GOLD STANDARDS MAY BIAS CLINICAL TEST VALIDATION. LOS PATRONES DE ORO PUEDEN SESGAR LA VALIDACIÓN DE LAS PRUEBAS CLÍNICAS

Prof. Manuel González de la Rosa, Marta González Hernández

Resumen

El uso de un Patrón de Oro puede resultar una opción aceptable cuando su precisión es bien conocida y no se espera que el procedimiento evaluado pueda ser más exacto. Pero cuando el Patrón de Oro es imperfecto o peor que el que se juzga, las conclusiones de la evaluación resultan generalmente inexactas. Su utilización no solo puede infravalorar el método evaluado, sino ofrecer unos resultados incorrectos, especialmente cuando se ha utilizado para la selección de las muestras.

It is essential in any area of science to establish rigorous and stable standards of reference. I remember from my youth the definition of meter as the ten-millionth of the distance between the pole and the equator, and of a gram as the weight of one cubic centimetre of water at the melting temperature of the ice. Both definitions were set in 1792 by the French Academy of Sciences. Those basic descriptions were latter materialized into physical patterns of meter and kilogram by the International System of Weights and Measures in 1889 as the length of a bar and the weight of a cylinder of iridium-platinum that are kept in the International Bureau of Weights and Measures, in the basement of the Breteuil Pavilion on the outskirts of Paris, to be precise.

By 1983, and to avoid relying on the instability or deterioration of such patterns, the meter was redefined as the distance travelled by light in 1/299,792,458 seconds and from May 30th, 2019, the kilogram became based on the mass unit known as Planck constant (6,626,070 15 x 10-34), which describes a physical constant that is the quantum, which relates the energy carried by a photon to its frequency, taking into account that mass and energy are equivalent.

Likewise, in Medicine, we keep trying to optimize diagnostic methods, from the simple thermometer for measuring body temperature to the most sophisticated digital imaging systems. Certain indicators may be very accurate in pointing out the onset of a disease, but unfortunately many diagnostic methods available often produce contradictory results, making it difficult for us to achieve real confidence to support our decisions, especially at the initial stages of these disease processes.

Sensitivity, specificity, ROC area, intra-class correlation coefficient of reproducibility, kappa index of diagnostic agreement and many other statistical tools help us in the evaluation of diagnostic indices that emerge from experience or technology, but do not describe the absolute truth. They are, as we have said, useful mathematical arguments, but are not entirely free from possible biases or methodological errors.

Comparing these diagnostic procedures, so as to identify which are most effective in diagnosing a disease, requires applying them to both healthy and pathological patient samples. This is where we face a difficult problem. How could we be sure that subjects are sick or healthy if our dilemma is precisely to select the best diagnostic criteria? Most clinical studies assume that the best currently known method should be chosen for the recruitment of samples. This is

where the concept of the Gold Standard comes into play.

Unfortunately, in many diseases this standard is either non-existent or quite inaccurate. Therefore, great sophisms and false diagnostic conclusions may arise from their adoption. A good example is the published literature on glaucoma diagnosis. It will be a model for these considerations, which will certainly be useful for assessing similar scenarios in other pathologies.

Many editors and reviewers request researchers to divide their samples into normal, ocular hypertensives, suspected glaucoma, and confirmed glaucomas. But the critical problem in diagnosing glaucoma is determining the onset of the disease. Studying normals, suspects and glaucomas separately assumes that there is no such diagnostic problem. That would be assuming that subjects with ocular hypertension certainly do not suffer from glaucoma and that suspected and confirmed glaucomas are situations with a well-defined division.

Nevertheless, using kappa analysis to interpret the strength of the agreement of multiple morphological, functional or perfusion diagnosis indexes, it is clear that they do not easily agree, far from it. Consequently, the diagnostic problem holds its relevance. So which Gold Standard should we use to identify ocular hypertension, suspected glaucoma and confirmed glaucoma in order to assess different diagnostic procedures? Well, it simply does not exist. If it did, we would already have an adequate solution, which would mean there would be no need to keep publishing about it.

A Gold Standard is useful when there is already a very accurate procedure available but another cheaper, handier or faster one is to be evaluated. However, it should not be used when its accuracy is questionable. This is even more so, when the one to be evaluated might be more accurate than the current one. The inherent bias may be so significant that if the method being evaluated was better than the Gold Standard, it could never be actually proven. And this is because a better method than the Gold Standard will inevitably look statistically worse than itself. And this is only because mistakes made by the Gold Standard would be interpreted as failures of the evaluated procedure, even if the latter was right. It is a mere Logic problem. If we took a tailor's tape measure as Gold Standard, we would misinterpret the Paris platinum-iridium scale to be wrong.

