EIDON, nº 50
diciembre 2018, 50: 35-47
DOI: 10.13184/eidon.50.2018.35-47

Ética del valor p: una oportunidad para nuevas metodologías

P-value Ethics. An Opportunity for New Methodologies

Guillermo Droppelmann [1, 2, 3], Rolf Nickel [1], Felipe Feijoo [4]

1. **Área de Investigación, Unidad Académica, Clínica MEDS, Santiago, Chile.**
2. **Facultad de Ciencias, Universidad Mayor, Santiago, Chile.**
3. **Programa de Doctorado en Ciencias mención Estadística Avanzada, Universidad Católica San Antonio de Murcia, Murcia, España.**
4. **Escuela de Ingeniería Industrial, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.**

Email: guillermo.droppelmann@meds.cl

Resumen

La significación estadística ha sido el gran protagonista en la comunicación de resultados en las investigaciones biomédicas. Sin embargo, las ciencias computacionales han desarrollado nuevas metodologías. Se describen las vías por las que la significancia estadística podría ser reemplazada, impactando en la ética de la validez científica. Resulta necesario conocer las limitaciones de las herramientas estadísticas tradicionales e incorporar los nuevos elementos para el análisis de datos.

Palabras clave: Ética, Metodología, Significación estadística, Valor-p.

Abstract

Statistical significance has been the great protagonist in the communication of results in biomedical research. However, computational sciences have developed new methodologies. It is described that statistical significance could be replaced by impacting the ethics of scientific validity. It is necessary to know the limitations of traditional statistical tools and incorporate the new elements for data analysis.

Keywords: Ethics, Methodology, Statistical significance, P-value.

1. Introducción

Diversas guías, códigos, declaraciones, normas e instancias desarrolladas por organismos internacionales han propuestos los principios éticos para la realización de investigación biomédica en seres humanos (World Medical Association, 2013; CIOMS, 2002; UNESCO, 2000). Cualquier propuesta científica que involucre a personas requerirá necesariamente de una exhaustiva evaluación de los aspectos legales, éticos y metodológicos por parte de los integrantes de un Comité de Ética de la Investigación (CEI) (Rodríguez, 2002), a fin de asegurar el valor científico del estudio, único modo de lograr un aporte sustancial al conocimiento (Council for International Organizations of Medical Sciences, 2016).

En la actualidad existen diversas alternativas para evaluar la factibilidad ética de un proyecto de investigación en seres humanos. Uno de los documentos que se ha caracterizado por su utilidad práctica es el desarrollado por Ezequiel Emanuel (Emanuel, 2000), quien a través de siete requisitos aborda los principales temas que a su entender deben ser considerados para que una investigación posea viabilidad desde el prisma de la ética. Dentro de lo que propone, el punto que representa mayor alcance metodológico es la denominada validez científica, la cual constituye un requisito fundamental para alcanzar la excelencia.

También se han propuesto una serie de interminables aspectos metodológicos que debieran ser considerados en el momento de evaluar un proyecto de investigación (Droppelmann, 2018), entre ellos la definición de los objetivos, el planteamiento de la hipótesis, la descripción de los procedimientos, la descripción del universo y la obtención de la muestra, las técnicas, los procedimientos, las formas de registro y las pruebas estadísticas utilizadas para el análisis de los datos (González, 2010).

Desde el punto de vista estadístico, el concepto de significación estadística adquiere un particular interés por parte del grupo de investigadores y clínicos, ya que este valor permite determinar objetivamente si la intervención de interés a través del grupo experimental (sustancia activa, nuevo tratamiento) presenta un cambio sustancial al compararlo con el grupo control (sustancia inactiva, tratamiento estándar). Además de la importancia que tiene en sí de este valor, es preciso tener en cuenta el tamaño de la muestra, la potencia seleccionada y el tamaño del efecto, ya que si un estudio tiene una baja potencia, es decir, si presenta una probabilidad menor de un 80% de producir un valor de p menor a 0,05 con un escaso tamaño de efecto, resultaría poco ético que los participantes aceptaran algún tipo de riesgo, además de las incomodidades propias de participar en un estudio (Bacchetti et al., 2005). Cumplir con un valor de p menor a lo establecido por la comunidad de expertos no es suficiente para demostrar que la hipótesis que se encuentra bajo estudio tiene validez científica.

