Journal Club

پنجمین جلسه

ژورنال کلاب در دانشگاه علوم پزشکی بابل

مرکز آموزشی درمانی کودکان امیرکلا

جلسه پنجم: پنجشنبه ۲۲ فروردین ۱۳۹۸

موضوع:
مطالعات ارزش تشخیصی

طراحی، تدوین و تدریس: دکتر عباس کشتکار
عضو هیات علمی
دانشگاه علوم پزشکی تهران، دانشکده بهداشت

برنامه راه‌اندازی
ژورنال کلاب اختصاصی
متخصصین و اعضای هیات علمی بالینی

A Keshtkar MD, MPH, PhD / abkeshtkar@gmail.com

دانشگاه علوم پزشکی بابل

دانشگاه علوم پزشکی بابل

معاونت تحقیقات و فناوری

جلسه پنجم ژورنال کلاب

Journal Club

ژورنال کلاب اختصاصی

متخصصین و اعضای هیات علمی بالینی

موضوع این جلسه: **مطالعه ارزش تشخیصی در فیلد کودکان**

A Keshtkar MD, MPH, PhD / abkeshtkar@gmail.com

آمار افزار

ارائه مقاله جلسه پنجم ژورنال کلاب

European Journal of Pediatrics
https://doi.org/10.1007/s00431-019-03362-3

ORIGINAL ARTICLE

Comparison of diagnostic accuracy of screening tests ALT and ultrasound for pediatric non-alcoholic fatty liver disease

Laura G. Draijer[1] · Sana Feddouli[1] · Anneloes E. Bohte[2] · Olga vd Baan Slootweg[3] · Tammo H. Pels Rijcken[4]
Marc A. Benninga[1] · Jaap Stoker[2] · Bart G. P. Koot[1]

Draijer, Laura G., et al. "Comparison of diagnostic accuracy of screening tests ALT and ultrasound for pediatric non-alcoholic fatty liver disease." *European journal of pediatrics* (2019): 1-8.

تعریف «مطالعه ارزش تشخیصی»

- مطالعه ارزش تشخیصی، یک مطالعه پژوهشی است که هدف آن ارزیابی اعتبار / کارآیی یا ارزش تشخیصی حداقل یک «آزمون» یا «تست» در جداسازی افراد مبتلا به یک «بیماری» یا یک «پیامد سلامتی»، از افراد بدون آن بیماری / پیامد است.
- در یک مطالعه ارزش تشخیصی، نتیجه سنجش یک «آزمون» (تست، مارکر، نشانگر یا ... که Index Test نام دارد) با وضعیت ابتلاء به یک بیماری (Condition)، مورد ارزیابی قرار میگیرد. بطور معمول، تایید ابتلاء یا وقوع بیماری / پیامد مورد نظر، از طریق یک آزمون بسیار معتبر (تا حد امکان) که به Reference Test (استاندارد طلایی) مشهور است، صورت میگیرد.
- گاهی تایید ابتلاء به بیماری مورد نظر، از طریق یک آزمون صورت نگرفته و میتواند از طریق یک فرایند بالینی باشد (براساس تعریف بالینی بیماری و نظر متخصص مربوطه). فلذا «استاندارد طلایی» هواره یک آزمون بیولوژیک نیست.

مولفه‌های مهم در یک «مطالعه ارزش تشخیصی» (۱)

- **آزمون مرجع (Reference Test, Reference Standard, Gold Standard):** آزمونی است که تا زمان ارزیابی، به معتبرترین شکل ممکن پیامد / بیماری مورد مطالعه را نشان میدهد.
- **آزمون مورد ارزیابی (Index Test):** آزمون، نشانگر یا تست یا آزمونهایی هستند که در فرایند اجرایی یک مطالعه ارزش تشخیصی، میخواهیم آن آزمون / آزمونها را ارزیابی کنیم.
- **کاربرد آزمون یا Intended Use of the TEST:** هدف از سنجش یا بکارگیری آزمون مورد نظر در شرایط عملی است. برای کاربردهای پزشکی (سلامت)، معمولاً این کاربردها به موارد زیر طبقه‌بندی میگردند:
  - تشخیص Diagnosis
  - غربالگری Screening
  - پیش آگهی (پروگنوز) Prognosis
  - پایش بیماری Monitoring
  - طبقه‌بندی Staging
  - پیش‌بینی رخداد بیماری Prediction

