Los estudios ecológicos como parte de la investigación de una epidemia en terreno

Victor M. Cárdenas

Resumen. La descripción de una epidemia en persona, lugar o tiempo puede involucrar comparaciones de grupos de individuos que, con la debida atención a los asuntos estadísticos e inferenciales a nivel de grupo o individuo, hagan que nuestro ejercicio de medición pueda elevarse a la estatura de un estudio ecológico. Con varios ejemplos ilustramos porque este es uno de los diseños de estudio más utilizado, aunque menos reconocidos por los epidemiólogos de los servicios de salud pública.

Abstract.

The description of an epidemic in person, place, or time may involve comparisons of groups of individuals that, with due attention to statistical and inferential issues at the group or individual level, can elevate our measurement exercise to the stature of an ecological study. With several examples we illustrate why this is one of the most widely used, though less recognized by public health epidemiologists, study designs.

Introducción

Una de las áreas menos explicadas en la formación didáctica de los epidemiólogos de campo es el lugar que tiene el análisis de los datos de la vigilancia de salud pública que muchas veces ya existe sobre el evento que ocupa a la investigación de una epidemia. La división misma entre lo que es un estudio epidemiológico, y lo que es un análisis descriptivo de los datos en persona, lugar y tiempo, es también un área gris. Una revisión del concepto estudio epidemiológico es conveniente antes de introducir el concepto de estudio ecológico.

¿Qué es un estudio epidemiológico?

La investigación epidemiológica por muchos años, digamos hasta la sistematización de métodos modernos en los 1970´s no hacia una gran distinción entre el análisis de datos de lo que hoy en día llamamos vigilancia de salud pública (morbilidad notificada o encuestas de salud y mortalidad) de los estudios planeados en que se reclutan individuos tradicionalmente con enfermedades crónicas como los de casos y controles o individuos de acuerdo a sus exposiciones como estilos de vida en los estudios de cohortes, o el análisis de datos de encuestas como en los estudios transversales. Inclusive los textos más antiguos cómo los de Fox, Hall y Elveback [1], y Lilienfeld y Lilienfeld [2], dedicaban una buena parte de su texto al estudio de los patrones de

morbilidad y mortalidad. El primero de ellos, un libro especialmente pensado para los epidemiólogos practicantes de salud pública dedica 300 de las 362 páginas a la descripción de la enfermedad en la población. No hay realmente una distinción entre la epidemiología descriptiva y la analítica.

Debemos preguntarnos qué diferencia real existe entre ambos sin desdeñar el hecho fundamental que los datos primarios son superiores a los datos secundarios [3, 4]. Ambos deben ser entendidos como la aplicación de un enfoque sistemático que atienda a los asuntos de validez y precisión de las mediciones y comparaciones, esto es, las metas de la estrategia de la investigación epidemiológica vista como "*un ejercicio de medición con exactitud como meta final*" según Rothman [5] cuando se planea e implementa un estudio, y se recolectan y analizan los datos o se interpretan los análisis de ellos. Abundaba Rothman que la reducción tanto de errores sistemáticos como aleatorios logran alcanzar la exactitud deseada en la medición.

Esta falta de distinción debe tenerse en mente cuando debido a la explosión de las fuentes de datos ahora disponibles por el crecimiento acelerado del internet y la así llamada era digital [6], que crea oportunidades crecientes para el uso de tales fuentes en el quehacer del epidemiólogo en general, incluyendo la o el practicante de esta ciencia cuando ocupan puestos de servicios de salud pública.

Una categorización de los estudios epidemiológicos que utilizamos comúnmente es la que distingue las mediciones por sus fines y la presencia o no de consideraciones de precisión, en particular las relacionadas con la "significancia" estadística o prueba de hipótesis. Los estudios descriptivos se dice que no prueban hipótesis, acaso las generan o tamizan, mientras que los estudios analíticos hacen pruebas de hipótesis, pero tal distinción ha sido cuestionada [7]. Una prueba de hipótesis trata de responder la pregunta de que tanto los datos observados son consistentes con la hipótesis nula. Una hipótesis nula o de nulidad es aquella que declara que no existen diferencias en la ocurrencia de enfermedad o la condición de salud o riesgo de interés entre expuestos o no expuestos o por niveles de exposición. La prueba de hipótesis se basa en una declaración de la probabilidad de los datos si la hipótesis nula fuese cierta, es decir el valor de *P*. En general preferimos usar el intervalo de confianza (las más de las veces del 95%) alrededor de un estimado ya sea una tasa u otra medida de frecuencia, o una razón de riesgos u otra medida de asociación, o aún mejor la función del valor de *P* porque además de hacer una declaración acerca de la hipótesis nula, da información sobre la dirección más probable y la magnitud del estimado de ocurrencia o asociación [8] y de cualquier otra hipótesis. Además, la tendencia actual del

pensamiento epidemiológico es hacia la preponderancia de la evaluación de la validez y no solamente del error aleatorio, es decir no pensar solamente en el valor de *P* o el intervalo de confianza o la función de *P*, sino que tan plausible es que los resultados sean debidos a sesgos o confusión utilizando el valor de la evidencia o valor -*E*, una medida mínima de la fuerza de asociación que podría ser explicada alternativamente por un sesgo y que ha sido propuesto por VanderWeele [9].

En realidad, uno puede al describir un brote epidémico o una epidemia determinar las tasas de ocurrencia, las tasas, estimar los intervalos de confianza, y al comparar las tasas locales a través del tiempo ajustarlas digamos por edad y probar la hipótesis de si las tasas han aumentado significativamente desde el punto de vista estadístico. Al recolectar los datos primarios en terreno y al revisar exhaustivamente los datos existentes, hay suficiente información para incluir análisis que pueden o no ser considerados estudios ecológicos como describiremos a continuación.

En resumen, cuando los recursos, principalmente nuestro tiempo, y nuestra capacidad lo permite podría tener el nivel de análisis de tal calidad, que el ejercicio de medición sea de la talla de un estudio epidemiológico descriptivo con análisis que llenan la definición de un estudio ecológico.

El estudio ecológico

Otra categorización de estudios epidemiológicos que se usa también es la que distingue entre estudios de individuos y estudios de grupos, y llama *estudios ecológicos a aquellos en que la unidad de análisis son grupos de individuos* [10]. Morgenstern y Wakefield aseguran que la diferencia entre un análisis descriptivo de los datos de la vigilancia y un estudio ecológico estribaría en la atención que se le da a los asuntos estadísticos e inferenciales. Aquí tratamos de proveer una revisión para dar mayor validez y precisión a la investigación de campo y al análisis de los datos de la vigilancia.

Los estudios ecológicos los dividen Morgenstern y Wakefield por su nivel de inferencia a nivel biológico, o a nivel ecológico o de grupo. El primero se refiere a que la inferencia entre la relación de la variable de exposición o independiente (X) y la de efecto o dependiente (Y) se haga a nivel individual, mientras que la segunda se hace a nivel de grupo [10]. Los efectos a nivel de grupo dependen del grado en que en los grupos se distribuya la exposición. Por ejemplo, uno puede estudiar la relación entre la atención que se presta en los servicios de salud

y digamos la mortalidad materna. Digamos que, en vez de medir la utilización de los servicios, utilicemos el tener o no seguridad social, es decir estar o no asegurado, como variable de accesibilidad a los servicios. Sin embargo, el uso o utilización de los servicios disponibles no es lo mismo que tener seguro lo cual puede introducir una mala clasificación, es decir un sesgo de información, dando una medición errónea de la asociación entre la mortalidad materna y el papel de los servicios por utilizar la accesibilidad de los servicios de salud como variable próxima. Asimismo, las prácticas de salud pueden ser bastante diferentes de un lugar a otro aun teniendo cobertura bajo la misma institución de seguridad social. Un estudio de alternativo de individuos podría ser aquel en que esta evaluación averiguaría entre las muertes maternas el antecedente de utilización de los servicios de salud, como el número de consultas prenatales y que tan temprana y adecuadamente se proporcionaron los servicios entre las mujeres que fallecieron (una especie de auditoría) comparados con los antecedentes entre un grupo de mujeres que no fallecieron y estuvieron embarazadas en un período semejante, o controles. Este último no es un estudio ecológico, sino un estudio de casos y controles.

Los estudios ecológicos de acuerdo con Morgenstern y Wakefield [10], además se pueden dividir entre exploratorios y etiológicos, según el grado de la validez de contenido y constructo de la medición: los estudios etiológicos se basan en mediciones sobre niveles promedios de exposición mientras que los exploratorios utilizan otras variables que solamente aproximan una medición de la exposición como la residencia o simplemente fechas en que se infiere ocurrió la exposición. Los estudios además se dividen según los autores en diseño de grupos múltiples, diseños de tendencias temporales y diseños mixtos. Para efecto de simplicidad, entendiendo que la mayoría de las veces los estudios ecológicos serán parte de una investigación de campo que es suplementada por el análisis descriptivo de los datos de vigilancia, serán considerados exploratorios.