> La validez científica constituye un aspecto fundamental para alcanzar la excelencia.

Por otra parte, las nuevas tecnologías de la información, los nuevos avances metodológicos provenientes de las ciencias matemáticas y computacionales y el uso de herramientas procedentes de la inteligencia artificial para el tratamiento de los datos crecen día a día de forma exponencial, siendo ampliamente utilizados en diferentes áreas (Kourou et al., 2015; Tang et al., 2009; Weng et al., 2017), y transformando la forma tradicional de analizar la información, de reportar los

resultados, lo que a su vez obliga a replantear algunos capítulos de la ética de investigación.

El presente artículo tiene como objetivo analizar cómo el uso de la significación estadística puede verse reemplazado por otros métodos de análisis de datos, lo que tiene consecuencias no solo metodológicas sino también éticas.

2. Significación estadística

Para comprender el concepto de significación estadística (valor p), resulta necesario introducir el término de docimasia de hipótesis, ya que el análisis de estudios tanto clínicos como epidemiológicos presenta como propósito la comparación entre diferentes grupos, intervenciones, aplicaciones o elementos de interés que se ven sometidos a un determinado procedimiento (Aedo, 2009), debiendo contrastar estadísticamente dos supuestos para comprobar la veracidad de la hipótesis planteada. Se considerará hipótesis estadística una afirmación que contenga uno o más parámetros que cumplan con las características de la distribución de probabilidad (Aedo, 2008).

Como el investigador se encuentra inicialmente en un escenario de incertidumbre y debe probar siempre su hipótesis, debe controlar al máximo la probabilidad de cometer algún tipo de error estadístico. Tal es el origen del error tipo I, la probabilidad de rechazar la hipótesis nula (H0) siendo esta verdadera, y el error tipo II, que se produce al no rechazar H0 siendo esta falsa.

Para poder contrastar una hipótesis debe tenerse presente que H0 será verdadera mientras que la evidencia contrastada no demuestre lo contrario. Cuando esta demostración no se produce, estamos ante el error tipo I el cual trata siempre de minimizarse al máximo. En este preciso momento es donde surge alfa (α), más conocida como significación de la dócima, que condiciona la veracidad de H0 cuando el resultado es menor que α, siendo el valor de “p” el tamaño del error (Cavada, 2009).

Para cuantificar la fuerza de la evidencia que se está construyendo a partir de las hipótesis de investigación o hipótesis alternativa (H1) en contra de la H0, se ha determinado por consenso que el valor de p debe ser menor a un 5% de significación, es decir “p<0,05”, como un nivel estándar de que no existe evidencia en contra de H1 (Fisher, 1950). Este valor se ha aceptado ampliamente en la comunidad biomédica y científica como referencia para determinar la significación estadística de las investigaciones. El establecimiento de un punto de corte para determinar la existencia de significación en las hipótesis es un asunto controvertido, ya que el valor p es influenciado por múltiples factores (Dahiru, 2008).

3. Consideración ética del valor p

La construcción del conocimiento biomédico y de ciencias de la salud se ha basado en resultados que se expresan a través de valores p (Dorey, 2010), como evidencia científica que condiciona o determina las conductas terapéuticas, pese a que el valor p aún resulta poco entendido por parte de los mismos profesionales (Sedgwick, 2014).

Esta construcción del saber resulta cuando menos arbitraria, al aceptar por consenso la significación estadística (valor de p) como un valor predeterminado (Hoekstra et al., 2006), asumiendo el rechazo de la hipótesis nula si la cifra es menor a un 5% (0,05), y otorgándole el rótulo de no significativo si los valores son superiores. El test de significación cobra así un carácter casi mágico, que algún autor ha calificado de "evidencitis" (Chia, 1997), que puede condicionar la significación clínica del estudio, sobre todo en aquellos investigadores poco experimentados, ávidos de encontrar valores que demuestren alta significación estadística.

> Las nuevas tecnologías y herramientas provenientes de la ciencia de datos han transformado la forma tradicional de analizar la información y reportar los resultados en el área de biomedicina.

