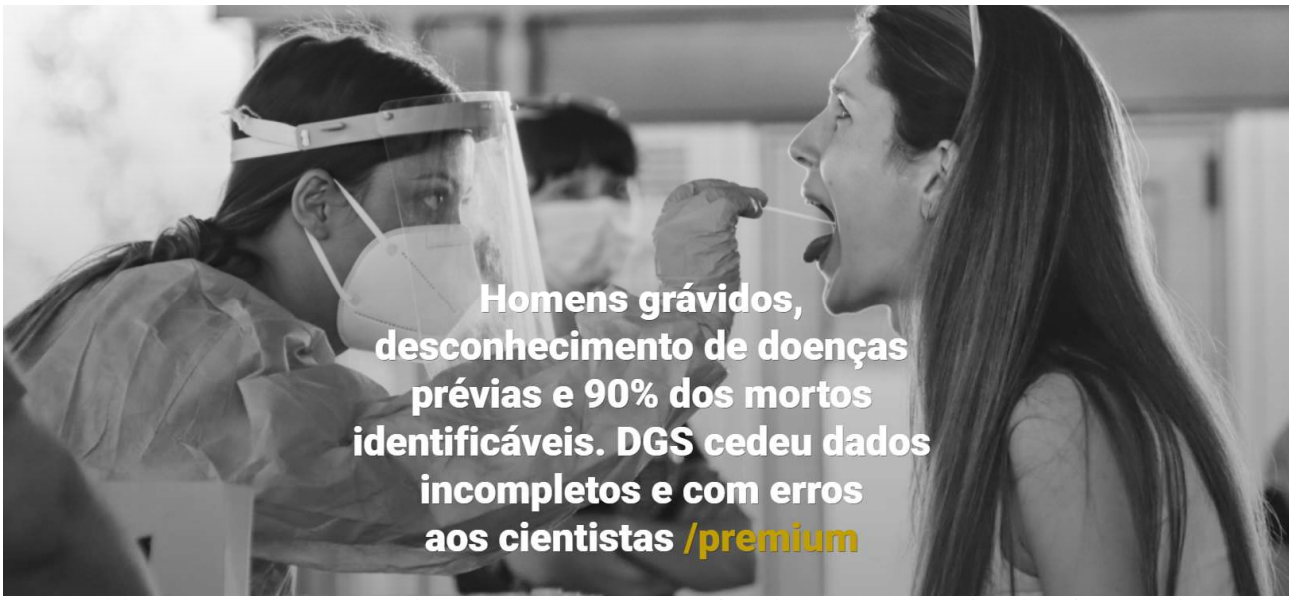
Data: 07-08-2020

Título: Homens grávidos, desconhecimento de doenças prévias e 90% dos mortos identificáveis. DGS cedeu dados incompletos e com erros ac

Pub: OBSERVADOR

Tipo: Internet Secção: Nacional



▲Os médicos que assistem os doentes com Covid-19 têm de preencher três plataformas de dados diferentes NICOLAS ECONOMOU/NURPHOTO VIA GETTY IMAGES

Vera Novais Texto

Os investigadores insistiram e tiveram acesso aos dados dos doentes com Covid-19. Mas há homens grávidos, falhas graves na proteção da identidade e artigos científicos a sair com informação errada.

07 ago 2020, 20:5973

Dados "com caráter provisório", que "poderão ser ainda alvo de validação" e que podem "não coincidir com aqueles reportados pelo boletim diário da DGS". São alguns dos pontos focados na descrição que acompanha os dados de vigilância epidemiológica da Covid-19 cedidos pela Direção-Geral da Saúde aos cientistas. Uma base de dados incompleta e com erros graves, desde homens grávidos (até uma criança de cinco anos) e desconhecimento de doenças prévias. **Mas também uma base de dados que não conseguiu proteger a identidade dos doentes — mais de 90% dos mortos são potencialmente identificáveis**, acusam alguns dos investigadores. A DGS garante que não.

