12. Psychologická metodologie v praxi klinického neuropsychologa

Štěpán Bahník, Eva Rubínová

Ačkoli se může zdát, že metodologie není v praxi klinického neuropsychologa důležitá, opak je pravdou. Pro správný výběr diagnostické metody je potřeba znát psychometrické vlastnosti různých metod a vědět, jaký mají praktický význam. Pro stanovení diagnózy je nutné rozumět možnostem interpretace testů a umět rozpoznat faktory, které interpretaci ovlivňují. Po stanovení diagnózy může být vhodné využít terapii či rehabilitaci. Pro výběr vhodného postupu je potřeba rozpoznat efektivní metodu od metody neefektivní a k tomu je nutné umět kriticky zhodnotit dostupné výzkumy těchto metod a rozumět jejich interpretaci. Tato kapitola se postupně věnuje všem zmíněným oblastem. Jelikož diagnostika tvoří převážnou část práce klinického neuropsychologa (Rabin, Barr & Burton, 2005), je jí věnována většina této kapitoly.

Diagnostika

Výběr diagnostické metody

Diagnostický proces vždy začíná výběrem metody, která je pro daný účel nejvhodnější. V případě klinického neuropsychologa se nejčastěji jedná o zhodnocení kognitivního profilu, rozsahu kognitivního deficitu, případně jeho progrese nebo zlepšení. Odborný výběr metody na základě její relevance k posuzované problematice, spolehlivosti a využitelnosti jejích výsledků je odpovědností každého psychologa.

Základními předpoklady kvalitního testu jsou standardizace a existence norem použitelných v populaci, se kterou se psycholog v praxi setkává. V případě, že je na výběr z většího množství standardizovaných testů[1] s vhodnými normami, měli bychom se orientovat podle dvou základních charakteristik – reliability a validity. Reliabilita se váže k spolehlivosti měření daného testu a v souladu s cílem diagnostiky je žádoucí, aby chyba měření byla co nejnižší. Validita, která je při volbě testu klíčová, určuje oblasti diagnostiky či predikce, pro které je vhodné jej použít.

[1] Ačkoli je v této části často řeč o psychologických testech, popsané principy lze aplikovat i při využití jiných diagnostických metod. Termíny diagnostická metoda a test jsou tedy používány zaměnitelně.

Standardizace a normy

Cílem neuropsychologické diagnostiky je obvykle zjistit úroveň výkonu jedince a porovnat ji s hodnotami srovnatelné populace. Aby to v praxi bylo možné, metoda musí být administrována v jednotném formátu, se stejnými instrukcemi a předem daným postupem. Právě proces standardizace vede k vytvoření těchto pravidel, která následně umožňují srovnání výsledku jedince s populačními normami, tedy výkonem srovnatelné skupiny lidí.

Každá psychometrická metoda by měla mít důkladně popsanou standardní proceduru administrace a měla by upozorňovat zejména na ty problematické oblasti, které mohou získané skóry nejvíce ovlivnit. Pokud nejsou tato pravidla dodržována, získané výsledky budou pravděpodobně zatíženy chybou a není pak možné spolehnout se na deklarované hodnoty reliability ani na validitu metody (Fischer & Milfont, 2010). Například u administrace paměťového testu může výsledek ovlivnit rychlost čtení materiálu nebo způsob zaznamenávání odpovědí při čtené rekognici, kdy verbálním (např. hodnocením „správně“) i neverbálním chováním (např. úsměvem) můžeme poskytovat zpětnou vazbu, která ovlivní následné testové chování.

Specifickým problémem standardizace je přejímání zahraničních metod, které je nutné přeložit a adaptovat na naše kulturní prostředí. Častý postup je přímý překlad pokud možno „slovo od slova“, který však ignoruje lingvistická specifika jazyků, což může v důsledku ztížit srozumitelnost a v krajním případě posunout celý význam přeloženého textu. V optimálním případě by proto měl překlad vycházet z odborné znalosti daného konceptu, specifické pro cílové kulturní prostředí.

Kulturní specificita se také může odrážet v mnohem komplexnější problematice, kterou je srovnatelnost s normami. Můžeme použít normy většinové populace pro hodnocení výkonu jedince z určité menšiny? Odpověď závisí na tom, zda daná charakteristika je nebo není závislá na znacích, ve kterých se dané skupiny liší (např. jazyk, přístup ke vzdělání, jeho hodnota pro danou kulturu). Pokud nemáme adekvátní normy k dispozici, mělo by k této skutečnosti být přihlíženo při interpretaci výsledků.

Kvalitu norem lze obecně ohodnotit podle toho, nakolik je vzorek, ze kterého pocházejí, reprezentativní vůči populaci, pro niž je test zamýšlen. Normy by měly reflektovat odlišnosti rozložení skórů dle klíčových charakteristik, které je ovlivňují – k nim patří obvykle základní demografické charakteristiky jako věk, pohlaví a vzdělání, ale někdy také třeba inteligence. Například u paměťových testů s věkem klesá výbavnost a mírně se zvyšuje variabilita skórů (zejména u nejstarší části populace, viz např. Mitrushina et al., 2005). Rozdíly mezi pohlavími

ve smyslu lepšího výkonu žen se objevují nejen v typicky uváděných jazykových testech, jako je verbální fluence, ale také ve verbálních paměťových testech (Herlitz, Nilsson & Bäckman, 1997). S vyšším vzděláním a inteligencí se často také zvyšují průměrné dosažené skóry a celková konzistence výkonu. Při interpretaci bychom měli brát tyto faktory v úvahu i v případě, že nemáme odpovídajícím způsobem rozvrstvené normy k dispozici. Budeme-li např. v praxi hodnotit výkon vysokoškolského profesora, jehož skór se bude pohybovat v blízkosti jedné standardní odchylky pod průměrem pro danou věkovou kategorii, pravděpodobně budeme odhadovat narušení v dané oblasti i přesto, že stanovená hranice pro kognitivní deficit nebyla překročena. Zcela jinak bychom takový výkon hodnotili u jedince se základním vzděláním, kde výsledek může reprezentovat normální výkon.

Reliabilita

Jakýkoli diagnostický nástroj bude vykazovat určitou chybu měření. V klasické testové teorii se skládá celkový skór z pravého skóru a chyby. Výsledek v testu tedy vždy obsahuje část variability tzv. pravého skóru, část variability pak můžeme přičíst chybě měření. Reliabilita vyjadřuje, do jaké míry je celkový skór prostý chyby, a koeficient reliability měření pak značí podíl pravé variance na celkové varianci (Peter, 1979):

$$r_{xx} = \frac{V_{pravá}}{V_{pozorovaná}}$$

Jelikož pravou varianci nemůžeme znát, koeficient reliability se obvykle odhaduje z chybové a celkové variance jako:

$$r_{xx} = \frac{V_{pozorovaná} - V_{chybová}}{V_{pozorovaná}}$$

Chybová variance se nejčastěji odhaduje pomocí opakovaného zadání testu, zadání alternativní verze testu, variability výsledků v jednotlivých položkách testu či variability hodnocení různých hodnotitelů. Tyto různé způsoby se pak obvykle označují jako re-testová reliabilita, reliabilita alternativních forem, split-half reliabilita (nebo obecněji vnitřní konzistence) a inter-rater reliabilita.

Re-testová reliabilita se odhaduje zadáním stejného testu stejným lidem v různém čase. Problémem u re-testové reliability je závislost výkonu na předchozí administraci testu – při opakování testu, a to i s delším odstupem, jsou skóry ovlivněny efektem učení. Re-testová reliabilita také výrazně závisí na době, která odděluje obě zadání testu. Kratší doba bude spjatá s menší odlišností výsledků při obou zadáních, a bude tedy naznačovat vyšší reliabilitu.

