بررسی مقایسه‌ای روش‌های ارزیابی تست‌های تشخیصی بدون استاندارد طلایی: *مقاله مروری*

فاطمه مصائبی[۱]
فرید زایری[۲*]
ملیحه نصیری[۱]
مهدی عزیزمحمد لوها[۱]

۱- گروه آمار زیستی، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران.
۲- گروه آمار زیستی، مرکز تحقیقات پروتئومیکس، دانشکده پیراپزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران.

[*] نویسنده مسئول: تهران، میدان قدس، خیابان دربند، دانشگاه علوم پزشکی شهید بهشتی، دانشکده پیراپزشکی.
تلفن: ۰۲۱-۲۲۷۰۷۳٤۷
E-mail: fzayeri@gmail.com

دریافت: ۱۳۹۷/۰۲/۲۱ ویرایش: ۱۳۹۷/۰۲/۲۸ پذیرش: ۱۳۹۷/۱۱/۲۰ آنلاین: ۱۳۹۷/۱۱/۳۰

چکیده

با پیشرفت علوم پزشکی استفاده از تست‌های تشخیصی به‌منظور تمایز افراد بیمار از سالم گسترش فراوانی پیدا کرده است. بنابراین بررسی دقت تست‌های تشخیصی تحت ارزیابی به‌منظور تعیین توانایی آن‌ها در جداسازی افراد بیمار از سالم از اهمیت ویژه‌ای برخوردار است. دقت یک تست را میزان تطابق نتایج آن تست با نتایج استاندارد طلایی تعیین می‌کند و این میزان براساس معیارهای حساسیت، ویژگی، ارزش اخباری مثبت، ارزش اخباری منفی و مساحت زیرمنحنی مشخصه‌ی عملکرد بیان می‌شود. استاندارد طلایی یک روش دقیق و بدون خطا برای تشخیص است که در برخی از بیماری‌ها به دلایلی مانند هزینه‌بر بودن وجود ندارد. در این وضعیت، استاندارد مرجع به‌عنوان بهترین روش جایگزین قابل دسترس برای تشخیص بیماری، مورد استفاده‌ی پزشکان قرار می‌گیرد. نتایج روش استاندارد مرجع لزوما بدون خطا نیست و گاهی نیز به‌دلیل داشتن حالت تهاجمی نمی‌توان آن‌را روی همه‌ی افراد مورد مطالعه اجرا کرد، همه‌ی این موارد به شرایطی اشاره می‌کند که در آن استاندارد طلایی موجود نیست. استفاده از استاندارد مرجع دارای خطا منجر به تفکیک درست افراد بیمار از سالم نمی‌شود و نمی‌تواند یک معیار مقایسه برای بررسی دقت تست‌های تشخیصی باشد. ازجمله روش‌هایی که در شرایط فاقد استاندارد طلایی به‌کار گرفته می‌شود، روش جانهی، روش تصحیح استاندارد مرجع ناقص و روش ساخت استاندارد مرجع است. هر کدام از این روش‌ها با توجه به ویژگی، مزیت و محدودیت‌هایی که دارند برای ارزیابی و تعیین دقت تشخیصی تست‌ها مورد استفاده قرار می‌گیرند. در مقاله کنونی به مرور و مقایسه‌ی این روش‌ها و نقاط ضعف و قوت آن‌ها پرداخته شد.

کلمات کلیدی: دقت، تست تشخیصی، استاندارد مرجع.

با پیشرفت علوم پزشکی استفاده از تست‌های تشخیصی به‌منظور تمایز افراد بیمار از سالم گسترش فراوانی یافته است. بنابراین ارزیابی دقت تست‌های تشخیصی به‌منظور تعیین توانایی آن‌ها در جداسازی افراد بیمار از سالم از اهمیت ویژه‌ای برخوردار است.[۱] دقت یک تست را می‌توان از میزان تطابق نتایج آن با نتایج استاندارد طلایی سنجید و میزان تطابق را براساس معیارهای حساسیت، ویژگی، ارزش اخباری مثبت، ارزش اخباری منفی و مساحت زیرمنحنی مشخصه‌ی عملکرد بیان کرد.[۲] این درحالی است که برای خیلی از بیماری‌ها، استاندارد طلایی که دقیق‌ترین و بدون خطاترین روش تشخیصی است، در عمل به دلایلی مانند هزینه‌بر بودن وجود ندارد. در این شرایط پژوهشگر از بهترین روش در دسترس استفاده می‌کند که به این روش، استاندارد مرجع گفته می‌شود.[۳] گاهی در ارتباط با استاندارد مرجع وضعیت‌های خاصی به‌وجود می‌آید که در آن نتایج استاندارد مرجع برای همه‌ی افراد مورد مطالعه موجود نیست و یا دارای نقص است، در این حالت، استفاده از استاندارد مرجع منجر به تفکیک درست افراد بیمار از سالم نمی‌شود.[۳] همه‌ی حالت‌های بیان‌شده‌ی فوق به شرایطی اشاره می‌کند که در آن استاندارد طلایی موجود نیست.[۳و٤] در این شرایط، روش‌هایی مانند روش‌های جانهی،