O ficheiro recebido pelos investigadores a 4 de agosto tem os dados dos doentes com Covid-19, como idade, sexo, se teve de ser hospitalizado, que cuidados recebeu e se recuperou ou não da doença, até ao dia 30 de junho. É o segundo ficheiro que os investigadores recebem — o primeiro chegou em abril —, apesar de todos esperarem atualizações mais frequentes. "Tínhamos informação que os dados iriam ser atualizados semanalmente", diz ao Observador Cristina Santos, investigadora no Centro de

Investigação em Tecnologias e Serviços de Saúde (Cintesis). "Só que as atualizações semanais nunca chegaram."

O volume de trabalho na DGS e a exigência na preparação das bases obrigou esta autoridade de saúde a optar por atualizações mensais, como se lê no site onde podem ser requisitados os dados. Mas até ao momento também não conseguiram cumprir com esta periodicidade.

""Uma das maiores preocupações que tenho em relação a isto é que sejam publicados artigos científicos com base nestes dados, por investigadores que vão tirar conclusões com base em dados pouco credíveis, mas que aparentemente têm muita credibilidade, pois são fornecidos pela DGS."

Cristina Santos, Cintesis / Universidade do Porto

A Luís Antunes, no entanto, bastou-lhe a primeira versão, ainda que incompleta, para perceber que **a Direção-Geral da Saúde tinha cometido erros básicos na anonimização dos dados** — ou seja, no processo que permite manter a integridade dos dados científicos, mas impedindo a identificação dos doentes. "Conseguimos ver que 38% das pessoas que morreram podiam ser identificadas", diz ao Observador o investigador da Universidade do Porto. A nova base de dados, em vez de corrigir o problema, só o agravou: potencialmente mais de 90% dos mortos podem ser identificados cruzando a base de dados com elementos externos, conta o investigador.

Como se não bastasse os problemas com a anonimização dos dados, Luís Antunes acrescenta que **"a qualidade dos dados é miserável" e põe em causa a credibilidade dos mesmos**. Cristina Santos optou por não avançar com o projeto que tinha. Além dos erros, entende que os "dados estão extremamente incompletos".

Apesar de tudo, nem todos os investigadores encaram os dados com uma perspetiva tão crítica. A Faculdade de Medicina da Universidade de Lisboa, por exemplo, já tem um artigo científico publicado com estes dados. "Vamos fazendo o melhor que se consegue", diz ao Observador Paulo Nogueira, investigador no Instituto de Saúde Baseada na Evidência.

Erros na base de dados pode invalidar as conclusões dos cientistas

Cristina Santos queria perceber como é que as doenças pré-existentes podiam agravar a situação dos doentes com Covid-19, mas considerou que os dados eram tão fracos que desistiu. Ao economista Pedro Pita Barros interessava-lhe saber como é que as organizações se ajustaram à situação. Obteve alguns resultados, mas precisa de uma amostra maior (que chegou na semana passada) para os confirmar. Paulo Nogueira diz que os "dados lhe pareceram bastante coerentes" e acredita que "as pistas encontradas são válidas", por isso as publicou num artigo científico, mas não descarta que haja limitações.

Estas são as histórias de três investigadores que queriam dedicar parte do seu tempo a ajudar o SNS a gerir melhor os recursos em tempos de pandemia e que se viram obrigados a fazer estudos com os dados insuficientes ou errados.

Hospitais com todos os doentes ou doentes em vários hospitais?

"É regra do investigador que os dados nunca são suficientes, queremos sempre mais", diz ao Observador Pedro Pita Barros, professor na Faculdade de Economia da Universidade Nova de Lisboa. O tom era de brincadeira, mas assume com seriedade que "face ao que era possível ter naquela altura [abril], mesmo não sendo perfeito, os dados eram suficientes".

"Nos dados iniciais, até ao final de abril, encontrei que o impacto da idade interagia de uma forma não óbvia com a mortalidade — e que nem todas as condições crónicas (as que vinham identificadas na base de dados) tinham o mesmo efeito."

Pedro Pita Barros, professor na Faculdade de Economia da Universidade Nova de Lisboa

Um dos objetivos principais do economista era perceber se ao longo do tempo as unidades de saúde tinham tido a capacidade de reduzir o número de mortos, mesmo quando as condições da unidade e o estado de saúde dos doentes era equivalente. Os dados iniciais não permitiram ter essas respostas. Mas permitiram outras: como o impacto da mortalidade e das doenças crónicas na mortalidade. **Com este tipo de resultados, o investigador espera ser possível sinalizar os casos potencialmente mais complicados.**

"Nos dados iniciais, até ao final de abril, encontrei que o impacto da idade interagia de uma forma não óbvia com a mortalidade — e que nem todas as condições crónicas (as que vinham identificadas na base de dados) tinham o mesmo efeito." Pedro Pita Barros **admite o grau de incerteza e assume que o número de casos era pequeno (1.746)**, por isso comunicou as conclusões preliminares à DGS, mas não as tornou públicas.