La relevancia de esta última va más allá de la realización de cálculos y la obtención de valores, ya que centra su atención en la magnitud del problema investigado, las diferencias obtenidas, la vulnerabilidad, y por supuesto los costos involucrados (Manterola y Pineda, 2008).

Se han planteado diversos cuestionamientos e interrogantes con respecto al valor que se debe tener en cuenta para demostrar la significación de las investigaciones (Rothman, 1986). En primer lugar, ¿por qué los investigadores debieran de considerar el valor de 0,05 como significación estadística y no otro valor como el 0,04 o el 0,06? Seguramente esto se debe a la amplia aceptación del valor por parte de la comunidad científica, de modo que si se escoge otro valor, este debe ser justificado por el grupo de investigadores debido a las características de la investigación. En segundo lugar, ¿por qué debe dicotomizarse el valor de p?, es también, ¿presenta o no presenta significación el punto de corte? Esto último es importante, porque la dicotomización siempre restringe información en un sistema continuo, lo que podría facilitar la manipulación *post hoc*. En tercer lugar, el valor de p trae consigo cierta vulnerabilidad implícita, debido a que presenta grandes variaciones según la potencia estadística preestablecida, el tamaño de la muestra, el tamaño del efecto y la variabilidad de los datos obtenidos (Kirby, 2002), pudiendo el investigador experimentado manipular estas consideraciones con el propósito de demostrar estadísticamente la validez de su hipótesis, a fin de obtener diferencias donde no las había (Droppelmann, 2016).

Es así como el uso del valor de p como referente de validez científica ha sido reportado hace años como una mala práctica de investigación (Berkson, 1942), justamente debido a su alta vulnerabilidad, siendo ampliamente cuestionado por diferentes autores (Gelman y Loken, 2014; Gigerenzer y Marewski, 2015), alguno de los cuales ha llegado a hablar de "falacia" (Goodman, 1999). A pesar de las criticas existentes, el concepto de valor p sigue siendo ampliamente utilizado por la comunidad de clínicos, investigadores e incluso por los mismos estadísticos (LeCoutre, Poitevineau y Lecoutre, 2003). Esto trae consigo una pobre contribución a sus respectivas disciplinas, ya que podrían aceptarse hipótesis de investigación únicamente porque se ajustan a un valor determinado y no porque la interrogante que se quiera develar posea valor en sí misma.

La significación estadística como un referente en investigación biomédica se considera una de las herramientas estadísticas más abusadas en la actualidad (The Editors, 2001). Por este motivo, múltiples grupos editoriales de gran renombre y pertenecientes

a importantes revistas científicas desaconsejan su uso desde hace varios años (Gardner y Altman, 1986), lo que no impide que su utilización siga vigente.

Debe tenerse claro en todo momento que el valor p exige una mirada reflexiva. Esto explica la importancia que progresivamente va adquiriendo el intervalo de confianza en vez del valor p (Goodman y Berlin, 1994). La sobrevaloración del valor p perdería importancia si el grupo de investigadores presentara una actitud crítica y reflexiva ante sus resultados, viendo si se responde, a través del uso de métodos estadísticos específicos, a las necesidades reales que se pretenden responder con la investigación (Mendoza, 2006).

Es así como de los 7 requisitos éticos de la investigación en seres humanos propuestos por Ezequiel Emanuel, cobra relevancia el principio número 2, denominado validez científica, según el cual los profesionales que investigan deben de poseer las competencias necesarias para poder ejecutarla, además del lugar y las condiciones donde para que la realización del estudio sea posible, que la recolección de datos y registros sea correcta, que los resultados correspondan a lo que realmente se quiso investigar, etc. Ni que decir tiene que el estudio, además de ser original, ha de poseer significancia científica (Emanuel, 1999).

En este último punto, debe considerarse que la significación científica vaya más allá de la mera significación estadística, lo que a todas luces podría resultar confuso si el investigador o el clínico con interés en investigar centrara el propósito de su intervención en la obtención de resultados más que en producir un impacto en la mejora de la calidad de vida de las personas. Debe tenerse claro que la clasificación de los resultados desde la estadística en “significativa” y “no significativa” puede traer consigo una mala interpretación por parte del lector, llevándole a evaluar erróneamente los resultados de una investigación (Greenland et al., 2016).