روش تصحیح استاندارد مرجع ناقص و روش ساخت استاندارد مرجع مورد استفاده قرار می‌گیرند. روش جانهی در شرایطی مورد استفاده قرار می‌گیرد که نتایج استاندارد مرجع مناسب به دلایلی از جمله هزینه‌بربودن برای همه‌ی افراد مورد مطالعه موجود نیست.[۳] در مطالعات تشخیصی وجود داده‌ی گمشده در نتایج استاندارد طلایی نشان‌دهنده‌ی تایید جزیی می‌باشد. در برخورد با داده‌های گمشده دو استراتژی را می‌توان در نظر گرفت. اولین استراتژی که به‌دلیل سهولت آن مورد توجه بسیاری از پژوهشگران قرار گرفته است، در نظر نگرفتن داده‌های گمشده و تعیین برآوردهای حساسیت و ویژگی براساس مجموعه داده‌ی کامل است. این‌گونه از تحلیل‌ها همراه با اریبی خواهند بود.[۵] استراتژی دیگر نیز روش جانهی است که به جایگذاری مقادیر گمشده می‌پردازد.[۳] Rubin در سال ۱۹۷۶ سه الگو برای گمشدگی داده‌ها در نظر گرفت.[۶] گمشدگی به‌طور کامل تصادفی، گمشدگی تصادفی و گمشدگی غیرتصادفی. چنانچه احتمال گمشدگی به هیچ‌یک از داده‌های مشاهده‌شده و مشاهده‌نشده بستگی نداشته باشد گمشدگی کاملاً تصادفی رخ می‌دهد. زمانی‌که احتمال گمشدگی فقط به داده‌های مشاهده‌شده بستگی داشته باشد با مکانیزم گمشدگی تصادفی مواجه هستیم. همچنین درصورتی‌که احتمال گمشدگی به مقادیر مشاهده‌شده و مشاهده‌نشده بستگی داشته باشد مکانیزم گمشدگی غیرتصادفی خواهیم داشت.[۷] انتخاب روش‌های جانهی تا حد زیادی به الگو و مکانیزم داده‌های گمشده بستگی دارد. در هنگام به‌کارگیری روش جانهی به حجم نمونه و مکانیزم گمشدگی داده‌ها نیز باید توجه نمود. حجم نمونه‌ی کوچک و عدم شناسایی دقیق مکانیزم گمشدگی باعث ایجاد اریبی در برآوردهای حساسیت و ویژگی خواهد شد.[۳] روش تصحیح استاندارد مرجع ناقص، روش دیگری است که در این قسمت به توضیح آن می‌پردازیم. حساسیت و ویژگی یک تست تشخیصی بیان‌گر توانایی یک تست در طبقه‌بندی درست افراد سالم و بیمار است. به بیان دیگر با توجه به جدول ۱ حساسیت، نشانگر نسبت افراد بیماری است که توسط تست تشخیصی، بیمار تشخیص داده شده‌اند ((a/(a+b)) و همچنین ویژگی نشانگر نسبتی از افراد سالم است که توسط تست تشخیصی، سالم تشخیص داده شده‌اند ((d/(d+c)).[۸] در واقع این دو معیار بر اساس مقایسه‌ی نتایج تست تشخیصی با یک استاندارد مرجع تعیین می‌شوند.[۹]

جدول ۱: حساسیت و ویژگی تست تشخیصی

وضعیت بیماری	نتیجه‌ی تست تشخیصی	
	منفی	مثبت
بیمار	b	a
سالم	d	c

a: تعداد مواردی که نتیجه‌ی هر دو تست مثبت است.

b: تعداد مواردی که تست تحت ارزیابی منفی و استاندارد مرجع مثبت است.

c: تعداد مواردی که تست تحت ارزیابی مثبت و استاندارد مرجع منفی است.

d: تعداد مواردی که نتیجه‌ی هر دو تست منفی است.