Análisis descriptivos de la variable tiempo

Los datos de vigilancia tienen entre sus atributos las fechas de inicio (enfermedades de notificación obligatoria) o de defunción por causas específicas o al menos de diagnóstico (cáncer). La escala de medición es continua, pero la variable la agregamos en intervalos predeterminados como horas, días, semanas, años. Los datos forman una así llamada *serie de tiempo*. La metodología ha sido desarrollada por estadísticos con aplicaciones para los mercados de valores que guían las inversiones, el clima con tal confiabilidad que nos lleva a tomar o no el paraguas y en la salud pública, permitiendo definir la estacionalidad y detectar epidemias. Algunos de los

métodos más conocidos son los promedios móviles, la desestacionalización y la destendencialización que utiliza métodos llamados de diferenciación [11], la regresión lineal y en la escala logarítmica para obtener porcentaje de cambio por unidad de tiempo, con extensiones de métodos de simulaciones para obtener segmentos de una recta que se ajusten mejor a los datos, o regresión de puntos de inflexión [12] y los métodos de promedios móviles regresivos integrados o ARIMA [11]. Veamos brevemente estos métodos.

Análisis de tendencias

El análisis de una serie de tiempo a menudo tiene como objetivo principal identificar si los casos van en aumento o en descenso. Por mucho tiempo se ha utilizado la regresión lineal simple para estimar la tendencia de casos o mejor de tasas ajustadas por edad, digamos por año, e informar de un cambio expresado en términos de la pendiente (b) en la ecuación conocida:

$$Y = \alpha + \beta_x + E$$

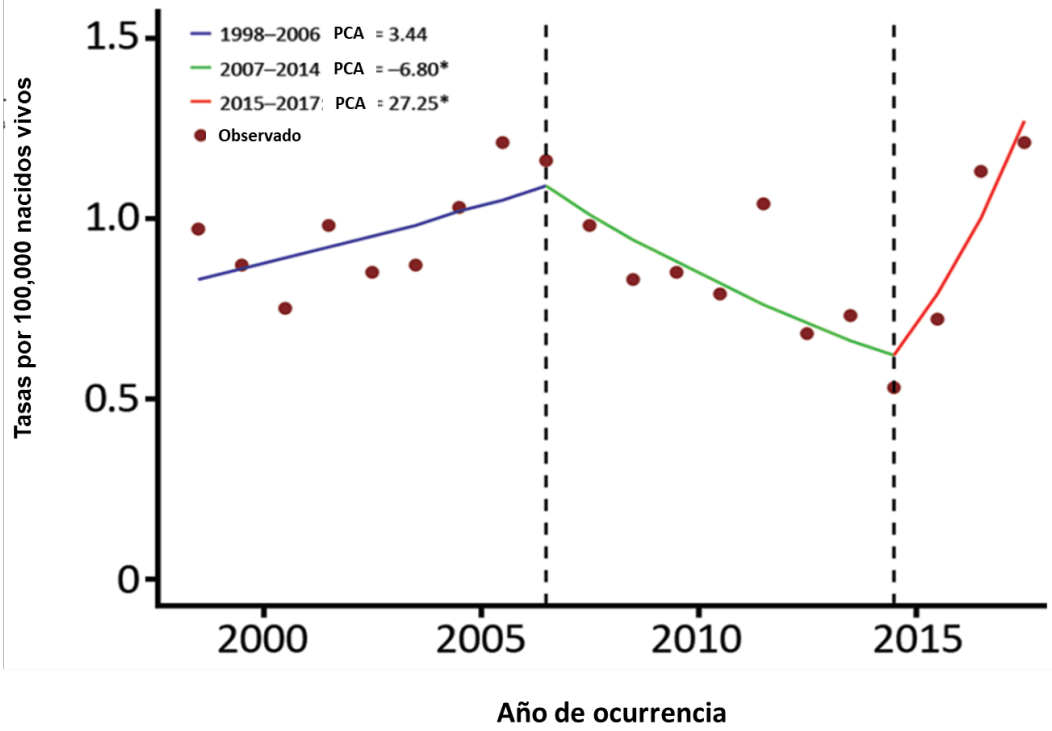
dónde Y son las tasas, x son los períodos digamos años, y α el intercepto, β es la pendiente, y E es un componente de error aleatorio, y este último generalmente se omite quedando $Y = \alpha + \beta_x$. Un problema con este método es que a menudo hay variaciones dentro de la recta lineal y no describe adecuadamente el fenómeno. Una extensión de este método es la regresión polinomial, una extensión de la regresión lineal con términos cuadráticos y la descomposición de una curva aplicando funciones de senos y arcosenos utilizando derivadas de Fourier como se explica por el Dr. Knox [13]. Un método desarrollado por Hyune-Ju Kim y cols. de la Universidad de Syracuse en el 2000 [12], disponible en el sistema Joinpoint (https://surveillance.cancer.gov/joinpoint/) desarrollado en el Instituto Nacional de Cáncer de los EE.UU., ha encontrado gran acogida para el estudio de tendencias. Este método llamado método de análisis de puntos de inflexión (*joinpoint analysis* en inglés), identifica y estima segmentos que describen una serie. El método utiliza simulación Monte Carlo para introducir permutaciones de posibles puntos de inflexión alternativos a modelos con n-1 nuevos segmentos. Un aspecto atractivo del método es la utilización de una transformación logarítmica de las tasas de manera que el estimado de la pendiente β, puede interpretarse en términos de porcentaje de cambio ascendente o descendente. La forma de la ecuación es:

$$Log(T_y) = \alpha + \beta_x$$

Por lo que el porcentaje de cambio anual (PCA) (o en la unidad de tiempo especificada) es $(e^{\beta_x} - 1) \times 100$. Como puede haber diferentes tendencias sobre un periodo observado, y podría interesar cual es la tendencia promedio, el programa también estima el porcentaje promedio de cambio anual (PPCA).

Para ilustrar la utilización de este método presento los resultados de una evaluación de las tendencias en registro de microcefalias entre menores de un año en México entre 1998 y el 2017 [14].

Gráfica 1. Defunciones infantiles por microcefalia en México, 1998-2017



Fuente: [14].

Los datos muestran que dentro de la serie se pueden identificar tres segmentos con tendencias diferentes significativamente desde el punto de vista estadístico denotadas por los tres colores con dos puntos de infección en 2007 y 2015. La tendencia que había en el segundo segmento era descendente y en el tercero fue ascendente representando un cambio del 27.3% mayor que los cambios habidos antes, de 3.4% y -6.8%. Los autores inferimos que el aumento del 2015 al 2017 representaron la llegada de la pandemia regional de Zika a México y estos eran algunos de los casos del síndrome congénito por Zika (SCZ), y comentamos que las

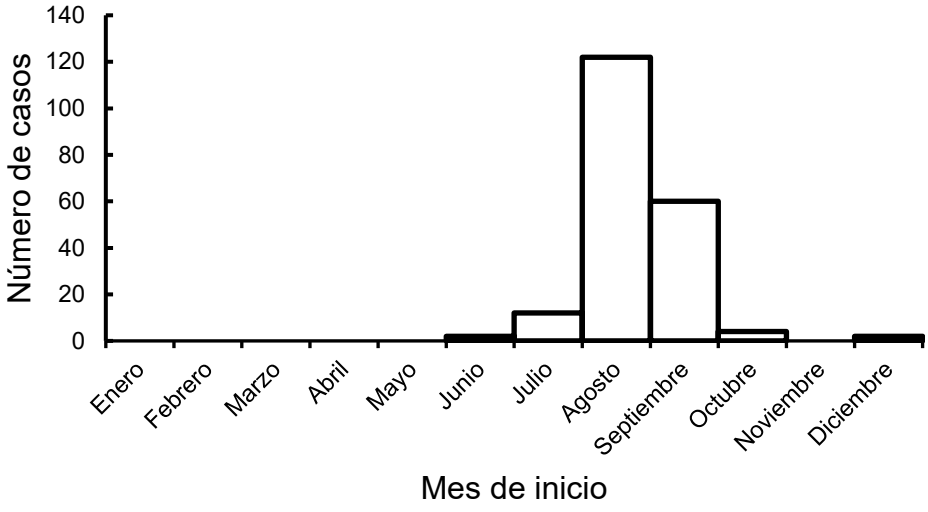
defunciones en números totales de SCZ que uno pudiera atribuir a la pandemia, representaban 17 defunciones comparados con un total de solamente 11 casos de SCZ notificados en México [14].

Otra aplicación de estos métodos es la comparación de las tendencias antes y después de una intervención, como lo presentó hace años el mismo Dr. Morgenstern [15], en lo que se denomina análisis de series de tiempo "interrumpidas" porque la tendencia cambia, es decir se interrumpe, por la intervención que ocurre [20]. Además de los modelos descritos antes como la regresión en escala logarítmica antes descrito, para obtener un estimado del cambio porcentual, hay otros modelos como el de series de tiempo que permite acomodar problemas como el de la estacionalidad como se discute en el tutorial sobre el tema [16].

Conglomeración estacional

La descripción de la estacionalidad no requiere necesariamente métodos sofisticados, sino simplemente la agrupación de los casos por semanas epidemiológicas o meses a lo largo de los años. En la gráfica 2 que se muestra a continuación se observa una clara estacionalidad de la enfermedad neuro-invasiva por el virus del Nilo Occidental (VNO) en El Paso, Texas, de hecho, circunscrita a los meses del verano tardío.