Mas o que o especialistas em políticas sociais quer mesmo perceber é se existe um efeito de escala (que destaca a importância da massa crítica de profissionais de saúde e equipamentos) e/ou de aprendizagem (com partilha de conhecimento) e o que os origina. Estes são dados importantes para a organização do SNS, por exemplo. Porque "se houver fortes efeitos de escala e/ou aprendizagem, então, tratar os doentes Covid-19 no mesmo local [um único hospital para uma determinada região] é mais vantajoso para os doentes — têm menor risco de morte —, podendo a proximidade ser sacrificada a favor desse menor risco", explica. "Se estes efeitos não forem grandes, então ter os doentes em vários hospitais, com as devidas zonas dedicadas à Covid-19, não tem riscos acrescidos para a mortalidade."

Fazer o que se pode com aquilo que se tem

Pedro Pita Barros não foi o único a analisar o impacto da idade e das doenças pré-existentes no aumento do risco de morte. A equipa da Faculdade de Medicina da Universidade de Lisboa, por exemplo, concluiu que a idade era o fator que tinha mais peso no risco de morte por Covid-19 e que, entre as doenças crónicas, as doenças cardíacas e renais eram aquelas em que se verificava o maior aumento desse risco. Os resultados foram publicados na revista científica Journal of Clinical Medicine.

Paulo Nogueira, responsável pelo projeto, admite que os primeiros dados "são de uma fase muito precoce" e que **se a base de dados tivesse sido atualizada regularmente como estava previsto, teriam "mais certeza sobre os dados"**. "Se calhar tínhamos produzir evidência diferente. Com mais dados e melhores, podíamos ter feitos trabalhos diferentes", diz o investigador. Ainda assim, prefere olhar para o copo meio cheio: "Era importante que os dados nos fossem dados e foram."

Cristina Santos, da Universidade do Porto, não está totalmente de acordo. Ficou satisfeita quando os dados lhe foram enviados em abril, mas **considerou, desde o primeiro momento, que aquela base de dados não tinha qualidade para ser publicada**. Avançou com os dados que tinha para poder testar os modelos estatísticos que desenvolveu e quando recebesse a atualização dos dados, podia publicar os resultados com base em dados corretos e fiáveis. Mas depois de receber o segundo ficheiro percebeu que isso não ia acontecer. "Perdi qualquer esperança de fazer alguma coisa de útil com dados tão maus."

Na primeira conversa com o Observador, o investigador do Instituto de Saúde Baseada na Evidência mostrou-se menos crítico dos dados. "Sabemos que há limitações e que temos de lidar com essas limitações", diz. **"Mas assumimos que o que estava feito, estava bem feito."** Agora, depois de olhar para a segunda base de dados diz que pode não ser possível repetir a análise que tinha feito antes — ou seja, não é replicável. "Há efetivamente diferenças na forma como os dados foram disponibilizados nos dois momentos."

"Nenhum" é nenhum, não devia ser "não sei"

A DGS disse ao Observador que se tem esforçado por enviar a melhor informação aos investigadores. "Todas as bases de dados necessitam de ajustes e foi o que fizemos e faremos sempre que for necessário." E é isso que, à partida, terão feito com a nova base de dados enviada.

"A DGS não está a tentar ludibriar os investigadores."

Direção-Geral da Saúde

Um erro básico identificado facilmente são os homens ou idosas grávidos. É certo que a responsabilidade inicial não é da DGS. Mas com os médicos assoberbados com

a quantidade de doentes que tinham para ver e a ter de preencher três plataformas distintas, com os mesmos dados, **as falhas humanas são praticamente inevitáveis**. Exceto se, como sugeriu Luís Antunes, as próprias plataformas tivessem sinais de alerta: como um homem não poder estar grávido ou existirem dados fundamentais em falta.