اما گاهی استاندارد مرجعی که روی افراد اعمال شده است یک استاندارد ناقص است و بنابراین نتایج دارای خطا هستند. تست استاندارد مرجع دارای ۱۰۰٪ حساسیت و ویژگی نیست و تفکیک افراد بیمار از سالم توسط این تست به‌درستی انجام نمی‌شود.[۹] اگر معیارهای حساسیت و ویژگی بدون توجه به رده‌بندی اشتباه نتایج استاندارد مرجع تعیین شوند به‌طور مسلم دارای مقداری اریبی خواهند بود. در این حالت از روش تصحیح استاندارد مرجع ناقص استفاده می‌کنیم. Gart و Buck در سال ۱۹۶۶ برآوردهایی را برای معیارهای دقت تست تحت ارزیابی براساس حساسیت و ویژگی استاندارد مرجع ناقص به‌صورت زیر معرفی کرده‌اند.[۱۰]

$$\hat{Se}_N = \frac{Gs_{p_R} - b}{n(S_{p_R} - 1) + e} \qquad \hat{Sp}_N = \frac{hse_R - c}{nse_R - e} \qquad (۱)$$

در روابط (۱) S_{p_R}، S_{p_R-b}، $\hat{Se}_N$، $\hat{Sp}_N$ به‌ترتیب حساسیت و ویژگی استاندارد مرجع ناقص و برآورد حساسیت و ویژگی تست تحت ارزیابی می‌باشند. تعداد نتایج مثبت هر دو تست با g، تعداد نتایج منفی هر دو تست با d، تعداد مواردی که تست تحت ارزیابی مثبت و استاندارد مرجع منفی تشخیص داده است با c، مواردی که تست تحت ارزیابی منفی و استاندارد مرجع مثبت است با b و مواردی که استاندارد مرجع ناقص، مثبت تشخیص داده است با e نشان داده می‌شود. همچنین تعداد افراد مورد مطالعه را n در نظر می‌گیریم. در برآوردهای فوق مقدار اریبی موجود تصحیح شده است.[۱۰] همچنین به‌کارگیری این روابط مستلزم برقراری فرض استقلال شرطی بین نتایج استاندارد مرجع و تست تحت ارزیابی است.[۳] در شرایطی که معیارهای دقت تست استاندارد مرجع در دسترس نیست می‌توان مدلی که توسط Hu و Walter در سال ۱۹۸۰

معرفی شده است را به‌کار گرفت. در این مدل برآوردهای نااریب براساس ماکسیمم نسبت درست‌نمایی تعیین می‌شوند و نیز وجود فرض استقلال شرطی بین نتایج الزامی است.[۱۱]

برخلاف روش پیش در این حالت نتایج استاندارد مرجع برای همه‌ی افراد موجود است و دیگر نیاز به جایگزینی مقادیر گمشده نیست.[۳] اگر استاندارد مرجع موجود یک استاندارد ایده‌آل باشد مقادیر حساسیت و ویژگی تست تحت ارزیابی بدون توجه به‌میزان شیوع بیماری، ثابت خواهند ماند. در غیر این صورت وجود نتایج ناقص سبب وابستگی این مقادیر به‌میزان شیوع بیماری خواهد شد. به‌طوری‌که بیشترین مقدار ویژگی در کمترین میزان شیوع و بیشترین مقدار حساسیت در بیشترین میزان شیوع اتفاق می‌افتد. Boyko و همکارانش این نتایج را با انجام یک شبیه‌سازی نشان دادند (جدول ۲). در این شبیه‌سازی حساسیت و ویژگی استاندارد مرجع ۹۰٪ در نظر گرفته شده است.[۹]

یک محدودیت مهم روش‌های فوق، وجود فرض استقلال شرطی است که برای خیلی از بیماری‌ها برقراری این شرط امکان‌پذیر نیست. در این صورت برای در نظر گرفتن همبستگی می‌توان از مدل اثرات تصادفی در تحلیل کلاس پنهان استفاده کرد.[۱۲]