انواع طبقه‌بندی در «مطالعات ارزش تشخیصی» (۱)

- **طبقه‌بندی از نظر هدف / کاربرد مطالعه:** منظور دسته بندی این مطالعات از نظر Intended Use of Test میباشد. هر چند همانطور که ذکر شد، از این منظر کاربردهای آزمونها یا نشانگرها، متعدد است، لیکن حداقل سه کاربرد زیر در این مطالعات از بقیه مهمتر و پرکاربردتر است:
  - تشخیص Diagnosis
  - غربالگری Screening
  - پیش آگهی (پروگنوز) Prognosis

- **طبقه‌بندی از نظر طراحی مطالعه:** از نظر Design یا طراحی، این دسته از مطالعات، بیشتر مشابه یا نزدیک به مطالعات مشاهده‌ای – تحلیلی میباشند (حتی تا قبل از سالهای ۳–۲۰۰۲، این مطالعات زیرمجموعه مطالعات مشاهده‌ای گروه بندی میشدند. اما بهر حال این مطالعات، از طراحیهای زیر استفاده میکنند:
  - مقطعی Cross-sectional (معمولاً شایعترین طراحی هستند)
  - مورد شاهدی Case-Control (معمولاً ضعیفترین طراحی هستند)
  - کوهورت Cohort

انواع طبقه‌بندی در «مطالعات ارزش تشخیصی» (۲)

- **طبقه‌بندی از نظر ماهیت Index Test:** ماهیت تست / آزمون مورد ارزیابی در این مطالعات، میتواند دارای انواع زیر باشد:
  - آزمونها / تستهای بیولوژیک
  - مدل پیش‌بینی یا تصمیم‌گیری (ترکیب خطی یا غیرخطی از متغیرهای متعدد مرتبط با بیماری / پیامد مورد مطالعه)

- **طبقه‌بندی از ماهیت (نوع) متغیر Index Test:** در این مطالعات، صرف نظر از اینکه ماهیت ایندکس تست چه باشد، از نظر نوع متغیرها، ایندکس تست میتواند به یکی از انواع زیر دسته بندی گردد:
  - کمّی: آزمون یا نشانگر بیولوژیک و یا ترکیب خطی (در مدلهای پیش بینی) دارای مقادیر یا ارزش عددی برای ایندکس تست مورد ارزیابی است (غالب موارد)
  - کیفی (طبقه بندی شده): آزمون یا نشانگر بیولوژیک، دارای پاسخ عددی یا کمّی نبوده و بجای آن از دسته بندی (اسمی یا رتبه‌ای) این آزمون یا نشانگر استفاده میگردد.