U delší doby můžeme očekávat menší vliv efektu učení a zapamatování předchozích odpovědí, odhad chybové variability bude ale zkreslen variabilitou pravého skóru, který se také může měnit.

Využití alternativních forem testu alespoň částečně odstraňuje některé problémy odhadu re-testové reliability (např. efekt učení), odhad chybové variability však může být ovlivněn odlišnostmi jednotlivých forem. V určitých případech může být reliabilita alternativních forem odhadována oddáleným zadáním obou forem testu − pak můžeme předpokládat menší vliv efektu učení, ale také nižší odhadovanou reliabilitu. U komplexnějších metod (např. Wechslerova inteligenční škála) však obvykle nejsou alternativní formy k dispozici.

Split-half reliabilita se odhaduje z jednoho měření, a tak není ovlivněna efektem učení. Chybová variabilita se zde odhaduje pomocí variability ve výsledcích ze dvou polovin testu.[2] Protože existuje mnoho způsobů, jak test rozdělit na dvě poloviny, používá se pro odhad obvykle Cronbachovo α, které je rovno průměru split-half reliability získané ze všech možných dělení testu na poloviny (Cronbach, 1951).[3] Jelikož split-half reliabilitu lze zjistit z jednoho zadání testu − a Cronbachovo α je tak pro vyjádření reliability testu často používané −, je vhodné zmínit určitá jeho omezení a nepochopení s ním spjatá. Předně Cronbachovo α neukazuje, do jaké míry test měří pouze jeden psychologický atribut. Cronbachovo α může snadno nabývat vysokých hodnot, i pokud test měří dva odlišné atributy. Homogenitu testových položek je tedy nutné zjistit jiným způsobem (Schmitt, 1996). Cronbachovo α nezávisí pouze na vnitřní konzistenci testu, je také výrazně ovlivněno délkou testu. Nelze tedy tvrdit, že vyšší Cronbachovo α znamená nutně vyšší vnitřní konzistenci (Cortina, 1993; Streiner, 2003). Vyšší Cronbachovo α nemusí také nutně znamenat lepší test. K vyšší hodnotě např. povedou velice podobné položky, nicméně velice podobné položky budou pravděpodobně měřit příliš úzce vymezený psychologický atribut (Streiner, 2003). Takovýto test tedy bude mít nejspíše vysoce homogenní položky a vysoké Cronbachovo α, ale nízkou užitečnost.

Inter-rater reliabilita se zjišťuje pomocí vztahu mezi hodnocením různých hodnotitelů. Ze své povahy má smysl pouze u metod, kde je hodnocení závislé na úsudku hodnotitele.

---

[2] Jelikož s rostoucím množstvím položek lze očekávat vyšší reliabilitu testu, při výpočtu split-half reliability se obvykle velikost vztahu mezi oběma polovinami testu upraví o vliv sníženého množství položek tak, aby výsledný odhad reliability odpovídal celému testu, a ne pouze jeho polovině.

[3] Ve skutečnosti rovnost Cronbachova α průměru split-half reliabilit závisí na výpočtu split-half reliability a rovnosti standardních odchylek jednotlivých položek (Cortina, 1993).

Například u kresebných nebo projektivních metod může nespolehlivost úsudků výrazně snižovat reliabilitu.

Lze si všimnout, že různé formy reliability značí odlišné formy zobecnitelnosti výsledků v testu. Například vysoká re-testová reliabilita může naznačovat zobecnitelnost testového výsledku v čase, split-half reliabilita může naznačovat zobecnitelnost mezi možnými položkami a inter-rater reliabilita naznačuje zobecnitelnost výsledků z hlediska různých hodnotitelů. Reliabilita alternativních forem pak může naznačovat jak zobecnitelnost výsledků v čase, tak zobecnitelnost mezi možnými položkami. Z tohoto pozorování vychází tzv. teorie zobecnitelnosti (generalizibility theory; Peter, 1979; Shavelson, Webb, & Rowley, 1989), která místo různých forem reliability klade důraz na zobecnitelnost testových výsledků. Bližší popis této teorie je nad rámec tohoto textu, nicméně rozpoznání faktu, že chybová variabilita může být způsobena více faktory, a tudíž je třeba dbát na všechny formy zobecnitelnosti a nepoužívat jen vybrané formy reliability, si zaslouží zmínku i zde.

Validita

I v případě, že je test reliabilní, má vytvořené normy a existuje standardizovaný způsob jeho použití, nemusí být validní a v důsledku toho ani užitečný. Reliabilita a standardizace testu tak jsou nutnými, nikoli však dostačujícími vlastnostmi validního testu. Validitu můžeme považovat za vlastnost značící, zda test měří to, co má měřit. Předpokladem validity testu je, že měřený atribut (např. verbální fluence nebo úzkostnost) existuje a jeho variance ovlivňuje výsledek měření (Borsboom, Mellenbergh & van Heerden, 2004).[4]

U validního testu očekáváme, že naměřené výsledky budou pozitivně korelovat s výsledky jiných testů měřících stejný atribut (konvergentní validita)[5] a že nebudou korelovat s výsledky testů, které měří jiný atribut (diskriminační validita; Foster & Cone, 1995). Lze tak např. předpokládat, že lidé budou mít podobné skóry v různých validních testech inteligence. Na druhou stranu nebudeme očekávat silný vztah mezi skóry v testech inteligence a úzkostnosti.

Validní test určitého atributu by měl měřit pouze daný atribut, a to zcela (obsahová validita). Od validního testu inteligence tak očekáváme, že nebude měřit schopnosti pod inteligenci

[4] Existují i jiná pojetí testové validity, která nesouhlasí se zde zmíněnou definicí a zde popsaným pojetím validity (viz např. Cizek, 2012; Newton, 2012; Strauss & Smith, 2009).

[5] Ačkoli jsou zde zmíněny termíny běžně používané při dělení validity, je vhodné podotknout, že odpovídají spíše různým formám validizace než formám validity. Je tedy přesnější tvrdit, že validitu testu podporuje jeho korelace s jinými testy měřícími stejný atribut, než že test vykazuje konvergentní validitu (Borsboom et al., 2004).

nespadající (např. pozornost) a že bude pokrývat celou škálu schopností, které pod inteligenci spadají. Validní test by měl být z hlediska měřeného atributu také reprezentativní, což znamená, že by měl pokrývat měřený atribut proporcionálně. Pokud se např. určitá porucha vyznačuje několika rysy, měl by být všem těmto rysům v testu přidělen význam do té míry, do jaké jsou pro tuto poruchu charakteristické (Haynes, Richard & Kubany, 1995).

Pro praktické účely jsou klíčovými vlastnostmi validního testu schopnost rozpoznat určitý stávající stav (souběžná validita; např. rozpoznat poranění mozku podle nízkých skórů v testu) a schopnost předpovídat budoucí chování či stav (prediktivní validita). U testu exekutivních funkcí můžeme např. očekávat, že bude s jeho pomocí do určité míry možné předpovídat možnost návratu do zaměstnání po úrazu postihujícím frontální lalok. Z hlediska využitelnosti testu v praxi je však nutné zmínit možné problémy odhadu prediktivních schopností testu.

Validizační studie mohou probíhat za jiných podmínek, než jaké jsou v klinické praxi, a tak se skutečná a ve studii zjištěná prediktivní schopnost testu mohou lišit. Důvodem může být např. odlišnost populace použité ve studii a populace běžné v klinické praxi. Pokud jsou tedy do studie vybráni lidé, kteří jsou jednoznačně nemocní, a lidé, kteří jsou jednoznačně zdraví, může být prediktivní schopnost testu značně nadhodnocená. Například studie zjišťující schopnost testu rozlišit mezi zdravými lidmi a pacienty s mírnou kognitivní poruchou (mild conitive impairment – MCI) nebo demencí (Yamamoto et al., 2004) bude pravděpodobně nadhodnocovat rozlišovací schopnosti testu v praxi, kde je klíčové rozlišovat pouze mezi zdravými lidmi a lidmi s deficitem na úrovni MCI (viz např. Nunes et al., 2008).

