avaliação de políticas – etapa em que se analisa os resultados e impactos de um programa – os registros administrativos se revelam especialmente poderosos. Como contêm informações longitudinais (ao longo do tempo) e frequentemente identificadores pessoais (sob restrições de acesso), eles possibilitam acompanhar trajetórias de indivíduos ou entidades antes e depois de uma intervenção. Isso abre caminho para avaliações quase-experimentais ou acompanhamento de coortes inteiras de beneficiários. Por exemplo, para avaliar o impacto de um programa de transferência de renda na saúde das crianças, pode-se ligar a base de beneficiários do programa aos registros de saúde (como hospitalizações ou mortalidade) e comparar indicadores de quem recebeu o benefício com quem não recebeu, controlando fatores pertinentes. Essa abordagem permite estudos observacionais robustos com populações inteiras, algo difícil de se obter com pesquisas amostrais tradicionais. Os registros administrativos também são úteis para medir o alcance das políticas (quantos foram atendidos, quem ficou de fora) e para identificar possíveis efeitos não intencionais (por exemplo, um cadastro de desempregados pode revelar se após um programa de qualificação as pessoas conseguiram emprego formal).

Em suma, as aplicações dos RA em políticas públicas podem ser resumidas em algumas grandes categorias de uso: (i) servirem de subsídio à formulação de políticas, oferecendo diagnóstico sobre o problema a ser enfrentado; (ii) funcionarem como instrumento para orientar a implementação, permitindo ajustes e alocação eficiente de recursos; (iii) atuarem no monitoramento contínuo das ações e seus resultados imediatos; (iv) apoiarem fiscalização e controle ao fornecer dados para auditorias e verificação da execução físico-financeira; e (v) servirem à transparência e prestação de contas, permitindo que cidadãos e órgãos de controle acompanhem os resultados e gastos de programas públicos[7][8]. Poucos registros isolados cumprem todas essas funções ao mesmo tempo, mas o conjunto do ecossistema de dados administrativos de um país pode abranger todos esses usos. No Brasil, por exemplo, a Portal da Transparência disponibiliza registros detalhados de despesas e receitas públicas em formato aberto, tornando possível tanto monitorar gastos (implementação) quanto exercer controle social e transparência (função v). Já bases como o CadÚnico fornecem insumos riquíssimos para formulação de políticas sociais (função i), ao mapear o público potencialmente elegível, e servem também para monitorar a inclusão de famílias (função iii).

Cabe destacar que pesquisadores acadêmicos também aproveitam registros administrativos para investigar questões sociais mais amplas que vão além da operação imediata de programas governamentais. Esses dados têm permitido estudos aprofundados sobre dinâmica populacional, mercado de trabalho, educação, mobilidade social, saúde, entre outros temas, muitas vezes revelando padrões que seriam invisíveis em amostras limitadas. No campo das ciências sociais, a possibilidade de cruzar diferentes registros tem levado a descobertas importantes, como veremos nos estudos de caso adiante. Em políticas públicas,

cada vez mais se reconhece que ampliar o uso inteligente desses registros pode melhorar o desenho e a avaliação de intervenções, baseando decisões em evidências sólidas[5][6]. Ao mesmo tempo, discute-se a necessidade de cautela ética e capacitação técnica para esse uso, tópicos que abordaremos nas seções sobre governança e perguntas frequentes.

Estado, políticas públicas e a produção de evidências

O aproveitamento de registros administrativos em pesquisa e avaliação traz uma série de vantagens significativas em comparação com dados coletados exclusivamente via surveys ou censos esporádicos. Primeiramente, a amplitude de cobertura costuma ser muito maior. Como os RA derivam de obrigações ou serviços universais (ou quase universais), eles frequentemente abrangem toda a população relevante ou, pelo menos, um conjunto muito amplo de casos. Isso elimina erros amostrais e viabiliza análises de subgrupos populacionais menores e áreas geográficas específicas com alto nível de detalhe. Por exemplo, a RAIS, que registra todos os vínculos formais de trabalho declarados pelas empresas, permite tabular informações de emprego formal em qualquer município do país, algo impossível de se obter com precisão em uma pesquisa amostral que visite apenas algumas casas em cada cidade.

Uma segunda vantagem é o caráter longitudinal e dinâmico dos registros. Muitos RA acompanham continuamente os indivíduos ao longo do tempo – seja registrando a permanência deles em um programa, seja atualizando sua situação anualmente. Isso possibilita a construção de séries históricas e o seguimento de trajetórias individuais (painel de dados). Em vez de uma "foto" estática, os RA oferecem um "filme" do processo social, permitindo estudar transições (por exemplo, entrada e saída do mercado de trabalho, mobilidade entre faixas de renda, progressão educacional, migração entre cidades, etc.). Essa característica temporal fortalece análises de causabilidade e tendências, pois podemos observar o antes e depois de políticas ou eventos na mesma pessoa ou família.

Outra vantagem chave é o baixo custo incremental de usar dados já coletados. Conduzir grandes surveys ou censos é caro e trabalhoso; em contraste, quando um registro administrativo de qualidade já existe, reutilizá-lo para pesquisa tem custo marginal muito menor. Há economia de recursos públicos ao evitar duplicidade de esforços – por exemplo, se os dados de emprego da RAIS já existem administrativamente, o governo economiza em não precisar coletar as mesmas informações via pesquisa amostral. Roberto Olinto, ex-presidente do IBGE, destaca que o uso de RA pode reduzir custos, evitar esforços duplicados, integrar bases de dados e diminuir a carga sobre os informantes (cidadãos e empresas), acelerando também a disseminação de estatísticas atualizadas[9]. Em outras palavras, aproveitar registros existentes é uma forma de obter mais conhecimento com menos investimento adicional, tornando a produção de estatísticas e evidências mais eficiente.

Além disso, registros administrativos frequentemente contêm detalhes e variáveis não disponíveis em pesquisas padrão. Como são formulados para a gestão, eles podem incluir informações muito específicas (por exemplo, cada procedimento médico realizado em um hospital público, ou cada nota escolar de um aluno, ou o histórico contributivo completo

de um trabalhador) que dificilmente seriam levantadas em entrevistas domiciliares. Essa riqueza de detalhes abre novas frentes de investigação. Por exemplo, os microdados do Imposto de Renda (quando acessíveis sob sigilo) têm informações detalhadas sobre fontes de renda, deduções e patrimônio, permitindo estudar distribuição de renda e riqueza com muito mais precisão no topo da distribuição do que pesquisas como a PNAD conseguem.

Por fim, uma vantagem pragmática: muitos registros administrativos são atualizados em tempo quase real ou com periodicidade curta (mensal, trimestral, anual), fornecendo dados frescos para análise. Isso é valioso para monitoramento conjuntural e resposta rápida – por exemplo, acompanhar mensalmente o número de beneficiários de seguro-desemprego durante uma crise econômica, algo possível com RA, enquanto uma pesquisa amostral poderia levar meses para coletar e processar informações semelhantes. Essa capacidade de resposta rápida com RA tornou-se evidente durante a pandemia de COVID-19, quando registros como os de notificações de casos e óbitos e os de pagamentos de auxílios emergenciais foram fundamentais para acompanhamento em tempo real, ao passo que surveys convencionais enfrentaram dificuldades logísticas.