It is relatively easy to determine the degree of normality of a sample. If we admit that the frequency of glaucoma in the adult population is between 2% and 6%, and exclude from our sample subjects with ocular hypertension, visual field defects, suspicious-looking optic nerves or a family history of glaucoma, then the specificity of our selection criteria will most likely be above 99%.

But it is much more difficult to ascertain that a subject with ocular hypertension and the rest of the parameters apparently normal is merely an "ocular hypertensive", thus ruling out any possibility of glaucoma. Among other things, we usually do not know the previous medical history, so the visual field may be within normal limits, but worse than 2 years ago. Neither do we know whether the optic nerve, despite being within normal limits, has lost axon thickness at the neuro-retinal rim. In these cases, progression may be the only criterion considered by any expert to confirm the case as glaucoma. Some researchers are determined to argue that there are procedures able to improve early diagnosis, even if they are not currently accepted by so-called Evidence-Based Medicine. There is no single test accepted at first by Evidence-Based Medicine. Any method must be given a time for different scientific groups to test it and then luckily publish. So, against all suspicions, one can only argue with the axiom that Science is not democratic, much less Medical Science.

Moreover, as far as glaucoma suspects are concerned, what is the difference between a confirmed and a suspected glaucoma? By waiting for several matching signs of the disease, the diagnosis would obviously be delayed. This is why specialists do not agree on which signs precede which others. The most widespread idea is that morphological signs that measure the atrophy of ganglion cells and their axons precede functional ones, i.e. visual defects. But such a paradigm is counterintuitive. In a spinal cord section, for instance, paralysis precedes motor neuron atrophy. But the difference between these two situations may hinge on the chronicity of the process or on the lack of sensitivity of the functional methods available or chosen for evaluation.

Many papers that evaluate diagnostic approaches, particularly those in recent times with the rise of Artificial Intelligence, use samples of glaucoma with strict diagnosis. Thus, it is common to find that the method of recruitment is the subjective observation of a clear abnormal optic nerve, with the presence of edge splinter hemorrhages, or/and vertical cup/disc ratios greater than 0.7, neuro-retinal rims with evident asymmetric thinning areas, etc. When programming and evaluating automatic diagnostic algorithms based on Deep Learning techniques, basing nerve selection

on such criteria, wonderful sensitivity figures and auROC close to one are easily reached, as found in the literature. It is very simple to deceive readers with results that do not represent the reality of the issue at all. The only credible conclusion from these studies is that such an expert system, is able to detect advanced glaucoma with a similar performance of a specialist with an ophthalmoscope. But if ophthalmoscopes were sensitive and specific enough to detect glaucoma, we would not need perimeters, tonometers or optical coherence tomographs (OCT), and certainly not Artificial Intelligence. Many of these methods are actually much less useful than they try to suggest. And the conclusions of such papers are often a mirage that is not always naive.

Another common example of Gold Standard bias: If samples are chosen using or giving more weight to morphological procedures, then functional ones will be at a disadvantage in the conclusions, and vice versa. A very usual model of this type of project would be the following: "Out of 100 OCT-defined glaucomas, 50% had normal visual fields and progressed over time, proving that OCT provides early diagnosis." This is a false rationale, because it might just as well be that: "Out of 100 perimetrically defined glaucomas, 50% had normal OCT and it got worse over time." The bias of such an approach leads to an apparently elegant (or interested) outcome, but one that is totally distorted by a misconception.

Our group's research activity has been focused on the design of diagnostic procedures and the evaluation of the evolution of glaucoma. For years we have been working on a methodological proposal to compare the results of different examination methods, based on these previous arguments.

Our approach, avoiding a Gold Standard, is clinical test validation comparing two groups: One of them with minimal likelihood of being glaucomatous, as mentioned above, and the other comprising confirmed glaucomas as well as ocular hypertensives and glaucoma suspects, who would fall on either side (normal or pathological) depending on the selection criteria. The idea is to include any subject, for whatever reason, in order to try to avoid biasing the sample as much as possible. Therefore, it will integrate severe ocular hypertension, moderate ocular hypertension if there is a family history of glaucoma, patients with characteristic visual field defects, anomalous optic nerves, sectorial fibre atrophy, subjects that seem to have progressed, etc.