Přestože je prediktivní schopnost testu v praxi klíčová, test nelze považovat za validní měřítko atributu pouze na základě jeho korelace s tímto atributem. Hmotnost člověka např. může silně korelovat s výškou, to ale neznamená, že můžeme považovat váhu za validní měřítko výšky (Borsboom et al., 2004). Důsledkem by mohlo být, že objevení pozitivní korelace mezi hmotností a rizikem úmrtí na kardiovaskulární choroby by bylo bráno jako důkaz vztahu mezi výškou a rizikem úmrtí na tyto choroby.

Nakonec je zde dobré zmínit specifický problém odhadu prediktivní schopnosti testu, pokud jsou jeho položky vybírány na základě vztahu s predikovaným jevem. Pokud je učiněn odhad užitečnosti ze stejných dat, podle kterých byly vybírány položky, bude nutně nadhodnocen (Cureton, 1950). Analogický problém je charakteristický pro studie využívající funkční magnetickou rezonanci (fMRI), genomové studie či studie hledající fyziologické markery onemocnění. Obecně jde o studie zjišťující velké množství potenciálních vztahů s vybraným ukazatelem bez existence předchozí hypotézy o konkrétním vztahu. Například pokud bychom

se snažili zjistit vztah BOLD signálu (blood oxygen level dependent – signálu závislého na obsahu kyslíku v krvi) měřeného fMRI a výskytem depresivních myšlenek a odhalili bychom, že určitá předem neurčená oblast koreluje silně a statisticky významně s námi vybraným ukazatelem, neznamená to, že můžeme pomocí BOLD signálu v této oblasti spolehlivě předpovídat výskyt depresivních myšlenek. Jelikož statisticky signifikantní budou v takovémto výzkumu pouze velmi silné vztahy (z důvodu korekce na mnohonásobná srovnávání), půjde obvykle pouze o vztahy, které silně nadhodnocují reálný efekt. Zjištěná velikost efektu tedy nemůže sloužit jako odhad skutečného efektu (Vul et al., 2009).

Usuzování z diagnostických metod

Poté, co vybereme test a použijeme jej, je nutné výsledky interpretovat a dospět k diagnóze. V této části je představena Bayesova věta, kterou můžeme po zjištění výsledků použít k výpočtu pravděpodobnosti existence poruchy. Kromě Bayesovy věty je dále popsána ROC analýza – formální postup, jenž může napomoci při určení prahu pro stanovení diagnózy. V některých případech nám nicméně mohou scházet informace nutné k použití těchto metod. V takových případech lze diagnózu stanovit v případě odlišnosti výsledku v testu od dřívějšího výsledku či v případě, že výsledek je abnormálně nízký. Všechny tyto zde popsané metody jsou příklady tzv. mechanických postupů úsudku (Dawes, Faust & Meehl, 1989). Druhý způsob, jak lze úsudku dosáhnout, je pomocí klinického postupu spočívajícího na úsudku psychologa, jeho zkušenostech a intuici. Tento postup je v praxi běžný, ale mnoho studií potvrzuje, že vede k horším výsledkům (Grove et al., 2000). Důvodem jsou např. usuzovací zkreslení (judgmental biases) či nereliabilita lidského úsudku (Gilovich, Griffin & Kahneman, 2002). Klinické prostředí v psychologii také obvykle neposkytuje příležitosti k učení se z chyb, které jsou klíčové pro účinnost klinického postupu (Kahneman & Klein, 2009).

Bayesova věta

Představme si zjednodušenou situaci, se kterou se můžeme setkat v rámci diagnostického procesu. U vyšetřovaného (72 let) předpokládáme kognitivní deficit na úrovni MCI a pro ověření této domněnky použijeme Paměťový test učení. Z ilustračních důvodů se bude jednat o test, ve kterém lze dosáhnout pouze dvou výsledků – nízkého nebo vysokého skóru.[6] K tomu, abychom mohli odhadnout, jaká je pravděpodobnost, že vyšetřovaný má MCI, potřebujeme znát dvě vlastnosti testu: a) jaká je pravděpodobnost, že vyšetřovaný bude mít

[6] Obecnějším případem testů využívajících kontinuální škálu se zabývá další část kapitoly.

nízký skór, pokud MCI má (senzitivita testu); b) jaká je pravděpodobnost, že vyšetřovaný bude mít nízký skór, pokud MCI nemá (pravděpodobnost tzv. falešně pozitivního výsledku nebo také 1 - specificita testu).[7] První pravděpodobnost zapíšeme formálně jako $P(D|H)$, druhou pak jako $P(D|\neg H)$, přičemž D značí důkaz (nízký skór) a H značí naší hypotézu (pacient má MCI). Předpokládejme první pravděpodobnost rovnou 0,76 a druhou pravděpodobnost rovnou 0,16 (hodnoty podle norem ze studie Estévez-González et al., 2003). Necelých osm z deseti vyšetřovaných s MCI tedy bude mít nízký skór v testu. U lidí bez MCI lze očekávat nízký skór u šestnácti ze sta vyšetřovaných.

Pacient v testu neuspěl a nás zajímá, jaká je pravděpodobnost, že MCI má, tj. $P(H|D)$. K výpočtu použijeme Bayesovu větu:

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}$$

Pravděpodobnost $P(D|H)$ již známe, zbývá tedy zjistit pravděpodobnost $P(H)$ a $P(D)$. $P(H)$ značí pravděpodobnost hypotézy, což je pravděpodobnost, s jakou jsme mohli očekávat MCI u vyšetřovaného před jeho testováním, tedy přirozený výskyt tohoto onemocnění v dané populaci (prevalence). Význam této pravděpodobnosti je klíčový a v usuzování často opomíjený (Kahneman & Tversky, 1973). Tato pravděpodobnost se totiž může značně lišit podle testové situace, a tak má výrazný vliv na náš výsledný odhad. Například pokud by se k nám vyšetřovaný dostavil na doporučení praktického lékaře, který u vyšetřovaného již delší dobu pozoruje zapomínání, můžeme očekávat, že tato pravděpodobnost bude vyšší, než pokud by k nám přišel vyšetřovaný na základě informací z veřejné vzdělávací kampaně. V našem případě přišel vyšetřovaný spíše z preventivních důvodů. Budeme tedy předpokládat, že nemáme příliš indikací pro MCI, a tak stanovíme $P(H)$ dle prevalence v běžné populaci na 0,08 (Hänninen, Hallikainen, Tuomainen, Vanhanen & Soininen, 2002). MCI tedy očekáváme zhruba u osmi ze sta lidí ve věku 70–76 let, které testujeme v podobné situaci.