No entanto, juntamente com esses benefícios, é crucial reconhecer os limites e desafios do uso de registros administrativos. Uma primeira limitação decorre justamente de seu propósito original: os RA não são coletados com objetivos de pesquisa, mas sim administrativos. Isso significa que problemas de qualidade podem existir – dados podem ser preenchidos de forma incompleta ou incorreta pelos informantes ou agentes administrativos, especialmente em campos que não afetam diretamente a concessão de um serviço. Por exemplo, em um formulário de matrícula escolar, o foco do gestor escolar pode ser confirmar a identidade e série do aluno (informações essenciais), mas campos como renda familiar informada podem ter qualidade inconsistente, pois não impactam diretamente a matrícula. Portanto, o pesquisador precisa avaliar criticamente precisão, validade e confiabilidade dos campos de um RA antes de usá-los[10][11]. Conceitos de controle de qualidade de dados – detecção de outliers, consistência longitudinal, cruzamento com outras fontes – tornam-se fundamentais ao trabalhar com RA.

Um segundo desafio é a heterogeneidade metodológica e mudanças ao longo do tempo. Diferentes órgãos coletam dados de maneiras distintas, e mesmo dentro de um mesmo cadastro as definições podem mudar. Um programa social pode alterar critérios de elegibilidade ou formulário de cadastro de um ano para outro; a classificação de doenças em sistemas de saúde pode mudar com novas normas; até mesmo variáveis básicas como "escolaridade" podem ser registradas com categorias diferentes após reformas educacionais. Essa heterogeneidade dificulta comparações históricas e integração de múltiplas fontes – requer esforço de harmonização e documentação das mudanças. Conforme apontado por Janine Mello (Ipea), fatores limitantes para uso confiável dos RA incluem a falta de padronização nas metodologias de coleta ao longo do tempo e entre órgãos, o que pode introduzir quebras de série e inconsistências[12][13]. Em suma, o pesquisador muitas vezes precisa "limpar" e padronizar os dados antes da análise, investindo tempo para compreender o contexto administrativo por trás de cada variável.

Outra limitação importante é a cobertura seletiva. Apesar de, em teoria, muitos registros serem universais (cobrir toda a população alvo), na prática sempre há segmentos que ficam de fora porque não interagem com o sistema formal. Por exemplo, a RAIS não inclui

trabalhadores informais, pois registra apenas empregos com carteira assinada declarados; logo, para estudar plenamente o mercado de trabalho, um pesquisador precisaria complementar com outras fontes os dados daqueles excluídos do registro administrativo. De forma semelhante, o CadÚnico cobre majoritariamente famílias de baixa renda inscritas para programas – mas famílias extremamente vulneráveis, como moradores de rua ou populações isoladas, podem não estar cadastradas, gerando lacunas de cobertura. Assim, ao usar RA, é preciso ter clareza sobre "quem está presente e quem está ausente" nos dados, para não tirar conclusões enviesadas sobre a realidade geral. Em outras palavras, os registros refletem a população administrada, não necessariamente a população total – e a diferença entre uma e outra pode ser substantiva.

Um desafio não trivial é o acesso e questões legais/éticas. Muitos registros administrativos contêm informações pessoais sensíveis (como CPF, nome, endereço, saúde, renda) protegidas por sigilo legal, o que impede sua liberação irrestrita. Pesquisadores esbarram em barreiras legais (por exemplo, a Lei Geral de Proteção de Dados – LGPD no Brasil, Lei nº 13.709/2018) e operacionais para obter microdados individualizados. Mesmo quando o acesso é permitido, ele pode ocorrer sob condições controladas – como em salas seguras ou mediante termos de confidencialidade – o que nem sempre é simples para iniciantes ou instituições sem convênios. A desatualização tecnológica de alguns órgãos também limita o acesso: há bases de dados sem sistemas fáceis de extração ou sem documentação pública. Embora haja um movimento crescente de dados abertos, estes tendem a ser dados agregados ou anonimizados, o que às vezes não atende a todas as necessidades de pesquisa. Voltaremos a este ponto ao falar de acesso programático.

Por fim, vale mencionar que registros administrativos, por melhores que sejam, não substituem totalmente as pesquisas amostrais e censos em todos os casos. Há tipo de informação que não é coletada administrativamente simplesmente porque não há motivo operacional – por exemplo, opiniões, atitudes ou mesmo alguns indicadores de renda e emprego (no setor informal) precisam de inquéritos amostrais para serem conhecidos. Além disso, quando se deseja inovar em um levantamento de dado não registrado em nenhum sistema (por exemplo, medir vitimização criminal em população geral, ou levantar uso de tempo diário em atividades), não há registro administrativo prévio a recorrer. Portanto, a abordagem mais poderosa costuma ser o uso complementar: combinar registros administrativos para aquilo que eles têm de melhor (cobertura ampla e dados objetivos registrados continuamente) com pesquisas amostrais onde for necessário captar informações específicas ou qualitativas. Essa complementaridade maximiza o conhecimento disponível, reduzindo custos e aumentando a qualidade das inferências[14].

Em síntese, os registros administrativos oferecem vantagens inegáveis de escala, detalhe e frequência, possibilitando análises ricas e políticas baseadas em evidências mais sólidas. Entretanto, o pesquisador ou gestor que os utiliza deve estar atento aos limites de qualidade, cobertura e acessibilidade, adotando estratégias para mitigar esses problemas – como limpeza de dados, triangulação com outras fontes e cuidados éticos na proteção das informações pessoais. Na próxima seção, examinaremos algumas metodologias práticas que ajudam a extrair valor dos RA enquanto enfrentamos esses desafios.

Diferenças entre registros administrativos e pesquisas amostrais

Registros administrativos existem para gerir políticas e programas (pagamentos, elegibilidade, controle operacional) e não são coletados com fins primários de inferência estatística; já as pesquisas amostrais são desenhadas para representar a população segundo um plano amostral explícito (estratificação, conglomeração, pesos), com vistas a produzir estimativas válidas para múltiplos temas. Por isso, é esperado que estimativas oriundas de pesquisas amostrais divirjam dos totais "oficiais" dos RA, pois as fontes têm propósitos, coberturas e processos diferentes. No caso brasileiro, mostro que parte substantiva dessas divergências decorre de duas fontes gerais: viés de representatividade (quando a amostra não reflete bem a distribuição territorial do fenômeno) e viés de captação (quando, mesmo nos estratos amostrados, o fenômeno é mal captado por desenho de questionário, erro de declaração ou dificuldade de alcançar subpopulações).

Para tornar concreto: a PNAD 2001–2009 selecionava todos os grandes centros (RMs e municípios autorrepresentativos) e apenas uma fração dos demais (NAR), com probabilidade proporcional ao tamanho; se o programa a ser estudado se concentra justamente nos pequenos municípios não incluídos, a PNAD tende a subestimar seu alcance, ainda que funcione bem para outros temas. Nesse cenário, diferencio formalmente os componentes "representatividade" e "captação" e decomponho o erro de inferência em três parcelas: viés de representatividade, viés de captação e a interação entre ambos.

Aplicando a metodologia a dois casos críticos, encontro padrões distintos com consequências práticas. Para o Bolsa Família (PBF), cerca de 40% da diferença PNAD × registros administrativos se explica pelo desenho amostral — isto é, por viés de representatividade associado à subcobertura de pequenos municípios — e outros 57% por captação; mesmo com "captação perfeita" no desenho antigo, a PNAD permaneceria aquém do total oficial. Isso recomenda cautela ao usar a PNAD para inferências substantivas sobre programas fortemente interiorizados e sugere ganhos ao combinar PNAD com RA e, quando possível, redesenhos amostrais que ampliem a cobertura municipal. Já para o BPC, o desenho amostral tenderia, se algo, a superestimar; aqui a diferença observada decorre essencialmente de captação (inclusive confusão de declaração com aposentadorias no período pré-2004), o que desloca o foco para melhoria de questionários e treinamento.