Prediction Model / Risk Prediction Model Synonyms

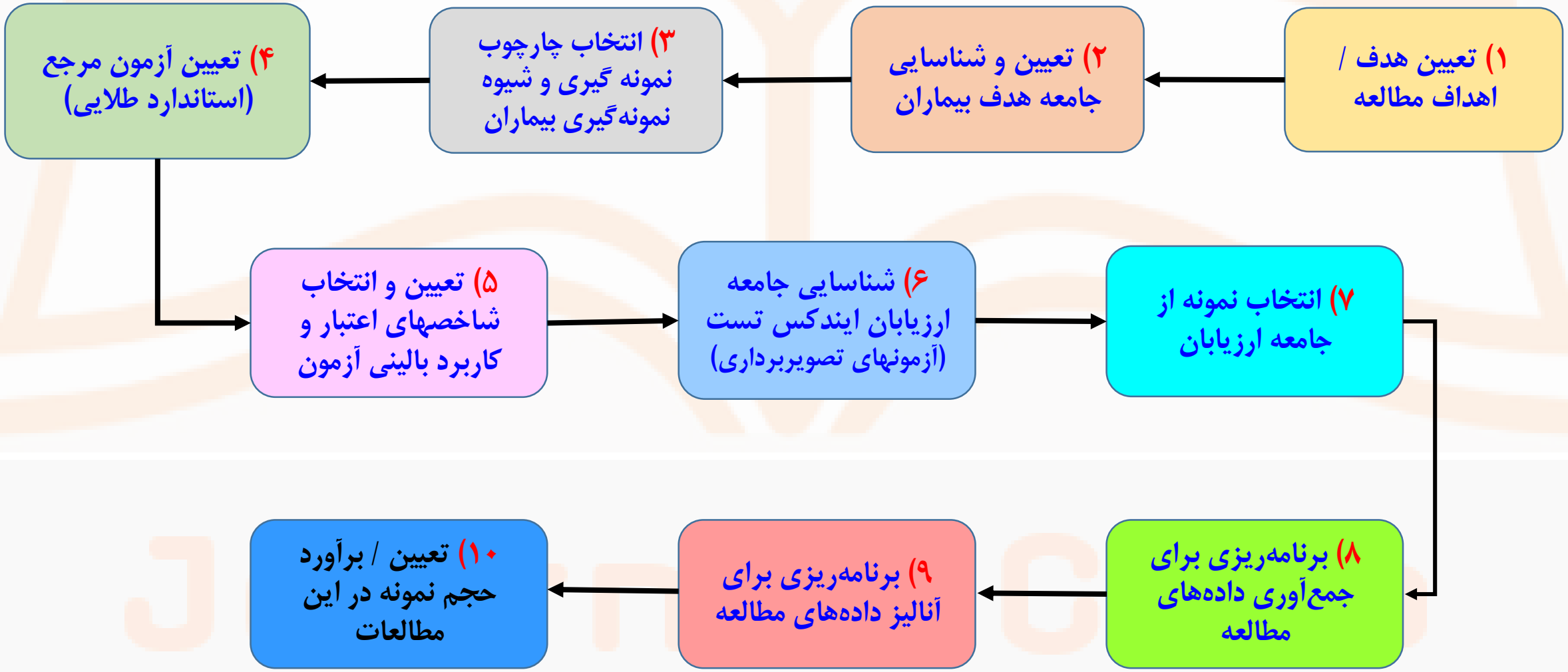
- Clinical Prediction Rule (CPR)
- Clinical Decision Support System (Clinical DSS)
- Risk Scoring System (RSS)
- Risk Scoring Prediction Rule (RSPR)
- Clinical Decision Rule (CDR)
- Clinical Decision System (CDS)
- …

Framingham Coronary RPM was 1st RPM in a chronic disease field that published 1976. (Kannel WB et al. *Am J Cardiol* 1976; 38: 46-51)

چرا طراحی و کاربرد مدلهای پیش بینی خطر در کنار موضوع «مطالعات ارزش تشخیصی» مطرح میشود؟

فرایند اجرایی این دو دسته مطالعات، انواع خطاها و سوگراییها، تجزیه و تحلیل آماری، شاخصهای Performance یا کاربرد آزمونها یا مدلها و جایگاه استفاده یا بهره برداری از آنها بسیار به یکدیگر نزدیک میباشند.

۱۰ گام اساسی در طراحی «مطالعات ارزش تشخیصی»