K výpočtu nám zbývá znát již jen pravděpodobnost důkazu $P(D)$, což je pravděpodobnost, že vyšetřovaný bude mít nízký skór. Tuto pravděpodobnost přitom můžeme vypočítat z dříve specifikovaných údajů jako součet pravděpodobnosti, že vyšetřovaný dosáhne nízkého skóru a bude mít MCI, a pravděpodobnosti, že vyšetřovaný dosáhne nízkého skóru a nebude mít MCI. Formálně tuto větu vyjádříme jako $P(D) = P(D \wedge H) + P(D \wedge \neg H)$. Pravděpodobnost, že vyšetřovaný dosáhne nízkého skóru a zároveň bude mít MCI, můžeme jinak také vyjádřit jako

[7] Specificita testu vyjadřuje pravděpodobnost, že vyšetřovaný bude mít vysoký skór, pokud poškození nemá.

násobek pravděpodobnosti výskytu MCI před vyšetřením (tj. prevalence) a pravděpodobnosti nízkého skóru u lidí s MCI (tj. senzitivity), formálně $P(D \wedge H) = P(H) \times P(D|H)$. Tento závěr si můžeme číselně představit následujícím způsobem. Pokud bude mít MCI 8 ze 100 podobně vyšetřovaných (vycházíme z $P(H) = 0{,}08$) a přibližně 6 z 8 vyšetřovaných s MCI bude mít nízký skór (předpokládané $P(D|H) = 0{,}76$), pak z původních 100 vyšetřovaných bude mít 6 současně nízký skór i MCI, tj. pravděpodobnost $P(D \wedge H)$ je 0,06. Podobně můžeme zapsat pravděpodobnost, že vyšetřovaný bude mít nízký skór, a přitom nebude mít MCI, jako $P(D \wedge \neg H) = P(\neg H) \times P(D|\neg H)$, kde pravděpodobnost absence MCI je $P(\neg H) = 1 - P(H) = 0{,}92$. Nyní získané rovnice dosadíme do původního vyjádření Bayesovy věty a dostaneme:

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(H) \times P(D|H) + P(\neg H) \times P(D|\neg H)}$$

Pokud dosadíme námi vybraná čísla, obdržíme:

$$P(H|D) = \frac{0{,}76 \times 0{,}08}{0{,}08 \times 0{,}76 + 0{,}92 \times 0{,}16} = \frac{0{,}0608}{0{,}0608 + 0{,}1472} \cong 0{,}29$$

Pravděpodobnost, že vyšetřovaný má MCI, tedy bude pouze 0,29, a to i přesto, že dosáhl nízkého skóru a test měl přiměřené psychometrické vlastnosti – senzitivitu 0,76 a specificitu 0,84.

Konkrétní příklad využití Bayesovy věty sloužil pouze pro demonstraci několika obecnějších principů tvorby úsudku z testu. Nejpodstatnějším z nich je důležitost pravděpodobnosti výskytu poškození v relevantní populaci. Tato pravděpodobnost je totiž obvykle v praxi neznámá a psycholog ji tedy musí odhadovat na základě dostupných dat, své zkušenosti a svého úsudku. Z dostupných dat mu může napomoci především prevalence poškození (či obecně jevu) v relevantní části populace. Tato hodnota může být využitelná např. v případě preventivního screeningu, kde neočekáváme selektivní vyšetřování lidí s poškozením. Povětšinou je ji však potřeba upravit, neboť vyšetření bude probíhat spíše u lidí, u nichž je nějaké poškození indikováno. V takových případech nás zajímá prevalence u vyšetřovaných lidí, nikoli prevalence v celé populaci. Stejně tak by neměl náš odhad spočívat pouze na výsledku testu, ale měl by přikládat váhu věku vyšetřovaného i dalším faktorům ovlivňujícím pravděpodobnost onemocnění u vyšetřovaného.

Na našem příkladu je také možné pozorovat důležitost psychometrických vlastností využité diagnostické metody. Pokud bychom např. počítali se senzitivitou 0,65 nebo 0,90 místo 0,76, bude vypočítaná pravděpodobnost 0,26, resp. 0,33 místo námi zjištěné 0,29. Pokud bychom

počítali se specificitou 0,75 a 0,95 místo 0,84, dostali bychom pravděpodobnost 0,21, resp. 0,57. Je tedy zřejmé, že s rostoucí specificitou a senzitivitou roste také užitečnost využití testu. Tu přitom můžeme vidět ve zvýšení pravděpodobnosti poškození z původních 0,08 na 0,29 v případě nízkého skóru, který jsme v příkladu předpokládali (užitečnost testu při vysokém skóru by byla vidět ve snížení pravděpodobnosti poškození).

Lze dodat, že využití Bayesovy věty se neomezuje na případy, kdy využíváme pro diagnostiku jediný test. Pokud bychom použili další test, brali bychom jako pravděpodobnost výskytu poškození $P(H)$ námi vypočítanou pravděpodobnost (vypočítaná pravděpodobnost 0,29 by tedy hrála při použití dalšího testu roli původní pravděpodobnosti 0,08).[8]

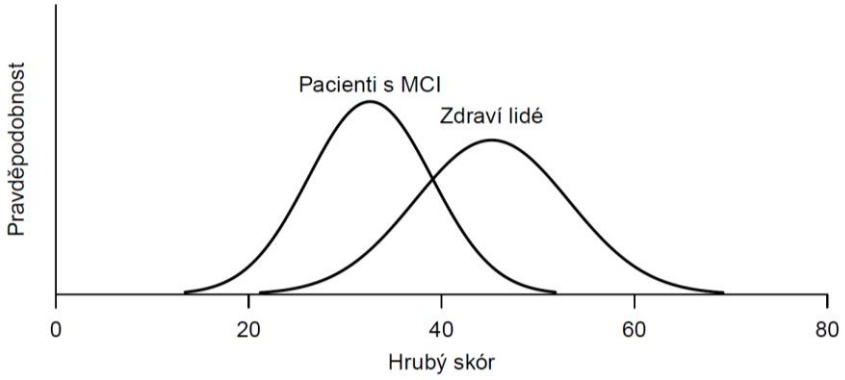
Křivka ROC (receiver operating characteristic)

Bayesova věta je při rozhodování o diagnóze užitečná, ale častěji než s testy, které mají pouze dva možné výsledky, se v neuropsychologii setkáme s testy, jejichž výsledkem je skór, který může nabývat většího množství hodnot. Příklad z minulé části můžeme tedy lehce upravit. Opět vyšetřujeme pacienta, u kterého je podezření na MCI, častou preklinickou fázi Alzheimerovy nemoci. V Paměťovém testu učení (v součtovém skóru prvního až pátého pokusu) lze nyní dosáhnout výsledků na škále od 0 do 75. V tomto případě specificita a senzitivita testu závisí na stanoveném prahu pro diagnózu.

Užitečnost diagnostické metody v klinické praxi vychází z toho, že s její pomocí můžeme rozlišit lidi s určitým znakem od lidí bez tohoto znaku. V našem příkladu je tímto znakem MCI a test je tedy užitečný na základě toho, že lidé s MCI se liší ve výsledcích v testu od těch bez MCI. Pacienti s MCI budou mít v našem případě horší paměťové funkce než zdraví lidé, a tedy budou mít nižší průměrný výsledek v zadaném testu. Příklad rozložení skóru lze vidět na obrázku 1.[9]

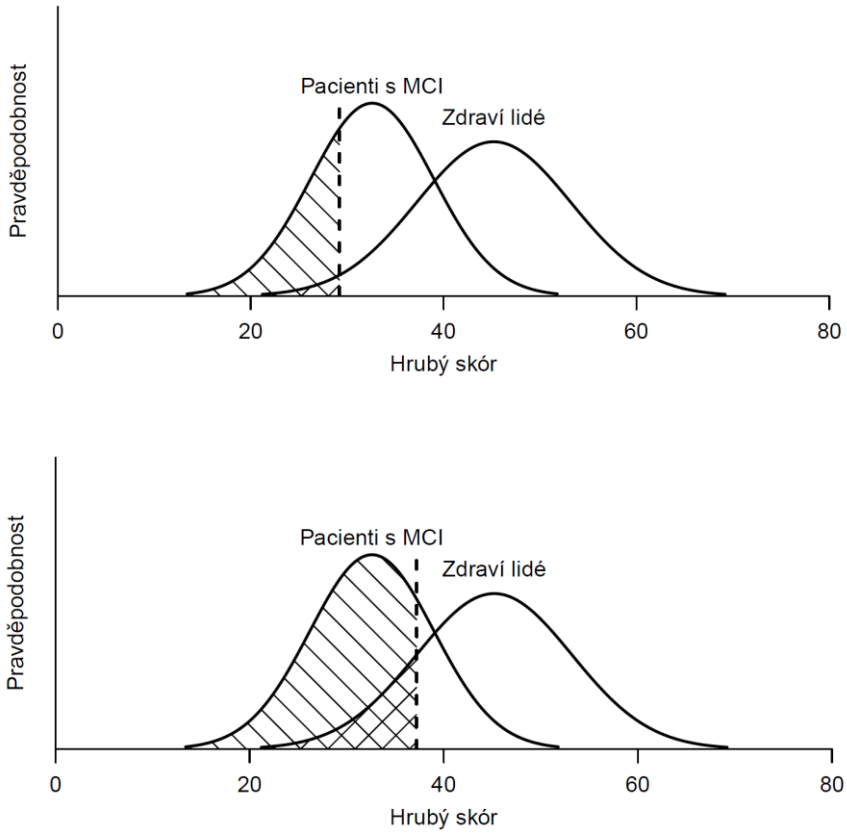
[8] Výpočet by se dále nelišil v případě, že jsou výsledky obou testů na sobě nezávislé. Pokud by byly výsledky testů na sobě závislé, je potřeba počítat s podmíněnými pravděpodobnostmi závislými na výsledcích prvního testu.