Em termos de agenda metodológica, duas implicações se seguem para pesquisas sociológicas: primeiro, é preciso incorporar o plano amostral e seus limites ao interpretar totais e prevalências, sobretudo para políticas "territorializadas"; segundo, convém triangular RA e surveys, explorando cada fonte conforme sua vocação — RA para totais administrativos e séries por município; surveys para variáveis não observadas em RA e para análises multitemáticas, sempre com pesos e desenho apropriados. Essas recomendações alinham-se às diretrizes técnicas do IBGE sobre uso de planos amostrais complexos e à transição para a "Amostra Mestra" que ampliou substantivamente a cobertura municipal, mitigando o viés de representatividade observado no desenho anterior.

Em síntese: RA e surveys são complementares. Onde houver forte interiorização do fenômeno ou riscos de erro de declaração, a estratégia sociológica prudente é: (i) explicitar

o componente de representatividade e de captação que pode afetar a inferência; (ii) usar RA como "denominador" de referência; e (iii) documentar limitações e sensibilidade das conclusões à fonte de dados utilizada.

Registros administrativos enquanto evidências

Explorar registros administrativos em pesquisa exige não apenas conhecimento substantivo do tema estudado, mas também a aplicação de métodos e técnicas específicas de ciência de dados e estatística. Aqui, apresentamos de forma acessível algumas noções básicas e boas práticas fundamentais: record linkage (ligação de registros), anonimização e proteção de dados pessoais, uso de dados sintéticos e a importância de documentação e metadados adequados.

Record Linkage (Linkagem de Registros): Uma das operações mais poderosas ao trabalhar com RA é a capacidade de combinar diferentes bases de dados que se referem aos mesmos indivíduos, famílias ou entidades, ampliando enormemente o escopo da análise. Record linkage refere-se justamente ao processo de ligar registros correspondentes de fontes distintas. Por exemplo, podemos querer vincular o registro de uma pessoa no CadÚnico (assistência social) com seu registro no SIM (mortalidade) para estudar como a participação em programas sociais afetou sua sobrevivência ou causas de morte. Para realizar essa ligação, é necessário um identificador comum ou quase comum. No Brasil, o CPF (Cadastro de Pessoa Física) é um identificador único nacional para indivíduos que, se presente nas bases, simplifica muito o linkage – basta fazer a junção exata pelo CPF. Em muitos casos, porém, nem todas as bases possuem CPF ou um identificador padronizado compartilhado. Nesses casos, recorre-se a métodos de linkage probabilístico ou aproximado, usando um conjunto de informações como nome, data de nascimento, nome da mãe e outras características para estimar a correspondência entre registros. Centros de pesquisa brasileiros desenvolveram algoritmos para isso, como o CIDACS-RL, uma ferramenta de código aberto criada pelo CIDACS/Fiocruz para vincular grandes bases de saúde e assistência social, gerando um escore de similaridade entre possíveis pares de registros[15]. Quando há um identificador administrativo como o NIS (Número de Identificação Social, usado em programas sociais) ou o PIS/PASEP (identificador trabalhista), o linkage pode ser determinístico – por exemplo, ao conectar beneficiários do Bolsa Família aos dados do próprio programa via NIS, obtém-se correspondência exata[16]. Em outros casos, usa-se linkage probabilístico com softwares especializados, seguido muitas vezes por revisão manual em amostras para validar a precisão do emparelhamento[17]. Para pesquisadores iniciantes, a mensagem principal é: combinar bases expande o potencial analítico, mas requer cuidados para evitar erros de ligação (falsos positivos ou perda de pares verdadeiros). Documentar critérios de linkage e eventualmente calcular medidas de qualidade (como porcentagem de registros ligados ou taxa de erro) faz parte das boas práticas.

Anonimização e Proteção de Dados: Dado que registros administrativos frequentemente contêm dados pessoais identificáveis, é fundamental aplicar técnicas de anonimização antes de utilizá-los em pesquisa (a não ser em ambientes controlados). Anonimizar significa remover ou transformar elementos que possam identificar diretamente alguém (como nome, CPF, endereço exato) e também tomar medidas para reduzir o risco de reidentificação

indireta (ou seja, combinar informações que, em conjunto, acabem apontando para um indivíduo único). Por exemplo, datas de nascimento podem ser agregadas em faixas etárias; nomes podem ser substituídos por códigos; endereços podem ser truncados para nível de bairro ou município apenas. A LGPD no Brasil estabelece que dados pessoais sensíveis só devem ser usados para pesquisa se houver bases legais e preferencialmente de forma anonimizada quando possível. Instituições frequentemente liberam microdados anonimizados dos registros – ou seja, os dados detalhados porém sem identificadores pessoais. Um exemplo é a disponibilização de microdados da RAIS para pesquisa: o governo tradicionalmente libera um arquivo público onde cada trabalhador tem um código aleatório no lugar do CPF, impossibilitando a identificação direta da pessoa (ainda que contenha informações como idade, sexo, ocupação, salário). Contudo, mesmo após a anonimização, existe o risco de ataque de reidentificação se o dataset tiver entradas muito específicas (por exemplo, alguém com combinação única de idade avançada, cargo raro e salário muito alto numa pequena cidade poderia ser identificável). Para lidar com isso, adota-se também supressão ou borramento de algumas variáveis em casos extremos, ou a geração de dados sintéticos (como veremos adiante) em vez dos dados reais. Em suma, a anonimização é um processo crucial e deve ser planejada caso a caso, balanceando utilidade analítica e proteção de privacidade. Um arcabouço internacionalmente reconhecido para pensar essa questão é o modelo dos Five Safes (Cinco Seguranças): acesso a dados para pesquisa deve garantir projeto seguro, pessoas seguras (pesquisadores confiáveis), dados seguros (anonimizados ou controlados), ambientes seguros (plataformas de acesso protegidas) e saídas seguras (resultados checados para não revelar indivíduos)[18]. Esse modelo, proposto por Desai, Ritchie e Welpton (2016), vem sendo adotado por órgãos estatísticos mundo afora e serve de guia para programas de acesso a microdados sensíveis no Brasil também.

Dados Sintéticos: Uma estratégia interessante que tem ganhado espaço para conciliar privacidade e transparência é a geração de dados sintéticos. Dados sintéticos são conjuntos de dados artificiais gerados por modelos estatísticos ou de aprendizado de máquina, que reproduzem as principais características estatísticas do conjunto de dados real sem conter registros verdadeiros de indivíduos. Em outras palavras, é como criar um "sósia"do banco de dados original: as distribuições, correlações e padrões gerais são mantidos, mas todos os registros pertencem a ficções estatísticas, não a pessoas reais. Esses dados podem ser liberados publicamente sem risco de expor alguém, pois mesmo que um usuário conseguisse "identificar"um registro sintético, ele não correspondiria a nenhuma pessoa de verdade – no máximo indicaria um perfil plausível estatisticamente. Órgãos como o IBGE e alguns centros de pesquisa internacionais experimentam fornecer microdados sintéticos para pesquisadores explorarem e desenvolverem seus códigos de análise; posteriormente, a análise pode ser validada nos dados reais restritos dentro de um ambiente seguro. Para iniciantes, a noção de dado sintético pode parecer estranha, mas é análoga a criar simulações fiéis à realidade. Por exemplo, imagine que um registro real tenha 52% de mulheres e 48% de homens, com idades distribuídas de certa forma e renda correlacionada à escolaridade; um algoritmo de síntese de dados tentaria gerar outro conjunto de registros que preserve esses percentuais e relações, sem copiar nenhum indivíduo exato. Essa técnica depende de modelos sofisticados (como algoritmos de imputação múltipla, modelos generativos tipo redes neurais, etc.), e não substitui completamente a necessidade de acesso a dados

reais para análises finais, mas é uma ferramenta promissora para ampliar o acesso inicial a dados semelhantes aos administrativos, ajudando pesquisadores a desenvolver métodos e divulgar resultados reprodutíveis sem infringir privacidade.

