

Certo, certissimo, anzi probabile *(cit.)*

Verga MC¹

¹ Pediatra di famiglia, ASL Salerno

Il valore di p

Il p -value entra ufficialmente nella vita dei ricercatori nel 1925, con la pubblicazione di *"Statistical Methods for Research Workers"* di Ronald Fischer, grande biologo e statistico del primo '900. È un parametro che si utilizza quando non ci si limita a semplici descrizioni, ma quando si fanno confronti per rilevare eventuali differenze tra gruppi.

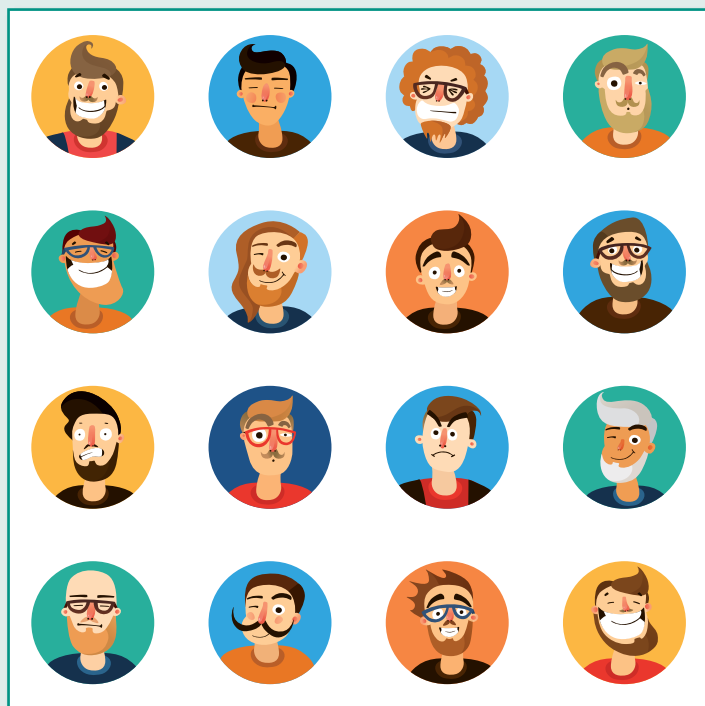
La ricerca ai tempi del *p-value*

Consideriamo il fumo di sigaretta e il cancro al polmone.

In uno studio puramente descrittivo, i risultati ci diranno che il cancro al polmone ha un tasso di incidenza (su 100.000 abitanti in 11 anni) di 11,1 tra i non fumatori, 45,7 tra i fumatori e 43,2 tra gli ex fumatori.¹

In uno studio analitico, i risultati ci diranno se ci sono differenze tra i gruppi associate direttamente al fumo di sigaretta o dovute semplicemente al caso.

Ma cosa significa *se ci sono differenze*? Quanto devono essere grandi le differenze per poter considerare i 3 gruppi realmente diversi e attribuire un'associazione tra lo sviluppo di cancro al polmone e la condizione di non fumatore/fumatore/ex fumatore?



Quanto devono essere grandi le differenze?

In termini strettamente matematici i numeri sono diversi anche per una semplice cifra decimale: 11,11 è differente da 11,10.

In epidemiologia e nella pratica clinica, invece, differenze minime possono essere dovute al caso, *in primis* per la naturale variabilità individuale.

I ricercatori, quindi, stabiliscono a priori quanto devono essere grandi queste differenze, stabiliscono cioè una *soglia di significatività* α , di solito = 0,05, in base alla quale le differenze nei tassi di incidenza tra i gruppi sono giustificate al 95% dal fumo e solo per il 5% dal caso.

Il *p-value* che si ricava dall'analisi statistica dei risultati, quindi, è un parametro pratico che ci indica solo se le differenze osservate superano o meno la soglia di significatività stabilita, se

sono, quindi, dovute al caso ($p \geq \alpha$, $p \geq 0,05$), oppure al fattore considerato, il fumo ($p < \alpha$, $p < 0,05$). In quest'ultimo caso le differenze sono definite "statisticamente significative".

Bisogna specificare che "statisticamente significativo" non significa necessariamente "rilevante", cioè clinicamente importante.² Per ogni condizione, le differenze considerate rilevanti possono essere di diversa entità: per esempio, per patologie lievi, di bassissimo impatto, consideriamo rilevanti differenze (di durata o di incidenza) molto grandi, mentre per condizioni rare, gravi, invalidanti o mortali, anche piccole differenze possono essere considerate rilevanti.



Mod. da 2

La rilevanza non è quindi determinata dalla statistica, per cui è assolutamente sbagliato far scaturire conclusioni o raccomandazioni per la pratica clinica solo sulla base del *p-value*, soprattutto se i risultati non sono stati riprodotti e confermati almeno da un secondo studio, in base al principio della *riproducibilità*.