[9] Ačkoli v testu lze dosahovat pouze celočíselných skórů, na obrázcích je výsledek v testu pro jednoduchost brán jako spojitá proměnná.



Obr. 1. Příklad rozložení skórů v testu u pacientů s MCI a zdravých lidí

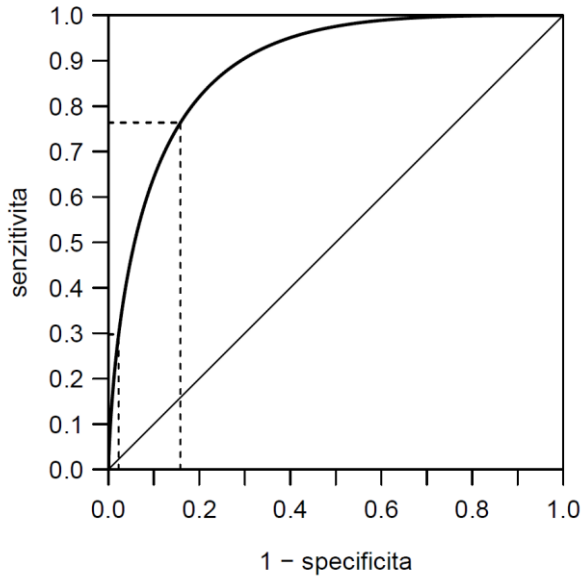
Pacienti s MCI mají v tomto případě průměrný skór 32,6 (SD = 6,4), zdraví lidé 45,2 (SD = 8,0) (Estévez-González et al., 2003). Vliv volby prahu pro stanovení poškození na specificitu a senzitivitu testu je znázorněn na obrázcích 2a a 2b. Na obrázku 2a je zvolen jako práh skór 37 (1 SD pod průměrem zdravé populace) a na obrázku 2b je práh 29 (2 SD pod průměrem zdravé populace). Šrafováním jsou znázorněné případy, u kterých usoudíme, že vyšetřovaný má MCI. Je vidět, že pokud zvýšíme práh, stoupne množství případů, kdy se rozhodneme pro to, že vyšetřovaný má diagnózu, a to jak u pacientů s MCI, tak u zdravých lidí. To znamená, že se zvýšením prahu se zvyšuje senzitivita testu (s větší pravděpodobností řekneme, že má vyšetřovaný diagnózu, když ji skutečně má) a snižuje se specificita (s větší pravděpodobností řekneme, že má vyšetřovaný diagnózu, když ji ve skutečnosti nemá).



Obr. 2. Vliv volby prahu pro stanovení diagnózy na počet pacientů s MCI a zdravých lidí označených diagnózou

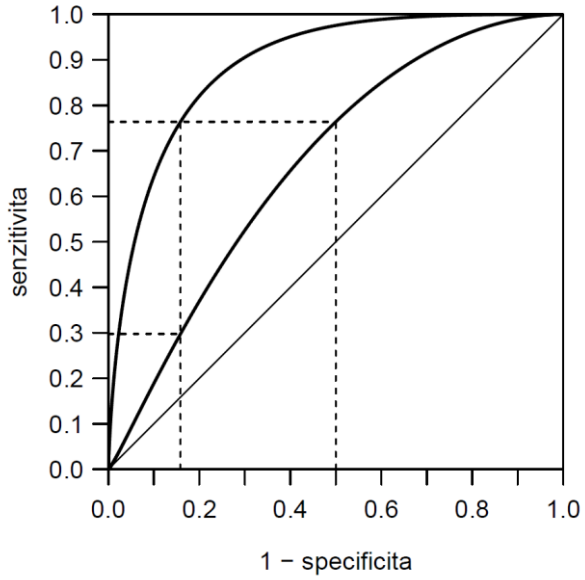
Pro znázornění vlivu všech možných hodnot prahu na specificitu a senzitivitu testu se používá tzv. ROC křivka (Fawcett, 2006; Swets, Dawes & Monahan, 2000). ROC křivku pro náš příklad můžeme vidět na obrázku 3.

V grafu jsou přerušovanými čarami znázorněny dva uvedené prahy (29 nalevo a 37 napravo). Z grafu lze tedy např. vyčíst, že práh 37 vede k senzitivitě 0,76 a specificitě 0,84. Kombinací specificity a senzitivity, které leží nalevo nahoře od křivky, nemůžeme s využitím diagnostického nástroje dosáhnout a kombinace pod křivkou jsou z hlediska využitého testu neefektivní. Linie propojující body (0, 0) a (1, 1) odpovídá hádání. Pokud budeme např. v náhodně vybraných 20 % případů hádat diagnózu MCI, dosáhneme senzitivity 0,2 a specificity 0,8. Bod (0, 0) značí případ, kdy se nikdy nerozhodneme pro stanovení diagnózy, a bod (1, 1) odpovídá případu, kdy vždy diagnózu určíme. Nejdůležitější praktický poznatek by měl být, že senzitivita a specificita testu jsou závislé na stanoveném prahu a že při stanovování prahu vede navýšení jedné z těchto vlastností k poklesu druhé. Tento vztah nicméně neplatí, pokud máme možnost zvolit jiný test.



Obr. 3. ROC křivka a odpovídající senzitivita a specificita testu pro dva vybrané prahy pro stanovení diagnózy

Co by se stalo, kdybychom zvolili méně validní test, je zobrazeno pomocí nové ROC křivky na obrázku 4. Lze vidět, že původní křivka leží nad novou křivkou a za použití původního testu lze tedy dosáhnout vyšší specificity při zachování stejné senzitivity a naopak.



Obr. 4. Vztah validity, senzitivity a specificity ilustrovaný ROC křivkami dvou odlišných metod

K porovnání vlastností diagnostických metod se někdy používá obsah oblasti pod křivkou (area under curve – AUC). Je vidět, že AUC je u původní křivky jasně větší než u nové, tj. z hlediska schopnosti rozlišovat mezi zdravými lidmi a pacienty s MCI je první test lepší než

druhý. AUC má také praktickou interpretaci. Pokud bychom vybrali náhodně jednoho člověka ze skupiny s MCI a jednoho ze skupiny bez poškození, pak AUC odpovídá pravděpodobnosti, s jakou by člověk bez MCI měl vyšší skór než člověk s MCI (Fawcett, 2006). Je tedy vidět, že AUC je rovna 0,5 v případě testu neposkytujícího žádnou spolehlivou informaci, což odpovídá linii spojující body (0, 0) a (1, 1), a nabývá maximální hodnoty 1, která odpovídá testu dokonale rozlišujícímu obě skupiny lidí.