Documentação e Metadados: Por fim, mas não menos importante, o trabalho com RA requer muita atenção à documentação. Metadados – isto é, dados sobre os dados – são indispensáveis para compreender e usar corretamente qualquer base administrativa. Diferente de surveys bem documentados em relatórios metodológicos, alguns registros administrativos podem vir com documentação escassa. É essencial buscar informações como: definição de cada campo/variável (o que exatamente significa, unidades, categorias possíveis), população coberta, periodicidade de atualização, mudanças de formato ao longo do tempo, códigos e abreviações utilizadas, e procedimentos de coleta. Muitas vezes, essa documentação não está reunida num único manual, exigindo ao pesquisador vasculhar portarias, manuais operacionais internos dos sistemas ou artigos técnicos produzidos por quem já utilizou a base. Sempre que possível, deve-se catalogar os metadados de interesse: por exemplo, elaborar uma tabela dicionário de dados com o nome de cada variável do RA, sua descrição, tipo (numérico, texto, etc.), possíveis valores e significados (códigos de categorias). Isso facilita não apenas o entendimento próprio durante a análise, mas também a reprodutibilidade do trabalho por outros. Documentar as transformações e filtros aplicados também é parte das boas práticas: por exemplo, se o pesquisador decide excluir registros com certo código de erro ou limitar a análise a determinada região e período, esses critérios devem ficar claros em anotações ou nos scripts de processamento de dados. Instituições públicas no Brasil têm avançado na publicação de metadados – por exemplo, o IBGE mantém catálogos de variáveis e classificações, alguns ministérios publicam manuais de sistemas (o Ministério da Saúde, por exemplo, divulga dicionários de dados para os sistemas DATASUS). Entretanto, persistem lacunas, e por isso a iniciativa individual de investigar o contexto do registro é imprescindível. Uma dica prática: quando obter acesso a um RA, procure se há portarias normativas ou leis que descrevem aquele sistema (muitas vezes elas trazem anexos com formulários e definições), leia artigos ou relatórios de outros pesquisadores que já utilizaram a base (eles costumam relatar desafios e decisões metodológicas), e não hesite em contatar técnicos do órgão responsável se for possível, para esclarecer dúvidas de interpretação. Em resumo, tratar registros administrativos com rigor científico requer tão rigorosa documentação quanto qualquer outro dado – isso garantirá que as análises sejam confiáveis e que outros possam entendê-las e reproduzi-las se necessário. Estudos de Caso Ilustrativos

Para concretizar as possibilidades proporcionadas pelos registros administrativos, analisemos brevemente dois estudos de caso. O primeiro exemplo vem do campo da mobilidade social intergeracional, ilustrando como dados fiscais administrativos permitiram inovações no entendimento da desigualdade de oportunidades. O segundo caso é a Coorte de 100 Milhões de Brasileiros (CIDACS), um enorme projeto nacional que integra registros sociais e de saúde para avaliar políticas públicas de forma pioneira.

Usos e funções dos registros administrativos no Brasil

Um dos desafios clássicos das ciências sociais é medir o grau em que as condições socioeconômicas dos pais influenciam a renda e oportunidades dos filhos na vida adulta – ou seja, qual é a mobilidade intergeracional em determinada sociedade. Tradicionalmente, esses estudos dependiam de surveys que perguntavam simultaneamente aos filhos sobre a renda/educação dos pais, ou então comparavam estatísticas de diferentes gerações em censos. Entretanto, essas abordagens sofriam com dados limitados e potenciais vieses de memória. Foi quando, na última década, economistas começaram a explorar registros administrativos fiscais para esse propósito. Nos Estados Unidos, um trabalho marcante de Raj Chetty e colaboradores revolucionou o campo ao usar os registros do imposto de renda (IRS) de milhões de indivíduos, vinculando declarações de pais e filhos ao longo de décadas para calcular medidas precisas de mobilidade[19][20]. O estudo "Where is the Land of Opportunity?"(Chetty et al., 2014) analisou uma coorte inteira de crianças nascidas no início dos anos 1980 nos EUA, cuja renda adulta por volta de 30 anos pôde ser determinada nos registros fiscais, e comparou com a renda de seus pais quando essas crianças eram adolescentes[21][22].

Os resultados revelados por meio desses registros foram reveladores: longe de haver uma única resposta sobre a mobilidade nos EUA, descobriu-se enorme variação regional. Em algumas cidades e áreas, as perspectivas de ascensão para filhos de famílias pobres eram relativamente altas – a ponto de certos locais apresentarem mobilidade comparável à de países escandinavos altamente móveis – enquanto em outras regiões poucas crianças escapavam da pobreza, com mobilidade pior que a observada em qualquer país desenvolvido[23]. Por exemplo, o estudo mostrou que um filho de família de baixa renda (25º percentil da distribuição) criado na cidade de Seattle alcançava, em média, um patamar de renda na vida adulta equivalente ao de um filho de família de renda mediana criado em Atlanta[24]. Ou seja, crescer em Seattle trazia uma "vantagem de mobilidade"tão grande que posicionava o jovem muito acima do que seu equivalente em Atlanta conseguiria. Cidades como Salt Lake City e San José destacaram-se com índices de mobilidade ascendente comparáveis aos de países líderes em igualdade de oportunidades, enquanto lugares como Atlanta e Milwaukee apresentaram mobilidade extremamente baixa, inferior a de qualquer país rico documentado[25].

Esses achados só foram possíveis graças à abrangência dos dados administrativos: ao compilar milhões de registros anônimos de renda, os pesquisadores conseguiram estimar taxas de mobilidade com granularidade municipal e identificar correlações importantes. Por exemplo, correlacionando os dados de mobilidade com características locais, encontraram fatores associados à maior mobilidade (menor desigualdade local, melhores escolas, maior capital social, estruturas familiares mais estáveis, entre outros)[26][27]. Embora o estudo em si tenha foco nos EUA, metodologicamente ele inspirou pesquisas em todo o mundo, inclusive no Brasil.

No contexto brasileiro, por muitos anos a falta de dados vinculando gerações dificultou medições precisas de mobilidade intergeracional de renda. Pesquisadores baseavam-se em inquéritos amostrais (como PNAD) que têm informações limitadas sobre renda dos pais. Recentemente, porém, iniciaram-se esforços para replicar abordagens administrativas.

Uma iniciativa de destaque é o Atlas da Mobilidade Social lançado em 2025 pelo Instituto Mobilidade e Desenvolvimento Social (IMDS). Esse projeto inédito no país utilizou a vinculação entre registros de pais e filhos em bases administrativas para estimar a mobilidade intergeracional de renda de brasileiros nascidos na década de 1980[28].