> El valor de <0,05 se ha estandarizado ampliamente al interior de la comunidad biomédica y científica como referencia para determinar la significancia estadística de las investigaciones.

Al resultar el valor p tan familiar y de gran relevancia en el ambiente clínico para describir los resultados obtenidos en una investigación (Oakes, 1986), la ética de la investigación debiera de promover un marco que considere un uso crítico y reflexivo del valor p, para lo cual deben considerarse los siguientes puntos (Hoover y Ziegler, 2008; Mayo y Spanos, 2006; Oakes, 1986; Royall, 1997):

- Debe colocarse particular atención en el uso e interpretación del valor p.
- El valor p debe ir siempre acompañado de intervalos de confianza, para minimizar la incertidumbre.
- La significación estadística no es necesaria ni suficiente para determinar la importancia científica o práctica de un conjunto de observaciones.
- Debe incentivarse el uso de otras medidas de resultados, cuando la naturaleza de los estudios sea clínica.
- Es preciso reportar los sesgos y su manejo.
- Atender a la significación estadística sin hipótesis alternativas explícitas puede producir un problema de inferencia.
- Encontrar muchísima evidencia que rechaza la H0 puede significar que existe, más que un cambio, un gran error.

- No debieran usarse pruebas de significación sin considerar las posibles hipótesis alternativas.
- Cualquier método estadístico es susceptible de malas interpretaciones.
- En la comunicación de resultados y conclusiones intervienen factores subjetivos, dependientes del interés de los investigadores.
- Importancia de aplicar herramientas estadísticas que permitan exponer los resultados de una investigación sin el uso del valor p.

El problema lo plantea el que aún se considere por parte de la comunidad de investigadores que el valor p es un elemento determinante en el proceso de construcción del conocimiento, existiendo una creencia generalizada de que los estudios carecen de ética cuando ciertos factores influyen en el valor de la significación estadística, tales como poseer escasos tamaños muestrales o bajo poder estadístico (Bacchetti, 2005). El reportar diferencias significativas cuando no lo son, debe considerarse una práctica inmoral o poco ética (Marco y Larkin, 2000). Nuestro objetivo no es negar la importancia de p, sino precisar su significado a fin de que pueda seguir siendo importante en la construcción del conocimiento.

4. Nuevas metodologías que no consideran el valor p

Es fundamental reconocer la relevancia de los aspectos metodológicos de los proyectos de investigación, ya que sustentan su validez científica, de modo que tanto los sujetos que realizan un proyecto como quienes los evalúan, deben estar preparados para los nuevos desafíos metodológicos que están surgiendo debido al exponencial desarrollo de las ciencias biomédicas (Penn State Statistics, 2018).

Hasta aquí nos hemos ocupado de las consideraciones que sobrevaloran la significación estadística del valor p en las investigaciones, sesgando el análisis, debido a que los datos poseen una cierta distribución frecuentista y a que pueden ser sometidos a contrastes de hipótesis que permiten la obtención de resultados. Sin embargo, resulta interesante exponer que la ética de la investigación, al abordar el ítem de validez científica, debe conocer las nuevas formas metodológicas que están sustentando el nuevo conocimiento sin el uso de este controversial indicador estadístico, p. A continuación, se presentan dos metodologías utilizadas en la actualidad en las ciencias biomédicas que entregan resultados válidos, pero sin el uso del valor p para el análisis de datos.

> La significación estadística trae consigo cierta vulnerabilidad implícita debido a que presenta grandes variaciones debido a que es susceptible a múltiples factores, desaconsejando su uso por grupos editoriales.

4.1. Métodos bayesianos

La presencia de modelos alternativos a los tradicionalmente utilizados en investigación clínica es ya una realidad, debido al paradigma bayesiano. Este se utiliza desde hace algunos años en temas relacionados con ensayos clínicos, pero es tan poco común que aún se considera una metodología nueva, al menos si se compara con la estadística clásica basada en el valor p. Su uso ha ido en aumento debido a que los modelos frecuentistas presentan diversas limitaciones (Austin, Brunner y Hux, 2002).

