Contents

Carátula

MAESTRÍA EN CIENCIA DE DATOS

Rosario

Cohorte 2021 - 2022

Ing. Emiliano Olivares

36227254

emiliano.olivares@unc.edu.ar[1]

Presentado a fin de cumplimentar con el Final Integrador

Materia: Analisis Inteligente de Datos

Fecha: 2021-09-06

Realizado utilizando R Studio Versión: 4.1.0

[1] mailto:emiliano.olivares@unc.edu.ar

Introducción

Seleccion de datos

¿Como se obtuvieron los datos?
Se obtuvieron de la pagina del gobierno de Argentina, https://datos.gob.ar/, que engloba set de datos obtenidos mediante diferentes estudios y metodos. En su gran mayoria corresponden a proyectos de analítica y/o censado estatales.
Nuestro dataset forma parte del paquete de datos relacionados a estudios sanitarios y epidemicos denominado "Área de Salud". Específicamente corresponden al tipo secundario e incluyen información sobre "Casos de infecciones respiratorias agudas en territorio argentino, por localidad, desde 2018 hasta primer trimestre de 2021". Para ello, se extraen y mergean dos dataset que incluyen informacion de 2018/2019 y 2020/2021, respectivamente.
Las infecciones respiratorias agudas (IRAs) se clasifican en: 1) Las ETI o enfermedades tipo influenza son aquellos procesos agudos que incluyen fiebre y tos/dolor garganta sin causa aparente.
2) Neumonía: se repiten los síntomas sumando un proceso de infiltración lobar o segmentario y/o derrame pleural. Un cuadro de mayor complejidad y riesgo.
3) Bronquiolitis en menores de 2 años: definida como cualquier episodio de sibilancias que se acompañe de una infección viral, con o sin fiebre.

Interés

Personal: Relación directa con la profesión de grado -Ingeniería Biomédica-. Impacto sanitario: Las infecciones respiratorias agudas (IRAs) están asociadas a procesos de comorbilidad y evolución en enfermedades complejas.
Además, en el marco de la pandemia por COVID-19 su estudio y vigilancia se vuelven fundamentales para seguir comprendiendo el fenómeno pandémico.

Algunos interrogantes interesantes

¿Existe una distribución uniforme de los casos de IRAs en nuestro territorio?
¿Están los casos positivos de IRAs asociadas a una época (estacionalidad)?
¿Atento a la pandemia por COVID-19, los casos de IRAs según las series de tiempo se vieron modificados?
¿Afectan las IRA principalmente a un grupo etario?

Limpieza y procesamiento de datos

El dataset se encuentra en formato largo, no requiere modificación de formato.

Table 0.1: Muestra Dataset final

departamento_id	departamento_nombre	provincia_id	provincia_nombre	anio	semanas_epidemiologicas	fecha	evento_nombre	grupo_edad_id	grupo_edad_desc	cantidad_casos
43	FRAY JUSTO SANTA MARIA DE ORO	22	Chaco	2018	32	2018-08	Enfermedad tipo influenza (ETI)	7	15 a 19	2
42	GENERAL ROCA	62	Rio Negro	2020	8	2020-02	Neumonia	4	2 a 4	1
28	CONCEPCION	18	Corrientes	2018	20	2018-05	Enfermedad tipo influenza (ETI)	2	6 a 11 m	1
63	SAN PEDRO	38	Jujuy	2020	37	2020-09	Enfermedad tipo influenza (ETI)	3	12 a 23 m	2
84	PARANA	30	Entre Rios	2019	19	2019-05	Enfermedad tipo influenza (ETI)	7	15 a 19	2