Embora os microdados detalhados não sejam públicos, o Atlas apresenta visualizações e indicadores calculados a partir de dados fiscais e trabalhistas vinculados: os pesquisadores conectaram indivíduos aos seus pais por meio de identificadores administrativos e compararam as rendas familiares dos pais (observadas entre 1990-2010 via registros tributários e trabalhistas) com as rendas dos filhos por volta dos 30 anos (entre 2015-2020)[29]. Para contornar a ausência de dados formais sobre renda de pais que trabalhavam no setor informal, técnicas de modelagem foram empregadas a fim de estimar rendimentos informais a partir de pesquisas domiciliares, complementando os registros oficiais[30].

O Atlas evidenciou um quadro de baixa mobilidade: cerca de dois em cada três filhos provenientes das famílias mais pobres permaneceram na metade inferior da distribuição de renda na vida adulta, e menos de 2% conseguiram ascender ao grupo dos 10% mais ricos[31]. Também foram encontradas fortes disparidades regionais – por exemplo, no Norte e Nordeste, aproximadamente 75% dos filhos de famílias pobres continuavam nos estratos de renda mais baixos, ao passo que no Sul essa proporção, embora ainda alta, era cerca de 40%[32].

Esses números reforçam que o Brasil tem mobilidade intergeracional muito limitada, confirmando análises anteriores, mas agora com um nível de detalhe e robustez muito maior graças aos registros administrativos integrados. Em suma, tanto o caso norte-americano de Chetty et al. quanto a experiência brasileira recente demonstram o poder dos RA fiscais: eles fornecem o vínculo empírico entre gerações necessário para quantificar a "transmissão"de vantagens e desvantagens socioeconômicas, algo que antes ficava no campo da especulação teórica ou de evidências fragmentadas.

Este estudo de caso ilustra para o iniciante como o uso criativo de grandes bases governamentais pode revolucionar nossa compreensão de problemas sociais complexos. Ao mesmo tempo, deixa lições importantes: a necessidade de preservar o sigilo (no caso de Chetty, todos os dados foram anonimizados e acessados via acordo com o fisco; no caso brasileiro, resultados foram divulgados apenas de forma agregada por região), a importância de infraestrutura computacional para lidar com milhões de registros, e o valor de parcerias entre órgãos detentores dos dados e pesquisadores para viabilizar esse tipo de análise.

Coorte de 100 Milhões de Brasileiros

Nosso segundo estudo de caso explora um uso de registros administrativos na fronteira entre saúde pública e avaliação social: a Coorte de 100 Milhões de Brasileiros, desenvolvida pelo Centro de Integração de Dados e Conhecimentos para Saúde (CIDACS), ligado à Fiocruz. Trata-se de um projeto ambicioso que criou, essencialmente, uma gigantesca coorte histórica de metade da população brasileira, ao integrar diversos registros administrativos para pesquisa epidemiológica e de políticas sociais[33]. A iniciativa partiu do reconhecimento de que o Brasil dispõe de muitas bases de dados de alta qualidade – porém dispersas

entre setores – e da oportunidade de vinculá-las para responder a perguntas cruciais sobre determinantes sociais da saúde e efeitos de intervenções em larga escala[34].

A base central da Coorte de 100 Milhões é o Cadastro Único (CadÚnico), que engloba informações socioeconômicas de famílias de baixa renda inscritas em programas sociais. O foco foi nos cadastrados entre 2001 e 2015, correspondendo a aproximadamente 114 milhões de pessoas – número que confere o nome à coorte e que representa quase 50% da população do país nesse período[33]. Esses indivíduos formam o "baseline" da coorte, com seus dados demográficos e de condição de vida registrados no momento da primeira inscrição no CadÚnico (que inclui idade, sexo, composição familiar, escolaridade, renda, moradia, etc.). A riqueza deste cadastro é enorme para pesquisa, pois delineia a situação social inicial de um enorme contingente populacional majoritariamente pobre.

A inovação veio ao conectar essa coorte inicial com outros sistemas administrativos de saúde e programas, de modo a acompanhar os resultados ao longo do tempo. Por exemplo, a coorte foi linkada com o Sistema de Informações de Mortalidade (SIM) para identificar óbitos ocorridos entre os participantes, e com o Sistema de Nascidos Vivos (SINASC) para coletar informações sobre os bebês nascidos das mães da coorte (peso ao nascer, prematuridade etc.)[35]. Além disso, integrou-se o histórico de benefícios do programa Bolsa Família, cruzando os indivíduos da coorte com a folha de pagamentos do Bolsa Família (BFP) para saber quem efetivamente recebeu a transferência de renda e em que período[35]. Essas vinculações usaram tanto métodos determinísticos (por exemplo, o NIS de cada pessoa para achar seu registro no pagamento do Bolsa Família) quanto métodos probabilísticos (para relacionar a coorte com registros de óbitos e nascimentos, usando nomes, datas e outros identificadores via o algoritmo CIDACS-RL)[15][36].

Todo o processo foi realizado respeitando confidencialidade: os dados nominativos ficaram em ambiente seguro e somente bases pseudoanônimas (com identificadores substituídos por códigos) foram usadas nas análises, garantindo que os pesquisadores não tivessem acesso a nomes ou CPFs individuais[37]. Com a coorte montada e enriquecida por essas ligações, abriu-se uma oportunidade única de realizar estudos de avaliação de impacto e de determinação em saúde em escala massiva. Até o momento, a Coorte de 100 Milhões já foi utilizada para investigar, por exemplo, os efeitos do programa Bolsa Família sobre desfechos de saúde. Um estudo publicado em 2021 examinou se crianças beneficiárias do Bolsa Família tinham menores taxas de mortalidade infantil e encontrou evidências de redução significativa de risco de óbito entre 1 e 4 anos de idade nas famílias atendidas[38]. Outro trabalho avaliou o impacto do Bolsa Família na incidência de hanseníase (lepra) – uma doença negligenciada associada à pobreza – e concluiu que a expansão da cobertura do programa contribuiu para diminuir casos da doença, possivelmente por melhorar as condições de vida e acesso a cuidados básicos[38]. Além das avaliações de programas, a coorte vem sendo usada para mapear determinantes sociais da saúde em geral: pesquisadores exploraram como fatores demográficos e socioeconômicos (capturados no CadÚnico) influenciam indicadores como baixo peso ao nascer, desfechos de tratamento de doenças infecciosas, entre outros[38].

A importância desse caso de uso de RA reside em vários pontos. Primeiro, demonstra a viabilidade de integrar dados de múltiplas fontes em um país de dimensão continental como o Brasil, superando barreiras burocráticas e técnicas para montar um recurso de pesquisa de altíssimo valor. Segundo, exemplifica o ganho em poder estatístico e granularidade:

com dezenas de milhões de pessoas acompanhadas, é possível detectar efeitos que seriam sutis demais para amostras pequenas e analisar subpopulações específicas (por exemplo, efeitos de programa em crianças indígenas da região Norte, ou diferenças entre áreas urbanas e rurais) com significância. Terceiro, reforça a sinergia entre políticas sociais e saúde – algo que só fica evidente quando se combinam registros tradicionalmente separados: constatou-se, por exemplo, que intervenções de redução da pobreza refletem diretamente em indicadores de saúde coletiva, justificando abordagens intersetoriais. Finalmente, o caso do CIDACS enfatiza a questão da governança ética: para acessar e vincular esses dados, foram estabelecidos protocolos claros de autorização com os ministérios responsáveis (Desenvolvimento Social, Saúde etc.), e a operação foi cercada de salvaguardas de segurança e comitês de ética analisando os projetos derivados. Todo o trabalho se apoia no conceito de stewardship de dados – a Fiocruz/CIDACS atuando como uma guardiã confiável que recebe os dados sensíveis e os utiliza para pesquisa em prol do interesse público, sem comprometer a privacidade dos cidadãos.