Gráfica 2. Casos de enfermedad por Virus del Nilo Occidental según fecha de inicio, El Paso, Texas, 2003-2010



Fuente: Departamento de Salud Pública de El Paso, Texas

Si hubiera duda sobre la estacionalidad o agrupamiento en el tiempo, digamos que los casos fuesen menos o se tratara de una enfermedad sin que se conociera el proceso subyacente, claro en el caso del VNO se debe

a las temperaturas del verano propicias para la emergencia de *Culex tarsalis*, vector principal en el oeste de EE.UU., y quizá la migración de aves que encuentran en los canales de riego y los traspatios irrigados por inundación [17], uno puede hacer una prueba de hipótesis de la probabilidad de que se agrupen los casos en una fracción, seis, de los períodos observados (n=12 meses), por ejemplo el estadístico de escaneo (*scan statistic* en inglés), que tiene varias versiones, una de ellas basada en la distribución Poisson calcula la probabilidad de que se agrupen los casos en elipses de intervalos [18], y el número esperado es aproximadamente 202 casos observados dividido entre 12 meses o 16.8, como hay dos meses en que se concentran 182 casos, agosto y septiembre la razón de observados sobre los esperados es $O \div E = \frac{182}{16.8 \times 2} = 45.5$. El valor de *P* del estadístico de prueba es menor de 1 en mil.

Series de tiempo

Como adelantamos, una serie de tiempo o serie temporal es una serie consecutiva de observaciones a intervalos fijos y se presentan en la práctica del epidemiólogo de campo en la forma de defunciones generales o por causas específicas, o casos de notificación de cada condición notificable, por intervalos fijos como días, meses, trimestres, semestres, o años. Su análisis es semejante al que se usa en la meteorología o la economía, siguiendo el trabajo pionero de Box y Jenkins [19].

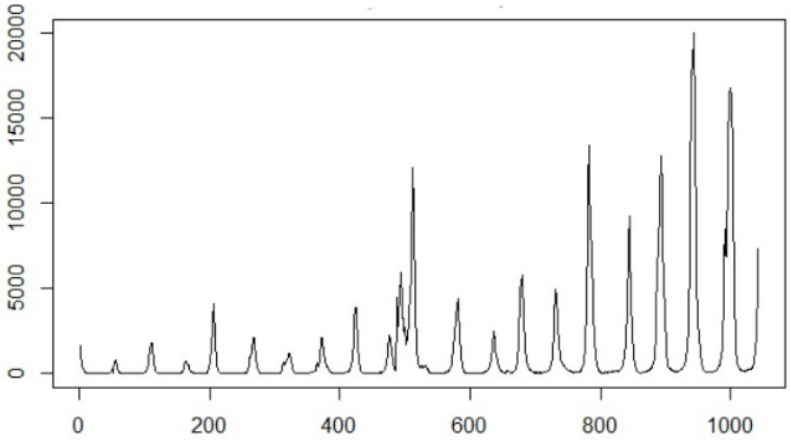
Las series de tiempo permiten identificar estacionalidad, cambios cíclicos o cambios en las tendencias separando cada uno de ellos. Una serie de tiempo puede pensarse que es una combinación de los siguientes componentes: $O_i = T_i \times E_i \times C_i \times I_i$ donde O_i son los eventos esperados, T_i es la tendencia, C_i es el componente cíclico, E_i la estacionalidad, e I_i es el componente irregular o error aleatorio.

El análisis de series de tiempo tiene como fines la estimación para tratar de entender la naturaleza de los cambios y la predicción o pronóstico. Usar varias técnicas siendo la más popular la de modelo autorregresivo integrado de promedios móviles o ARIMA por sus siglas en inglés. En el análisis de series de tiempo se tiene que explorar si la serie es “estacionaria” o no, es decir si la media y la varianza (esto es, el cuadrado de la desviación estándar) son constantes en el tiempo. Estacionario se refiere pues a una característica de los datos no a la estacionalidad. Esta evaluación se hace por inspección de las series removiendo el componente de estaciones, desestacionalizadas y removiendo las tendencias o destendencializadas, y examinando los

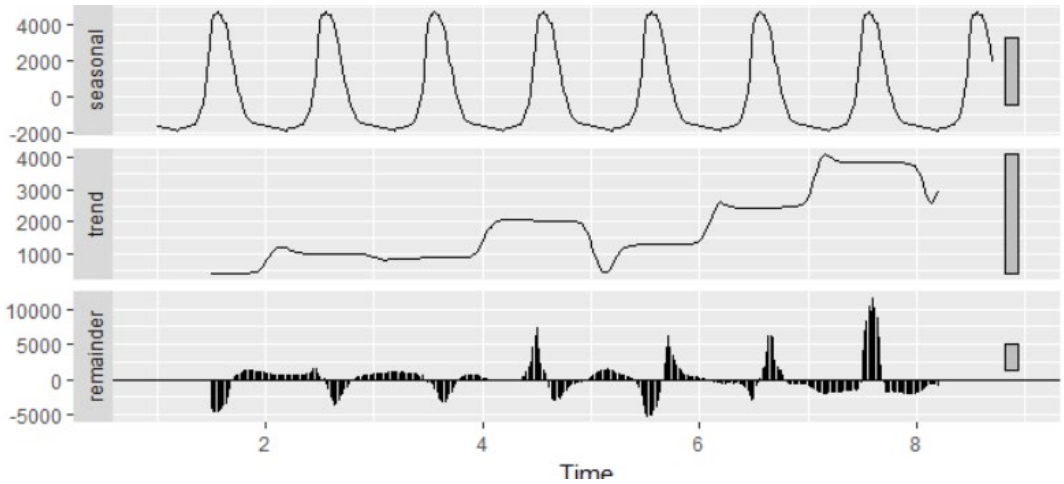
correlogramas con las funciones de autocorrelación y autocorrelación, o una prueba formal con el estadístico de Dickey-Fuller. Si la serie no es “estacionaria”, necesita transformarse para lograr la estimación. El siguiente paso de la regresión ARIMA estriba en identificar el modelo subyacente, lo cual se logra con grafica de las bandas de los límites de confianza de función de autocorrelación para conocer los promedios móviles, mientras que la gráfica de los límites de confianza de autocorrelación parcial nos permite conocer el orden del proceso autorregresivo. Los métodos gráficos de la función de correlación simple y parcial permiten identificar rezagos o retrasos que deben especificarse en los modelos. Estos rezagos deben ser validados, cerciorándose que pasan unas pruebas de bondad de ajuste, antes de obtener los parámetros de regresión finales o hacer un pronóstico con base en el modelo. Por supuesto que estos análisis son mas sofisticados y requieren de la participación de personas preparadas en estadística y con manejo de programas de cómputo. La gráfica 3 muestra el análisis de los datos de muertes por influenza de los Centros para el Control y Prevención de Enfermedades (CDC) de los EE.UU. y claramente distingue los componentes de tendencias, de las alzas estacionales.

Gráfica 3. Descomposición de la serie de tiempo de influenza A en los Estados Unidos de América, 2000-19

A. Serie Original



B. Componentes



Fuente: https://rpubs.com/Nikotino/flu_a_ts Visitado el 22 de febrero, 2024

Claramente se observa en el gráfico 3 que hay un componente estacional y las tendencias, mientras el componente irregular o aleatorio se muestra en la tercera fila. Hay un grupo de entusiastas que ha desarrollado toda clase de programas en el software R para los epidemiólogos de campo, incluyendo el análisis de series de tiempo, que pueden consultar en este sitio: https://epirhandbook.com/es/time-series-and-outbreak-detection.html.

Diseños de grupos múltiples

En este tipo de estudios ecológicos se comparan las medidas de frecuencia como incidencia, riesgo o prevalencia de la condición o factor de riesgo de interés por grupos de individuos según atributos que pueden o no ser homogéneos, digamos al comparar áreas con distintos niveles de ingresos. Según el conocimiento de la epidemiología del problema que uno estudia, el valor de la inferencia puede variar como veremos más adelante. Por ejemplo, uno sabe que, en cuanto a los ingresos por barrio o área, hay una distribución dentro de cada grupo. La o el epidemióloga (o) de campo sabe que el análisis de datos de vigilancia a menudo, aunque este basado en datos de individuos uno finalmente analiza datos ya sea agrupados por edad por municipios, barrios, puede incluso no agruparlos y basar la comparación en datos individuales como la residencia o el trabajo, y puede comparar los números de casos presumiendo que las tasas de enfermedad a través de los distintos

grupos de edad. Incluso podemos agregar los estudios de migrantes y el análisis de tasas por año calendario, período y cohortes de nacimiento. Vamos brevemente a ilustrarlos.

La mayoría de los análisis descriptivos basados en tasas de densidad de incidencia, o llamada simplemente incidencia $tasa = \frac{casos\ nuevos\ en\ un\ período}{tiempo-persona\ en\ el\ periodo} \times k$ o mortalidad $\frac{defunciones\ en\ un\ período}{tiempo-persona\ en\ el\ periodo} \times k$. El tiempo-persona de exposición en la práctica de salud pública se obtiene de los censos o de estimaciones intercensales. Los casos o las defunciones generalmente se suman a lo largo de un año, y el denominador es la población estimada al 30 de junio del año correspondiente. Cuando se obtienen tasas sobre varios años, se deben sumar los casos en tales años y sumar las poblaciones de todos esos años también. mientras que a veces podría usar uno otros registros disponibles como los registros de vacunación para obtener mejores denominadores de las tasas de incidencia o mortalidad, digamos para evaluar la efectividad de la vacuna. Las tasas de morbilidad o mortalidad se pueden comparar utilizando razones o diferencias entre ellas. Por razones de simplicidad y atributos estadísticos, la mayoría de las veces preferimos utilizar las razones, más que las diferencias al comparar tasas. El cuadro 1 reproduce un análisis de las tasas de mortalidad por COVID-19 según la edad en Honduras durante el primer año de la pandemia. Nótese que el número de defunciones entre los mas jóvenes se tuvo que colapsar para tener un grupo referente (0 a 19 años) más estable estadísticamente hablando.