Il *p-value* ai tempi dei big data



Big data – confronti multipli

Proprio la riproducibilità dei risultati ha indotto un gruppo di esperti a mettere in discussione l'importanza del *p-value* e,

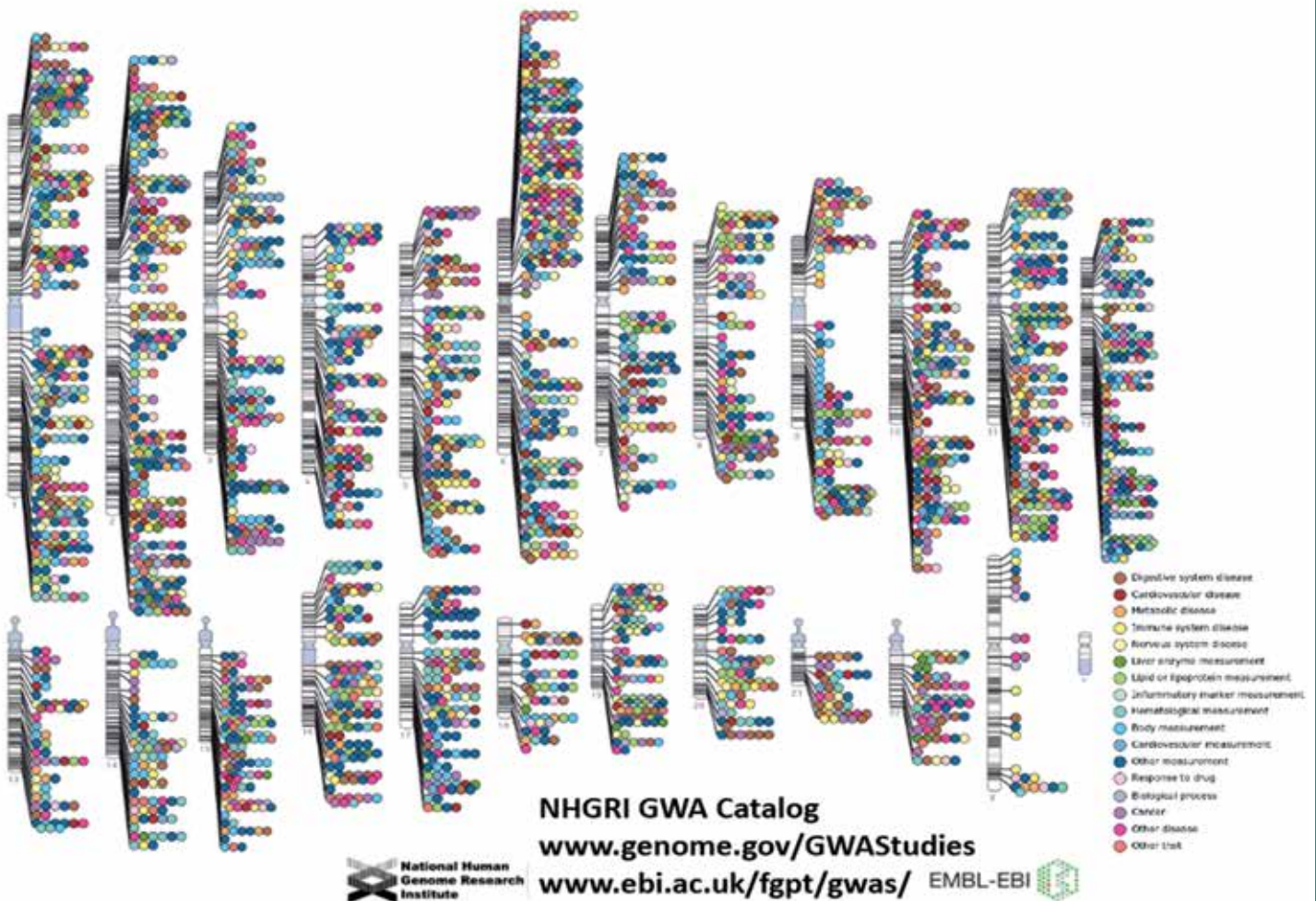
specificamente, della soglia = 0,05.³ In letteratura ci sono infatti troppi falsi positivi, troppi risultati che vengono confutati da studi successivi, soprattutto quando si hanno a disposizione molti dati e si fanno confronti multipli: è stato calcolato che, negli ultimi anni, circa l'85% dei risultati della ricerca pre-clinica non è stato poi confermato. Poiché più piccolo è il valore di p e minore è la probabilità che il risultato sia dovuto al caso, quindi minore è il rischio di falsi positivi, Benjamin et al. hanno proposto, solo per le nuove scoperte, di abbassare la soglia di significatività da 0,05 a 0,005. Del resto, già da un decennio, nei *Genome-Wide Association Studies*, o GWAS, in cui si confrontano centinaia di migliaia di varianti di DNA per cercare le differenze tra le persone con una malattia e le persone senza, i genetisti hanno adottato una soglia di significatività bassissima = 0,00000005.

Naturalmente andranno anche considerate altre cause di scarsa validità, oltre alla soglia α , come i *publication bias*, la bassa numerosità, o altri bias.