Mas a base de dados inicial tinha outro problema que Cristina Santos não conseguiu explicar à partida, mas que a deixou intrigada. Nos mais de 20 mil doentes, só 16,6% apresentava doenças pré-existentes (ou seja, 83,4% eram saudáveis). E isto apesar de cerca de 25% destas 20 mil pessoas ter mais de 66 anos — idade suficiente para apresentar já uma ou outra doença crónica.

A resposta às suas dúvidas chegou com a segunda base de dados: **o campo preenchido com "none" (nenhuma comorbilidade) no primeiro ficheiro juntava, na verdade, os que não têm realmente outras doenças, os que não se sabe se tem ou não e aqueles cujo campo está simplesmente em branco**.

Table 1. Characteristics of the SARS-CoV-2 infected individuals and corresponding lethality in Portugal on 21 April 2020.

Variable		*n* (%)	Deaths (%)	*p*-Value *
Outcome	Recovered	1244 (6.1%)		
	Died COVID-19	502 (2.5%)		
	Ongoing Treatment	18,524 (91.3%)		
	Unknown	23 (0.1%)		
Sex	Female	11,903 (58.7%)	253 (2.13%)	<0.001
	Male	8390 (41.3%)	249 (2.97%)	
Age	(0,18)	711 (2.5%)	0 (0%)	<0.001
	(19,35)	4153 (20.5%)	0 (0%)	
	(36,45)	3259 (16.1%)	3 (0.09%)	
	(46,55)	3653 (18.1%)	10 (0.27%)	
	(56,65)	3046 (15.1%)	30 (0.98%)	
	(66,75)	1926 (9.5%)	78 (4.05%)	
	(76,85)	1864 (9.2%)	177 (9.50%)	
	86+	1618 (8.0%)	204 (12.61%)	
Region	North	12,211 (60.2%)	319 (2.61%)	<0.001
	Algarve	472 (2.3%)	6 (1.27%)	
	Center	2817 (13.9%)	97 (3.44%)	
	Lisbon Metropolitan Area	4264 (21.0%)	74 (1.74%)	
	Alentejo	391 (1.9%)	6 (1.53%)	
	Madeira	48 (0.2%)	0 (0%)	
	Azores	90 (0.4%)	0 (0%)	
Hospitalization	No	15,697 (77.4%)	126 (0.80%)	<0.001
	Unknown	1623 (8.0%)	46 (2.83%)	
	Yes	2973 (14.7%)	330 (11.10%)	
Intensive Care	No	15,697 (77.4%)	126 (0.80%)	<0.001
	Unknown	4335 (21.4%)	475 (10.96%)	
	Yes	261 (1.3%)	27 (10.34%)	
Respiratory Support	No	1315 (6.5%)	156 (11.86%)	<0.001
	Oxygen	59 (0.3%)	0 (0%)	
	Ventilator	26 (0.1%)	0 (0%)	
	Unknown	18,893 (93.1%)	346 (1.83%)	

O número de doentes com mais de 66 anos, na base de dados inicial, segundo publicado no artigo da Journal of Clinical Medicine

Com o novo ficheiro, Cristina Santos verificou que **em 46% dos casos não é possível saber se o doente tinha ou não uma doença prévia** (40% dos quais porque a informação ficou por preencher) e que em 32% estava confirmado que não tinham. Os dados usados inicialmente pelos investigadores, que consideraram "none" como sem doenças prévias estão, assim, errados. A DGS garantiu que procede à retificação dos dados sempre que os problemas são detetados e que "não está a tentar ludibriar os investigadores".

"Uma das maiores preocupações que tenho em relação a isto é que sejam publicados artigos científicos com base nestes dados, por investigadores que vão tirar conclusões com base em dados pouco credíveis, mas que aparentemente têm muita credibilidade, pois são fornecidos pela DGS", lamenta Cristina Santos.

Anonimizar os dados não é só tirar o nome

"Os dados, destinados a investigação científica serão devidamente anonimizados, impossibilitando a identificação do respetivo titular, encontrando-se sob a responsabilidade da Direção-Geral da Saúde", lê-se numa das comunicações que a DGS estabeleceu com os investigadores antes do envio do primeiro ficheiro. A DGS nem o imaginava, mas a **promessa de dados anonimizados serviu de gatilho** e desafiou Luís Antunes, professor do departamento de Ciência de Computadores, a ver se esse processo tinha sido realmente bem sucedido. Mas, segundo ele, não foi.