Los modelos bayesianos utilizan ciertos aspectos conocidos por el método científico tradicional, como el planteamiento de una hipótesis tras la recolección de evidencia. De ese modo, la hipótesis presenta un carácter flexible, ya que se basa en una distribución inicial que va variando a medida que aumenta el conocimiento, siendo un planteamiento diametralmente opuesto al habitual en los investigadores cuando concentran sus esfuerzos en demostrar una hipótesis de trabajo preconcebida. Tal es la razón de que se haya convertido rápidamente en una herramienta de la estadística inferencial capaz de ser utilizada en diversos campos científicos (Ghosh, 2010).

> Los investigadores y revisores de proyectos deben estar preparados para evaluar e incorporar nuevas metodológicas.

Estos métodos permiten realizar diferentes tipos de actividades que el valor p solo permite plantear como potenciales hipótesis, tales como predicciones, agrupando la evidencia de múltiples fuentes y diseñando estudios con el propósito de encontrar las mejores soluciones posibles a las interrogantes planteadas por los investigadores (Spiegelhalter, 2000). Estos diseños ya están siendo utilizados por investigadores provenientes de las ciencias biomédicas (Louis, 2005).

Lo interesante de este tipo de métodos es que no utilizan al valor p como un elemento determinante del tipo de resultados que entregarán, ya que el factor de Bayes compara básicamente el valor relativo dado por dos hipótesis construidas a través de los datos, en contraste con el valor p, que se calcula en función de la hipótesis nula. Esta propiedad diferencial le permite producir una evaluación diferente de la fuerza de la evidencia, por lo que la discusión de la importancia de los resultados radica principalmente en el grupo de expertos que finalmente toman las decisiones, pudiendo facilitar la incorporación de sus aportes de forma importante (Goodman, 2005), fomentando un proceso reflexivo en la construcción del conocimiento científico.

4.2. Aprendizaje automático (*machine learning*)

Existen nuevos métodos provenientes de las ciencias computacionales que abordan de forma distinta las problemáticas relacionadas con el área de la salud, ya que utilizan herramientas, métodos y estrategias que necesariamente impactarán en la forma tradicional de hacer investigación, incentivando un cambio en las metodologías actuales, ya que la incorporación de nuevas tecnologías permitirá acceder a otras formas en la construcción conocimiento distintas a las del valor p.

Estas nuevas estructuras científicas que colaboran en el avance de la frontera del conocimiento utilizan ciertos elementos de las estadísticas convencionales, pero se caracterizan por el uso de modelos matemáticos y computacionales que aprenden de los datos ya existentes (Koohy, 2017). Múltiples son las áreas provenientes de la salud que recientemente ya las han incorporado, existiendo evidencia reportada en cáncer (Gründne et al., 2018; McDonald, 2018; Wang et al., 2018; Wong y Yip, 2018), aparato respiratorio (Nishio et al., 2018), dermatología (Li y Shen, 2018), endocrinología (Sollini et al., 2018) y patología (Bejnordi, Litjens y van der Laak, 2018; van Smeden, Van Calster y Groenwold, 2018), entre otras.

Destacan, principalmente, dos modelos de aprendizaje, uno supervisado y otro no supervisado. El primero permite predecir ciertos valores de variables dependientes a partir de variables independientes. Destaca por su poder de clasificación y predicción continua, incluso mayor que el de las regresiones logísticas, los modelos multivariados, los métodos de riesgo como Hazard, o los modelos proporcionales de Cox. El aprendizaje no supervisado, por su parte, no busca distinguir las variables sino que trata de encontrar una estructura que explique de la mejor forma los datos. Se caracteriza por no presentar resultados predictivos, sino buscar patrones o diferentes agrupaciones que surjan naturalmente al interior de los datos (González, 2015; Rahul, 2015).

Los modelos bayesianos y de machine learning utilizan ciertos aspectos conocidos por el método científico tradicional. Sin embargo, no utilizan el valor p como un elemento para el análisis de resultados.

Existen otros tipos de modelos más complejos, también provenientes del aprendizaje automático, tales como el aprendizaje profundo *(deep learning)*, consistentes en algoritmos complejos que son capaces de combinar diversas características, lo que les hace particularmente apropiados para el análisis de datos provenientes de disciplinas biomédicas, debido a su complejidad (Ching, Himmelstein y Beaulieu-Jones, 2018), prometiendo un cambio definitivo en la forma de analizar la información.