Jelikož senzitivita a specificita závisí vždy na prahu a rozložení skórů pouze jedné ze skupin, ROC křivka není nijak závislá na velikosti obou skupin. Pokud bychom chtěli určit pravděpodobnost poškození u konkrétního člověka, museli bychom opět aplikovat Bayesovu větu. S pomocí ROC analýzy můžeme nicméně určit optimální práh pro stanovení diagnózy. Ten přitom závisí kromě původní pravděpodobnosti jevu také na benefitech ze správného rozhodnutí a na nákladech vyplývajících z rozhodnutí špatného. Sklon křivky v bodě odpovídajícím optimálnímu prahu můžeme potom vypočítat pomocí následujícího vzorečku (Swets et al., 2000):

$$S(optimální) = \frac{P(\neg H)}{P(H)} \times \frac{B(\neg H \wedge \neg R) + N(\neg H \wedge R)}{B(H \wedge R) + N(H \wedge \neg R)}$$

S značí sklon křivky, kde vyšší sklon odpovídá nižšímu prahu, *R* značí diagnostikování poruchy, ¬*R* pak značí rozhodnutí o absenci poruchy. Benefity a náklady jsou značené jako *B*, resp. *N*. Skutečné poruše odpovídá *H* a skutečné absenci poruchy odpovídá ¬*H*. Ze vzorečku je tedy vidět, že práh budeme v našem případě snižovat, pokud budou růst benefity ze správného zamítnutí diagnózy a náklady na falešně pozitivní stanovení diagnózy. Práh budeme naopak volit vyšší při vyšších benefitech ze správné diagnózy a vyšších nákladech na nestanovení diagnózy v případě poruchy. To znamená, že např. v případě, kdy na poruchu neexistuje účinná léčba či terapie, můžeme zvolit nižší práh, neboť benefity ze správného zamítnutí diagnózy budou spíše nízké. Naopak v případě, že existuje účinná léčba, kterou je potřeba včas využít, práh bychom měli volit vyšší, neboť náklady na chybné zamítnutí poruchy i benefity ze správné diagnózy jsou vyšší. Ze vzorečku je také vidět, že práh by měl záviset na poměru lidí zdravých a s poškozením ve vyšetřované populaci. V případě většího podílu zdravých lidí bychom měli volit práh spíše nižší. Vztahy jsou samozřejmě opačné, pokud vyšší skóry v testu indikují poruchu a nižší skóry absenci poruchy (např. skóry ve škálách deprese).

Hodnocení výkonu při dlouhodobém sledování

Využití ROC analýzy vychází ze znalosti rozložení skórů u zdravých lidí a lidí s poruchou. Tyto skóry jsou nicméně známy pouze u některých poruch a některých diagnostických nástrojů. Pokud těmito informacemi nedisponujeme, musíme kritérium pro poruchu stanovit jiným způsobem. Jednou z možností je, že poruchu diagnostikujeme na základě výrazného zhoršení výkonu v testu oproti dřívějšímu stavu. Předpokladem tohoto postupu je, že známe výsledek dřívějšího vyšetření nebo můžeme minulý stav spolehlivě odhadnout. Následující text se věnuje tomu, jak zhodnotit zhoršení výkonu jedince srovnáním dvou vyšetření.

Pro ilustraci postupu si uveďme následující příklad. Hodnotíme longitudinální sledování pana N. ve skórech Paměťového testu učení (pro výpočty jsou použity normy ze studie Knight et al., 2007). Při prvním vyšetření uvedl celkově 52 slov. Rok poté v re-testu s použitím alternativní formy uvedl celkově 39 slov. Došlo tedy k poklesu o 13 slov. Nyní chceme zhodnotit, zda je tento pokles spolehlivý a reflektuje progresi kognitivního deficitu, nebo odráží pouze chybu měření. Pro výpočet použijeme následující vzoreček pro výpočet standardní chyby měření rozdílu:

$$SEM_{x2-x1} = \sqrt{sd_{x1}^2 + sd_{x2}^2 - 2\sqrt{r_{xx}}sd_{x1}sd_{x2}}$$

Jako sd_{x1} a sd_{x2} jsou označeny směrodatné odchylky populačních skórů při prvním a druhém vyšetření. Obvykle budou tyto směrodatné odchylky přibližně stejné, i když se mohou měnit např. v případech, kdy se variance skórů liší podle věku vyšetřovaného. Jako r_{xx} je zde označován koeficient re-testové reliability pro daný test, který ve výpočtu zohledňuje závislost obou měření. Použita by při tom měla být hodnota získaná ze studie, která používala oddálení přibližně odpovídající době mezi měřeními u vyšetřovaného.

V našem případě sd_{x1} i sd_{x2} jsou 9,85, korelace mezi výsledkem v dvou zadáních testů je 0,79, r_{xx} je tedy $0{,}79^2 = 0{,}624$. Počítáme-li se stejnými směrodatnými odchylkami, vzoreček lze zjednodušit na:

$$SEM_{x2-x1} = \sqrt{2sd_x^2 \times \left(1 - \sqrt{r_{xx}}\right)}$$

Po dosazení dostáváme:

$$SEM_{x2-x1} = \sqrt{2 \times 9{,}85^2 \times \left(1 - \sqrt{0{,}624}\right)} \cong 6{,}38$$

Ze standardní chyby měření rozdílu můžeme následně vypočítat index reliabilní změny (reliable change index – RCI) tím, že hodnotu $SEM_{x2–x1}$ vynásobíme 1,96 či případně jinou hodnotou podle toho, jakou pravděpodobnost falešně pozitivní diagnózy v případě reálné absence zhoršení jsme ochotni tolerovat (hodnota 1,96 odpovídá pravděpodobnosti 2,5 %, hodnota 1,645 pravděpodobnosti 5 %, další hodnoty lze zjistit z charakteristik standardizovaného normálního rozdělení). Pokles výkonu o 13 slov hodnotu RCI (12,5) překračuje, a u pana N. tedy spolehlivě indikuje pokles výkonu. Vzhledem k tomu, že hodnota populačního průměru je 42,34, při pouhém srovnání s hodnotami v populaci by i druhé měření bylo v rámci normy (výsledek odpovídá zhruba 37. percentilu). Je zřejmé, že pomocí výpočtu indexu reliabilní změny můžeme zjistit poruchu, kterou by samotné srovnání s populačními normami neodhalilo.

Regrese k průměru a testové baterii

V běžné praxi často nejsou dřívější výsledky vyšetřovaného k dispozici, a je tedy potřeba učinit závěr o diagnóze z jediného vyšetření. V takovém případě se budeme spoléhat pouze na nízký skór vyšetřovaného vzhledem k demograficky vázané populační normě. Diagnostická kritéria pro stanovení určité poruchy jsou pak často formulována vzhledem k obecnému výkonu v testu určitého typu (např. pro stanovení amnestické MCI je kritériem výkon v paměťovém testu s oddáleným vybavením snížený pod hranicí 1 nebo 1,5 standardní odchylky pod průměrem; Petersen et al., 1999). S těmito kritérii jsou nicméně spojené určité problémy. Protože se kritéria vztahují k určité kognitivní doméně a nejsou formulována vzhledem k výsledku v konkrétním testu, je třeba vzít v potaz, že testy používané pro diagnostiku neměří tuto doménu zcela přesně. U méně reliabilního testu bychom tedy měli zvolit přísnější práh pro stanovení diagnózy. Díky regresi k průměru lze očekávat, že vyšetřovaný, který je v testu y směrodatných odchylek pod populačním průměrem, bude reálně v měřené dovednosti $y \times r$ směrodatných odchylek pod průměrem, kde r je korelace mezi výsledkem testu a reálnou úrovní dané dovednosti (Bland & Altman, 1994). Vzhledem k tomu, že r většinou neznáme, je možné alespoň určit jeho horní mez pomocí koeficientu reliability. Potom r nabývá maximálně hodnoty odpovídající $\sqrt{r_{xx}}$. Pokud tedy chceme diagnostikovat poruchu u lidí, kteří jsou z hlediska dané dovednosti z směrodatných odchylek pod průměrem, měli bychom diagnostikovat poruchu pouze u lidí, kteří v testu měřícím tuto dovednost skórovali alespoň níže než $z / \sqrt{r_{xx}}$ směrodatných odchylek pod průměrem.

