Recomendaciones respecto al nuevo portal de datos de COVID-19 del Departamento de Salud de Puerto Rico

Por la Berenjena Anónima
27 de julio del 2021

Primero quiero resaltar que el nuevo portal es una enorme mejora respecto al anterior. Y entiendo que hay algunos logros en este que merecen resaltarse más; un ejemplo que me fijé por mi cuenta es que el nuevo portal recorta por un día el tiempo entre recepción de datos en Bioportal y la publicación de cifras correspondientes.

Diseño de las medidas y visualizaciones

Enfatizar PCR + antígeno

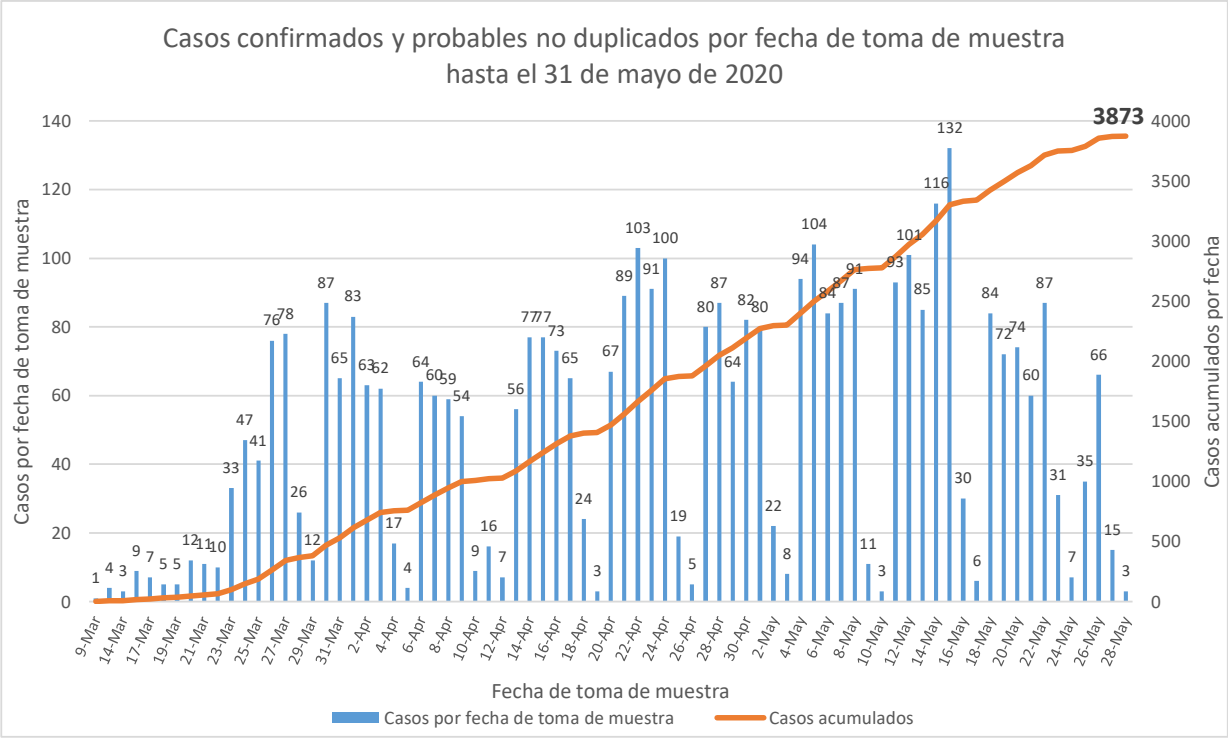
El portal separa sistemáticamente los denominados casos “confirmados” (= que cuentan con una prueba molecular positiva) de los “probables” (= que cuentan con una prueba de antígeno positiva pero no una molecular). Pero se me hace difícil creer que a la gran mayoría de la audiencia de este portal le interese o sirva esta distinción, y que les ayudaría mejor presentar como primera opción cifras y curvas combinadas de casos por ambos tipos de prueba, ya que:

1. Ambas detectan la presencia corriente del virus;
2. Ambos tipos de prueba gozan de alta especificidad y valor predictivo positivo.