روش ساخت استاندارد مرجع روش رایج دیگر در مطالعات تشخیصی فاقد استاندارد طلایی است که در آن نتایج تست‌های مختلف با هم ترکیب می‌شوند. روش‌های ترکیب نتایج برای حالت‌هایی که در آن نتایج استاندارد طلایی دارای نقص است و یا برای همه‌ی افراد مورد مطالعه موجود نیست (وضعیت‌هایی که به‌عنوان شرایط بدون استاندارد طلایی شناخته می‌شوند) متفاوت است.[۳] برای برخی از بیماری‌ها یک استاندارد طلایی مناسب موجود است اما به دلایلی از جمله تهاجمی بودن آن قابل اجرا برای همه‌ی افراد مورد مطالعه نیست.[٤] در تایید افتراقی به‌طور دقیق شرایطی مشابه روش جانهی برقرار است. اما در این حالت، از داده‌های گمشده چشم‌پوشی نمی‌شود و جایگزینی مقادیر گمشده رخ نمی‌دهد بلکه از یک استاندارد مرجع دوم استفاده می‌شود.[۳] برآورد معیارهای دقت حاصل از این روش نیز بدون اریبی نیست و باعث بیش‌برآوردی مقدار حساسیت و کم‌برآوردی مقدار ویژگی می‌شود.[۵] اما در چنین شرایطی بهترین رویکرد، به‌کارگیری یک استاندارد مرجع دیگر است. استاندارد مرجع اول نسبت به استاندارد مرجع دوم دارای حالت تهاجمی و هزینه‌بر است.[۳] به‌طور معمول استاندارد مرجع اول روی افرادی که نتیجه‌ی تست تحت ارزیابی آن‌ها مثبت است اجرا می‌شود و استاندارد دوم روی افرادی اعمال می‌شود که نتیجه‌ی تست آن‌ها منفی است.[۵] به‌عنوان نمونه در تشخیص سرطان مری فقط افرادی که نتیجه‌ی تست توموگرافی گسیل پوزیترون (PET) آن‌ها مثبت است توسط استاندارد مرجع مناسب که حالت تهاجمی دارد (نمونه‌برداری) بررسی می‌شوند.[۳]

روش تایید افتراقی زمانی‌که چندین تست استاندارد قابل قبول در دسترس باشد، رویکرد مناسبی خواهد بود. اما برآوردها در مقایسه با شرایطی که فقط یک استاندارد مرجع داریم دارای مقداری اریبی خواهند بود.[۳و۵] به اریبی حاصل از این روش، "اریبی تایید افتراقی" گفته می‌شود.[۳] جدول ۳ و مقایسه نتایج مراحل ۱ و ٤ نیز بیش‌برآوردی حساسیت و کم‌برآوردی ویژگی را نسبت به حالتی که فقط یک استاندارد مرجع داریم نشان می‌دهد.

جدول ۲: نتایج شبیه‌سازی برای مقادیر حساسیت، ویژگی و شیوع بیماری

شیوع بیماری	حساسیت	ویژگی
۰/۰۱	۰/۲٤	۰/۸۰
۰/۱۰	۰/٤٥	۰/۷۹
۰/۳۰	۰/٦۰	۰/۷۸
۰/۵۰	۰/٦٥	۰/۷۵
۰/۷۰	۰/٦۸	۰/۷۰
۰/۹۰	۰/٦۹	۰/۵۵
۰/۹۹	۰/۷۰	۰/۳٤

جدول ۳: مراحل روش تایید افتراقی

مراحل	نتیجه‌ی تست	وضعیت بیماری	
		سالم	بیمار
۱	مثبت	۸۰	۱٦۰
	منفی	۷۲۰	٤۰
۲	مثبت	۸۰	۱٦۰
	منفی	----	----
۳	مثبت	----	----
	منفی	۷٤۰	۲۰
٤	مثبت	۸۰	۱٦۰
	منفی	۷٤۰	۲۰

همچنین در رویکرد تایید افتراقی، بهتر است مقادیر پیشگوی مثبت و پیشگوی منفی به‌جای حساسیت و ویژگی به‌عنوان معیارهای دقت بیان شوند.[۳و۵]

گاهی عدم وجود استاندارد طلایی برای یک بیماری خاص سبب به‌کارگیری استاندارد مرجع ناقص می‌شود.[۳] اگر معیارهای دقت براساس نتایج استاندارد مرجع ناقص تعیین شوند همواره دارای مقداری اریبی خواهند بود.[٤] پیش‌تر در چنین شرایطی برای کاهش میزان اریبی به تصحیح نقص استاندارد مرجع پرداختیم. حال در این بخش براساس روش تحلیل مغایرت می‌خواهیم میزان اریبی موجود را کاهش دهیم.