Cuadro 1. Defunciones y tasas de mortalidad por COVID-19 por grupo de edad y sexo, Honduras, 2020- 2021

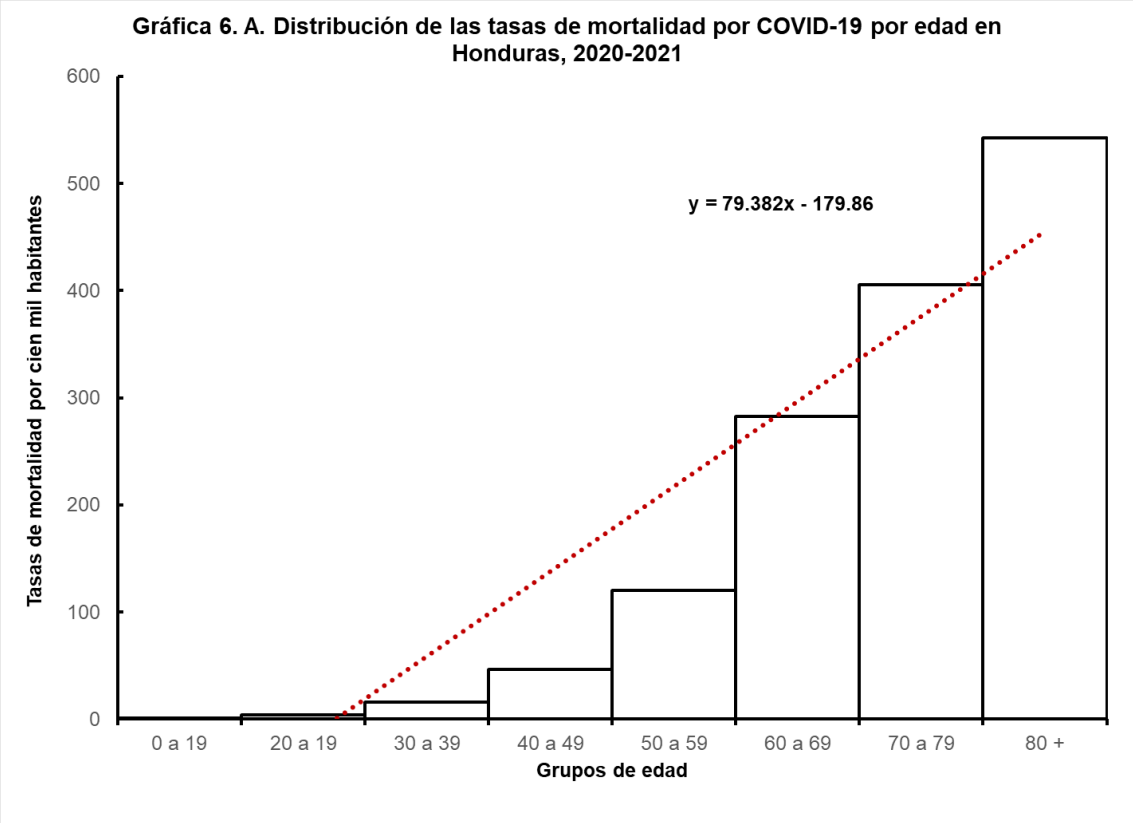
Grupo de edad	Sexo				Total		*RT	IC 95%
	Masculino		Femenino					
	Defunciones	Tasa*	Defunciones	Tasa	Defunciones	Tasa		
0 – 9	5	0.5	18	1.9	23	1.2	1	Referente
10 – 19	7	0.7	9	0.9	16	0.8		
20 – 29	34	4.1	40	4.5	74	4.3	4.3	2.9, 6.3
30 – 39	143	22.1	77	10.8	220	16.2	16.1	11.5, 22.9
40 – 49	283	64.6	152	30.8	435	46.7	46.5	33.5, 64.5
50 – 59	487	163.3	279	82.7	766	120.5	119.9	86.9, 165.4
60 – 69	730	374.6	449	201.5	1,179	282.3	280.8	265.1, 297.4
70 – 79	588	544.9	367	287.4	955	405.4	403.3	292.8, 555.5
80 +	365	698.2	273	417.9	638	542.5	539.7	390.6, 745.7
Total	**2,642**	**58.3**	**1,664**	**34.8**	**4,306**	**46.3**		
RT**	1.7 (IC 95% = 1.6, 1.8)							

*Por 100,000 **RT: razón de tasas de mortalidad; IC: intervalo de confianza
Fuente: [20]

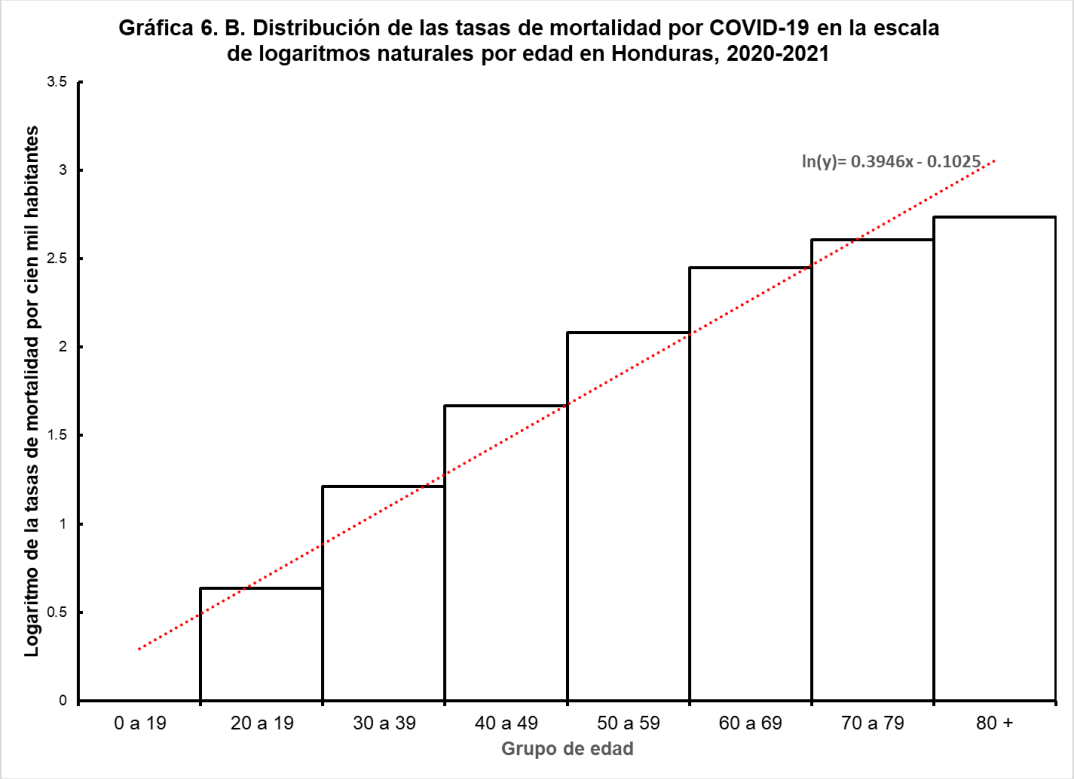
Los datos de Rivera y cols. [20] arriba presentados en el cuadro 1, muestran claramente que las tasas de mortalidad por COVID-19 se cuadruplican por cada 10 años de los 20-29 y 30-39 y luego por 2 a 3 veces por cada incremento de diez años hasta los 70 años, después ya el riesgo no se eleva tanto. Los incrementos son significativamente diferentes con respecto a la mortalidad entre menores de 20 años y entre la mayoría de los grupos etarios respecto de cada uno. El cuadro también muestra que hay un exceso de 70% y estadísticamente significativo de la mortalidad por COVID-19 en varones. En este ejemplo, las epidemiólogas tenían información individual de la edad de cada defunción, pero los datos se tienen que agrupar y utilizar los estimados de población por grupo de edad como denominador para estimar las tasas y razones de tasas. No todos los días se lee de una razón de tasas de 540 veces, cómo se observó en las tasas de mortalidad por COVID-19 entre las personas de uno y otro extremo de la vida. Nótese que hay poco traslape entre los IC 95% de las razones de riesgo de un grupo etario al otro.

Una observación en que un grupo de individuos contribuye con observaciones sobre una exposición (X) y su resultado (Y) puede graficarse en una gráfica de dispersión a los que puede ajustarse una recta de regresión lineal, o logarítmica. Es usual que los coeficientes de correlación (r) y regresión o pendiente (β_1) sean computados simultáneamente. Mientras que la correlación indica el grado y signo de la relación entre las dos variables, y va de -1 a +1, el coeficiente de regresión nos indica la cantidad de cambio en Y por unidad de cambio en X, De ahí que, si se utilizan tasas de incidencia como variable dependiente (Y), y el coeficiente de regresión puede interpretarse como cambio en la tasa por unidad de cambio en la exposición de interés (X), mediante la fórmula $RT = 1 + \frac{\beta_1}{\beta_0}$, en donde *RT* es la razón de tasas, β_1, es la pendiente y β_0 es el intercepto, si se ajusta por regresión lineal. Si se han transformado los datos de las tasas a la escala logarítmica natural (ln), la $RT = e^{\beta_1}$. A menudo la distribución de los datos es diferente a la normal, de manera que la regresión lineal no es adecuada y los datos pueden ser mejor descritos por la regresión exponencial, o polinómica, tal como lo advierten Morgenstern y Wakefield, es mejor usar una transformación logarítmica en la natural (ln) o usar otros métodos referidos en su texto [10]. De hecho, a menudo la regresión lineal puede dar resultados absurdos fuera del rango o implausibles.