Measures of Test Accuracy

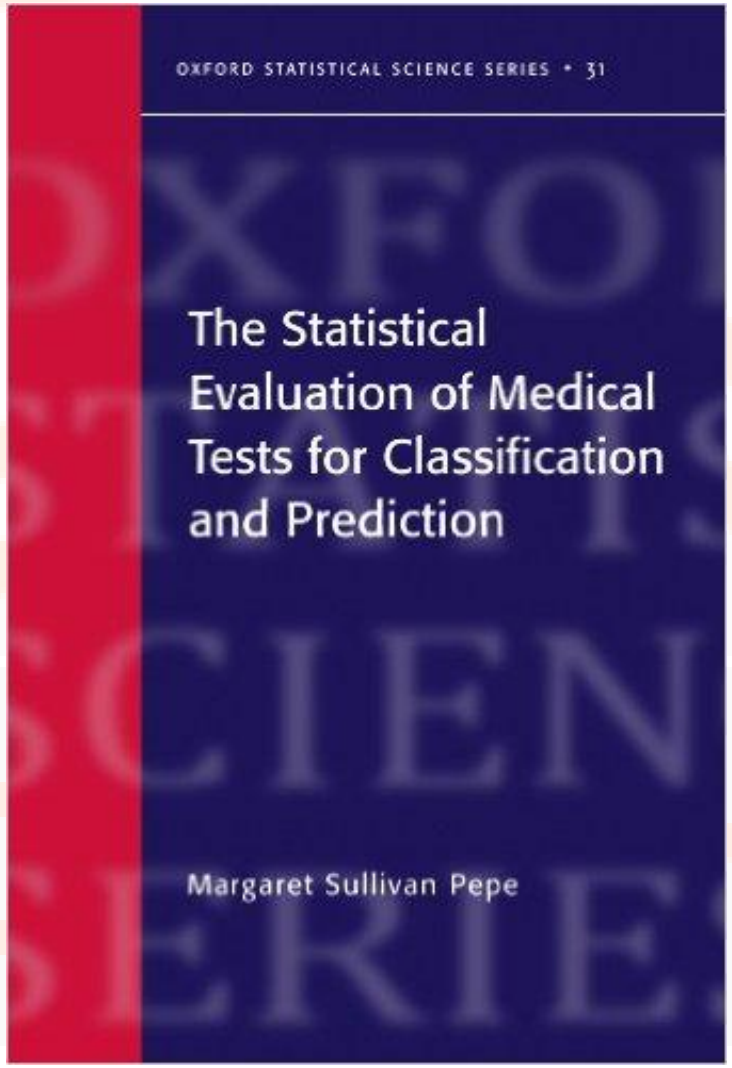
- **Dual Measures:**
  - Extrinsic Measures:
    - PPV (Positive Predictive Value)
    - NPV (Negative Predictive Value)
  - Intrinsic Measures:
    - Sensitivity
    - Specificity
    - LR+ (Positive Likelihood Ratio)
    - LR- (Negative Likelihood Ratio)
- **Single Measures:**
  - DOR (Diagnostic Odds Ratio)
  - AUC / AUROC (Area Under Curve / Area Under RO Curve)
  - Youden's Index

	with disease	no disease	total
test +	True Positive (a)	False Positive (b)	450
test -	False Negative (c)	True Negative (d)	550
total	500	500	1000

Measures of Test Accuracy

- **Dual Measures:**
  - Are indicated to the both components of + and – test
  - Are appropriate for **clinical judgments**
  - Are categorized to the **EXTRINSIC** and **INTRINSIC** measures (based on the Disease **PREVALENCE** dependency)
- **Single Measures:**
  - are combined the above both components
  - Are appropriate for:
    - Combining the test accuracy measure in **META-ANALYSIS**
    - Comparing the accuracy of **different TESTS** of any given disease
    - Comparing the accuracy of any given test between **different STUDIES**

منبع اصلی



Margaret Sullivan Pepe,
Professor of Biostatistics
Fred Hutchinson Cancer Research Center
University of Washington,
School of Public Health and Community Medicine

A Practical Book that the examples were prepared based on STATA modules and commands, edition 2003

سوءبرداشتها در مطالعات ارزش تشخیصی

Researchers' Challenges in Designing, Conducting & Interpreting DAS (1)

۱) دو واژه یا ترم «رابطه» و «قدرت تشخیصی» بطور نابجا، مترادف پنداشته میشوند:

- ***Scenario A:*** **Glut-1 as a prognostic biomarker in oral squamous cell carcinoma**

BACKGROUND: Glut-1 is a glucose transporter protein, the expression of which is upregulated in malignant cells which show increased glucose uptake. …..

The objectives of the present study were to compare the expression of Glut-1 in normal persons and in patients with oral squamous cell carcinoma (OSCC), to correlate the expression of Glut-1 with respect to clinical staging of OSCC and to evaluate the expression of Glut-1 with respect to different histopathological grades of OSCC.

MATERIALS AND METHODS: 30 cases of OSCC were staged clinically and graded histopathologically. Immunohistochemical method was used to detect the expression of Glut-1 in OSCC and the same was compared with 30 normal subjects. The scores were compared using the chi-square test.

RESULTS: Glut-1 expression was detected in all grades of OSCC. A significant correlation with a P value of 0.00004 was found in immunostaining between normal and OSCC. The expression of Glut-1 was significant when compared with different clinical stages with significant P value of 0.0004 and in different histopathological grades of OSCC with a P value of 0.00001.