Em suma, a Coorte de 100 Milhões ilustra as possibilidades transformadoras dos registros administrativos para a pesquisa social e de saúde no Brasil. É um exemplo palpável de big data público bem aproveitado: ao invés de os dados coletados pelo Estado ficarem confinados em silos, eles foram integrados (com responsabilidade) para gerar evidências científicas sobre problemas complexos, como a efetividade de políticas de combate à pobreza na melhoria de condições de vida. Para estudantes iniciantes, esse caso evidencia o porquê de se investir em aprender a lidar com RA – as descobertas obtidas podem ter impacto real na formulação de políticas mais acertadas. Também serve de inspiração sobre a importância de interdisciplinaridade: profissionais de saúde, ciência de dados, estatística e políticas sociais uniram expertise para tornar esse projeto possível.

Governança e Ética no Uso de Registros Administrativos na Pesquisa em Políticas Públicas

O uso de registros administrativos em pesquisa traz não apenas desafios técnicos, mas também importantes considerações de governança, ética e segurança dos dados. Quando se lida com informações sobre pessoas – muitas vezes obtidas sem um consentimento específico para pesquisa, mas sim para finalidades administrativas – é fundamental adotar frameworks que garantam que esses dados sejam usados de forma responsável, protegendo os direitos dos indivíduos e ao mesmo tempo possibilitando o avanço do conhecimento e a melhoria das políticas públicas.

Um elemento central nesse debate é o já mencionado modelo dos Five Safes (Cinco Seguranças). Desenvolvido originalmente em órgãos estatísticos de países como Reino Unido e Austrália, esse modelo estabelece cinco dimensões a serem asseguradas em qualquer projeto que utilize microdados sensíveis: Safe Projects (Projetos Seguros) – a pesquisa deve ter uma finalidade legítima e socialmente benéfica, passando por avaliação ética e legal; Safe People (Pessoas Seguras) – somente pesquisadores qualificados e autorizados devem ter acesso aos dados, geralmente após treinamento em confidencialidade; Safe Data (Dados Seguros) – os dados fornecidos devem ser tratados para minimizar riscos

(por exemplo, já fornecidos anonimizados ou com detalhes sensíveis removidos); Safe Settings (Ambientes Seguros) – o acesso deve ocorrer em ambientes controlados (como data centers seguros, computadores sem acesso à internet, ou via VPNs com múltiplas camadas de autenticação), evitando extravasamento de informações; e Safe Outputs (Resultados Seguros) – qualquer resultado antes de publicação deve ser revisado para garantir que nenhuma informação individual possa ser inferida (por exemplo, tabelas que tenham células com poucos indivíduos devem ser suprimidas ou agregadas)[18].

Esse arcabouço vem sendo difundido também no Brasil. Por exemplo, o Ipea, em discussão sobre acesso a dados trabalhistas, destaca a importância do conceito de five safes para orientar a abertura de bases como RAIS e CAGED para pesquisa acadêmica[39]. A implementação prática pode envolver desde acordos de confidencialidade e penalidades legais para uso indevido, até a criação de laboratórios de acesso – espaços físicos ou virtuais onde pesquisadores podem analisar os microdados, mas sem poder baixar uma cópia integral, apenas extrair resultados já aprovados. O IBGE, por exemplo, opera um laboratório desse tipo (Centro de Dados para Pesquisa) para acesso a microdados censitários detalhados; modelos similares poderiam ser aplicados a registros administrativos.

Conectado a isso está o conceito de stewardship de dados, que podemos traduzir como zeladoria ou curadoria responsável dos dados. Em muitas situações, não é viável ou desejável que dados brutos de um registro administrativo fiquem livremente disponíveis. Então entra em cena o papel de instituições públicas ou parcerias de pesquisa que atuam como guardiãs: elas recebem os dados dos órgãos originais mediante acordos, limpam e documentam esses dados, e oferecem acesso regulado aos pesquisadores. O caso do CIDACS exemplifica essa função – a Fiocruz se tornou depositária dos dados do CadÚnico e outros, sob confiança do governo, e estabeleceu procedimentos para seu uso científico. Esse modelo de stewardship é crucial para viabilizar pesquisas complexas sem comprometer ética: garante-se que há alguém "tomando conta" dos dados, atualizando, controlando quem utiliza e para quê, e garantindo retorno dos resultados para a sociedade. Um bom stewardship também implica transparência sobre quais dados existem e como podem ser acessados. No Brasil, o avanço de portais de transparência e catálogos de dados abertos (como o dados.gov.br) ajuda nesse mapeamento, embora muitas bases administrativas complexas ainda não estejam listadas lá de forma útil ao pesquisador.

A fonte de financiamento e sustentabilidade dos bancos de dados administrativos é outro aspecto de governança. Manter e aprimorar registros exige recursos financeiros e humanos contínuos: sistemas de TI atualizados, servidores de armazenamento, equipes para garantir qualidade dos dados, etc. Muitas vezes, projetos de integração de bases (como grandes coortes) começam com verbas de editais ou doações internacionais, mas enfrentam o desafio de se tornarem permanentes. Idealmente, o próprio orçamento público deve incorporar o financiamento dessas infraestruturas de dados como parte da missão do Estado em produzir estatísticas e conhecimento. A argumentação aqui é que investir em sistemas de informação robustos e na abertura segura dos dados traz retornos indiretos enormes – melhores políticas, redução de fraudes, pesquisa inovadora. Entretanto, quando o financiamento é escasso, há risco de descontinuidades: bases podem ficar desatualizadas ou sistemas saírem do ar. Um exemplo positivo foi o esforço integrado no Censo 2022 e cadastros: apesar de restrições, houve investimento para digitalização da coleta censitária e tentativa de usar

cadastros para complementar informações, o que só foi possível porque esses registros (como o cadastro de endereços) foram financiados ao longo dos anos anteriores. Assim, parte da governança é planejar a interoperabilidade e perenidade dos dados – ou seja, que diferentes sistemas conversem entre si e continuem operacionais no longo prazo. A interoperabilidade merece destaque: é a capacidade de sistemas distintos trocarem dados de forma automática, eficiente e segura. Do ponto de vista tecnológico, envolve adoção de padrões comuns (por exemplo, padronizar códigos de municípios, padronizar identificação de indivíduos via CPF/NIS, usar formatos de dados compatíveis).

Do ponto de vista institucional, requer acordos e cooperação entre órgãos. No Brasil, historicamente cada ministério ou secretaria desenvolveu seus sistemas um pouco isoladamente, mas tem havido avanços. Por exemplo, a criação do número de CPF como identificação universal do cidadão (hoje presente desde certidões de nascimento) facilita que bases diferentes utilizem essa chave. Iniciativas como o projeto de lei do Cadastro Base do Cidadão propuseram integrar dados de vários órgãos para melhorar serviços, embora gerando debates sobre privacidade. Para a pesquisa, uma maior interoperabilidade significa que, com devidos processos, é menos custoso vincular dados de saúde, educação, assistência, etc., porque eles compartilham identificadores e vocabulário comum. A interoperabilidade também envolve documentação interoperável – ter glossários e definições harmonizadas. Um exemplo concreto: o Instituto Nacional de Estudos e Pesquisas Educacionais (Inep) e o Ministério da Educação integraram sistemas de educação básica, superior e profissional em uma plataforma unificada (Sistema Educacional Brasileiro – SEB), permitindo traçar percursos educacionais completos. Do lado da saúde, o Cartão Nacional de Saúde (CNS) tenta unificar o prontuário do paciente no SUS. Esses esforços ainda enfrentam obstáculos, mas apontam a necessidade de pensar dados de forma transversal, e não "feudos" isolados.