Lo anterior evidencia que la estadística convencional presenta algunas limitaciones cuando se busca desarrollar experimentos complejos donde cada sujeto contribuye con un sinnúmero de observaciones diferentes, como sucede al comparar el fenotipo de dos tejidos, donde el uso de estas nuevas técnicas resulta muy ventajoso respecto a las pruebas estadísticas tradicionales (Bzdok, Altman y Krzywinski, 2018). Lo cual demuestra, una vez más, que el uso del valor p es un recurso estadístico muy importante pero que puede quedar desplazado debido a una evolución natural de la metodología de la investigación.

5. Reflexiones finales

Actualmente sigue aún utilizándose el uso indiscriminado del valor p como factor determinante en la comunicación de resultados y, consecuentemente, en la construcción del conocimiento en las ciencias biomédicas. Un número importante de profesionales provenientes del área de la salud lo sigue utilizando y sobrevalorando, a pesar de su carácter poco sensible y altamente dependiente de la interpretación del lector.

Dentro de las consideraciones metodológicas que se evalúan al momento de ejecutar un protocolo de investigación, debería promoverse el uso de elementos adicionales respecto a los estadísticos clásicos. Debe considerarse que el cálculo del valor de p se fundamenta en la suposición (condición) de que la hipótesis nula (H0) es verdadera, lo que significa que un investigador no puede inferir el valor de p sí la hipótesis nula es verdadera o falsa. Al aceptar la H0, el valor que arroja p es nada más que la probabilidad de que el experimento realizado demuestre o no diferencias entre las muestras observadas (van Raaij et al., 2010). Sin embargo, las nuevas estrategias metodológicas provenientes de disciplinas matemáticas y computacionales entregan nuevas formas de analizar resultados, por lo que han sido incorporadas en disciplinas

pertenecientes a las ciencias biomédicas. Es así como el aprendizaje automático plantea nuevas formas de análisis de una investigación, haciendo necesaria la homogenización de criterios a través del consenso de expertos, ya que repercutirá en la forma de aceptar la validez científica de los estudios.

Debe quedar claro que, con independencia del uso de diferentes modelos metodológicos para el análisis de resultados, estos deberán utilizarse siempre y cuando creamos que captan la esencia del problema en estudio y que son capaces de contribuir a la construcción del conocimiento y el subsecuente mejoramiento de la calidad de la salud de las personas.

Es importante destacar que, pese a que el valor p aún juega un rol importantísimo en la demostración de resultados, tanto los clínicos como los evaluadores de proyectos de investigación deben estar al tanto de sus limitaciones y conocer los nuevos modelos metodológicos que se están incorporando aceleradamente. Debemos estar preparados para poder evaluar correctamente las fronteras de estas nuevas metodologías, lo que exige un entrenamiento y perfeccionamiento permanente de aquellos grupos interesados en ejecutar y revisar proyectos de investigación.

Bibliografía

Aedo, S. (2008). La bioestadística: una ciencia que reduce el azar en obstetricia y ginecología. *Revista de Obstetricia y Ginecología Hospital Dr. Luis Tisné Brousse*, 3(3), 241-2.

Aedo, S. (2009). Docimasia de Hipótesis. *Revista de Obstetricia y Ginecología Hospital Dr. Luis Tisné Brousse,* 4(1), 82-84.

Austin, P.C; Brunner, L.J y Hux, J.E. (2002). Bayeswatch: an overview of Bayesian statistics. *Journal of Evaluation in Clinical Practice*, 8(2), 277-86.

Bacchetti, P; Wolf, L; Segal, My McCulloch, C.E. (2005). Ethics and sample size. *American journal of epidemiology*, 161(2), 105-110.

Bejnordi, B.E; Litjens, G y van der Laak, J.A. (2018). Machine Learning Compared With Pathologist Assessment-Repl. *American Medical Association*, 319(16), 1726.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association,* 37, 325–335.

Bzdok, D; Altman, N y Krzywinski, M. (2018). Points of Significance: Statistics versus machine learning. *Nature Methods*, 15(4), 233-234.