Para ilustrar con un ejemplo de la vida real el uso de la regresión para estimar una relación entre una variable de un atributo epidemiológico y una medida de frecuencia de una enfermedad, revisitemos los datos de Rivera y cols. sobre las tasas de mortalidad por edad en Honduras entre marzo del 2020 y marzo del 2021. La gráfica 6 presenta dos paneles, en el primero presentamos la serie original tomada del cuadro 1, mostrado antes y el segundo la distribución en la escala logarítmica natural



Fuente: [20]



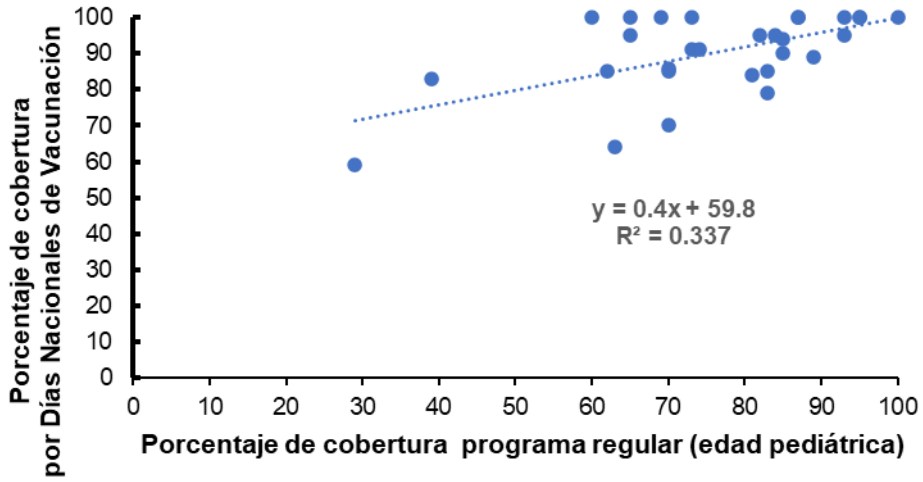
Fuente: [20]

Nótese que la distribución parece que sería mejor descrita por una curva exponencial, y que el ajuste de la regresión lineal da un valor menor de 0 para las tasas, lo cual es implausible, pues el valor menor es 0. Una opción que utilizaré para fines de ilustración es la transformación logarítmica que se muestra en el panel B de la gráfica. El cómputo de la RT es $= e^{0.3946} = 1.5$, y un estimado aproximado del intervalo de confianza del 95% es 1.4 y 1.6. Para su cálculo usamos los datos de las 4,306 defunciones, usamos el estimado del logaritmo de la varianza de RT es $\ln(\text{var})RT = \sqrt{\frac{1}{a} + \frac{1}{b}} = \sqrt{\frac{1}{2{,}153} + \frac{1}{2{,}153}}$, es decir la mitad de las defunciones para los expuestos y la otra para los no expuestos. La razón de tasas de 1.5, agrega la información de que las tasas de mortalidad por COVID-19 aumentaron en promedio 50% por cada década de incremento en la edad de los habitantes Honduras. Obviamente este número subestima la naturaleza del incremento del riesgo por entre las personas de la tercera edad, entre quienes las tasas de mortalidad aumentaron más de 500 veces comparadas con la mortalidad de las personas jóvenes, pero enfatiza el incremento continuo con la edad.

Muchas veces el sólo análisis gráfico de una gráfica de dispersión puede proporcionar más información que un análisis estadístico. En 1986 en México había iniciado el nuevo programa llamado Días Nacionales de

Inmunización, que había sido exitoso en Brasil para vacunar a todos los niños en una sola jornada, dos veces al año, asistiendo a puestos de vacunación disponibles en la calle con un ambiente festivo. Al poco tiempo se notificó de un brote de poliomielitis en Sinaloa, y fuimos enviados por dos diferentes agencias, mi amigo, el Dr. Hugo Vilchis, por la Dirección General de Medicina Preventiva a donde se habían ido los programas de control y yo por la Dirección General de Epidemiología (DGE), donde quedaba la vigilancia. Contamos con el apoyo de las epidemiólogas locales, y así estudiamos lo que resultó ser el último "brote" de poliomielitis paralítica en México [21]. Entre otras actividades, hicimos también una encuesta de cobertura en las áreas afectadas del estado de Sinaloa, usando la metodología de muestreo por conglomerados del Programa Ampliado de Inmunizaciones en que cada conglomerado era un grupo seleccionado al azar de siete menores de cinco años. La gráfica 7 muestra el gráfico de dispersión con la cobertura de los Días Nacionales de Inmunización en el eje de las ordenadas o Y en el eje vertical, y la cobertura por vacunación a la edad pediátrica, es decir antes de la nueva estrategia en el eje de las abscisas o X, en el horizontal.

Gráfica 7. Relación entre las coberturas de dos programas de vacunación contra la poliomielitis en niños de Sinaloa, 1986



Fuente: [21]

Independientemente de que la correlación no era muy fuerte, la sola inspección visual permite encontrar que algunas áreas no mejoraron con la nueva estrategia y que podrían rezagar el progreso en la eliminación de la enfermedad. El Dr. Ciro de Cuadros que asistió a nuestra presentación en la DGE tomó nuestro informe y se lo llevó a su reunión con el vice-secretario, el Dr. Jesús Kumate, quien era de Sinaloa. Al poco tiempo se instaló

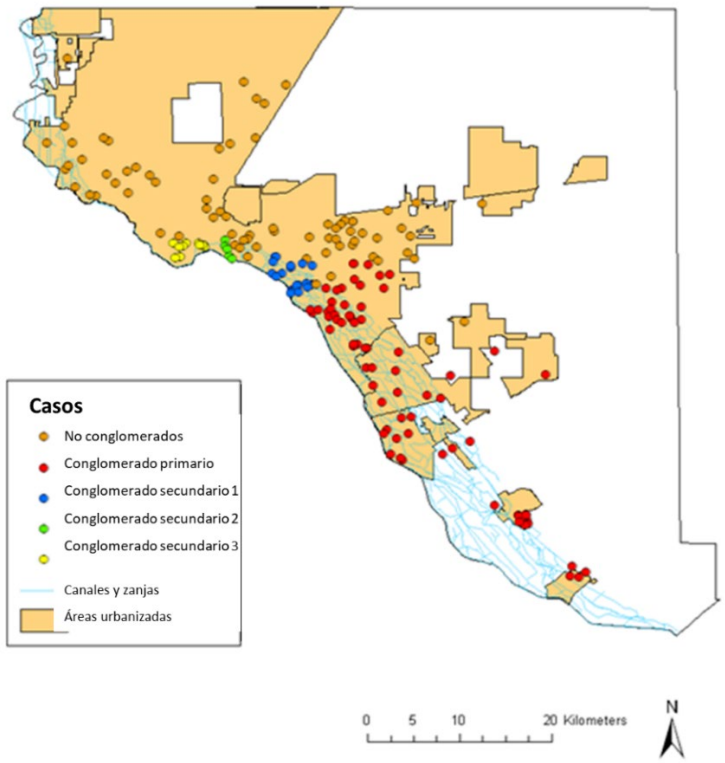
un nuevo programa para ir casa a casa, después de los Días Nacionales de Vacunación a áreas de difícil acceso. Este trabajo remedial se le llamó después operación Sinaloa o “barrido” o *mop-up operation*, en inglés. Esto ilustra una situación en la que la significancia estadística o la falta de una estimación de un efecto expresado en coeficientes de regresión o de razón de tasas no debe impedir el examinar detenidamente los datos.

El análisis espacial de casos y tasas puede ser visto como un análisis descriptivo y probar la ocurrencia de conglomerados de casos o tasas y puede ser considerado un estudio ecológico de grupos múltiples de tipo exploratorio. Hay muchas formas de análisis espacial. Describiré solamente una que utiliza el estadístico de escaneo y basada en la distribución Poisson que estima el número esperados de casos como función de la tasa (esto es, el número de casos dividido sobre la experiencia de tiempo-persona) de cada lugar utilizando una ventana de escaneo y realiza una prueba de permutaciones en relación con todas las ventanas de escaneo del mismo tamaño que sean posibles, vía una simulación Monte-Carlo. En el estudio de la epidemia de VNO en El Paso, Texas, utilizamos tal método para identificar áreas de alto riesgo con base en un mapa de puntos e identificamos conglomerados estadísticamente significativos (P <0.05) uno principal en rojo y otros tres de conglomerados secundarios (azul, verde y amarillo canario), todos ellos cercanos a los canales de riego.