Las provincias y los departamentos tienen nombres diferentes (por problemas de carga de datos por parte del área competente). Se implementa un procesamiento y limpieza de datos para:
1) Unificar IDs y nombres de provincias, evitar el uso de tildes.
2) Unificar los nombres de eventos (infección) que se han escrito de manera diferente para cada dataset, pero significan exactamente lo mismo.
3) Evitar conflictos de incompatibilidad de variables y realizaciones. Sorprende, además, la cantidad de diferencias entre set de datos aún cuando ambos provienen de una misma área gubernamental y responden a una misma medición en diferentes momentos pero con poca diferencia temporal.
Nuestro dataset se presenta en escala “semanas epidemiológicas”, un concepto asociado al sanitarismo.
Nos puede interesar observar y analizar las series de tiempo en formato tradicional (%Y%m). Por ello, agregaremos una columna que sea “fecha” con codificación %Y%m que se compone de una combinación de las columnas “anio” y “semanas_epidemiologicas”.
Guardamos el dataset final, al terminar el proceso de limpieza de datos. Para futuros trabajos y como referencia.
Mantenemos “anio” para evitar problemas de codificación a futuro. Además, las columnas se encuentran siguiendo el estándar de delimitar palabras con guión bajo y se entienden representativas de su contenido.

Materiales y métodos

Exploración y Presentación de Datos

Se hará uso de las herramientas de tidyverse buscando la implementación de pipelines claros, comentados y reproducibles.
Nuestro dataset tiene un total de 994498 observaciones individuales. Contamos con informacion sobre IRA respecto a los años 2018, 2019, 2020, 2021, sobre las 24 provincias argentinas (Buenos Aires, CABA, Catamarca, Chaco, Chubut, Cordoba, Corrientes, Entre Rios, Formosa, Jujuy, La Pampa, La Rioja, Mendoza, Misiones, Neuquen, Rio Negro, Salta, San Juan, San Luis, Santa Cruz, Santa Fe, Santiago del Estero, Tierra del Fuego, Tucuman).
Nos interesa primeramente conocer la distribución total de las infecciones. Podemos visualizar las frecuencias relativas de cada infección, agrupando por año.

Distribucion por infección

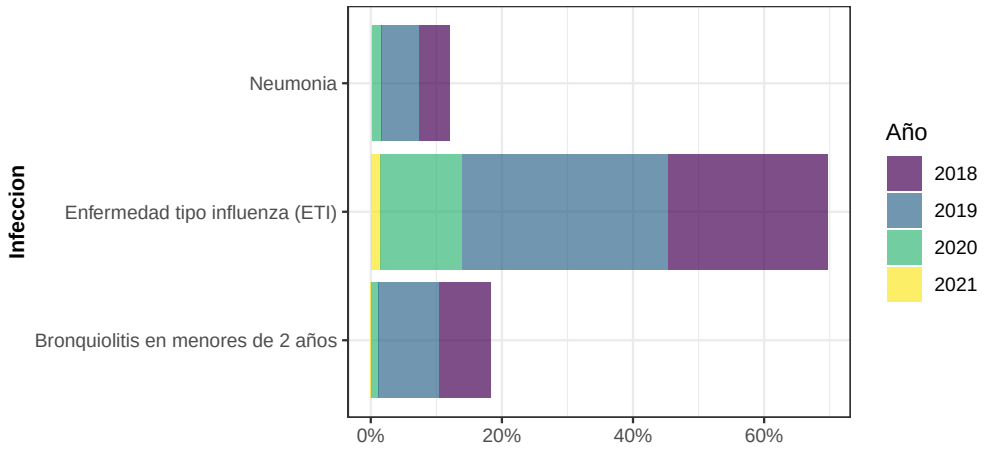

Figure 0.1: Distribución porcentual de casos positivos según infeccion para los años 2018, 2019, 2020 y 2021

Se observa a las infecciones determinadas como ETI como las mas frecuentes por su relacion directa con etapas agudas y su definicion de caso laxa que no requiere imagenología ni causa aparente. Además, se observa una diferencia en la distribución de casos respecto a los años. Esto se debe a que la información ofrecida por el gobierno de la Nación respecto al año 2018 se encuentra incompleta. Respecto al año 2021 apenas se tienen algunos casos correspondientes a las primeras semanas (es de esperarse, ya que nos encontramos en Septiembre de 2021). Resulta, en ese caso, interesante estudiar la distribución porcental de casos para los dos años que contamos con información completa: 2019 y 2020. El 100% corresponde al total de casos. Podremos ver como ese total se distribuye entre los dos años. Lo esperable, sería observar una distribución cercana al 50% en cada año (sin cambios significativos de uno a otro).