El Departamento de Salud de hecho solía publicar una tal gráfica combinada en las ediciones tempranas del informe diario de casos (ver próxima página). Entiendo que en aquel entonces esto era desacertado porque los casos que se denominaban “probables” eran por prueba serológica que no se deben mezclar con prueba diagnóstica, pero ahora con prueba de antígeno se torna acertado retomar esto.

No dudo que habrá personas que quieran adentrar en cuántos casos cuentan con prueba molecular positiva y cuántos no, pero este tipo de desglose me parece no debe ser a costa de un conteo de casos que integren ambos tipos de prueba.

Una forma de enfatizar esto es notar que el nuevo portal, en efecto, no dice en ningún lugar cuántos casos diarios se están detectando. Quien le interese contestar esa sencilla pregunta tiene que ponerse a sumar las cifras desglosadas, que probablemente no le interesan.



Terminología: “confirmados” vs. “probables”

La terminología de casos“probables,” aunque bien entiendo la establece el Consejo de Epidemiólogos Estatales y Territoriales de los EEUU, a mi entender es desafortunado aplicarla a casos positivos a prueba de antígeno porque da a entender al público que estos como “probables” son menos certeros que los “confirmados” cuando en realidad los anteriores son por pruebas de antígeno que gozan de altas tasas de especificidad y altos valores predictivos positivos.

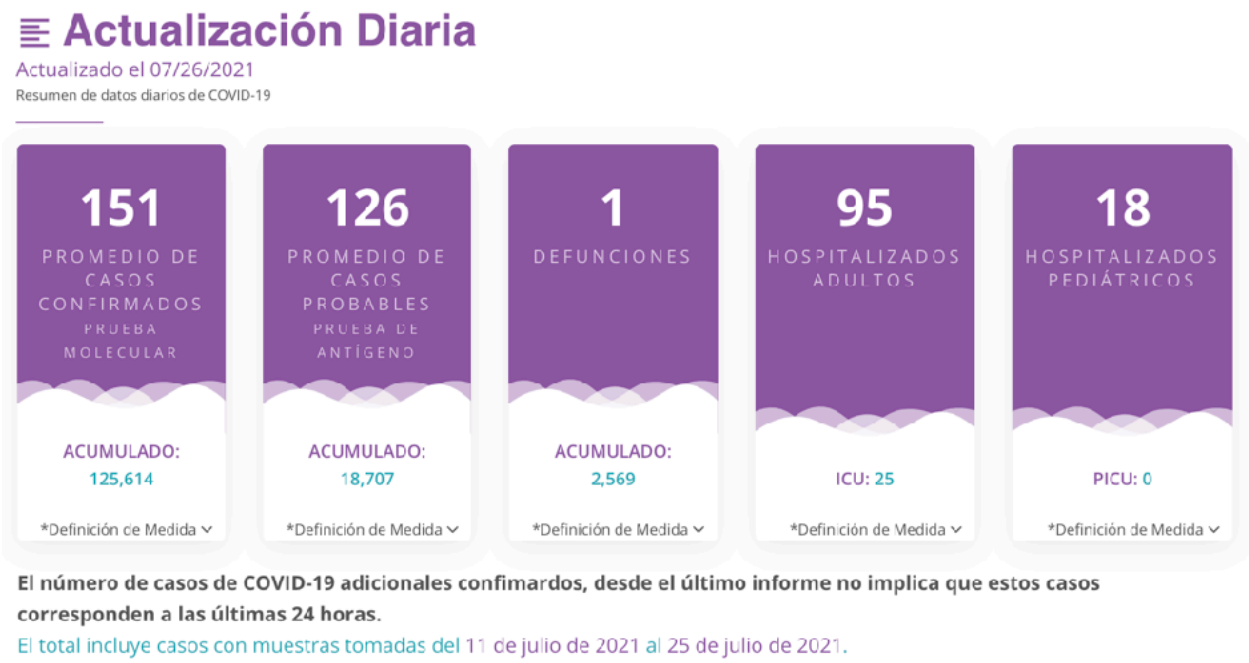
La consecuencia más negativa de esto es que muchas personas han adoptado la práctica de solo citar las cifras de “confirmados” que, si entendieran bien la ejecutoria de las pruebas de antígeno, preferirían usar una suma de ambas (como recomiendo arriba que debe enfatizarse).