CONCLUSION: Higher immunohistochemical staining scores were obtained with increased clinical staging and histopathological grades of OSCC. ***High expression of Glut-1 may be related to poor prognosis in OSCC***.

Harshani, J. M., Yeluri, S., & Guttikonda, V. R. (2014). Glut-1 as a prognostic biomarker in oral squamous cell carcinoma. *Journal of oral and maxillofacial pathology: JOMFP*, *18*(3), 372.

Researchers' Challenges in Designing, Conducting & Interpreting DAS (1)

۱) دو واژه یا ترم «رابطه» و «قدرت تشخیصی» بطور نابجا، مترادف پنداشته میشوند: (ادامه...)

- *Scenario A:* Glut-1 as a prognostic biomarker in oral squamous cell carcinoma

IHC scores	Normal	%	OSCC	%	Total
Score 0	18	60.00	1	3.33	19
Score 1	5	16.67	10	33.33	15
Score 2	4	13.33	14	46.67	18
Score 3	3	10.00	5	16.67	8
Total	30	100.00	30	100.00	60

Chi-square test=22.9332, df=3, P=0.00004, Chi-square test; P=0.00004, Significant, OSCC: Oral Squamous Cell Carcinoma

	OSCC	Normal	total
IHC score >= 2	19	7	26
IHC score < 2	11	23	34
total	30	30	1000

Sn= 19 / 30 = 63.3% Sp= 23 / 30 = 76.7%

PPV= 19 / 26 = 73.1% NPV= 23 / 34 = 67.7%

DOR= 19 × 23 / 7 × 11 = 5.68

Harshani, J. M., Yeluri, S., & Guttikonda, V. R. (2014). Glut-1 as a prognostic biomarker in oral squamous cell carcinoma. *Journal of oral and maxillofacial pathology: JOMFP*, *18*(3), 372.

Researchers' Challenges in Designing, Conducting & Interpreting DAS (1)

نتیجه‌گیری: (۱)

- یافتن رابطه معنی‌دار آماری بین یک «تست» یا هر نوع «مارکر زیستی»، با یک بیماری، پیامد سلامتی و یا هر نوع واقعه (پیامد نامطلوب در مقابل پیامد مطلوب، مرگ در مقابل بقاء و ...)، **نمیتواند** نشانگر ارزش تشخیصی (Diagnostic) یا پیش‌آگهی (Prognostic) باشد.
- مقالاتی مشابه سناریوی A در ژورنالهای علمی (حتی گاهی ژورنالهای معتبر و قوی) کم نیستند (بیش از ۱۰۰ مقاله فقط در PubMed یافته‌ام)، متاسفانه در دهه فعلی 2010s، این چالش (خطای فاحش تفسیری) به مقالات مرور سیستماتیک و متاآنالیز نیز گسترش یافته است (متاسفانه با موضوع یا هدف دو مقاله سناریوهای مذکور، مقالات مرور سیستماتیک و متاآنالیز مربوطه نیز منتشر شده است). به عبارت دیگر، خطای تفسیری به خطا در کاربرد یا Malpractice منجر خواهد شد.
- برای اینکه یک «تست» یا «مارکر زیستی»، قدرت یا ارزش تشخیصی / پیش‌آگهی (Diagnostic or Prognostic Value) داشته باشد، باید رابطه بسیار قوی با بیماری / پیامد مورد نظر داشته باشد. بدیهی است که با شاخصهای آماری از قبیل p value نمیتوان تفاوت بین قدرت رابطه در شرایط مختلف را تشخیص داد و باید از شاخصهای «قدرت رابطه» یا Effect Size استفاده نمود.