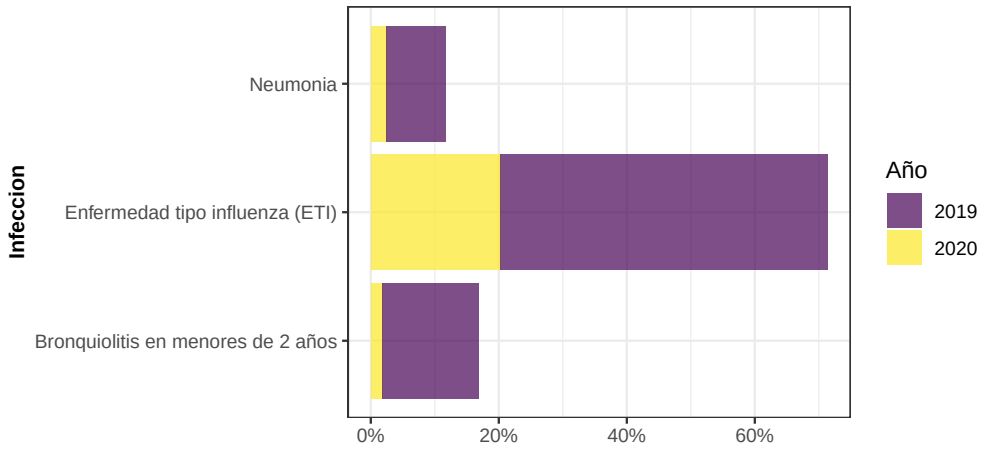

Figure 0.2: Distribución porcentual de casos positivos según infeccion para los años 2019 y 2020

Distribucion por provincia

Además, resulta de interés estudiar la distribución de casos por provincia para detectar varaciones geográficas. Nuevamente excluiremos los casos del 2021 que en la comparación total (suma total de casos por año) no resultan representativos. Ref: ver figura 0.3.