Pokud použijeme komplexnější test nebo testovou baterii, máme k dispozici větší množství skórů. S nárůstem počtu měření také stoupá pravděpodobnost, že objevíme abnormálně nízký

skór u zdravého člověka. Snadno se pak může stát, že nízkým skórům budeme přikládat příliš velkou váhu a náš závěr bude falešně pozitivní – přisoudíme diagnózu v případě, kdy ji pacient ve skutečnosti nemá.

Ke zpřesnění diagnostického úsudku pomůže klinickému neuropsychologovi znalost obecného rozšíření nízkých skórů v dané testové baterii u zdravé populace. Pak je možné stanovit úroveň a množství nízkých skórů, které budeme považovat za významné, resp. stanovit hranici, pod kterou jsou nízké a extrémně nízké skóry u zdravé populace málo pravděpodobné (objevují se např. v méně než 10 % případů). Odhad obecného rozšíření nízkých skórů lze získat z empirických dat administrací testové baterie velkému množství zdravých jedinců (pro skóry třetí edice Wechslerovy škály paměti tato data nabízí např. Brooks et al., 2008) nebo pomocí statistické simulace. Výsledky empirických studií lze získat přímo z originálních materiálů, podrobněji se tedy zaměříme na ukázku využití simulace.

Crawford, Garthwaite a Gault (2007) vytvořili volně dostupný program (http://homepages.abdn.ac.uk/j.crawford/pages/dept/PercentAbnormKtests.htm), který vyhodnocuje a) běžné rozšíření počtu abnormálně nízkých skórů, b) běžné rozšíření abnormálních rozdílů mezi jednotlivými skóry v testech, c) běžné rozšíření abnormálních rozdílů mezi průměrným skórem jedince a dalšími skóry v testech v dané baterii. Pro výpočet je nutné mít k dispozici korelační matici testů standardizované testové baterie. Takový požadavek v našem prostředí splňuje např.třetí. edice Wechslerovy inteligenční škály pro dospělé (Wechsler Adult Inteligence Scale – WAIS-III), jejíž indexové skóry verbálního porozumění (IVP), percepčního uspořádání (IPU), pracovní paměti (IPP) a rychlosti zpracování (IRZ)[10] využijeme v příkladu aplikace tohoto přístupu.

Z výsledků simulace pro prevalenci nízkých indexových skórů v české verzi WAIS-III u zdravé populace (tab. 1) vyplývá, že nízké skóry pod hranicí jedné standardní odchylky (zhruba 16. percentilu) nebo pod hranicí 10. percentilu jsou poměrně časté (výskyt u více než 20 % zdravé populace). Stanovení kognitivního deficitu na základě nalezení jednoho takového skóru tedy může často vést k falešně pozitivnímu závěru. Skóry pod hranicí 1,5 SD, případně pod 5. percentilem jsou již mnohem méně obvyklé. V této souvislosti Brooks, Iverson, Feldman a Holdnack (2009) navrhují hodnotit výkon jedince na základě stanovené hranice nízkých skórů tak, že pokud se nízký skór objevuje zhruba u 20 % zdravých jedinců,

[10] Jedná se o skóry, které tvoří základní faktorovou strukturu testu a jsou v klinické neuropsychologické praxi často používány.

má být výkon hodnocen jako *možný kognitivní deficit*, pokud se nízký skór objevuje zhruba u 10 % zdravých jedinců, má být výkon hodnocen jako *pravděpodobný kognitivní deficit*.

Tab. 1. Procenta očekávaných abnormálně nízkých výkonů v *j* nebo více indexových skórech WAIS-III u zdravé populace vzhledem ke zvolenému kritériu abnormality

Kritérium abnormality	**Procento výskytu *j* nebo více abnormálně nízkých indexových skórů WAIS-III**			
	1	**2**	**3**	**4**
< 15,8 % (1 SD)	33,4	17,5	9,2	3,5
< 10 %	22,9	10,6	5,0	1,6
< 6,6 % (1,5 SD)	16,3	6,7	2,9	0,9
< 5 %	12,7	4,9	2,0	0,5
< 2,28 % (2 SD)	6,3	2,0	0,7	0,2

Poznámka: SD = standardní odchylka. Byly vybrány nejčastěji užívané hranice definující abnormální výkon. Výsledky byly získány z programu, který tvoří součást článku Crawford, Gathwaite a Gault (2007).

Pro ilustraci nyní uveďme případ paní D., která při autonehodě utrpěla úraz hlavy. Při testování za využití WAIS-III dosáhla skórů IVP = 112, IPU = 100, IPP = 85 a IRZ = 92, přičemž skór IPP se nachází 1 standardní odchylku pod průměrem. Alespoň jeden takový skór se ale objevuje u celé třetiny zdravé populace, nebudeme ho tedy považovat za příliš významný. Dle manuálu k testu (tabulka B2, viz Černochová et al., 2010) bychom měli za abnormální považovat rozdíl mezi skóry IVP a IPP (27 bodů), který se objevuje u méně než 5 % zdravé populace. Výsledky simulace však ukazují, že alespoň jeden abnormální rozdíl mezi indexovými skóry se objevuje téměř u 20 % zdravé populace (dva a více abnormálních skórů se objevuje u necelých 8 % zdravé populace). Hodnocení tedy spadá do úrovně *možného* kognitivního deficitu. Finální interpretaci je však nutné učinit s přihlédnutím ke vzdělání nebo jinému odhadu premorbidních schopností, které bohužel program nezohledňuje. Jak ukázali Brooks et al. (2008), chceme-li při diagnostickém rozhodování redukovat chyby, je nutné při hodnocení výkonu brát v úvahu demografické charakteristiky, zejména inteligenci (resp. odhad její premorbidní úrovně) nebo nejvyšší dosažené vzdělání. Například u skupiny jedinců s průměrnou inteligencí byly alespoň tři podprůměrné skóry ve WMS-III běžné u 25 % vzorku, u skupiny jedinců s nadprůměrnou inteligencí to bylo pouze u 5 % (Brooks et al.,

2008). Bez respektování těchto rozdílů se významně zvyšuje riziko, že přeceníme význam nízkých skórů u jedinců s nižší inteligencí (přisoudíme falešně pozitivní diagnózy), v případě jedinců s vysokou inteligencí naopak můžeme význam nízkých skórů zanedbat (závěr bude falešně negativní).

Terapie a rehabilitace

Práce klinického neuropsychologa se neomezuje pouze na psychologické testování a diagnostiku, ale zahrnuje i jiné činnosti, jako jsou např. terapie a rehabilitace. Stejně jako je důležité pro diagnostiku vybrat test s dobrými psychometrickými vlastnostmi, je klíčové v rámci terapie a rehabilitace vybrat metody, které jsou účinné. Špatná volba terapeutické metody může vést nejen ke ztrátě času a prostředků, ale může pacientovi i uškodit (Lilienfeld, 2007). Schopnost zvolit metody, jejichž účinnost je vědecky prokázána, je tedy důležitou dovedností klinického neuropsychologa (Baker, McFall & Shoham, 2009).