Se puede entender el deseo de querer cumplir con esta terminología que viene de una autoridad externa, pero vale pensar también si se puede reducir la exposición del público a esta, que a mi parece tiende a confundir. Mi recomendación de priorizar cifras combinadas me parece que ayuda a reducir el énfasis en esta terminología.

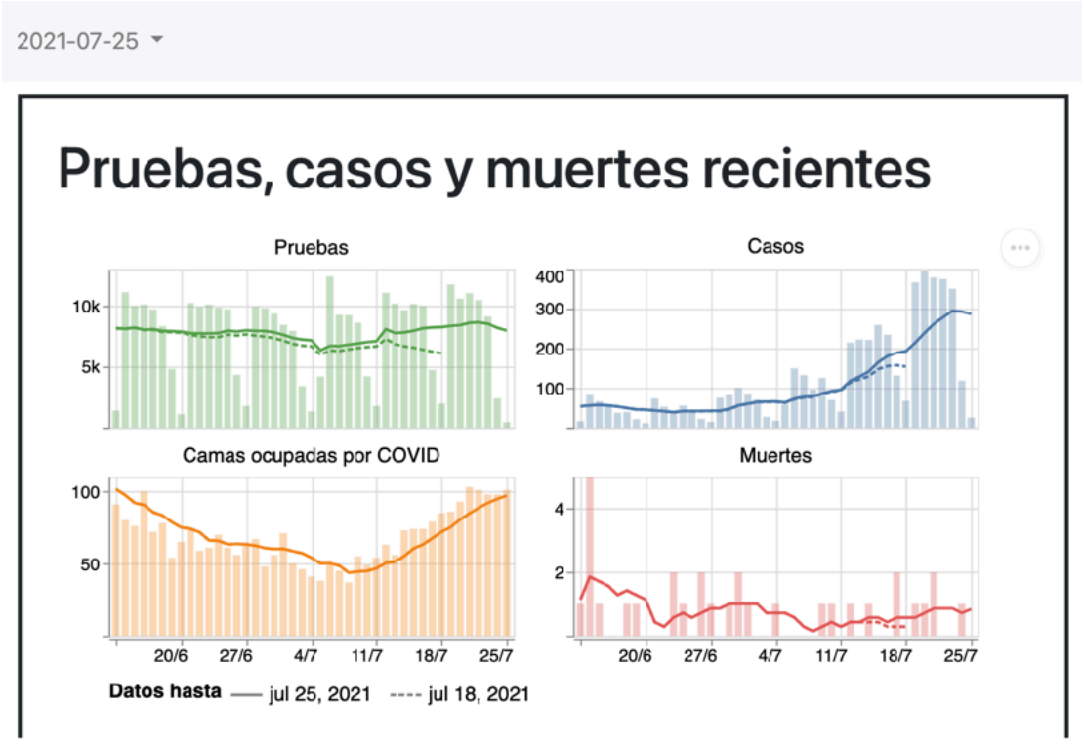
Una consideración secundaria pero que añado de todos modos es que esta terminología también prioriza la pregunta de cuáles casos se hizo o no prueba molecular, por encima de cuál prueba fue la que inicialmente detectó el caso. Cuando un caso se detecta inicialmente por antígeno, y días más tarde el paciente da positivo a molecular, estas definiciones exigen reclasificar el caso de “probable” a “confirmado.” Pero muchos malentenderán que las categorías se refieren a cuál prueba fue la que inicialmente detectó cada caso—un dato que, de hecho, me parece legítimo querer saber.

Uso ineficiente de espacio en la página frontal

En el espacio que ocupa esta actualización diaria…



…yo puedo meter gráficas breves con múltiples semanas de volumen de pruebas, casos, hospitalizaciones y defunciones. Lo digo porque de hecho hago algo así:



No digo que hay que adoptar este diseño preciso (por ejemplo, no incorpora las cifras promedio recientes y las acumuladas, y no distingue entre hospitalizaciones de adultos y pediátricas), pero entiendo que alguna variante de pequeñas gráficas ayudaría a comunicar de un vistazo no solo los valores actuales puntuales sino la tendencia reciente de estos.