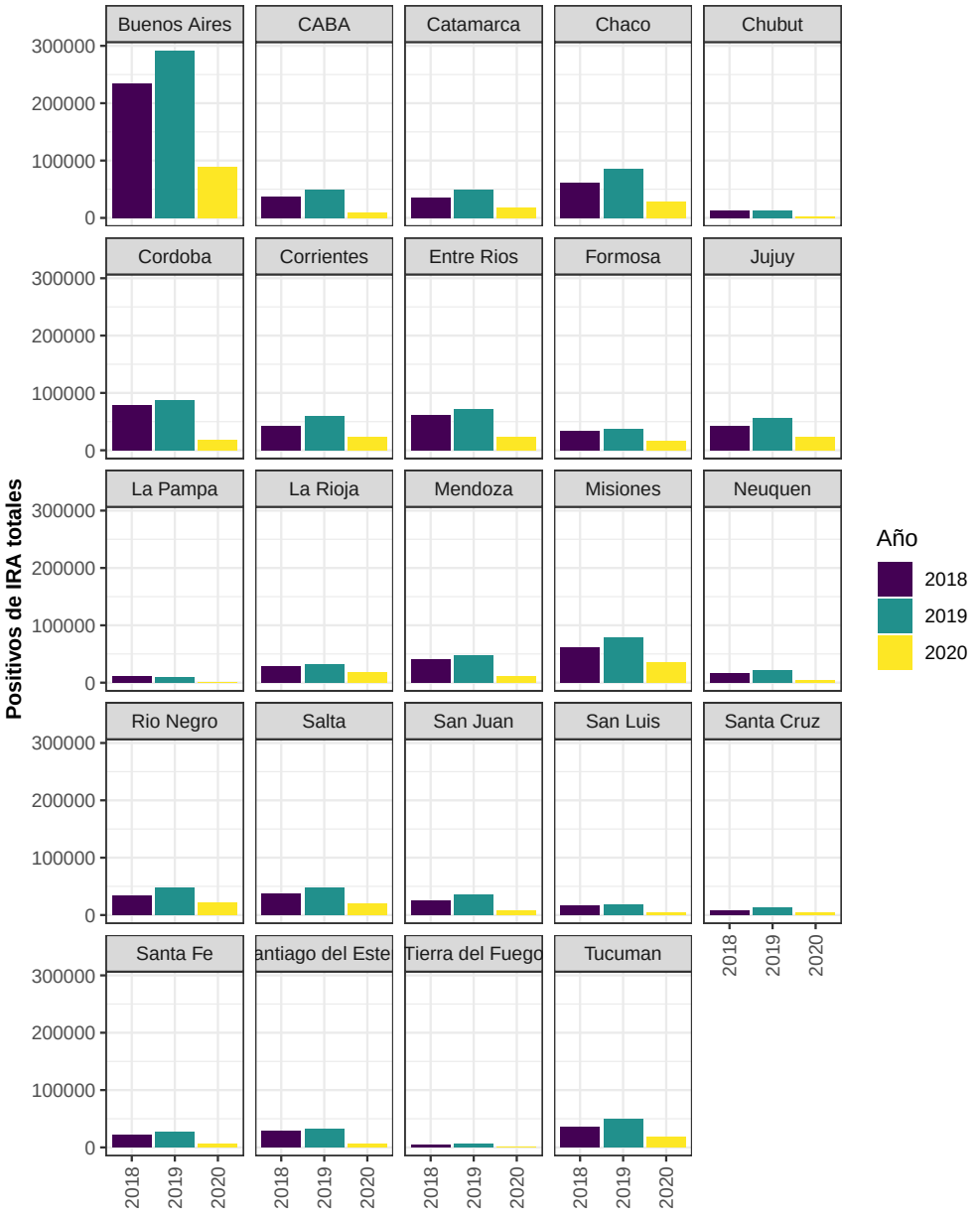

Figure 0.3: Cantidad de casos positivos anuales de IRA

La distribución de casos por provincia puede resultar engañosa. Un observador podría concluir que Buenos Aires se ve fuertemente afectado por enfermedades infecciosas y que nuestro gobierno nacional debería aumentar los fondos destinados a áreas de salud epidemiológicas en la provincia.
Para evitar esto, podemos estudiar la cantidad de casos por provincias según cantidad de habitantes. Los habitantes por provincia no se encuentran en nuestro dataset. Por eso y para mantener la consistencia de datos obtendremos esta información del mismo ente que ofrece nuestros datos originales.

Distribución por provincia cada cien mil habitantes

La media anualizada para el territorio argentino de contagios de IRA cada 100 mil habitantes para el año 2019 es: 4377,88, mientras que para el año 2020 es: 1534,099, Catamarca, por ejemplo, tiene una Incidencia Acumulada de: 12058,8391 y 4308,9462 respectivamente.
La proporción que hemos formalizado anteriormente puede definirse como Incidencia Acumulada o

Table 0.2: Incidencia Acumulada por provincia

Provincia	Incidencia Acumulada 2019	Incidencia Acumulada 2020
Santa Fe	756,1889	154,2804
CABA	1581,2938	297,5310
Buenos Aires	1663,3525	511,1640
Chubut	2005,5122	508,2440
Cordoba	2335,3056	492,7868
Mendoza	2362,7645	590,1510
La Pampa	2799,1675	480,7102
Tucuman	2868,5468	1026,3440
Tierra del Fuego	3117,0718	444,5546
Santiago del Estero	3246,3026	697,1184
Neuquen	3249,8716	628,1087
Salta	3360,3693	1358,9610
Santa Cruz	3452,3022	1313,6522
San Luis	3623,6446	768,0081
San Juan	4551,3602	1084,0778
Entre Rios	5214,1438	1658,4882
Corrientes	5265,5199	2119,3771
Formosa	6214,0507	2656,6732
Misiones	6294,8052	2763,6697
Rio Negro	6400,6634	2889,4745
Chaco	7089,3394	2352,0162
Jujuy	7216,0295	2999,0362
La Rioja	8342,6719	4715,0034
Catamarca	12058,8391	4308,9462

proporción de individuos sanos que desarrollan la enfermedad a lo largo de un periodo determinado. La medida de tiempo será anualizada y nos centraremos sobre los años 2019 y 2020. Utilizaremos una tabla de incidencia para demostrar la utilidad de la proporción definida. Mediante el analisis de la tabla buscaremos identificar zonas geográficas de mayor riesgo epidemiologico.
Más adelante, mediante el uso de otras herramientas -serie de tiempo-, incorporaremos los datos incompletos del año 2019 y 2021 a nuestro análisis.