Por fim, não se pode falar de ética em RA sem mencionar a importância de engajamento e transparência com o público em relação ao uso de seus dados. Embora a LGPD preveja a possibilidade de uso de dados pessoais pelo poder público para pesquisa, é recomendável que haja comunicação clara à população sobre como seus dados serão utilizados além da finalidade primária. Isso constrói confiança e legitimação social. Programas de pesquisa com RA geralmente publicam notas de privacidade, esclarecem que nenhum resultado individual será divulgado, e destacam os benefícios esperados (como aprimorar políticas de saúde, educação etc.). A ética também engloba a responsabilidade científica: usar RA implica lidar com grandes números e poder inferencial alto – logo, há que se evitar interpretações descuidadas que possam estigmatizar grupos ou inferir causalidades indevidas. Pesquisadores devem aplicar rigor metodológico, complementar análises quantitativas com compreensão qualitativa/contextual, e reconhecer limitações dos dados e métodos usados.

Em resumo, a governança do uso de registros administrativos deve assegurar um equilíbrio: de um lado, garantir segurança, privacidade e respeito aos direitos (via modelos como Five Safes, anonimização e stewardship); de outro, não desperdiçar o valor desses dados, permitindo que sejam utilizados para gerar bem público. Os países que mais avançaram nesse campo criaram estruturas dedicadas – no Reino Unido, por exemplo, consórcios nacionais de dados administrativos para pesquisa, com conselhos de ética; no Canadá e em outros, unidades estatísticas que integram dados sob estrito controle mas fornecem acesso a pesquisadores credenciados. O Brasil ainda está em fase de consolidar suas práticas,

mas experiências como a do CIDACS, os laboratórios de dados seguros e as diretrizes da LGPD indicam o caminho para uma cultura de uso ético e efetivo dos RA.

Considerações Finais

Um aspecto prático crucial, especialmente para estudantes e novos pesquisadores, é como acessar e trabalhar efetivamente com registros administrativos. Diferentemente de dados de surveys disponibilizados em arquivos para download, muitos RA exigem abordagens específicas de acesso – seja por meio de APIs, portais de dados abertos ou solicitações formais a órgãos públicos. Além disso, dado o volume e complexidade dos RA, é altamente recomendável adotar ferramentas computacionais (programação) para manipulação e análise, o que traz também enormes benefícios em termos de reprodutibilidade científica. Nos últimos anos, a expansão da política de dados abertos governamentais no Brasil trouxe melhorias no acesso a informações administrativas. Muitos órgãos passaram a publicar datasets em portais como o dados.gov.br ou em seções dedicadas em seus sites. Por exemplo, o IBGE disponibiliza séries históricas de estatísticas derivadas de registros (como as Estatísticas do Registro Civil, ou dados agregados do Censo Escolar) em seu portal, e mantém uma API de serviços de dados que permite que programas de computador façam consultas automáticas a tabelas de estatísticas (como via o sistema SIDRA).

Outro exemplo é o Portal da Transparência do governo federal, que fornece grande quantidade de dados administrativos sobre execução orçamentária, servidores públicos, convênios, e beneficiários de programas em formato aberto. O Portal da Transparência, em particular, implementou uma API de dados abertos que permite que usuários técnicos escrevam scripts para consultar informações diretamente, sem precisar usar a interface web manualmente. Como a própria documentação do Portal explica: "Através da API, usuários podem desenvolver programas que se conectam diretamente às máquinas do Portal da Transparência e selecionam os dados desejados"[40]. Isso significa que, por exemplo, um pesquisador pode escrever um código em Python ou R para extrair todos os gastos de um certo ministério em um dado ano, ou todas as entradas de um banco de dados de beneficiários, de forma automatizada. Para usar essas APIs, geralmente é necessário se registrar e obter uma chave de acesso, e ter algum conhecimento de chamadas web/HTTP, mas muitos tutoriais existem (a Controladoria-Geral da União, mantenedora do Portal, fornece guias, assim como a comunidade de desenvolvedores cívicos brasileiros, que compartilha dicas em blogs e fóruns).

Além do Portal da Transparência, diversos ministérios e instituições oferecem webservices ou downloads periódicos de dados administrativos. O Ministério da Saúde, através do DATASUS, permite baixar bases de dados de sistemas como SIM, SINASC e SIH em formato CSV ou DBF para vários anos (embora em muitos casos sejam dados agregados por município/mês, e não microdados individuais – os microdados completos exigem parcerias ou solicitações formais ao Datasus/Departamento de Informática do SUS). O Ministério do Trabalho, até alguns anos atrás, disponibilizava microdados da RAIS e CAGED anonimizados para pesquisa mediante solicitação. Com a substituição desses sistemas pelo eSocial recentemente, espera-se novas formas de acesso a esses dados integrados. A Receita Federal não libera microdados do Imposto de Renda, por questões legais de sigilo fiscal, mas

publica tabulações agregadas e alguns estudos (e pesquisadores às vezes conseguem acesso in loco via acordos especiais, sob confidencialidade). Já o Inep, para dados educacionais, disponibiliza microdados anonimizados anuais do Censo Escolar, Censo da Educação Superior, ENEM, ENADE, que são riquíssimos e integram o rol de RA educacionais.

Para quem está começando, uma recomendação é explorar os portais de dados abertos em busca de conjuntos de interesse e praticar o uso de APIs com exemplos simples. Por exemplo, tentar obter via API a série de população por município (dados do IBGE) ou listar os beneficiários de Bolsa Família em um determinado município (dados que já foram disponibilizados em transparência). Essa prática ajuda a desenvolver habilidades de coleta automatizada de dados – fundamental quando se lida com big data administrativo. Ferramentas como Python (com bibliotecas requests, pandas) ou R (com pacotes como httr, jsonlite) são comumente usadas para consumir essas APIs e montar bancos de dados atualizados localmente para análise.

No tocante à reprodutibilidade, o uso de programação também se alia ao rigor científico. Reprodutibilidade significa que outra pessoa poderia seguir os mesmos passos que você e chegar aos mesmos resultados, verificando sua validade. Quando lidamos com RA, que podem ser constantemente atualizados e muito volumosos, documentar e automatizar os passos analíticos é essencial. Por exemplo, em vez de baixar manualmente centenas de arquivos e ficar copiando e colando dados em planilhas (o que é propenso a erros e difícil de reproduzir), o pesquisador deve investir em escrever scripts que realizem todas as etapas: desde o download dos dados (ou consulta via API), passando pela limpeza e fusão das diferentes fontes, até as análises estatísticas e geração de gráficos/tabelas finais. Ferramentas como notebooks Jupyter ou scripts R Markdown permitem combinar código, descrição em texto e visualizações, facilitando tanto o processo de pesquisa quanto a posterior comunicação e replicação. Mesmo que um iniciante ainda não domine essas linguagens, vale a pena começar com pequenos passos – por exemplo, aprender a usar um software estatístico (R, Python, Stata) para ler um arquivo grande de dados e fazer um resumo simples. Aos poucos, evoluir para integrações mais complexas.