"Não por incapacidade técnica, mas por incúria", diz o informático. Que exemplifica: se **em vez de se usarem as idades exatas dos doentes se usasse um intervalo de idades**, o potencial de reindentificação diminuía imenso. "E do ponto de visto estatístico seria mesma coisa", reforça. Numa situação hipotética, é a diferença entre ter um único doente com 77 anos ou ter um conjunto deles entre os 75 e os 79 anos — já não é tão fácil saber qual era o de 77.

Anonimizar os dados é mais do que retirar o nome e os números de identificação (cartão de cidadão, NIF ou segurança social), porque muitas vezes a data de nascimento, localização geográfica, profissão ou grupo étnico podem ser o suficiente para chegar ao indivíduo.

"Na primeira versão, ainda que incompleta, conseguimos ver que 38% das pessoas que morreram podiam ser identificadas. A nova base de dados, em vez de corrigir o problema, só o agravou: potencialmente mais de 90% dos mortos podem ser identificados cruzando a base de dados com elementos externos."

Luís Antunes, Faculdade de Ciências da Universidade do Porto

Neste caso, os investigadores começaram por escolher um conjunto de atributos — idade, sexo, local de infeção (região) e resultado (desenlace do problema) — e encontraram mais de **1.400 pessoas (entre as 20.293 da base de dados) que tinham combinações**

únicas destes quatro atributos — logo, potencialmente identificáveis. É como ser o único a ter a chave que dá acesso ao primeiro prémio do Euromilhões, mas aqui com resultados menos positivos.

Quem detém uma base de dados que deve ser anónima, neste caso a DGS, deve **verificar se existe o risco de se voltarem a identificar os indivíduos** com base na informação disponibilizada por comparação com fontes externas. Se a anonimização estiver bem feita e a avaliação do risco for realizada, a probabilidade de reidentificação dos doentes é mínima. A DGS garante ao Observador que "fez a anonimização dos dados e uma avaliação do risco de reidentificação, tendo solicitado apoio de investigadores".

A 24 de março, durante um debate quinzenal, o próprio primeiro-ministro, António Costa, garantiu que o Governo já tinha "condições para anonimizar todos os dados que serão disponibilizados".

A equipa de Luís Antunes foi testar isso mesmo em relação ao primeiro ficheiro, o de abril. Os investigadores podem juntar diferentes atributos de forma a isolar os doentes com combinações únicas, mas no caso da equipa da Universidade do Porto o interesse estava nas 502 pessoas que tinham morrido. À referência de óbito, juntaram idade, sexo e região, e concluiram que **192 pessoas (cerca de 38%) podiam ser identificadas, porque tinham combinações únicas dos fatores**. Depois, bastou-lhes olhar para os jornais — para dois em concreto que não foram nomeados — para identificar rapidamente 12 pessoas. E nem sequer tentaram usar as redes sociais, onde certamente encontrariam mais informação.

O que o investigador não estava à espera é que a segunda base de dados, a de 4 de agosto, estivesse ainda pior, conforme disse ao Observador. "As alterações introduzidas só pioraram a situação. A percentagem de 'single-out' [pessoas que podem ser distinguidas das restantes] é bem maior e se considerarmos o subconjunto das pessoas que faleceram então é brutal." **A alteração? A base de dados agora tem a data do óbito, em vez de ter só se o doente tinha morrido ou não.** A consequência? Mais de 90% dos mortos são potencialmente identificáveis.

A Direção-Geral da Saúde, por sua vez, diz que os dados foram anonimizados e que agora foram fornecidos aos investigadores "não são passíveis de reidentificação".