Mostrar volumen de pruebas de antígeno y PCR lado a lado

En la pestaña de pruebas solo puede ver a la vez o moleculares o antígeno:

Hay muchísimo espacio ahí para mostrar simultáneamente las cifras de ambos números de pruebas. De hecho, me parece que se puede reducir el espacio dedicado a la cifra de acumuladas (o hasta quitarla), y poner gráficas de semanas recientes resumidas como las que sugiero arriba. Con la misma escala de una vez para comunicar de un solo vistazo la proporción entre estas.

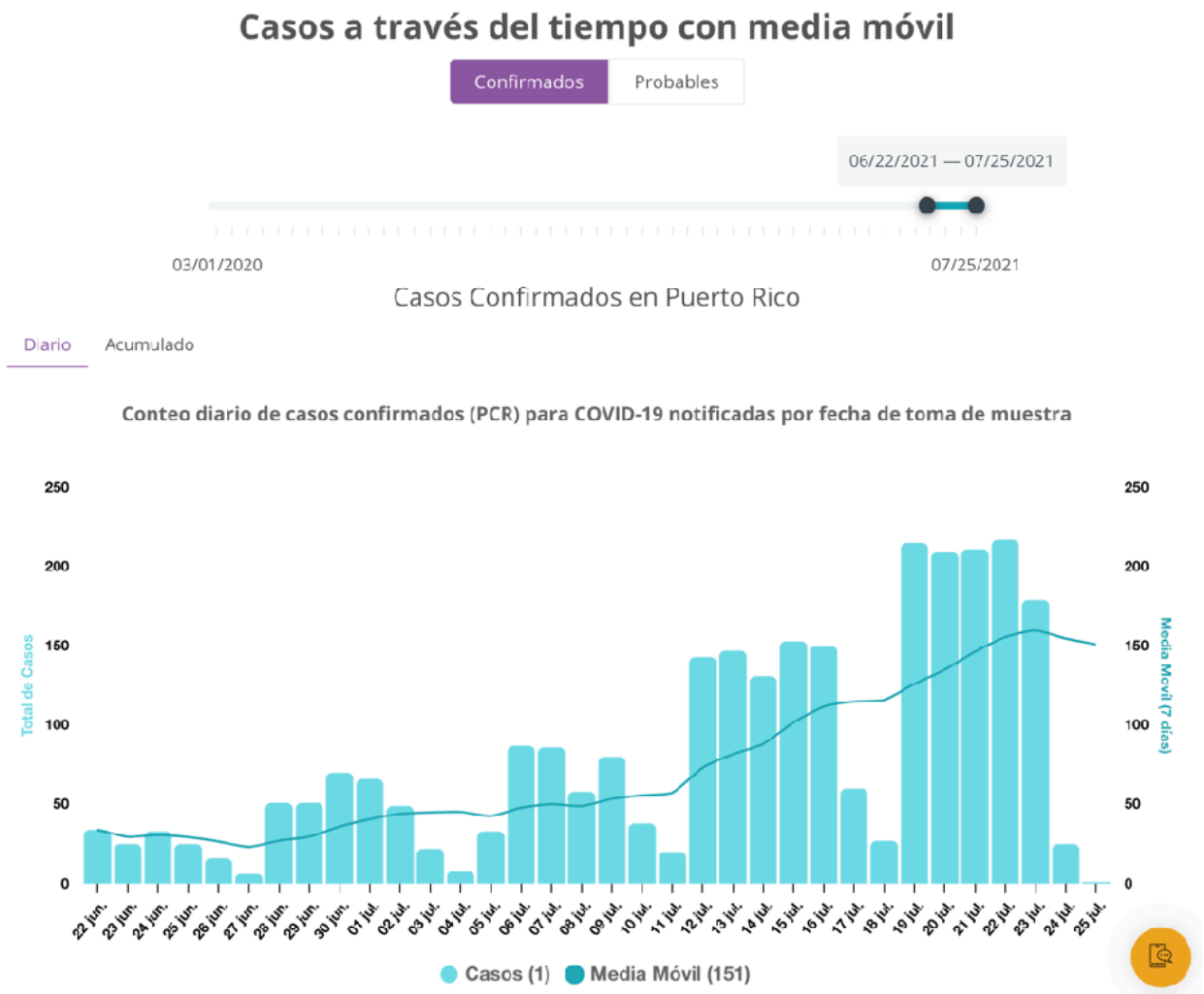
Esta recomendación de usar gráficas de hecho se puede generalizar también a las pestañas de defunciones y sistema de salud.