Análisis temporal: series de tiempo

Se presenta la serie de tiempo mensualizada para estudiar el comportamiento de los casos positivos a lo largo de los datos. Se observan diferencias entre las gráficas. Se recuerda, en este punto, que se tienen datos correspondientes al primer trimestre para el caso del año 2021. El dato puntual para el mes de Abril del año 2018 tampoco es válido y debe excluirse del análisis ya que se encuentra con informacion parcial del mismo. Se excluye del gráfico.

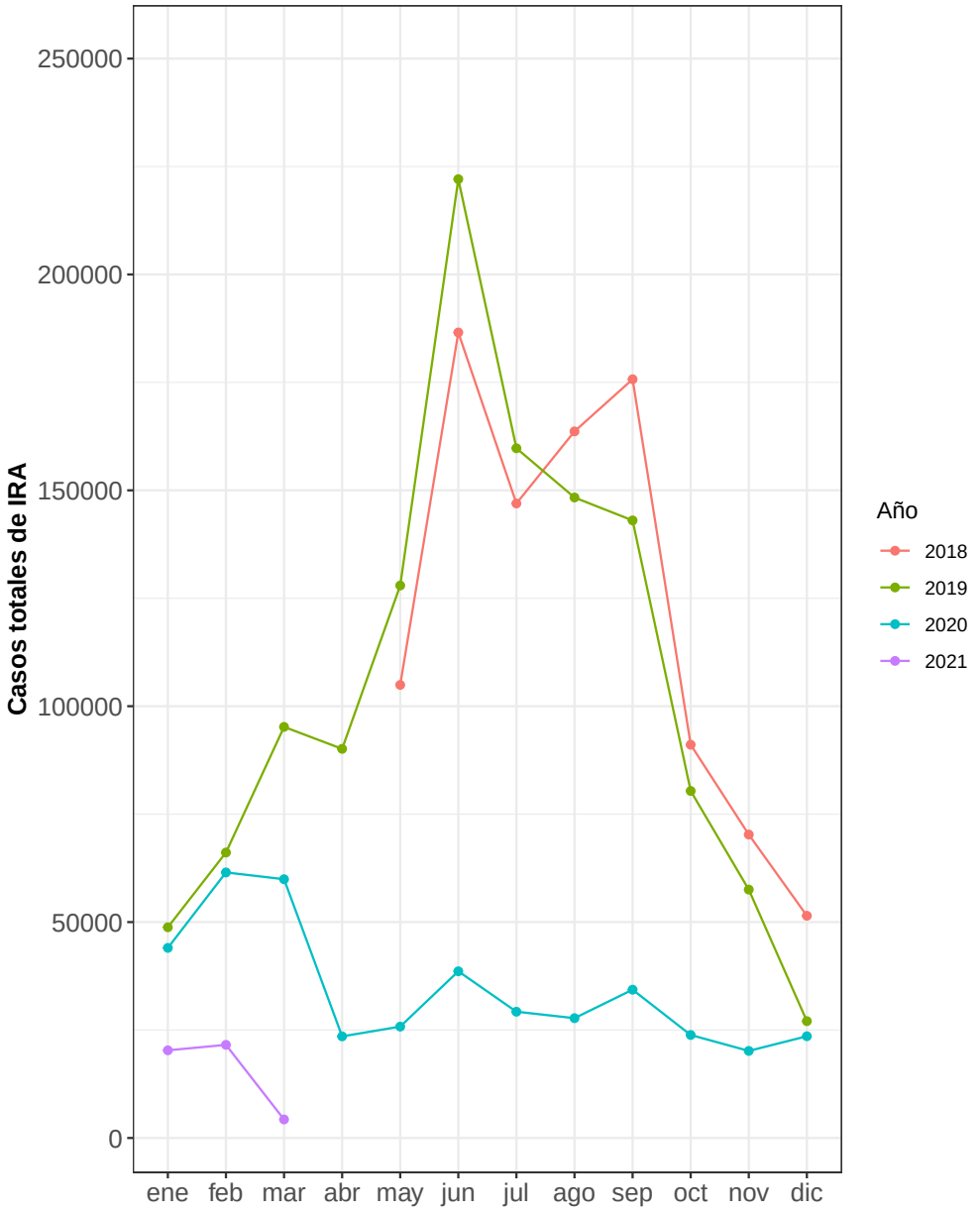

Figure 0.4: Series de tiempo. Casos positivos de IRA en Argentina.