Thus, if several diagnostic, morphological, functional or perfusion procedures are applied to two such groups, the one achieving the highest sensitivity to separate that heterogeneous sample from the normal one, while maintaining a very high specificity, will most likely be the most accurate, minimizing methodological bias.

This approach is likely to result in less impressive ROC curves and sensitivity and specificity figures than often found in the literature, but the comparison of performances will be less contaminated by the a priori use of a Gold Standard.

RECOMMENDED REFERENCES

1. République française. Loi relative aux poids et mesures: du 18 germinal, an 3.e de la République française, une et indivisible [7 avril 1795]. À Paris; France. De l'Imprimerie de la République, 1798. https://archive.org/details/loirelativeauxpo00fran/mode/2up

2. Ransohoff DJ, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299:926-930.

3. Hawkins DM, Garrett JA, Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. Stat Med. 2001;20:1987-2001.

4. Alonzo TA, Pepe MS. Assessing the accuracy of a new diagnostic test when a gold standard does not exist. UW Biostatistics Working Paper Series, 1998. Working Paper 156. http://biostats.bepress.com/uwbiostat/paper156

5. Knottnerus JA, van Weel Ch, Muris JWM. Evaluation of diagnostic procedures. BMJ. 2002;324:477-480.

6. Zhou X, Castelluccio P, Zhou C. Nonparametric Estimation of ROC Curves in the Absence of a Gold Standard. Biometrics. 2005;61:600-609.

7. Rutjes AWS, Reitsma JB, Coomarasamy A et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess. 2007;11:1-72.

8. Cronin AM, Vickers AJ. Statistical methods to correct for verification bias in diagnostic studies are inadequate when there are few false negatives: a simulation study. BMC Med Res Methodol 2008;8:75:1-9.

9. Reitsmaa JB, Rutjesa AWS, Khanb KS et al. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009;62:797-806.

En cualquier tipo de ciencia resulta fundamental establecer patrones de referencia rigurosos y estables. De mi formación juvenil recuerdo la definición de metro como la diezmillonésima parte de la distancia entre el polo y el ecuador, y la de gramo como el peso de un centímetro cúbico de agua a la temperatura de fusión del hielo. Ambas habían sido establecidas en 1792 por la Academia de Ciencias de Francia. Estas descripciones básicas se concretaron en unos patrones físicos de metro y kilogramo, que el Sistema Internacional de Pesos y Medidas estableció en 1889 como la longitud de una barra y el peso de un cilindro de iridio-platino que se encuentran guardados en la Oficina Internacional de Pesos y Medidas, concretamente en el sótano del Pabellón de Breteuil a las afueras de París.

Para evitar depender de la estabilidad o deterioro de estos patrones, en 1983 se redefinió el metro como "la distancia recorrida por la luz en 1/299.792.458 segundos" y desde el 30 de mayo de 2019 se ha establecido el kilogramo en base a la unidad de masa conocida como "constante de Planck" (6.626 070 15 x 10-34), que describe una constante física que es el quantum, la cual relaciona la energía transportada por un fotón con su frecuencia, teniendo en cuenta que la masa y la energía son equivalentes.

De la misma manera, en Medicina hemos ido tratando de perfeccionar los procedimientos diagnósticos, desde el simple termómetro que mide la temperatura corporal al más sofisticado de los sistemas de imagen digital. Determinados indicadores pueden ser muy precisos para señalar el comienzo de una enfermedad, pero lamentablemente en muchos casos los métodos diagnósticos de que disponemos producen frecuentes resultados contradictorios, que dificultan el adquirir una autentica certeza en la que podamos apoyar nuestras decisiones, especialmente al comienzo de estos procesos.

Sensibilidad, especificidad, área ROC, coeficiente de correlación intraclases de reproducibilidad, índice kappa de concordancia diagnóstica y otras muchas herramientas estadísticas nos ayudan en la evaluación de los índices diagnósticos que van surgiendo de la experiencia o de la tecnología, pero no describen la verdad absoluta. Son, como hemos dicho, argumentos matemáticos útiles, pero no están necesariamente exentos de posibles sesgos o errores metodológicos.

La comparación de estos procedimientos, para identificar los más eficaces en la tarea de diagnosticar una enfermedad, exige que sean aplicados a muestras de pacientes sanos y enfermos, y es aquí donde nos encontramos con un difícil problema. ¿Cómo adquirimos la certeza de que los sujetos son enfermos o sanos si nuestro dilema es, precisamente, seleccionar el mejor criterio diagnóstico? La mayor parte de las investigaciones clínicas presuponen que se elegirá, para reclutar a las muestras, el mejor método conocido hasta el momento. Y aquí nace el concepto de Patrón de Oro.