Otro comentario adicional: aunque arriba digo que se debe preferir combinar PCR + antígeno en las cifras de casos, mi recomendación respecto a volumen de pruebas y positividad es distinta:

- Volumen de pruebas a veces vale combinarlos, a veces separarlos.
- Positividad nunca debe mezclar las dos pruebas, ya que la sensibilidad de estas es bien distinta. La convención es solo usar moleculares para positividad.

Cifras calculadas a partir de datos incompletos

Existen varios lugares en la página en que me constato o sospecho que una cifra se calcula a partir de datos que aún están incompletos. El ejemplo más obvio son las leyendas de las gráficas con casos o pruebas por fecha de muestra, como la media móvil de 151 en esta:



En este caso me parece que se puede subsanar simplemente no incluyendo la cifra en la leyenda al fondo, y añadiendo alguna explicación visual de que las fechas más recientes adolecen de datos incompletos. Una convención común es una barra vertical gris de trasfondo en las fechas más recientes, que diga arriba “datos incompletos.”

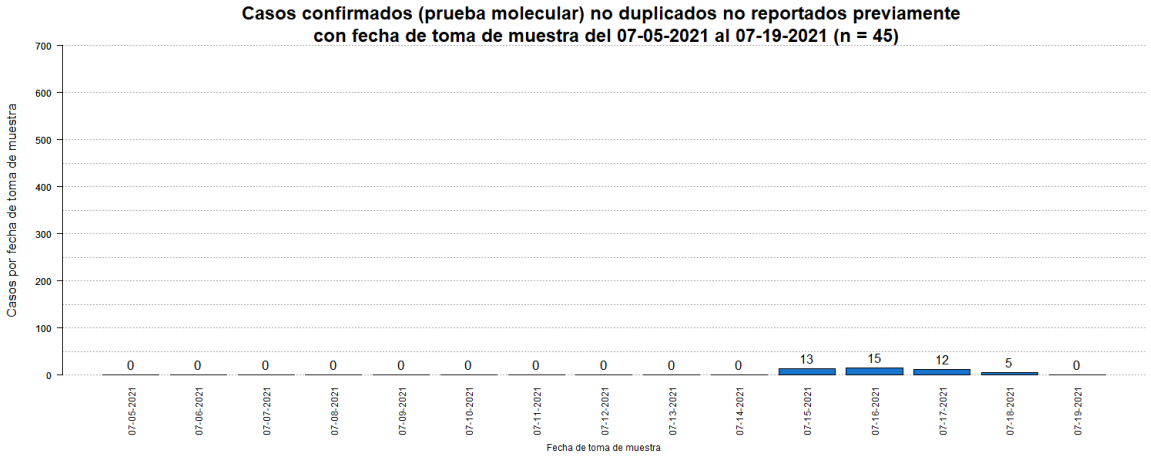
Existen otros lugares bien prominentes sin embargo en que menos obviamente se usan cifras que, según mis cotejos, parecen ser promedios móviles que incluyen fechas recientes aún incompletas. Por ejemplo al tope de la pestaña de casos sale ese mismo valor de 151, que coincide con ese promedio con datos incompletos:

Y me alarma que la pestaña de actualización diaria repite ese mismo promedio de 151 casos:

Se me ocurren dos alternativas a esto. El primero sería usar algún criterio para determinar cuál es la fecha más reciente que cumpla con algún requisito predefinido de completitud, y usar el promedio móvil por fecha de muestra de esa fecha.

Una segunda alternativa sería retomar una versión de cómo el viejo informe diario calculaba los casos adicionales diarios: mediante un criterio que combinaba la fechas tanto de muestra como de reporte del caso:

[1]Los casos confirmados son casos con una prueba molecular (RT-PCR) positiva. El número de casos confirmados adicionales desde el último informe no implica que estos casos corresponden a las últimas 24 horas. El total incluye casos con muestras tomadas del 06 de julio de 2021 al 20 de julio de 2021. La gráfica muestra la distribución de los 101 casos adicionales por la fecha de toma de la muestra.