Cavada, G. (2009). Docimasia de hipótesis. *Revista Chilena de Endocrinología y Diabetes*, 2(4), 256-257.

Chia, K. (1997). “Significant-itis”—an obsession with the P-value. *Scandinavian journal of work, environment & healtht,* 152-154.

Ching, T; Himmelstein, D.S y Beaulieu-Jones, B.K. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141).

Council for International Organizations of Medical Sciences (2002). Pautas éticas internacionales para la investigación biomédica en seres humanos. Consultada 27 de septiembre del 2018. Disponible en: http://www1.paho.org/Spanish/BIO/CIOMS.pdfn

Council for International Organizations of Medical Sciences. (2016). Ethical Guidelines: advancements and unsolved topics in 2016 upgrade. *Medwave*. 2018 25; 18(2): e7208.

Dahiru, T. (2008). Value, a true test of statistical significance? A cautionary note. *Annals of Ibadan Postgraduate Medicine*, 6(1), 21-26.

Dorey, F. (2010). In Brief: The P Value: What Is It and What Does It Tell You? *Clinical Orthopaedics and Related Research*,468(8), 2297-2298.

Droppelmann, G. (2016). Bioética en el tratamiento de datos. *Revista Española de Bioética,* 45, 96-101.

Droppelmann, G. (2018). La instrumentalización metodológica en la Ética de la Investigación. *EIDON*, 49, 102-114.

Emanuel, E. (2000). ¿What makes clinical research ethical? *JAMA*, 31; 283(20), 2701-11.

Fisher, R.A. (1950). Statistical methods for research workers. *Nigerian Journal of Paediatrics. London: Oliver and Boyd*, 80.

Gardner, M.J y Altman, D.G. (1986). Confidence intervals rather than pvalues: estimation rather than hypothesis testing. *British Medical Journal,* 292, 746-50.

Gelman, A y Loken, E. (2014). The statistical crisis in science: Data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American scientist*, 102, 460-465.

Ghosh, S.K. (2010). Basics of Bayesian methods. *Methods in Molecular Biology,* 620, 155-78.

Gigerenzer, G y Marewski, J.N. (2015). Surrogate science: the idol of a universal method for scientific inference. *Journal of Management,* 41, 421-440.

González, F. (2015). Modelos de aprendizaje computacional en reumatología. *Revista Colombiana de Reumatología*, 22(2), 77-78.

González, I. (2010). Partes componentes y elaboración del protocolo de investigación y del trabajo de terminación de la residencia. *Revista Cubana de Medicina General Integral,* 26(2), 387-406.

Goodman, S.N y Berlin, J.A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine,* 121, 200-6.

Goodman, S.N. (1999). Towards evidence-based medical statistics, I: the P-value fallacy. *Annals of Internal Medicine,*130,995-1004.

Goodman, S.N. (2005). Introduction to Bayesian methods I: measuring the strength of evidence. *Clinical Trials*, 2(4), 282-90, discussion 301-4, 364-78.

Greenland, S; Senn, S.J; Rothman, K.J; Carlin, J.B; Poole, C; Goodman, S.N y Altman D.G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology,* 31, 337-350.

Gründner, J; Prokosch, H.U; Stürzl, M; Croner, R; Christoph, J y Toddenroth, D. (2018). Predicting Clinical Outcomes in Colorectal Cancer Using Machine Learning. *Studies in Health Technology and Informatics*, 247, 101-105.

Hoekstra, R; Finch, S; Kiers, H y Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin &Review*, 13(6), 1033-1037.

Hoover, K y Siegler, M. (2008). The rhetoric of 'Signifying nothing': a rejoinder to Ziliak and McCloskey. *Journal of Economic Methodology*, 15, 57-68.

Kirby, A; Gebski, V yKeech, A.C. (2002). Determining the sample size in a clinical trial. *The Medical Journal of Australia*, 177, 256-7.

Koohy, H. (2017). The rise and fall of machine learning methods in biomedical research. *F1000Research,* 6.

Kourou, K; Exarchos, T; Exarchos, K; Karamouzis, M y Fotiadis D. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal,* 13, 8-17.

LeCoutre, M.P; Poitevineau, J y Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of null hypothesis tests. *International Journal of Psychology,* 38, 37–45.