Conglomeración por lugar

El análisis espacial de casos y tasas puede ser visto como un análisis descriptivo y probar la ocurrencia de conglomerados de casos o tasas. Hay muchas formas de análisis espacial. Describiré solamente una que utiliza el estadístico de escaneo y basada en la distribución Poisson que estima el número esperados de casos como función de la tasa (esto es, el número de casos dividido sobre la experiencia de tiempo-persona) de cada lugar utilizando una ventana de escaneo y realiza una prueba de permutaciones en relación con todas las ventanas de escaneo del mismo tamaño que sean posibles, vía una simulación Monte-Carlo [18]. En el estudio de la epidemia de VNO en El Paso, Texas, utilizamos tal método para identificar áreas de alto riesgo con base en un mapa de puntos e identificamos conglomerados estadísticamente significativos (P <0.05) uno principal en rojo y otros tres de conglomerados secundarios (azul, verde y amarillo canario), todos ellos cercanos a los canales de riego.

Gráfico 8. Mapa de puntos mostrando la conglomeración espacial de casos confirmados por laboratorio de enfermedad por Virus del Nilo Occidental por lugar de residencia, El Paso, Texas, 2003-2010



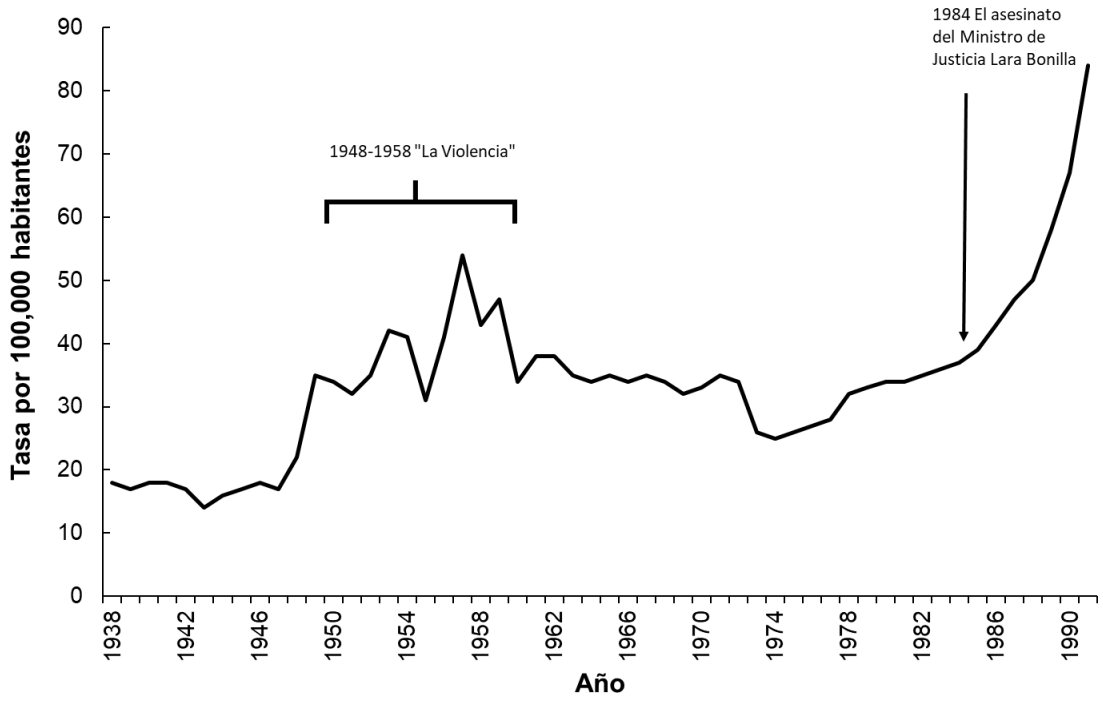
Fuente: [17]

Diseños mixtos

Los datos de mortalidad o morbilidad por edad a lo largo de décadas pueden servirnos para observar qué papel juegan los fenómenos subyacentes durante las circunstancias ambientales asociadas a determinados tiempos, o periodos, independientemente de la edad, o de las experiencias de generaciones de nacimiento o cohortes de nacimiento que reflejan exposiciones ambientales compartidas, también independientemente de la edad o el periodo. Este enfoque fue desarrollado por un epidemiólogo del servicio de salud pública que asumió la dirección del primer departamento académico de epidemiología en el mundo, el Dr. Wade H. Frost, de la Escuela de Salud Pública de la Universidad Johns Hopkins, y que aplicó al estudio de la tuberculosis [22]. Este tipo de estudios, llamados análisis de edad-período-cohorte, así como la modelación de series de tiempo que evalúan el impacto de otras variables son considerados en la categoría de diseños mixtos por Morgenstern y Wakefield [10].

Para ilustrar el uso de este método primero describimos la ocurrencia de homicidios en Colombia entre 1938 y 1991.

Gráfica 9. Tasa de mortalidad por homicidios, Colombia, 1938-1991



Fuente:[23]

En la gráfica 9 se muestran dos olas de homicidios en Colombia en el siglo XX: una a partir de 1948, marcada por el asesinato del líder liberal Jorge Eliécer Gaitán, terminando 10 años después, y la segunda marcada arbitrariamente para efectos didácticos por el asesinato del ministro Rodrigo Lara Bonilla en 1984 y que denominaré la guerra por el tráfico de drogas. Los datos están truncados hasta 1991, ya que eran los mas recientemente disponibles cuando lo publicamos [23].

Cuadro 2. Tasas de homicidios por 100,000 varones en Colombia por grupos de edad en años selectos, 1939-1989

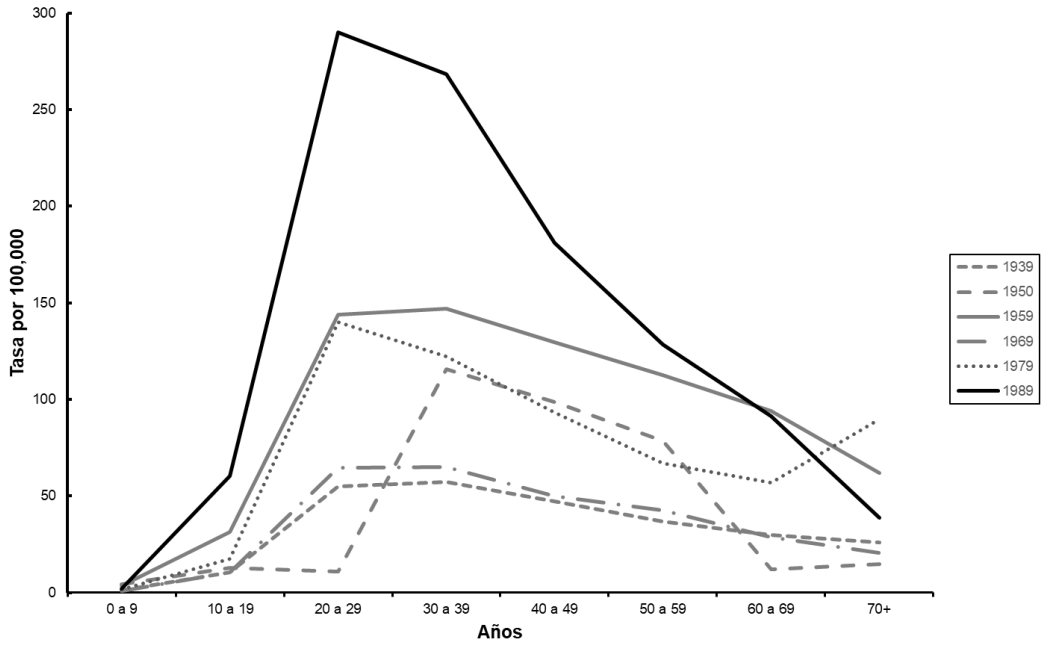
Edad	1939	1950	1959	1969	1979	1989
0 a 9	1.2	4.4	3.7	1.0	1.6	2.0
10 a 19	10.5	13.0	31.4	10.4	17.3	60.4
20 a 29	55	10.9	144	64.6	140	290
30 a 39	57.1	115.5	147	65	122.4	268.2
40 a 49	47.4	98.5	129.4	50	93.1	181.2
50 a 59	36.7	78.4	112.7	42.5	67.0	128.6
60 a 69	29.8	12.1	94.0	28.7	57.0	91.3
70+	26.1	14.9	62.0	20.7	89.6	38.6

Fuente: [23]

El cuadro 2 muestra la distribución de las tasas de homicidios exclusivamente entre varones, porque es entre ellos entre quienes ocurre el 90% o más de los homicidios.

Uno puede graficar de varias formas los datos del cuadro 2 para desplegar las tasas por edad en cada año o digamos década seleccionada, en el gráfico 10 se observa una V invertida o un solo pico que describe la ocurrencia típicamente en hombres jóvenes.