El viejo informe diario tenía el defecto que no hacía uso de un promedio móvil, pero esta técnica de como distinguir casos reportados por muestras recientes vs. otros adicionales por resultados recibidos a destiempo siempre me ha parecido excelente, y vale contemplar si vale retomarla pero en versión mejorada que tome el promedio móvil de este cálculo para las últimas siete fechas de reporte.

Casos adicionales diarios crudos por fecha de reporte

He visto varias personas quejarse que la página ya no dice cuántos casos nuevos se añadieron hoy y que opta por un promedio móvil. Siempre les explico que es mejor fijarse en un promedio móvil en vez de las cifras crudas que lo componen, pero me parece que debe preservarse alguna manera de acceder a las cifras crudas diarias por fecha de reporte.

Pestaña de vacunación

Yo propongo que las dos preguntas claves que debe contestar a primera vista son:

1. ¿Qué porciento de la población total tiene ya régimen completo de vacunas?
2. ¿Cuán rápido está avanzando la vacunación?

Y estimo que el diseño actual de la pestaña no contesta ninguna de estas dos preguntas.

El porciento de la población total con régimen completo es de primordial importancia porque es lo que cuenta para la inmunidad colectiva. Al virus no le importa cuáles son las personas corrientemente aptas para recibir la vacuna.

La velocidad de la vacunación es importante porque es lo que determina cuán rápido podemos alcanzar esa inmunidad colectiva. Y velocidad cual es, debe expresarse en una unidad que tenga tiempo como denominador, e.g., dosis per cápita por día.

No me place para nada las medidas que enfatiza la pestaña al tope:

Me he fijado sin embargo que existen aún obstáculos serios para medir la velocidad de la vacunación, especialmente el rezago entre la administración de dosis y la entrada del datos correspondiente.

El interfaz para ver datos municipales me confunde

La primera vez que le di clic a un municipio en este interfaz me confundí muchísimo. Acabo de darle clic a uno y el mapa no indica cuál:

La primera vez ni me fijé que apareció ese botón nuevo que dice “Puerto Rico.” Y me puse a moverme dentro de la página y es fácil meterte en una situación que en la pantalla sale un número que no dice que es un solo municipio ni cuál es. ¿Cuál es este?

También acabo de fijarme que puedo escoger un municipio distinto por pestaña.

En fin, yo no soy diseñador de interacción ni de usabilidad así que se me hace difícil recomendar concretamente cómo mejorar esto, pero sí doy fe que me confundí mucho la primera vez y pensé que la página se había “tilteado.”

Recomendaciones respecto a las descargas de datos

Ofrezcan tablas de sumas resumidas

El esquema actual de las descargas es que presentan un récord por cada prueba, caso, defunción o dosis administrada. Esto tiende a producir archivos enormes, que a mí que soy ingeniero de datos esto no me causa gran dificultad, pero muchas personas querrán usar hojas de cálculo como Excel y les será difícil. Y en realidad muchas personas lo que querrán es descargar los números que aparecen en cada una de las tablas y gráficas que aparecen en las páginas.

No propongo que se eliminen los archivos de récord por caso/prueba/dosis como tal, pero lo que digo es que estos son aptos más bien para usuarios avanzados y que vale mucho apoyar también a otras poblaciones.