با توجه به جدول ٤ این فرآیند در سه مرحله انجام می‌شود. مرحله ۱: ابتدا استاندارد مرجع ناقص روی همه‌ی افراد مورد مطالعه اعمال می‌شود و موارد بیمار و سالم تعیین می‌شوند. خانه‌های a و d خانه‌هایی هستند که در آن‌ها نتایج تست تحت ارزیابی و استاندارد مرجع یکسان هستند (مثبت حقیقی و منفی حقیقی)، این نتایج وارد مرحله دوم نمی‌شوند. مرحله ۲: در این مرحله نتایج خانه‌های b (مثبت کاذب) و c (منفی کاذب) که در آن نتایج دو تست (تست استاندارد مرجع ناقص و تست تحت ارزیابی) یکسان نیستند توسط استاندارد مرجع دوم بررسی می‌شوند. در این خانه‌ها مواردی وجود دارد که توسط استاندارد مرجع دوم مثبت و منفی واقعی تشخیص داده شده‌اند و به‌دلیل نقص استاندارد مرجع به اشتباه در این خانه‌ها قرار گرفته‌اند. مرحله ۳: حال، مواردی که در خانه‌های b و c توسط استاندارد مرجع دوم مثبت و منفی حقیقی تشخیص داده شده‌اند به خانه‌های a و d انتقال می‌یابند. به‌این‌ترتیب در مرحله ۳ نتایج دو تست باهم ترکیب می‌شوند و معیارهای دقت تست تحت ارزیابی تعیین می‌شود.[۱۳]

در این روش همه‌ی افراد مورد مطالعه توسط استاندارد مرجع دوم بررسی نمی‌شوند.[۳] بنابراین زمانی‌که استاندارد دوم یک تست هزینه‌بر و زمان‌بر باشد این رویکرد، مقرون به صرفه خواهد بود.[٤] استاندارد مرجع دوم باید یک تست استاندارد مرجع مناسب باشد و کمابیش هزینه‌برتر از استاندارد مرجع اول است.[۳] اگرچه در این روش از نتایج یک استاندارد مرجع مناسب هم استفاده می‌شود اما برآوردهای حاصل همراه با مقداری اریبی خواهند بود، زیرا تعدادی از افراد توسط استاندارد مرجع ناقص بررسی می‌شوند.[۳و۱۳] یک نقد اساسی که بر این روش وارد است این است که نتایج تست تحت ارزیابی نیز در تعیین وضعیت بیماری افراد دخالت دارند که این خود یک منبع ایجاد اریبی است.[۱٤و۱۵]

روش استاندارد مرجع ترکیبی رویکردی دیگر برای ارزیابی تست‌های تشخیصی است. گاهی برای تشخیص یک بیماری خاص یک تست که به‌عنوان استاندارد طلایی شناخته شده است، در دسترس نیست.[۳] اما چندین تست استاندارد ناقص وجود دارد که براساس روش استاندارد مرجع ترکیبی باهم ترکیب می‌شوند و یک استاندارد مرجع ساخته می‌شود.[۳] در این روش همه‌ی تست‌ها روی همه‌ی افراد مورد مطالعه انجام می‌شوند و نتایج تست‌ها بر اساس دو قانون AND و OR ترکیب می‌شوند.[۱٦] در قانون AND وضعیت بیماری فرد در صورتی مثبت تشخیص داده می‌شود که نتایج تمام تست‌های ناقص

جدول ٤: روش تحلیل مغایرت

مراحل	وضعیت استاندارد مرجع	تست تحت ارزیابی	وضعیت	
			سالم (منفی)	بیمار (منفی)
۱	مرجع ناقص	مثبت	(TP) a	(FP) b
		منفی	(FN) c	(TN) d
۲	مرجع ناقص	مثبت	(TP) a	($FP+TP_b$) b
		منفی	($FP+TN_c$) c	(TN) d
۳	استاندارد طلایی	مثبت	($TP+TP_b$) a	(FP) b
		منفی	(FN) c	($TN+TN_c$) d

TP: مثبت واقعی، TN: منفی واقعی، FP: مثبت کاذب، FN: منفی کاذب، TP_b: تعداد موارد مثبت حقیقی در خانه b، TN_c: تعداد موارد منفی حقیقی در خانه c.