Análisis etario

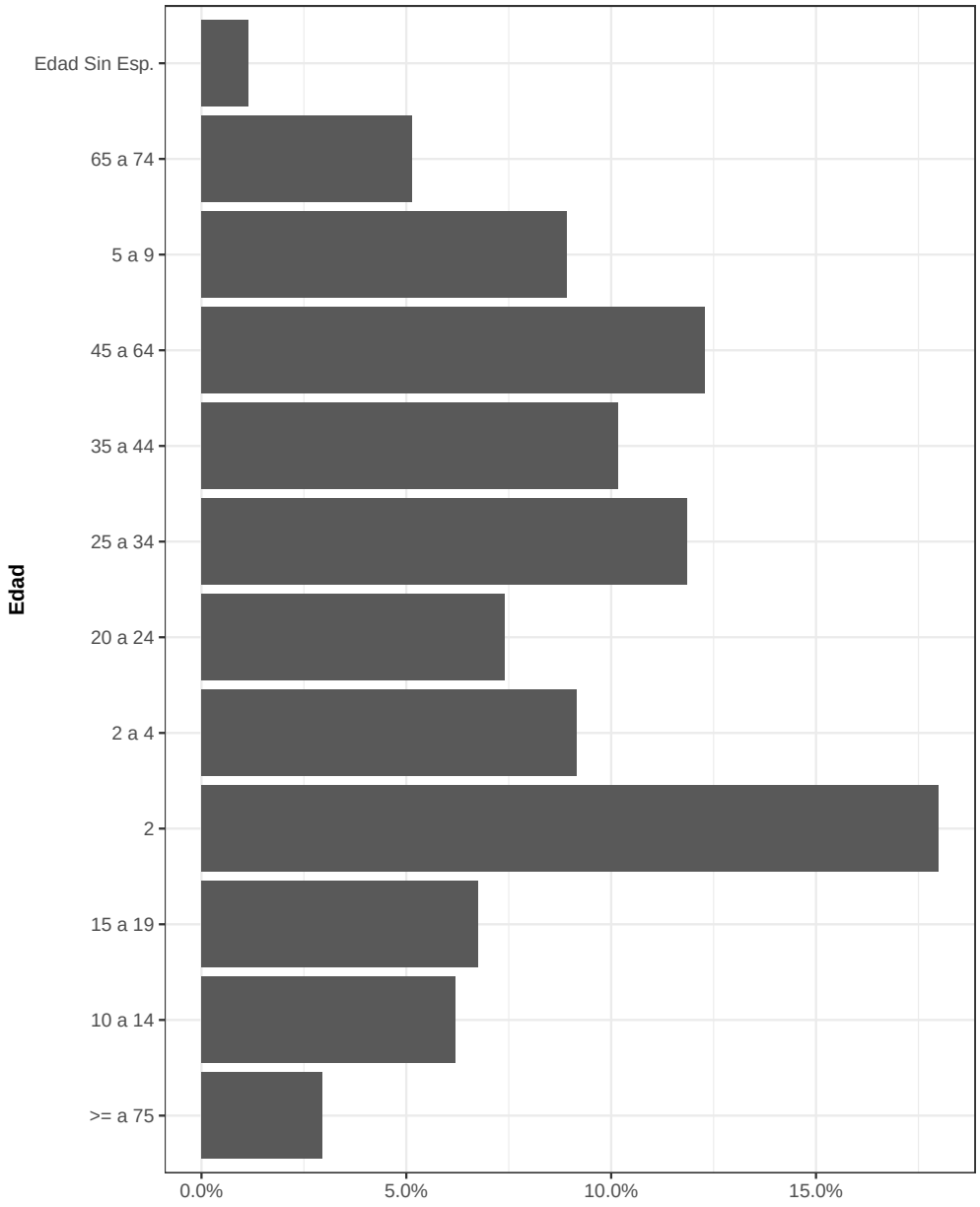

Figure 0.5: Semanas epidemiológicas. Casos positivos de IRA en Argentina.