Restaurar campos de fecha de reporte

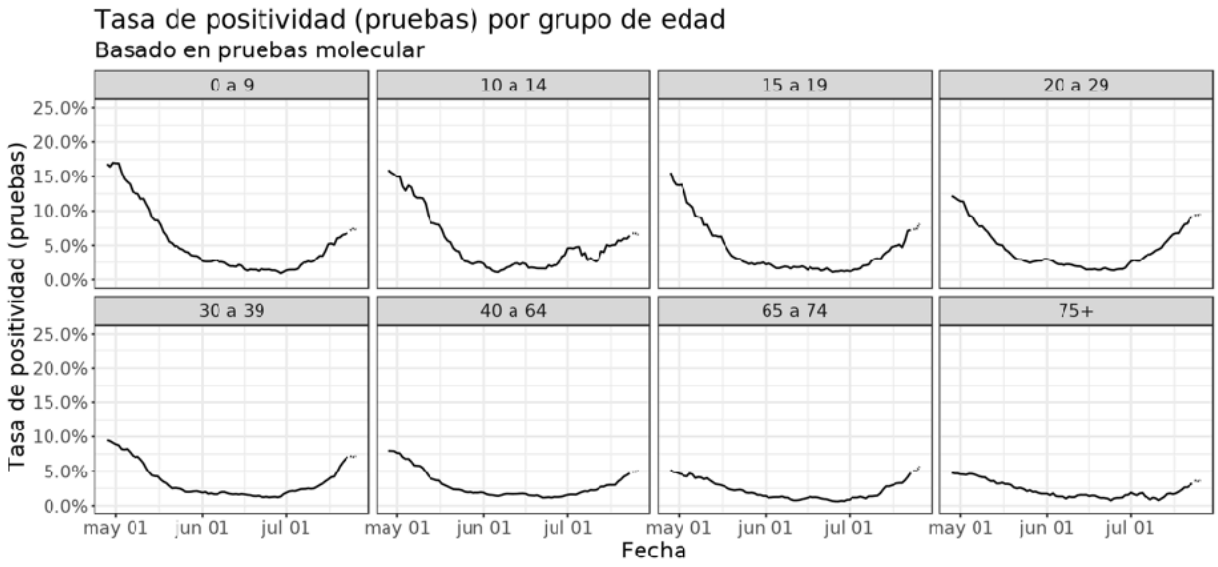
En las versiones iniciales de las descargas habían campos llamados `fe_reporte` que entiendo eran la fecha que el récord en cuestión se había añadido a los datos reportados en el portal. Luego se quitó, pero entiendo que es útil tenerlo para quien quiera calcular por ejemplo casos por fecha de reporte en vez de fecha de muestra, o por combinación de ambos criterios.

Esto cobra especial importancia con las vacunas que sufren de serios rezagos de reporte.

Categorías de edad

Varias de las descargas de datos clasifican las edades en incrementos de 10 años, e.g., 0 a 9, 10 a 19, etc. Esta clasificación no me parece la más apta, porque a veces estos grupos consisten de poblaciones que son bien distintas. El caso más notorio es los 10 a 19, que con datos del API de Bioportal (que ofrece granularidad de 5 años) se puede ver que los 10-14 y los 15-19 son bien distintos. Opciones:

1. Hacer como Bioportal y ofrecer categorías más granulares
2. Ofrecer categorías menos granulares pero mejor divididas. Por ejemplo el profesor Rafael Irizarry en sus páginas opta por estas:



Ingresos hospitalarios por COVID-19

La descarga de sistema de salud ofrece datos tales como camas ocupadas por COVID-19, ocupadas por otra causa, y disponibles. Que son importantes, pero valdría tener también el número de pacientes ingresados diarios por COVID-19 (sospechado o confirmado), que podría ser interesante analizarlos en conjunto con los casos diarios.

El Departamento de Salud y Servicios Humanos del gobierno federal publica unas descargas de datos que contienen esos campos, así que lo que entiendo que se está encuestando a los hospitales para esto. Ver esta descarga de datos:

- https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh

Municipios inválidos en vacunaciones

He observado problemas de calidad en las descargas de datos de vacunaciones. En una descarga temprana observo valores inválidos en el campo `co_municipio`:

	co_municipio	count
1	#239	1
2	.....	2
3	.MAYAGUEZ	2
4	00612	2
5	00623	2
6	00646	4
7	00692	1
8	00725	1
9	00727	2
10	00769	2
11	00778	1
12	00784	2
13	00794	2
14	00921	1
15	00923	2
16	00926	3
17	00927	1
18	00949	4
19	00950	1
20	00952	2
21	00953	3
22	00956	3
23	00959	1
24	00971	2
25	00979	1
26	00983	2
27	00985	3
28	0688	2
29	2604	2
30	3	4
31	4643_BLACKBERRY_I	2
32	4_3212_SAN_JOSE	2
33	65TH_INFANTRY	7
34	691	2
35	72	1
36	7876690465	2
37	8-42_KILOMETRO_1	1
38	A	1
39	AASCO	6
40	ABBOTSFORD	2
41	ABELA	1
42	ABERDEEN	3
43	ABIE	1
44	ABINGDON	2
45	ABRAS_DEL_COROZ(	2
46	ACCORD	2
47	ACWORTH	2
48	ADA	3
49	ADDISON	3
50	ADJIUNTAS	1
51	ADJUNATAS	2
52	ADJUNTAA	1
53	ADJUNTAS	59,140
54	AFTON	2
55	AGADILLA	1
56	AGAUDILLA	2
57	AGAWAM	7
58	AGIRRE	2
59	AGUADA	127,322
60	AGUADILLA	2
61	AGUADILA	6
62	AGUADILL	4
63	AGUADILLA	166,879
64	AGUADILLAS	23
65	AGUADILLA_DF-030(	24
66	AGUADILLA_PUEBLO	41
67	AGUADILLLA	6
68	AGUADLLA	4
69	AGUAD_BUENAS	2
70	AGUAFS	2
71	AGUAS	1
72	AGUASA_BUENAS	2
73	AGUASBUENAS	2
74	AGUAS_3GUENAS	2
75	AGUAS_3UEAS	2
76	AGUAS_3UENA	2
77	AGUAS_3UENAA	2
78	AGUAS_3UENAS	83,375
79	AGUAS_BUENAS.	4
80	AGUAS_BUENAS_	2
81	AGUAS_BUENAS_#4'	4
82	AGUAS_BUENAS_P.R	2
83	AGUAS_BUENAS_PR	3
84	AGUAS_BURNAD	2
85	AGUAS_BYENAS	2
86	AGUAS_CLARAS	2
87	AGUAS_NBUENAS	1
88	AGUAS_NUENAS	2
89	AGUAS__BUENAS	4
90	AGUA_BUENAS	3
91	AGUDA	4
92	AGUDADA	1
93	AGUDILLA	10
94	AGUGAS_BUENA	2
95	AGUIADILLA	1
96	AGUILITA	2
97	AGUIRR	2
98	AGUIRRE	5,865
99	AGUIRRE_	4
100	AGUIRRE_PR	2
101	AGUS_BUENAS	2
102	AGVUADILLA	2
103	AIBINITO	2
104	AIBOINTO	2
105	AIBONITO	97,703
106	AIBONITOA	2
107	AIBONITO_PR	2
108	AIBONITP	2
109	AIBONOTO	2
110	AIBONTIO	34
111	AIBONTO	4
112	AIIBONITO	2
113	AIKEN	3
114	AIOBONTIO	2
115	AIONITO	2
116	AIRMONT	2
117	AKRON	6

En una posterior ya no veo los valores problemáticos, pero mi análisis breve me hace pensar que solo se están excluyendo los récords con códigos que no corresponden a un municipio, y que se están excluyendo récords con códigos como `RIO_PIEDRAS` probablemente representan dosis en verdad administradas:

```
SELECT
    co_municipio "Código de municipio",
    count(*) "Número de récords",
    downloaded_at "Fecha que hice descarga"
FROM covid19datos_v2_etl.vacunacion
WHERE co_municipio IN (
    'SAN_JUAN',
    'BARRIO_OBRERO',
    'SANTURCE',
    'HATO_REY',
    'RIO_PIEDRAS'
)
GROUP BY co_municipio, downloaded_at
ORDER BY co_municipio, downloaded_at;
```

Results 1

SELECT co_municipio "Código de municipio", Enter a SQL expression to filter results (use Ctrl+S

	Código de municipio	Número de récords	Fecha que hice descarga
1	BARRIO_OBRERO	161	2021-07-24 17:24:51
2	HATO_REY	1,372	2021-07-24 17:24:51
3	RIO_PIEDRAS	8,128	2021-07-24 17:24:51
4	SANTURCE	2,683	2021-07-24 17:24:51
5	SAN_JUAN	358,991	2021-07-24 17:24:51
6	SAN_JUAN	359,540	2021-07-25 19:06:20
7	SAN_JUAN	359,773	2021-07-26 09:55:54

La limpieza de datos es tremendo pugilato, lo sé por harta experiencia personal, pero aquí parece que hay mangós bajitos que se pueden corregir fácil y cubren miles de dosis, mediante la preparación de una tabla de adjudicaciones de códigos frecuentes erróneos pero inteligibles (como los de mi ejemplo) al código correcto.