که روی فرد اعمال شده است مثبت باشد و در قانون OR اگر نتیجه‌ی یکی از تست‌ها مثبت باشد، وضعیت فرد مثبت تشخیص داده می‌شود.[۱۵] یک استاندارد مرجع ترکیبی زمانی‌که از ترکیب چندین تست ساخته شده باشد دقت بیشتری خواهد داشت، در مقایسه با زمانی‌که فقط یک تست در ساخت آن دخالت دارد.[۱۶] افزایش تعداد تست‌های ناقص می‌تواند باعث افزایش میزان حساسیت و یا ویژگی تست استاندارد مرجع ترکیبی شود.[۱۷]

اگر در ساخت استاندارد مرجع نتایج براساس قانون OR ترکیب شوند، افزایش تعداد تست‌های ناقص سبب افزایش حساسیت و کاهش ویژگی می‌شود. همچنین اگر قانون AND به‌کار گرفته شود افزایش تعداد تست‌ها سبب افزایش میزان ویژگی و کاهش حساسیت خواهد شد. افزون‌بر تعداد تست‌های ناقص، مقادیر حساسیت و ویژگی آن‌ها نیز نقش مهمی را در تعیین معیارهای استاندارد مرجع ترکیبی ایفا می‌کنند، به‌طوری‌که اگر این تست‌ها در هنگام به‌کارگیری دو قانون AND و OR دارای مقادیر بالای ویژگی و حساسیت باشند دیگر کاهش ویژگی و حساسیت را نخواهیم داشت.[۱۵و۱۶] رویکرد ساخت استاندارد مرجع ترکیبی در مقایسه با روش تحلیل افتراقی رویکردی دقیق‌تر و مناسب‌تر است، زیرا تست تحت ارزیابی که هنوز دقت تشخیصی آن مشخص نیست، هیچ نقشی در تعیین وضعیت بیماری افراد و ساخت استاندارد مرجع ایفا نمی‌کند.[۳] در این رویکرد باید فرض استقلال شرطی، بین نتایج تست تحت ارزیابی و استاندارد مرجع ترکیبی، برقرار باشد. در غیر این‌صورت، برآوردهای حساسیت و ویژگی تست تحت ارزیابی همراه با اریبی خواهند بود.[۱۶]

یک روش دیگر برای بررسی تست‌های تشخیصی تحلیل کلاس پنهان است که به‌عنوان یک روش مهم و پرطرفدار برای ارزیابی تست‌های تشخیصی فاقد استاندارد طلایی شناخته شده است.[۱۸] هنگامی‌که استاندارد طلایی برای یک بیماری در دسترس نباشد وضعیت واقعی بیماری افراد نامعلوم است. در این رویکرد متغیر وضعیت بیماری که به‌صورت بیمار و سالم تعریف می‌شود به‌عنوان متغیر پنهان در نظر گرفته می‌شود. متغیر پنهان متغیری است که به‌صورت مستقیم قابل مشاهده و اندازه‌گیری نیست و توسط متغیرهای آشکار و قابل اندازه‌گیری سنجیده می‌شود. متغیرهای آشکار در این روش نتایج دوحالتی تست‌های ناقص و تست تحت ارزیابی می‌باشد.[۳] مدل آماری که در این رویکرد مجموعه‌ی متغیرهای آشکار و پنهان را به‌هم مرتبط می‌سازد مدل تحلیل کلاس پنهان نام دارد.[۱۷] مدل تحلیل کلاس پنهان در سال ۱۹۵۰ توسط Lazarsfeld معرفی شد.[۱۸] مدل احتمالی که ایده اصلی مدل تحلیل کلاس پنهان است به‌صورت مدل (۲) تعریف می‌شود.[۲۰و۲۱]

$$P(Y=y)=\sum_{x=1}^{C}p(X=x)p(Y=y\mid X=x) \qquad (۲)$$

یک شرط اساسی برای این مدل برقراری فرض استقلال شرطی بین نتایج تست‌ها است.[۱۸] تحت این فرض داریم:

$$P(Y=y)=\sum_{x=1}^{C}p(X=x)\prod_{i=1}^{L}p(Y_i=y_i\mid X=x) \qquad (۳)$$

در مدل (۳) متغیرهای x، Y_i، i، C، به‌ترتیب نشان‌دهنده‌ی متغیر پنهان (وضعیت واقعی بیماری)، L امین متغیر آشکار (تست‌های ناقص و تست تحت ارزیابی)، تعداد تست‌های ناقص و آشکار، تعداد کلاس‌های پنهان (بیمار، سالم) می‌باشند شود.[۱۹و۲۰] در مدل (۳) پارامترهای $p(X=1)$، $p(Y_1=1\mid X=0)$ و $p(Y_1=1\mid X=1)$ بیان‌گر شیوع، نرخ مثبت کاذب و نرخ مثبت حقیقی می‌باشد.[۲۰] بنابراین با برآورد آن‌ها می‌توان شیوع بیماری، حساسیت و ویژگی تست تحت ارزیابی و تست‌های ناقص دیگر را تعیین کرد.[۳] یک روش برای برآورد پارامترهای مدل تحلیل کلاس پنهان، روش ماکسیمم درست‌نمایی است.[۲۱] شرط استفاده از روش ماکسیمم درست‌نمایی این است که درجه آزادی $(2^l - 1)$ بزرگ‌تر و یا مساوی تعداد پارامترهای مدل $(2l + 1)$ باشد.[۳و۲۰] در صورت عدم برقراری این شرط می‌توان با اعمال محدودیت روی پارامترها و یا با استفاده از رویکرد بیزی پارامترهای مدل را برآورد کرد. به‌طور نمونه فرض کنید افزون‌بر تست تحت ارزیابی تنها یک تست ناقص در مدل باشد $(l = 2)$، در این حالت پارامترهایی که باید برآورد شوند عبارت‌اند از حساسیت و ویژگی تست تحت ارزیابی، حساسیت و ویژگی تست ناقص و شیوع، بنابراین پنج پارامتر باید برآورد شود که این تعداد بیش‌تر از مقدار درجه‌ی آزادی (سه) است و این یعنی شرط لازم برای به‌کارگیری روش ماکسیمم درست‌نمایی وجود ندارد. در این حالت می‌توان روی پارامترها اعمال محدودیت کرد. به‌طور نمونه با در نظر گرفتن فرضیات موجود می‌توان حساسیت و ویژگی دو تست را یکسان در نظر گرفت (سه پارامتر). اما بهترین روش برای برآورد پارامترها استفاده از رویکرد بیزی است.[۳]

در میان همه‌ی روش‌های بیان‌شده رویکرد تحلیل کلاس پنهان از مطلوبیت بالایی برخوردار است.[۱۸] این روش در بسیاری از مطالعات از جمله مطالعات تشخیصی کاربرد فراوانی دارد.[۲۰] براساس مدل تحلیل کلاس پنهان و با استفاده از قوانین بیزی می‌توان وضعیت بیماری افراد مشخص کرد و آن‌ها را به دو کلاس سالم و بیمار تقسیم کرد.[۲۱] در این رویکرد در صورت عدم برقراری فرض استقلال شرطی حتما باید همبستگی در مدل در نظر گرفته شود، در غیر این‌صورت برآورد پارامترها دارای اریبی خواهند بود.[۱۴و۲۰] یک نقد وارد بر این روش این است که نتایج به‌طور کامل براساس مدل آماری بوده و هیچ تعریف بالینی در ارتباط با بیماری مورد بررسی در نظر گرفته نمی‌شود.[۱۴و۲۲و۲۳]