دانشگاه علوم پزشکی بابل
آمرافزار

Researchers' Challenges in Designing, Conducting & Interpreting DAS (1)

نتیجه‌گیری: (۲)

- بهترین گزینه بعنوان شاخص قدرت رابطه در ارزیابی یک «تست / بیومارکر»، شاخص **نسبت برتری تشخیصی** یا Diagnostic Odds Ratio (DOR) میباشد.
- **خصوصیات مثبت** شاخص نسبت برتری تشخیصی، عبارتند از:
  - ✓ با شاخص هم نام خود در مطالعات مشاهده‌ای تحلیلی Analytical OR ، از نظر محاسبه (در جدول ۲ در ۲)، یکسان است.
  - ✓ خوشبختانه مانند هر یک از زوج شاخصهای (Dual Indicator) مطالعات ارزش تشخیصی (حساسیت، ویژگی، ارزش اخباری مثبت، ارزش اخباری مثبت و منفی، نسبت درستنمایی مثبت و منفی)، واجد اطلاعات تک بعدی نبوده و اطلاعات زوج شاخصها را با یکدیگر ترکیب مینماید.
  - ✓ چون شاخص OR، پرکاربردترین شاخص ES در مطالعات پژوهشی سلامت و پزشکی است، پس محققین آشنایی قابل قبولی با این شاخص دارند.
  - ✓ در شرایطی که محققین نسبت به کاربرد یا ارزش تشخیصی یا پیش‌آگهی تستی تردید داشته باشند، میتوانند با استفاده از طبقه‌بندیهای قدرت رابطه در این شاخص، به تردید خود فایق آیند.

دانشگاه علوم پزشکی بابل
آمرافزار

Researchers' Challenges in Designing, Conducting & Interpreting DAS (1)

نتیجه‌گیری: (۳)

- اگر یک تست یا بیومارکر تشخیصی / پروگنوستیک مورد سوال، دارای **حساسیت و ویژگی** برابر با ۷۰ درصد باشد، ببینیم در این شرایط، چه مقداری از DOR، خواهیم داشت؟
- مقدار DOR در شرایط فوق، برابر با ۵.۴ میباشد. حال اگر وضعیت شاخصهای این تست یا بیومارکر بهبود یابد، مقادیر بزرگتری از شاخصهای حساسیت و ویژگی خواهیم داشت. در شرایط با حساسیت و ویژگی ۸۰ درصد، مقدار DOR برابر با ۱۶ خواهد بود.

	GS+	GS-	total
Test +	70	30	100
Test -	30	70	100
total	100	100	200

	GS+	GS-	total
Test +	80	20	100
Test -	20	80	100
total	100	100	200

Researchers' Challenges in Designing, Conducting & Interpreting DAS (1)

نتیجه‌گیری: (۴)

- از منظر دیگر نیز میتوان به این موضوع نگریست؛ همانطورکه میدانید، **علایم و نشانه‌های** بسیاری از بیماریها، روابط حتی قوی با بیماری مورد نظر دارند، اما متخصصین از این علام و نشانه‌ها برای تشخیص (بعنوان مارکر تشخیصی)، **استفاده نمیکنند** (مانند درد قفسه سینه در سندرم حاد قلبی).
- بر اساس منبع زیر، در زنان، حساسیت و ویژگی درد شانه ۴۵% و ۶۷% و درد بازو ۴۹% و ۶۹% برای سندرم حاد قلبی میباشد، حال با محاسبه شاخص DOR به ازای هر علامت (به ترتیب ۲.۵۳ و ۲.۱۵ بود)، میتوان به این واقعیت پی برد که مارکرهایی که مقدار عددی DOR آنها به این مقادیر نزدیک است، در این محدوده دارای قدرت تشخیصی هستند.

	GS+	GS-	total
Shoulder Pain+	45	33	78
Pain -	55	67	122
total	100	100	200

DeVon, Holli A., et al. *Journal of the American Heart Association* 3.2 (2014): e000586.

	GS+	GS-	total
Test +	49	31	80
Test -	51	69	120
total	100	100	200

دانشگاه علوم پزشکی بابل
آمرافزار

Researchers' Challenges in Designing, Conducting & Interpreting DAS (2)

۲) متخصصین در تفسیر یک «تست» یا «مارکر زیستی» در سطح فردی، دو شاخص «حساسیت» و «ویژگی» را مورد استفاده قرار میدهند (بجای دو شاخص «ارزش اخباری مثبت و منفی»):