Sin embargo, en muchas enfermedades este patrón no existe, o es bastante inexacto. Por ello de su elección pueden derivarse enormes sofismas y falsas conclusiones diagnósticas. La literatura referente al diagnóstico del glaucoma es un buen ejemplo de ello, y nos va a servir de modelo para estas reflexiones, que serán sin duda útiles para analizar situaciones similares en otras enfermedades.

Muchos editores y revisores solicitan a los investigadores que dividan sus muestras en normales, hipertensos oculares, sospechosos de glaucoma y glaucomas confirmados. Pero el problema esencial del diagnóstico del glaucoma es identificar el comienzo de la enfermedad. Estudiar por separado normales, sospechosos y glaucomas presupone que el problema diagnóstico no existe y que, se sabe perfectamente que el hipertenso no es un glaucomatoso, y que sospechosa de glaucoma y glaucoma confirmado son situaciones con una frontera bien definida.

Pero si hacemos un análisis kappa de coincidencia diagnóstica, de múltiples indicadores morfológicos, funcionales o de perfusión, veremos que no se ponen fácilmente de acuerdo, ni muchísimo menos. Por eso el problema diagnóstico mantiene su actualidad. ¿Entonces cuál es el Patrón de Oro que debemos utilizar para separar hipertensos, sospechosos de glaucoma y glaucomas confirmados y evaluar diversos procedimientos diagnósticos? Simplemente no existe, porque si existiese, significaría que ya tendríamos una solución aceptable, y que no tendría sentido seguir publicando al respecto.

Un Patrón de Oro estándar es útil, cuando se dispone de un procedimiento muy exacto y se quiere valorar otro más barato, más práctico o más rápido, pero no se puede utilizar cuando su exactitud es cuestionable y, especialmente, cuando el que se trata de evaluar puede ser incluso más exacto de los que ya existen. El sesgo que introduce el Patrón de Oro estándar puede ser tan importante que, si el método evaluado fuese mejor que el gold estándar, nunca podrá llegar a demostrarte. Y es que un método mejor, frente al Patrón de Oro, inevitablemente parecerá estadísticamente peor, porque los errores que cometa el segundo serán interpretados como fallos del procedimiento evaluado, aunque este último tenga la razón. Es un simple problema de Lógica. Si tomamos como patrón la cinta métrica de un sastre, interpretaremos falsamente que la barra de iridio-platino de París comete errores.

Establecer la normalidad de una muestra es un problema probabilístico relativamente fácil. Si admitimos que la frecuencia del glaucoma en la población adulta está entre el 2% y el 6%, y excluimos a los sujetos con presiones oculares elevadas, defectos del campo visual, aspecto sospechoso del nervio óptico o que presenten antecedentes familiares de glaucoma, la especificidad de nuestro criterio de selección superará fácilmente el 99%.

Pero mucho más difícil es establecer que un sujeto con presiones oculares elevadas y el resto de los parámetros aparentemente normales es simplemente un "hipertenso ocular", y excluir la posibilidad de que sea glaucomatoso. Entre otras cuestiones, generalmente no conocemos su pasado, y puede que el campo visual sea aparentemente normal, pero peor que hace 2 años. Tampoco sabemos si su nervio óptico, pese a ser supuestamente normal, ha perdido espesor de axones en el anillo neuro-retiniano. En estos casos la progresión puede ser el único criterio que admitiría cualquier experto para confirmar el caso como glaucoma. Algunos investigadores nos empeñamos en defender que existen procedimientos para adelantar el diagnóstico, aunque no sean admitidos por la llamada Medicina Basada en la Evidencia. Pero es que no hay ningún método para diagnosticar el glaucoma que sea admitido como perfecto por la Medicina Basada en la Evidencia. Contra sus suspicacias solo podemos contraponer el axioma de que la Ciencia no es democrática, y la Ciencia Médica bastante menos que otras.