Gráfica 10. Tasas de homicidios entre varones por 100,000 habitantes por décadas, Colombia 1938-1989

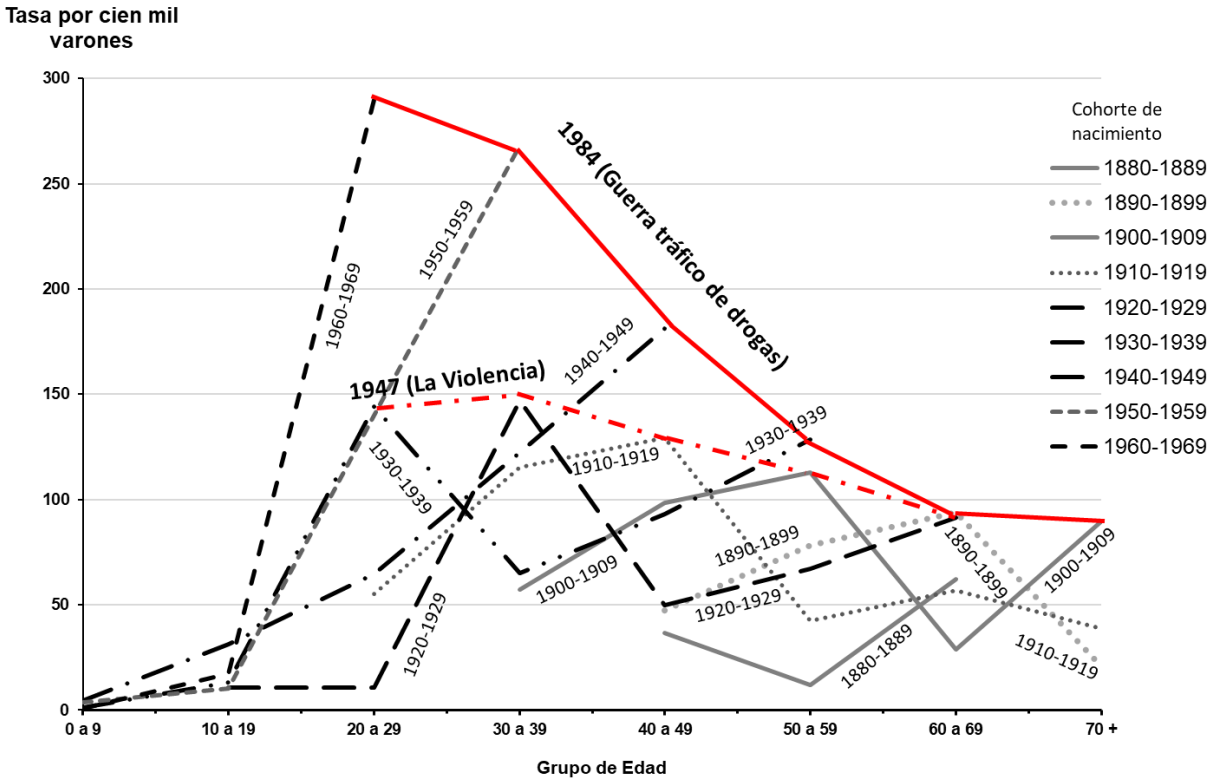


Fuente:[23]

Pero si uno lee en diagonal las celdas del cuadro 2, uno sabe que las personas nacidas en 1910-1919 tendrían 20 a 29 años en 1939, y su tasa fue de 55 por 100,000. Diez años después cuando tenían 30 a 39 su tasa fue de 115, es decir subió, y estuvo a ese nivel en 1959 hacia el final de “La Violencia”, disminuyendo de la sexta década de la vida en adelante. De las 48 celdas en el cuadro 1, excluimos aquellas que no contribuyen con al menos 3 observaciones por cohortes de nacimiento, contribuyendo solamente 42 de las celdas.

La experiencia de las diferentes cohortes y los dos períodos se muestran en la gráfica 11. Claramente las tasas son mas altas para cada cohorte de nacimiento entre los jóvenes. Sin embargo, las tasas de homicidios en hombres mayores aumentan después de los 30 años. Por ejemplo, los que nacieron en 1910-1919, vieron sus tasas de homicidio más altas cuando tenían 33 años durante el inicio de “La Violencia” y de nuevo a partir de 1984 durante la segunda ola vieron sus tasas subir en su séptima década de la vida, mientras que quienes nacieron entre 1920-1929 vieron sus tasas más altas en sus 30’s, aproximadamente en los 1950’s pero sus tasas se elevaron de nuevo en los 1980’s cuando estaban en sus 60’s. Este incremento fue parejo para todas las cohortes de nacimientos anteriores, se observa un crecimiento de sus tasas debido al período.

Gráfica 11. Tasas de homicidios en varones en Colombia por cohorte de nacimiento, 1938-1989



Fuente: [23]

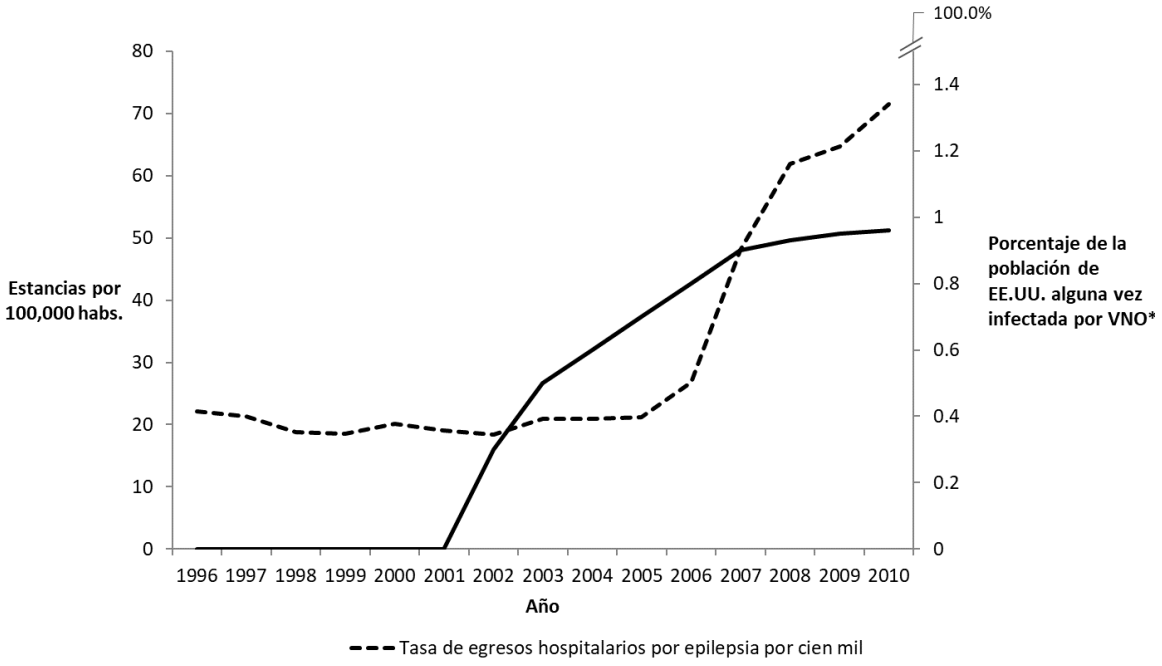
Inferencias y otras limitaciones de los estudios ecológicos

De los hallazgos de estudios epidemiológicos se pueden hacer inferencias a nivel de grupo, y no necesariamente a nivel de individuo. Además, como lo popularizó una crítica del Dr. Mervin Susser, en los 1960s [24], pueden estar sujetos a falacias ecológicas, que explicaba Susser "La falta de discriminación en los niveles de organización al hacer inferencias constituye la falacia ecológica" citando trabajos de sociólogos que ya habían tratado el tema. A su vez es significativo que el mismo Dr. Susser hacia mediados de los 1990s hubiese insistido en el poder de los estudios ecológicos para el estudio de los problemas actuales bajo el nuevo paradigma eco-social de la epidemiología de las "cajas chinas" en sustitución al paradigma de la "caja negra" [25] como acota también el Dr. Borja-Aburto en su revisión de los estudios ecológicos [26]. A su vez se ha demostrado que el enfoque de estudios ecológicos más que los de individuos tienen mayor poder analítico para identificar asociaciones e impactos importantes en la vida real [27]. Teniendo presente que deben entenderse las relaciones postuladas entre las variables de estudio, y cómo se relacionan en las diferentes capas o niveles de un fenómeno en estudio, que la o el epidemiólog(a)o puede seleccionar correctamente lo que debe analizarse. En tiempos en que se amasan gran cantidad de datos hay una propensión a desdeñar el trabajo de terreno y querer hacer todo desde el escritorio sin sofocarse, caminar y ensuciar su ropa y zapatos, sin entrevistar a nadie y solamente usar bases de datos disponibles en el internet.

Los datos disponibles a menudo sufren de idiosincrasias que el investigador debe conocer. Durante la pandemia del COVID-19, por ejemplo, mucha gente dejó de ir a los servicios de salud por temor a infectarse, y a su vez los centros de atención estuvieron cerrados o con disponibilidad parcial o restringida, resultando en un descenso artificial de los datos. En el 2011-2012 observé un incremento en las hospitalizaciones por epilepsia en los EE.UU. que me hizo sospechar de algún agente, y rápidamente pasó por mi cabeza el VNO, pues las encefalitis virales se ha documentado que resultan en epilepsia [28]. El examen minucioso de los datos, con la ayuda de los expertos, mi colega el Dr. Gustavo Román y el Dr. W. Allen Hauser, me permitió percatarme que se trataba de un artificio producido por un cambio en las políticas recomendadas por los neurólogos unos años antes por la que cualquier convulsión pasaba a diagnósticos de epilepsia. La gráfica 12 muestra como el incremento en la tasa de egresos por estancias cortas con epilepsia como diagnóstico principal de egresos y su relación