References

1. Linnet K, Bossuyt PM, Moons KG, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. *Clin Chem* 2012;58(9):1292-301.
2. Kumar R. Evaluation of diagnostic tests. *Clin Epidemiol Glob Health* 2016;4(2):76-9.
3. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 2007;11(50):iii, ix-51.
4. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *J Clin Epidemiol* 2009;62(8):797-806.
5. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140(3):189-202.
6. Toutenburg H. Rubin, DB: Multiple imputation for nonresponse in surveys. *Stat Pap* 1990;31(1):180.
7. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. *Test (Madr)* 2009;18(1):1-43.
8. Goroll AH, May LA, Mulley AG, editors. Primary Care Medicine: Office Evaluation and Management of the Adult Patient. 3rd ed. Philadelphia, PA: Lippincott-Raven; 1995.
9. Boyko EJ, Alderman BW, Baron AE. Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. *J Gen Intern Med* 1988;3(5):476-81.
10. Buck AA, Gart JJ. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. American Journal of Epidemiology. 1966;83(3):593-602.
11. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics. 1980 Mar 1:167-71.
12. Basso W, Hartnack S, Pardini L, Maksimov P, Koudela B, Venturini MC, et al. Assessment of diagnostic accuracy of a commercial ELISA for the detection of Toxoplasma gondii infection in pigs compared with IFAT, TgSAG1-ELISA and Western blot, using a Bayesian latent class approach. *Int J Parasitol* 2013;43(7):565-70.
13. Miller WC. Bias in discrepant analysis: when two wrongs don't make a right. *J Clin Epidemiol* 1998;51(3):219-31.
14. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med* 1999;18(22):2987-3003.
15. Miller WC. Editorial response: can we do better than discrepant analysis for new diagnostic test evaluation? *Clin Infec Dis* 1998;27(5):1186-93.
16. Schiller I, van Smeden M, Hadgu A, Libman M, Reitsma JB, Dendukuri N. Bias due to composite reference standards in diagnostic accuracy studies. *Stat Med* 2016;35(9):1454-70.
17. Smeden M. Diagnostic research in the absence of a gold standard. [Dissertation on the Internet]. Netherlands: Utrecht University; 2016 [cited 2019 Jun 15]. Available from: https://dspace.library.uu.nl/handle/1874/327738
18. Ling DI, Pai M, Schiller I, Dendukuri N1. A Bayesian framework for estimating the incremental value of a diagnostic test in the absence of a gold standard. *BMC Med Res Methodol* 2014;14:67.
19. Dey A. A review on applications of latent class analysis. *Indo-Asian J Multidiscip Res (IAJMR)* 2015;2(4):681-6.
20. Vermunt JK, Magidson J. Latent class cluster analysis. In: Hagenaars JA, McCutcheon AL, editors. Applied Latent Class Analysis. New York, NY: Cambridge University Press; 2002; P. 89-106.
21. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. *Stat Methods Med Res* 1998;7(4):354-70.
22. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. *Biostatistics* 2007;8(2):474-84.
23. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford: Oxford University Pres; 2003.

TUMJ Tehran University Medical Journal, February 2019; Vol. 76, No. 11: 708-714 Review Article

Comparing the evaluation methods of diagnostic tests in the absence of the gold standard: *review article*

Fatemeh Masaebi M.Sc.[1]
Farid Zayeri Ph.D.[2*]
Malihe Nasiri Ph.D.[1]
Mehdi Azizmohammad Looha M.Sc.[1]

1- Department of Biostatistics, Faculty of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
2- Department of Biostatistics, Proteomics Research Center, Faculty of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

* Corresponding author: Faculty of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Darband St., Qods Sq., Tehran, Iran.
Tel: +98 21 22707347
E-mail: fzayeri@gmail.com

Abstract

Received: 11 May 2018 Revised: 18 May 2018 Accepted: 09 Feb. 2019 Available online: 19 Feb. 2019

Considering the advancement of medical sciences, diagnostic tests have been developed to distinguish patients from healthy population. Therefore, Determining and evaluation of the diagnostic accuracy tests is of great importance. The accuracy of a test under evaluation is determined through the amount of agreement between its results with the results of the gold standard, and this test accuracy can be defined based on sensitivity, specificity, positive predictive value, negative predictive value and the area under the receiver operative characteristic curve (AUC). Gold standard is an accurate and error- free method to determine the presence or absence of disease of interest and classify patients, which is not available in some diseases and situations as this method is costly or invasive. In these cases, reference standard is a best available replacement method to be used by physicians to diagnostic disease. However, in some situation, the acceptable reference standard is invasive or costly and does not exist or unreliable. It can be imperfect and results of the reference standard method are not necessarily error- free and cannot be applied to everyone in the study; all these cases point to the conditions in which the gold standard is not available. The use of reference standard including error causes to incorrect separation of patients from healthy population and thus, it cannot be a comparing measure for other diagnostic tests and its results are inaccurate. Therefore, other alternatives methods are needed for evaluation and determine the diagnostic accuracy tests when the gold standard does not exist. Imputation method, correct imperfect reference standard method, the construct reference standard method, latent class models, differential verification, composite reference standard and discrepant analysis are of these alternative methods. Each of these methods, considering its features, advantages, and limitations can be used to evaluate the accuracy of diagnostic test in the absence of gold standard. The present study gave an overview of methods to evaluation of diagnostic accuracy tests when there is no gold standard and the focus of this study was on explain the concept of these solutions, review and compare them and their strengths and weaknesses.

Keywords: accuracy, diagnostic test, reference standards.