Por otra parte, ¿qué diferencia a un sospechoso de glaucoma de un glaucoma confirmado? Si se espera a detectar la presencia de varios signos coincidentes de la enfermedad, inevitablemente el diagnóstico será tardío. De aquí que los especialistas no se pongan de acuerdo sobre que signos preceden a cuáles otros. La idea más generalizada es que los signos morfológicos que miden la atrofia de las células ganglionares y de sus axones, preceden a los funcionales, es decir a los defectos visuales. Este paradigma es contrario a la lógica, que evidencia, por ejemplo, que la parálisis precede a la atrofia de la neurona motora en una sección medular, pero es posible que la diferencia entre ambos problemas dependa de la cronicidad del proceso o de la falta de sensibilidad de los métodos funcionales de que disponemos o que hemos querido evaluar.

Muchos trabajos que valoran métodos diagnósticos, especialmente en los últimos tiempos con el auge de la inteligencia artificial, utilizan una muestra de glaucomas absolutamente confirmados. Así, es frecuente que el método de selección sea la observación subjetiva de un nervio óptico claramente anormal, con presencia de hemorragias en astilla en su borde, o un cociente vertical Copa/Disco superior a 0.7, un anillo neuro-retiniano con evidentes zonas asimétricas de adelgazamiento etc. Al diseñar programas de diagnóstico automático basados en técnicas de Aprendizaje Profundo a imágenes de nervios ópticos seleccionados por expertos de esta manera, pueden encontrarse en la literatura cifras maravillosas de sensibilidad y curvas ROC próximas a la unidad. Es muy fácil engañar a los lectores con unos resultados que no representan en absoluto la realidad del problema. La única conclusión admisible de estos trabajos es que un sistema experto, diseñado de esta manera, es capaz de detectar glaucomas avanzados con una eficacia similar a un especialista que posea un oftalmoscopio. Pero si un oftalmoscopio fuese suficientemente sensible y específico para detectar un glaucoma no necesitaríamos perímetros, tonómetros o tomógrafos de coherencia óptica (OCT), ni mucho menos Inteligencia Artificial. La utilidad real de muchos de estos métodos es muchísimo menor de lo que tratan de aparentar, y las conclusiones de estos trabajos son, en realidad, un espejismo no siempre inocente.

Veamos otro ejemplo habitual de sesgo originado por el Patrón de Oro: Si se eligen o ponderan preferentemente procedimientos morfológicos, los funcionales quedarán en desventaja en las conclusiones, y al contrario si se seleccionan los casos por procedimientos funcionales. Un modelo muy habitual de trabajo podría corresponder a este esquema: "De 100 glaucomas definidos por OCT, el 50% tenían campo visual normal y empeoró con el tiempo, lo que demuestra que el OCT es precoz". Este es un falso razonamiento, porque igualmente podría encontrarse que "de 100 glaucomas definidos por perimetría, el

50% tenían OCT normal y empeoró con el tiempo". El sesgo del planteamiento conduce a resultados aparentemente elegantes (o interesados), pero totalmente contaminados por falta de lógica.

Nuestra actividad investigadora se ha orientado al diseño de procedimientos diagnósticos y de evaluación de la evolución de la enfermedad glaucomatosa. De los razonamientos previos ha surgido la propuesta metodológica que venimos aplicando desde años a nuestros trabajos para comparar los resultados de diferentes métodos de examen.

En nuestra opinión, la única solución viable es prescindir del Patrón de Oro enfrentando dos grupos: Uno de ellos con bajísimas probabilidades de ser glaucomatoso, como hemos señalado en un párrafo previo, y otra muestra en la que no se separan los hipertensos oculares, los sospechosos de glaucoma y los glaucomas confirmados, que caerían en uno u otro lado según el criterio de selección. Se trata de incluir en él a cualquier sujeto, por cualquier tipo de sospecha, para tratar de crear el menor sesgo posible en la muestra. Admitiremos en él, por lo tanto, hipertensos oculares severos, o moderados si tienen antecedentes familiares de glaucoma, pacientes con defectos de campo característicos, nervios ópticos de aspecto sospechoso, atrofias sectoriales de fibras, sujetos que aparentemente ha empeorado etc.

Si a dos grupos como estos, se les aplican varios procedimientos diagnósticos, morfológicos, funcionales o de perfusión, el que obtenga la mayor sensibilidad para diferenciar esa muestra heterogénea de la normal, manteniendo una alta especificidad, será con toda probabilidad el mejor, con el menor sesgo metodológico posible.

Este criterio posiblemente conducirá a cifras de sensibilidad y especificidad y curvas ROC menos espectaculares que las habituales en la literatura, pero la comparación de los resultados de los diferentes procedimientos estará menos contaminada por la utilización a priori de un Patrón de Oro.