A proteção da identificação dos doentes sempre foi uma das questões fundamentais para o Ministério da Saúde e para a DGS e um dos fatores que pode ter potenciado o atraso da sua entrega aos investigadores. Luís Antunes é claro ao dizer que não quer que estes resultados — que vai comunicar à DGS e publicar num artigo científico — sirvam de desculpa para deixar de fornecer dados para investigação científica. Muito pelo contrário. **A investigação é fundamental e a sua equipa propõe soluções para resolver estes problemas.**

O investigador, que trabalha na área da proteção de dados há muito tempo, começa logo com uma sugestão simples: primeiro encontrar os casos únicos potencialmente identificáveis, depois retirá-los da base de dados. A possibilidade de identificação fácil da pessoa são motivo suficiente para invocar a proteção destes dados. **Mas há outras soluções para estes casos, como intervalos de idades ou regiões geográficas mais abrangentes**, por exemplo. Os doentes ficam mais protegidos e a informação continua a ser válida em termos científicos.

Outra situação que deve ser evitada são datas corretas de acontecimentos, como a de hospitalização ou do óbito. Mas como as referências temporais para um indivíduo e mesmo para a comparação entre indivíduos podem ser importantes, os investigadores sugerem que se crie um novo conjunto de datas, que mantenham a relação entre si (e a mesma distância relativa entre elas).

Das 72 horas aos meses de espera por uma base de dados

Pouco tempo tinha passado da chegada da pandemia a Portugal e os investigadores já sabiam o que poderiam fazer com os dados que viessem a ser recolhidos pelos médicos em relação aos doentes infetados com o novo coronavírus. O objetivo era estudar a evolução do surto e encontrar formas de ajudar, por exemplo, com soluções de triagem online, para evitar a sobrecarga da linha SNS24 e das urgências — ideias de uma fase precoce da pandemia, quando até os sintomas eram mal conhecidos.

"Nenhum de nós está a fazer isto por uma razão de protagonismo individual ou institucional. Estamos aqui para ajudar."

Nuno Sousa, presidente da Escola de Medicina da Universidade do Minho

Uma carta aberta e uma petição com mais de cinco mil assinaturas mostrava à tutela a importância de ter acesso a dados mais específicos, para se poderem criar modelos mais robustos. "Nenhum de nós está a fazer isto por uma razão de protagonismo individual ou institucional. Estamos aqui para ajudar", disse, na altura, Nuno Sousa, presidente da Escola de Medicina da Universidade do Minho.

Na sequência dos apelos, **o primeiro-ministro garantiu num debate quinzenal do final de março que os dados seriam disponibilizados**. E a prorrogação do estado de emergência, no início de abril, incluía essa possibilidade. O artigo 39.º autorizava, assim, o "acesso a dados anonimizados do Sistema Nacional de Vigilância Epidemiológica para investigação científica". E a 10 de abril foi lançado o formulário que permitia aos investigadores pedir acesso aos dados. Idade, sexo, data do teste positivo, data da hospitalização, recurso a ventilação, doenças crónicas, estão entre os dados a que os investigadores poderiam ter acesso.

Na altura era dito que os investigadores teriam acesso aos dados até 72 horas depois do pedido. Mas nesse período o que receberam foi um pedido para preencherem um novo formulário, agora com a descrição do projeto e com nota da aprovação do diretor da instituição e de uma comissão de ética — um processo que normalmente atrasa os projetos, mas que os investigadores conseguiram em tempos recorde. O antigo subdiretor-geral da Saúde, Diogo Cruz, disse em conferência de imprensa que **até 20 de julho tinham recebido 400 pedidos de acesso aos dados, mas que só 50 tinham completado o processo até ao momento**.

Ainda durante o mês de abril, os investigadores receberam a primeira base de dados, com 20.293 doentes, mas **as prometidas atualizações semanais nunca chegaram**. "Verificámos, logo em abril, que os dados estavam muito incompletos", diz Cristina Santos. "Nessa altura, tínhamos informação que os dados iriam ser atualizados semanalmente. Sempre pensámos que esta seria apenas a primeira base de dados, mas que, com as atualizações, os dados ficariam mais completos e poderíamos começar a trabalhá-los melhor."

A gestão das bases de dados deve ter-se mostrado um trabalho maior do que a DGS podia suportar e no site alterou a periodicidade das atualizações de semanais para mensais, mas nem isso conseguiu cumprir. Os cientistas só voltaram a receber uma atualização no dia 4 de agosto. Ainda é cedo para dizer o que vão poder tirar destes dados, mas à primeira vista já deu para perceber que há erros que se mantém, como **um dos doentes grávidos que é um rapaz de cinco anos**.