Li, Y yShen, L. (2018). Skin Lesion Analysis towards Melanoma Detection Using Deep Learning Network. *Sensors (Basel),* 18(2), 556.

Louis, T.A. (2005). Introduction to Bayesian methods II: fundamental concepts. *Clinical Trials*, 2(4), 291-4, discussion 301-4, 364-78.

Manterola, C y Pineda, V. (2008). El valor de "p" y la "significación estadística": Aspectos generales y su valor en la práctica clínica". *Revista Chilena de Cirugía*, 60(1), 86-89.

Marco, C y Larkin, G. (2000). Research Ethics: Ethical Issues of Data Reporting and the Quest for Authenticity. *Academic Emergency Medicine*, 7(6), 691-694.

Mayo, D y Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction. *The British Journal for the Philosophy of Science*, 57, 323-357.

McDonald, J.F. (2018). Back to the future - The integration of big data with machine learning is re-establishing the importance of predictive correlations in ovarian cancer diagnostics and therapeutics. *Gynecologic Oncology*, 149(2), 230-231.

Mendoza, C. (2006). El valor p en epidemiología. *Revista Chilena de Salud Pública,* 10(1), 47-57.

Nishio, M; Nishizawa, M; Sugiyama, O;Kojima, R; Yakami, M; Kuroda, T y Togashi, K. (2018). Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PLoSOne*, 13(4), e0195875.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: Wiley.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. La declaración universal sobre el genoma humano y los derechos humanos. (2000). Consultado el 26 de septiembre del 2018. Disponible en: http://unesdoc.unesco.org/images/0012/001229/122990So.pdf

Penn State Statistics. Ethics and Statistics. [Consultado el 3 de julio del 2018]. Disponible en:https://onlinecourses.science.psu.edu/statprogram/ethics/

Rahul, D. (2015). Machine learning in medicine. *Circulation*, 132(20), 1920-1930.

Rodríguez, E. (2004). Comités de evaluación ética y científica para la investigación en seres humanos y las pautas CIOMS 2002. *Acta Bioethica*, 10(1), 37-48.

Rothman, K.J. (1986). Significance questing. *Annals of Internal Medicine*, 105, 445–447.

Royall, R. (1997). *Scientific Evidence: A Likelihood Paradigm.* London: Chapman& Hall.

Sedgwick, P. (2014). Understanding P values. *British Medical Journal*, 349, g4550.

Sollini, M; Cozzi, L; Chiti, A y Kirienko, M. (2018). Texture analysis and machine learning to characterize suspected thyroid nodules and differentiated thyroid cancer: Where do we stand?.*European Journal of Radiology,* 99, 1-8.

Spiegelhalter, D.J; Myles, J.P; Jones, D.R y Abrams, K.R. (2000). Bayesian methods in health technology assessment: a review. *Health Technology Assessment*, 4(38), 1-130.

Tang, J; Rangayyan, RM; Xu, J; El Naqa, I y Yang, Y. (2009). Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *IEEE Transactions on Information Technology in Biomedicine,*13(2), 236-251.

The Editors. (2001). The value of p. *Epidemiology*, 12(3), 286.

van Raaij, T.M; Reijman, M; Brouwer, R.W; Bierma-Zeinstra, S.M y Verhaar, J.A. (2010). Medial knee osteoarthritis treated by insoles or braces: a randomized trial. *Clinical Orthopaedics and Related Research,*468(7), 1926-32.

van Smeden, M; Van Calster, B y Groenwold, R. (2018). Machine Learning Compared With Pathologist Assessment. *American Medical Association*, 319(16), 1725-1726.

Wang, Z; Meng, Q; Wang, S; Li, Z; Bai, Y y Wang, D. (2018). Deep learning-based endoscopic image recognition for detection of early gastric cancer: a Chinese perspective. *Gastrointestinal Endoscopy,* 88(1), 198-199.

Weng, SF; Reps, J; Kai, J; Garibaldi, J y Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoSONE*, 12(4), e0174944.

Wong, D y Yip, S. (2018). Machine learning classifies cancer. *Nature,* 555(7697), 446-447.

World Medical Association Declaration of Helsinki (2013): Ethical principles for medical research involving human subjects, 27; 310(20): 2191-4.