Um ponto a mencionar é que a reprodutibilidade em contexto de dados administrativos às vezes esbarra na confidencialidade: se os dados não podem ser totalmente compartilhados publicamente, como replicar? Uma saída é compartilhar ao menos o código e instruções detalhadas, de modo que outros pesquisadores qualificados (que obtenham permissão aos mesmos dados) consigam reproduzir. Outra possibilidade é publicar resultados para múltiplas amostras ou subpopulações, conferindo robustez. Em projetos grandes, às vezes se criam dados sintéticos públicos como mencionado, para que analistas possam testar seus códigos e métodos, que depois são aplicados aos dados reais dentro de um ambiente seguro. De todo modo, a transparência do método – deixar claro o processo de tratamento do RA – é parte essencial da ciência aberta.

Também é importante notar que alguns órgãos começam a oferecer plataformas de análise remota: ao invés de enviar dados para o pesquisador, o pesquisador envia seu código para ser executado no servidor do órgão que hospeda os microdados confidenciais, e recebe de volta apenas os resultados agregados. Isso já ocorre, por exemplo, no acesso a dados sigilosos do IBGE via o DataLab (pesquisador submete scripts que rodam nos microdados do censo completo, retornando apenas as tabelas aprovadas). Essa modalidade deve crescer

para RA de outras naturezas, requerendo do pesquisador a capacidade de preparar códigos bem escritos e testados previamente em dados amostra ou sintéticos.

Resumindo, para o iniciante as dicas são: aproveite as fontes abertas existentes (portais de transparência, dados abertos, microdados públicos de censos escolares, etc.) para treinar; aprenda fundamentos de programação de dados (não precisa ser um expert imediatamente, mas conhecimentos básicos de manipulação de data frames, junção de tabelas, etc., lhe pouparão imenso tempo e permitirão lidar com RA); e documente seu trabalho (anote a versão da base usada, data de extração, scripts em anexo, para que você mesmo ou outros consigam futuramente entender de onde vieram os números). Isso tudo tornará sua pesquisa mais robusta e também conferirá credibilidade na hora de divulgar resultados – você poderá mostrar com confiança como chegou a determinada conclusão.

Por fim, vale ressaltar que comunidades online e recursos educacionais podem apoiar quem ingressa nesse campo. Há fóruns dedicados a dados governamentais (com desenvolvedores cívicos e cientistas de dados dispostos a ajudar), repositórios no GitHub com exemplos de análises usando RA brasileiros, cursos e workshops promovidos por universidades e órgãos públicos sobre uso de microdados. Manter-se engajado nessa comunidade ajuda a aprender mais rápido e a se atualizar sobre novas fontes de dados e ferramentas que estão surgindo.

Referências

BABBIE, Earl. *Métodos de Pesquisas de Survey.* Belo Horizonte: Editora UFMG, 1999.

BARRETO, M. L. *et al.* Cohort Profile: The 100 Million Brazilian Cohort. *International Journal of Epidemiology*, v. 51, n. 2, p. e27–e38, 2022.

BILOTTO, Antonella; GUERCIO, Maria. The management of corporate records in Italy: traditional practice and methods and digital environment. *Records Management Journal*, v. 13, n. 3, p. 136–146, 2003.

BUTZ, William P. The future of Administrative Records in the Census Bureau's demographic activities. Washington, DC: U.S. Bureau of the Census, 1982. mimeo.

CARDOSO, Onésimo de O. Comunicação Empresarial versus Comunicação Organizacional: Novos Desafios Teóricos. *Revista de Administração Pública*, v. 40, n. 6, Rio de Janeiro, nov./dez. 2006.

CARTWRIGHT, D. W.; ARMKNECHT, Paul A. Statistical uses of administrative records. *Journal of the American Statistical Association*, v. 78, n. 382, jun. 1983.

CHETTY, Raj; HENDREN, Nathaniel; KLINE, Patrick; SAEZ, Emmanuel. Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States. *Quarterly Journal of Economics*, v. 129, n. 4, p. 1553–1623, 2014.

CONTROLADORIA-GERAL DA UNIÃO (CGU). Portal da Transparência – API de Dados Abertos.

DESAI, T.; RITCHIE, F.; WELPTON, R. *Five Safes: designing data access for research.* Bristol: University of the West of England, 2016. (Economics Working Paper Series,

n. 1601).

FERREIRA, Frederico Poley M. Registros administrativos como fonte de dados. In: *Anais do IV SEGeT – Simpósio de Excelência em Gestão e Tecnologia*. Resende: AEDB, 2007.

HOPPEN, N.; MEIRELLES, F. S. Sistemas de informação: um panorama da pesquisa científica entre 1990 e 2003. *Revista de Administração de Empresas*, v. 45, n. 1, p. 24–35, jan./mar. 2005.

INSTITUTO MOBILIDADE E DESENVOLVIMENTO SOCIAL (IMDS). *Atlas da Mobilidade Social – Resultados e metodologia*. Rio de Janeiro: IMDS, 2025.

IWHIWHU, Enemute Basil. Management of records in Nigerian Universities. *Records Management Journal*, v. 23, n. 3, 2005.

LAI, V. S.; MAHAPATRA, R. K. Exploring the research in information technology implementation. *Information & Management*, v. 32, n. 4, p. 187–201, 1997.

LONG, John F. Demographic Applications of Administrative Records. Washington, DC: U.S. Bureau of the Census, 2001. mimeo.

MALIN, Ana Maria Barcellos. Gestão da Informação Governamental: em direção a uma metodologia de avaliação. *DataGramaZero: Revista de Ciência da Informação*, v. 7, n. 5, out. 2006.

MELLO, Janine. Produção estatal de evidências e uso de registros administrativos em políticas públicas. In: KOGA, N. M.; PALOTTI, P. L. M.; MELLO, J. (Org.). *Políticas públicas e usos de evidências no Brasil: conceitos, métodos, contextos e práticas*. Brasília: Ipea, 2022. p. 457–478.

NEWTON, C. Information and malformation. Records management in information systems. In: *Information '85: Using Knowledge to Shape the Future*. Bournemouth, 16–19 Sept. London: Library Association Publishing, 1986. p. 75–86.

OFFICE OF MANAGEMENT AND BUDGET (OMB). Management of Federal Information Resources. *Federal Register*, v. 5, n. 247, p. 252730–252751, 24 dez. 1985.

OLIVEIRA, Miriam; MAÇADA, A. Carlos Gastaud; GOLDONI, V. Análise da Aplicação do Método: Estudo de Caso na Área de Sistemas de Informação. In: *Anais do 30º Encontro da ANPAD*. Salvador, BA, 2006.

SENHRA, Nelson de Castro. A questão dos registros administrativos vis-a-vís a geração de estatísticas. *Revista Brasileira de Estudos Populacionais*, v. 13, n. 2, Campinas, SP, 1996.

SILVEIRA, Fernando Gaiger. Evidências e dados nas políticas públicas: o falso embate entre quantitativo e qualitativo. *Texto para Discussão*, n. 2324, Ipea, 2017.

SOUZA, Pedro Herculano Guimarães Ferreira de. Uma metodologia para explicar diferenças entre dados administrativos e pesquisas amostrais, com aplicação para o Bolsa Família e o Benefício de Prestação Continuada na PNAD. *Revista Brasileira de Estudos de População*, v. 30, n. 1, p. 299–315, 2013.

WILLEMIN, Georges. The International Committee of the Red Cross (ICRC) official

e-mail system: An example of records management. *Records Management Journal*, v. 16, n. 2, p. 82–90, 2006.

ZACHARIAS, Maria Luiza Barcellos. Cadastros Estatísticos de Empresas Construídos a partir de Registros Administrativos. ONU/CEPAL, mimeo. Segunda reunião da Conferência de Estatística das Américas, Santiago do Chile, 18–20 jun. 2003.