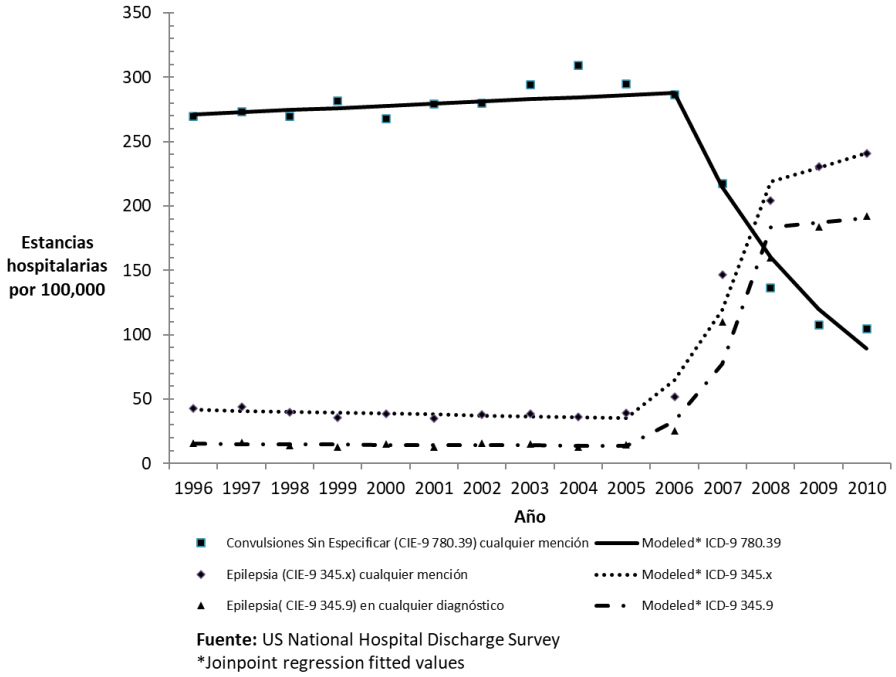
temporal con el incremento de infecciones por VNO. La gráfica 12 muestra los resultados de la indagación y la transferencia de los egresos de una categoría de trastornos convulsivos a la otra que explica el efecto aparente de la epidemia de VNO.

Gráfica 11. Tasas ajustadas por edad de hospitalizacion con epilepsia como diagnóstico principal de egreso y la prevalencia de infección por el VNO, EE.UU., 1996-2010



Fuentes: NHDS Public-USe Data Files* From Planitzer *et al*. estimados 2009-2010 por interpolación

Gráfico 12. Tasas de egresos con cualquier mención de convulsiones y diagnósticos de epilepsia, EE.UU. 1996-2010



Fuente: [28]

Otras limitaciones

Los estudios ecológicos tienen además del posible sesgo ecológico, que puede ser entendido como un caso de confusión, y de las limitaciones de los datos como la que ilustramos, otros problemas potenciales. Entre ellos, el ya mencionado brevemente, de la inadecuada especificación de la relación entre las variables, como ejemplificamos en el uso inadecuado de la regresión linear en la relación entre la edad y las tasas de mortalidad por COVID-19. Los estudios ecológicos requieren que se estudie bien las posibles variables de confusión de la relación estudiada y Morgenstern y Wakefield enuncian cuatro condiciones para que no haya sesgo ecológico. Puede haber problemas de ambigüedad temporal en series de tiempo si no se especifica bien el rezago que uno esperaría entre una exposición y el efecto en la población. Uno puede pensar en métodos que agreguen diferentes niveles de grupos e individuos para tratar de entender los niveles de organización y hacer inferencias que sean correctas.

Agradecimientos

El autor agradece a Fernanda Bruzadelli y Dora Rafaela Ramírez por su crítica a una versión previa de este manuscrito.

Referencias

1. Fox JP, Hall CE, Evelback LR. *Epidemiología. El hombre y la enfermedad*. México: La Prensa Médica Mexicana. 1975.
2. Lilienfeld AM, Lilienfeld DE. *Fundamentos de Epidemiología*. México: Fondo Educativo Interamericano. 1983.
3. Kahn HA, Sempos CT. *Statistical Methods in Epidemiology.* New York: Oxford University Press, 1989.
4. Hill AB. Observation and Experiment. *NEJM* 1953; 248 (24): 3-9. doi: 10.1056/NEJM195306112482401.
5. Rothman KJ. *Modern Epidemiology*. 1st edition. Boston: Little, Brown & Co. 1986.
6. Wise L, Hartge P. Chapter 10. Field Methods. In Lash T, VanderWeele TJ, Haneuse S, Rothman KJ. *Modern Epidemiology*. 4th edition. Philadelphia: Wolters Kluwer. 2021.
7. Rothman KJ, Lash T, Haneuse S, VanderWeele TJ. Chapter 1. The Scope of Epidemiology. In Lash T, VanderWeele TJ, Haneuse S, Rothman KJ. *Modern Epidemiology*. 4th edition. Philadelphia: Wolters Kluwer. 2021.
8. Rothman KJ, Lash T. Chapter 15. Precision and Study Size. En Lash T, VanderWeele TJ, Haneuse S, Rothman KJ. *Modern Epidemiology*. 4th edition. Philadelphia: Wolters Kluwer. 2021.
9. VanderWeele TJ. Chapter 12. Confounding and Confounders. En Lash T, VanderWeele TJ, Haneuse S, Rothman KJ. *Modern Epidemiology*. 4th edition. Philadelphia: Wolters Kluwer. 2021.

10. Morgenstern H, Wakefield J. Chapter 30. Ecologic studies and Analysis. In Lash T, VanderWeele TJ, Haneuse S, Rothman KJ. *Modern Epidemiology*. 4th edition. Philadelphia: Wolters Kluwer. 2021.
11. Tello Anchuela O, Amela Heras C, Pachón del Amo I, Martínez Navarro JF. Capitulo 24. Vigilancia de la Salud Pública. En Martinez Navarro JF, Antó JM, Castellanos PL, Gii M, Marset P, Navarro V. *Salud Pública*. Madrid: McGraw Hill-Interamericana. 1998.
12. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine* 2000; 19:335-351: (correction: 2001;20:655). doi: 10.1002/(sici)1097-0258(20000215)19:3<335::aid-sim336>3.0.co;2-z.
13. Knox EG. Chapter 7, Volume 2. Spatial and temporal studies in epidemiology. In Detels R, Holland WW, McEwen J, Omenn GS. *Oxford Textbook of Public Health*. New York: Oxford University Press, 1997.
14. Cardenas VM, Paternina-Caicedo AJ, Salvatierra EB. Underreporting of Fatal Congenital Zika Syndrome, Mexico, 2016-2017. *Emerg Infect Dis.* 2019 Aug;25(8):1560-1562. doi: 10.3201/eid2508.190106.
15. Morgenstern H. Uses of ecologic analysis in epidemiologic research. *Am J Public Health.* 1982 Dec;72(12):1336-44. doi: 10.2105/ajph.72.12.1336.
16. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int J Epidemiol.* 2017;46(1):348-355. doi: 10.1093/ije/dyw098.
17. Cardenas VM, Jaime J, Ford PB, Gonzalez FJ, Carrillo I, Gallegos JE, Watts DM. Yard flooding by irrigation canals increased the risk of West Nile disease in El Paso, Texas. *Ann Epidemiol.* 2011;21(12):922-9. doi: 10.1016/j.annepidem.2011.08.001.
18. Kulldorff M. and Information Management Services, Inc. SaTScan™ v10.0: Software for the spatial and space-time scan statistics. [www.satscan.org], 2021.

19. Box GEP, Jenkins GM. *Time series analysis*: *forecasting and control*. San Francisco: Holden-Day. 1970.
20. Rivera AC, Mendizábal Solé, R, Enamorado JA. Epidemiologic characteristics of the COVID-19 pandemic during its first year in Honduras. *American Journal of Field Epidemiology* 2023; 1(2): 11–21. doi.org/10.59273/ajfe.v1i2.7383.
21. Cárdenas-Ayala VM, Vilchis-Licón H, Stetler HC, Cabrera-Coello L, Koopman JS, Valdespino-Gómez JL, Ruíz-Matus C, Vega-Ramos RG, Muro-Amador M. Risk factors for the persistence of wild poliovirus transmission in Sinaloa, Mexico, 1984-1986. *Bull Pan Am Health Organ*. 1988;22(3):227-39. PMID: 2852044.
22. Frost WH. The age selection of mortality from tuberculosis in successive decades. 1939. *Am J Epidemiol*. 1995 Jan 1;141(1):4-9. doi: 10.1093/oxfordjournals.aje.a117343.
23. Lehtonen S, Suárez G, Morales A, Sanchez C, Cardenas VM. Homicidios en Colombia, 1938-1983. Instituto Nacional (Colombia) de Salud. *Boletín Epidemiológico*. 1994; 2 (4): 58-62.
24. Susser M. *Conceptos y estrategias en epidemiología: el pensamiento causal en las ciencias de salud.* México: Fondo de Cultura Económica. 1991.
25. Susser M, Susser E. Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology. *Am J Public Health.* 1996 May;86(5):674-7. doi: 10.2105/ajph.86.5.674.
26. Borja-Aburto VH. Estudios ecológicos. *Salud Publica Mex.* 2000;42(6):533-8. PMID: 11201582.
27. Koopman JS, Longini IM Jr. The ecological effects of individual exposures and nonlinear disease dynamics in populations. *Am J Public Health*. 1994 May;84(5):836-42. doi: 10.2105/ajph.84.5.836.
28. Cárdenas VM, Román GC, Pérez A, Hauser WA. Why U.S. epilepsy hospital stays rose in 2006. *Epilepsia.* 2014 Sep;55(9):1347-54. doi: 10.1111/epi.12719.