Se realiza un gráfico de barras sobre los casos confirmados de IRA respecto al año 2019 para estudiar la distribución respecto a edad. Las IRA son una fuerte causa de muerte en menores de 5 años. Además, afecta fuertemente a adultos mayores, quienes en general sufren los procesos infecciosos con menos posibilidades de recuperación/curación.

Resultados y discusión

Impacto de la pandemia por COVID-19

Se encontró un decrecimiento significativo de casos totales de IRA de 2019 a 2020. Por ejemplo, tomando como referencia la provincia de Córdoba, la incidencia acumulada para el año 2019 fue de 2335,3056 y de 492,7868 para el año 2020. El decrecimiento podría en principio vincularse a la pandemia por COVID-19.

De hecho, observando la gráfica en figura 0.4 podemos apreciar como la tendencia que indica un aumento en el mes de Abril, ya que los meses vinculados al clima frío presentan un aumento significativo de los casos de IRA, no se cumple para el año 2020. Ese mes resulta sumamente significativo ya que implica el comienzo del ASPO en nuestro país. Claramente, la pandemia y con ella las decisiones gubernamentales, afectaron la cantidad de casos positivos detectados y notificados de IRA por nuestro sistema de salud.
Existen dos situaciones a destacar. Por un lado los coronavirus son también virus respiratorios, lo que podría llevar a un submuestreo de los casos de IRA: muchos casos positivos de IRA pueden no haberse detectado por considerarse COVID-19. Además, el APSO puede haber jugado un rol fundamental para evitar la circulación de los virus que provocan IRA.
Por último, observamos que un aumento de casos significativo asociado al período invernal. Esta conclusión, extraída principalmente de la Serie de Tiempo, refuerza el argumento del impacto del APSO: en nuestro país se vivió un aislamiento estricto en la mayoría de las provincias durante ese período.

Distribución geográfica

Provincias como Catamarca, La Rioja, Jujuy y Chaco poseen una proporción de casos positivos mucho mayor a la media de Argentina. El estudio de tablas de incidencia puede mostrarnos zonas de riesgo: en nuestro caso detectamos al NOA como una zona geográfica fuertemente afectada por las IRA. Si bien las IRA están asociadas a procesos altamente contagiosos, por su mecanismo intrínseco, los conglomerados más densamente poblados de nuestro país no presentan una proporción de casos positivos mayor a la media, de hecho se encuentran entre las más bajas del país, tomando a Santa Fé como ejemplo.
Esto indica que debemos prestar atención no sólo a los mecanismos intrínsecos de la enfermedad, sino también a las condiciones sanitarias y al acceso a herramientas de prevención en salud. La inequidad sanitaria es más notoria en provincias con menores recursos absolutos y en específico destinados a salud, y esta afirmación juega un rol fundamental en el control y prevención de las IRA.

Distribución etaria

El gráfico asociado a la distribución etaria (figura 0.5) muestra el fortísimo impacto que tienen las IRA sobre la población menor a 5 años. Mejorar la respuesta sanitaria ante procesos infecciones y proteger a nuestros infantes en epocas invernales es necesario para dismunuir la cantidad de casos positivos en niños y lactantes.

Conclusiones

Se pudieron abordar correctamente los cuatro interrogantes planteados al comienzo del trabajo. Se evidencia el impacto de la pandemia de COVID-19 que afectó notoriamente los casos de IRA. Además, los interrogantes sobre regionalización y grupo etario de las IRA fueron resueltos: se evidenciaron zonas geográficas y edades de mayor riesgo epidemiológico.
Sin embargo, para encontrar las causas de estas afirmaciones se debería sumar información no presente en

nuestro dataset y buscar vínculos o correlaciones entre eventos y variables. A saber, a futuro, se propone: Estudiar la serie de tiempo de los contagios por coronavirus en conjunto con los casos confirmados de IRA. Acompañar el análisis geográfico de un análisis de recursos estatales/privados destinados a prevención en salud per capita, por provincia o región.
Identificar tasas de mortalidad de las IRA para refrendar el análisis de impacto etario.
Por su parte RMarkdown se presentó como una herramienta potente para llevar adelante los procesos de exploración, análisis y comunicación, permitiendo la incorporación de gráficos y tablas de manera directa sin necesidad de ninguna herramienta adicional.