- **تعریف حساسیت:** توانایی طبقه‌بندی درست بیماران بر اساس نتیجه تست.
  - حساسیت ۸۵ درصد در یک تست، یعنی اینکه ۸۵ درصد بیماران بر حسب نتیجه تست مورد نظر، درست طبقه بندی میگردند (۸۵ درصد بیماران، تست + هستند)
- **تعریف ویژگی:** توانایی طبقه بندی درست افراد سالم براساس نتیجه تست.
  - ویژگی ۹۰ درصد در یک تست، یعنی اینکه ۹۰ درصد افراد سالم، بر حسب نتیجه تست مربوطه، بدرستی طبقه بندی میگردند (۹۰ درصد افراد سالم، تست – هستند)
- **تعریف ارزش اخباری مثبت:** توانایی طبقه بندی درست افراد تست مثبت بر حسب وضعیت پیامد یا بیماری مورد مطالعه
  - ارزش اخباری مثبت ۶۰ درصد در یک تست، یعنی اینکه ۶۰ درصد افراد تست مثبت، بدرستی در جایگاه مورد نظر، طبقه بندی میشوند (۶۰ درصد تست مثبت ها، بیمار میباشند)
- **تعریف ارزش اخباری منفی:** توانایی طبقه بندی درست افراد تست منفی بر حسب وضعیت پیامد یا بیماری مورد مطالعه
  - ارزش اخباری منفی ۷۰ درصد در یک تست، یعنی ۷۰ درصد افراد تست منفی، بدرستی در جایگاه مورد نظر، طبقه بندی میشوند (۷۰ درصد تست منفی ها، سالم میباشند)

دانشگاه علوم پزشکی بابل
آمار افزار

Researchers' Challenges in Designing, Conducting & Interpreting DAS (3)

۳) متخصصین در استفاده از اطلاعات شاخصهای اعتبار یک آزمون از روی منابع (مطالعات سایر کشورها)، به نوع شاخص توجهی ندارند.

- **Diagnostic Accuracy Measures:**
  - **Extrinsic Measures:**
    - PPV (Positive Predictive Value)
    - NPV (Negative Predictive Value)
  - **Intrinsic Measures:**
    - Sensitivity
    - Specificity
    - LR+ (Positive Likelihood Ratio)
    - LR- (Negative Likelihood Ratio)

مثال: در یک مقاله، به اطلاعات ارزش تشخیصی یک مارکر در پیش آگهی سرطانهای کولون فامیلیال در یک مجموعه کلینیک ارجاعی فوق تخصصی با طراحی مورد شاهدی، در یکی از کشورهای اسکاندیناوی پرداخته بود. آیا میتوانیم این اطلاعات را بدون ملاحظه خاصی در شرایط کلینیکهای گوارش دانشگاه علوم پزشکی بابل استفاده نمود؟

Journal Club

Researchers' Challenges in Designing, Conducting & Interpreting DAS (3)

۳) متخصصین اطلاعات شاخصهای اعتبار یک آزمون در منابع را برای شرایط کاربردی بیماران خود، استفاده مینمایند

Mammography & Breast Ca Pathology

	Breast ca +	Breast ca -	total
mamm +	29	19	48
Mamm -	1	11	12
total	30	30	60

Prevalence = 50%
Sensitivity = 29 / 30 = 96.7%
Specificity = 11 / 30 = 36.7%

	Breast ca +	Breast ca -	total
mamm +	29	1881	1910
Mamm -	1	1089	1090
total	30	2970	3000

Prevalence = 1%
Sensitivity = 29 / 30 = 96.7%
Specificity = 1089/2970 = 36.7%

Researchers' Challenges in Designing, Conducting & Interpreting DAS (3)

۳) متخصصین اطلاعات شاخصهای اعتبار یک آزمون در منابع را برای شرایط کاربردی بیماران خود، استفاده مینمایند

Mammography & Breast Ca Pathology

	Breast ca +	Breast ca -	total
mamm +	29	19	48
Mamm -	1	11	12
total	30	30	60

Prevalence = 50%
Sensitivity = 29 / 30 = 96.7%
Specificity = 11 / 30 = 36.7%
PPV = 29 / 48 = 60.4 %
NPV = 11 / 12 = 91.7 %

	Breast ca +	Breast ca -	total
mamm +	29	1881	1910
Mamm -	1	1089	1090
total	30	2970	3000

Prevalence = 1%
Sensitivity = 29 / 30 = 96.7%
Specificity = 1089/2970 = 36.7%
PPV = 29 / 1910 = 1.5 %
NPV = 1089 / 1090 = 99.9 %

پایان
جلسه پنجم