Un altro prestigioso gruppo di ricercatori, però, ha espresso molte perplessità su questa proposta, per almeno 3 motivi:⁴

1. non vi sono prove sufficienti del fatto che lo standard attuale sia una "principale causa di non riproducibilità";
2. i problemi correlati al valore di $p < 0,05$ (molti falsi positivi non confermati da ulteriori studi) non giustificano l'utilizzo immediato e diffuso della nuova soglia ($p < 0,005$);
3. un valore soglia inferiore avrebbe probabilmente conseguenze negative, come la necessità di aumentare la numerosità campionaria e rendere la ricerca molto più difficile e onerosa.

Published Genome-Wide Associations through 12/2012 Published GWA at $p \leq 5 \times 10^{-8}$ for 17 trait categories



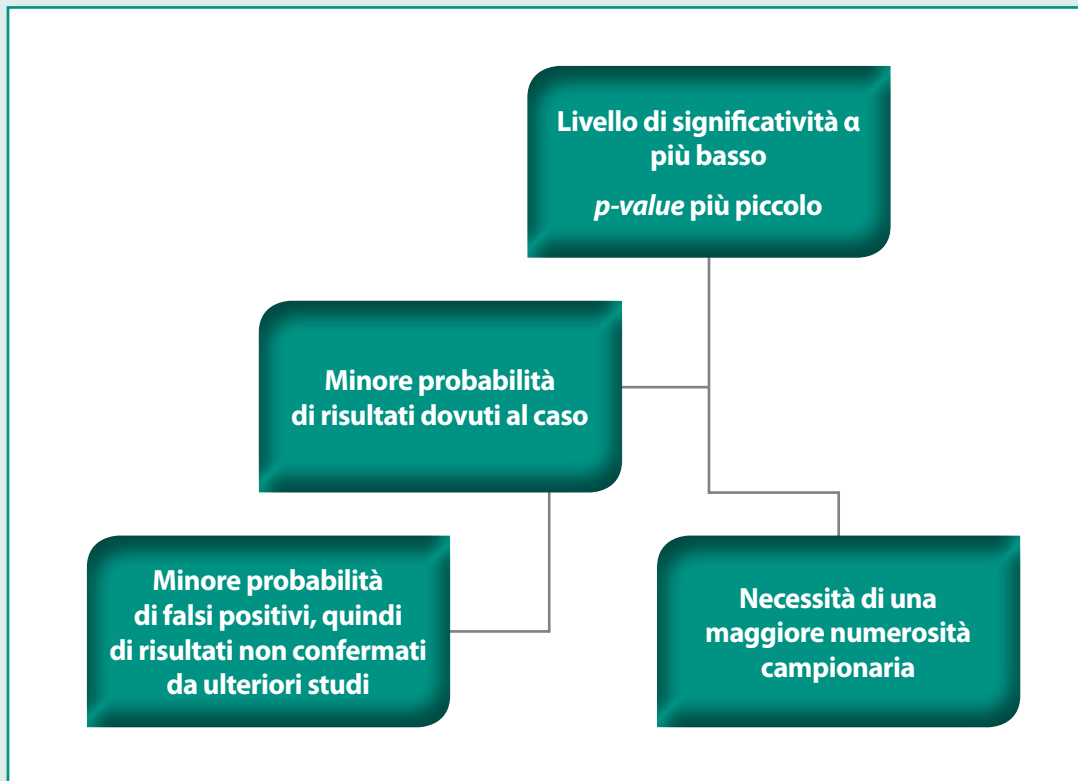


Figura 1. Diminuire il valore del *p-value*: pros and cons.

Essi concludono che il termine "statisticamente significativo" non dovrebbe più essere usato e suggeriscono che i ricercatori debbano riferire e giustificare in modo trasparente tutte le scelte che fanno quando disegnano uno studio, incluso il livello α .

Il problema è quindi complesso e ancora si discute per una possibile soluzione. Tutti però concordano sul fatto che la validità dei risultati non

dipende solo dagli strumenti disponibili ma dall'uso che se ne fa, per cui qualunque metodo statistico potrà risultare fallace in mani poco esperte o poco "trasparenti", come quelle di alcuni ricercatori che cercano a tutti i costi un qualunque risultato statisticamente significativo pur di aumentare le *chances* di pubblicare la loro ricerca.

Ringrazio le compagne e i compagni di EBM per gli opportuni e preziosi consigli.

Bibliografia

1. <http://www.registri-tumori.it/cms/it/system/files/Impatto%20del%20fumo%20sui%20tumori%20del%20polmone%20in%20provincia%20di%20Sondrio.pdf>
2. http://www.quadernodiepidemiologia.it/epi/assoc/pro_sig.htm
3. Benjamin DJ, Berger JO, Johannesson M et al. Redefine statistical significance. *Nature Human Behaviour* 2018;2:6-10.
4. Lakens D, Adolfs FG, Albers CJ et al. Justify Your Alpha. *Nature Human Behavior*. In Press.

Per contattare l'Autore **Maria Carmen Verga**: vergasa@virgilio.it