Pro zjištění vlivu terapie či rehabilitace je potřeba provést znáhodněné kontrolované studie (Garske & Anderson, 2003). Studie musí být znáhodněné, neboť v jiném případě si nemůžeme být jistí, zda pokusné osoby v kontrolní skupině jsou srovnatelné s pokusnými osobami ve skupině, které je terapie poskytnuta. Nelze tedy pouze sledovat, zda se stav pacientů podstupujících terapii liší od stavu pacientů bez terapie. Obě skupiny se mohou lišit v původním stavu, v motivaci k léčbě, v sociální podpoře, v socioekonomických charakteristikách apod. a zjištěný rozdíl mezi skupinami pak může být chybně interpretován jako efekt terapie.

Studie musí být kontrolované z důvodu možnosti zjištění kauzálního vlivu terapie na zlepšení pacienta. Pokud bychom např. srovnávali stejnou skupinu pacientů před terapií a po terapii, nemůžeme si být jistí, že případné zlepšení lze přisoudit vlivu terapie. Jelikož stav pacienta je obvykle variabilní a terapie je indikována spíše v případě zhoršení stavu, lze očekávat zlepšení v průběhu času pouze na základě regrese k průměru. Ke zlepšení stavu může také vést přirozený vývoj nemoci. Kontrolní skupinou může být např. srovnatelná skupina, která na terapii čeká. Preferovanou kontrolní skupinou je nicméně skupina, která podstupuje jinou formu terapie či léčby. Touto léčbou může být určitá forma podpůrné terapie, u níž můžeme očekávat pouze nespecifický efekt vycházející z očekávání zlepšení a sociální podpory. Studie, kterou lze využít pro volbu terapeutické metody, by měla nicméně používat kontrolní skupinu, které je poskytnuta alternativně použitelná forma terapie. Pokud bychom totiž zjistili pouze to, že zkoumaná terapie je lepší než podpůrná terapie, a nesrovnali bychom ji s jinými

účinnými formami terapie, nemohli bychom vědět, zda tyto formy terapie nejsou účinnější, a v důsledku pro praxi vhodnější než námi studovaná terapie.

V rámci výzkumu terapie je potřeba zvolit vhodný způsob posuzování účinnosti. Pokud budeme měřit účinek terapie nebo rehabilitace na škále, která bude mít nízkou externí validitu, může se stát, že sice získáme statisticky signifikantní rozdíl mezi experimentální a kontrolní skupinou, ale dopad na fungování v běžném životě bude mizivý (Kazdin, 2008). Právě zobecnitelností výsledků terapie, resp. jejich přenositelností do jiných kontextů se zabývá výzkum efektivity (effectivity research), na rozdíl od výzkumu účinnosti (efficacy research), jehož hlavní otázka je, zda se nějaký účinek objevuje. Cílem terapie je zejména pozitivně ovlivnit běžné fungování člověka, proto by to mělo být ve výzkumech operacionalizováno a dlouhodobě sledováno.

Výzkumy terapie obvykle využívají testování hypotéz, které ověřuje, zda existuje pouze malá pravděpodobnost, že bychom našli nalezenou velikost efektu, pokud by žádný efekt reálně neexistoval. Pokud je tato pravděpodobnost (označovaná jako *p*) nižší než stanovená hodnota (hladina významnosti označovaná jako α – v psychologii obvykle 0,05), říkáme, že je efekt statisticky významný (Cohen, 1990, 1994). V klinické praxi nám však jde o klinickou významnost efektu terapie, nikoli o její statistickou významnost. Samotná znalost statistické významnosti efektu tak není dostatečná pro rozhodnutí o využití dané metody terapie. Stejný problém se může projevit také při srovnávání různých forem terapie. Pokud existují dvě studie, z nichž jedna ukazuje, že určitá forma terapie má statisticky významný vliv na léčenou poruchu, a druhá ukazuje, že jiná forma terapie statisticky významný vliv na léčenou poruchu nemá, nemůžeme říci, že první forma terapie je účinnější než druhá (Gelman & Stern, 2006). Důvodem je, že velikost efektu nutná pro statistickou významnost klesá s růstem zkoumaného vzorku. První studie tedy mohla pouze využívat více pokusných osob, a tak se jako statisticky signifikantní mohl projevit menší efekt. Ve skutečnosti tak může být druhá terapie účinnější a dokonce jsme u ní mohli objevit větší efekt. Podobný problém může nastat i v rámci jedné studie, pokud nesrovnává efekty terapií přímo, ale hodnotí je na základě statistické významnosti změny vůči stavu před terapií či je srovnává pouze s kontrolní skupinou. Jakékoli závěry o účinnosti terapie tedy musí vycházet ze znalosti velikosti efektu terapie, a nikoli pouze ze statistické významnosti tohoto efektu v rámci určité studie. Důležité je na tomto místě zdůraznit, že to neznamená, že bychom měli statistickou významnost opomíjet. Velký efekt u studie s malým počtem pokusných osob může být snadno dán pouze náhodou.

Závěr

Znalost psychologické metodologie je podstatná pro všechny činnosti, kterým se klinický neuropsycholog v praxi věnuje. Pro spolehlivou diagnostiku je potřeba znát informace o psychometrických vlastnostech jednotlivých metod a umět je využít nejen při výběru metod, ale i při interpretaci výsledků. Diagnostické metody, které nejsou standardizované, nemají aktuální normy pro relevantní populaci nebo nejsou reliabilní a validní, nelze obecně doporučit pro využití v praxi. Bez norem nemůžeme vědět, zda je výsledek vyšetřovaného abnormální, či zda lze podobný výsledek očekávat i u zdravého jedince. Reliabilita určuje, do jaké míry je získaný skór skutečným odrazem výkonu vyšetřovaného, a na validitě závisí množství informace, kterou nám výsledek může poskytnout. Díky znalosti těchto vlastností můžeme při diagnostice využít formální postupy usuzování, které jsou obvykle výrazně přesnější než pouhý klinický úsudek. Metodologické znalosti jsou nezbytné také při výběru terapeutické či rehabilitační metody. Psycholog by měl vědět, jaké závěry lze činit z dostupných výzkumů a jaká jsou jejich omezení, a na základě těchto znalostí by se měl rozhodovat tak, aby v zájmu klienta vždy zvolil nejefektivnější metodu.

Literatura

Baker, T. B., McFall, R. M., & Shoham, V. (2009). Current status and future prospects of clinical psychology: Toward a scientifically principled approach to mental and behavioral health care. *Psychological Science in the Public Interest, 9*, 67–103.

Bland, J. M., & Altman, D. G. (1994). Regression towards the mean. *British Medical Journal, 308*, 1499.

Borsboom, D., Mellenberg, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.

Brooks, B. L., Iverson, G. L., Feldman, H. H., & Holdnack, J. A. (2009). Minimizing misdiagnosis: Psychometric criteria for possible or probable memory impairment. *Dementia and Geriatric Cognitive Disorders, 27*, 439–450.

Brooks, B. L., Iverson, G. L., Holdnack, J. A., & Feldman, H. H. (2008). Potential for misclassification of mild cognitive impairment: A study of memory scores on the Wechsler memory scale-III in healthy older adults. *Journal of the International Neuropsychological Society, 14*, 463–478.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17,* 31–43.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997–1003.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104.

Crawford, J. R., Garthwaite, P. H., & Gault, C. B. (2007). Estimating the percentage of the population with abnormally low scores (or abnormally large score differences) on standardized neuropsychological test batteries: A generic method with applications. *Neuropsychology, 21*, 419–430.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement, 10,* 94–96.

Černochová, D., Goldmann, P., Král, P., Soukupová, T., Šnorek, V., & Havlůj, V. (Eds.). (2010). *WAIS-III-Wechslerova inteligenční škála pro dospělé*. Praha: Hogrefe – Testcentrum.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1673.

Estévez-González, A., Kulisevsky, J., Boltes, A., Otermín, P., & García-Sánchez, C. (2003). Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: Comparison with mild cognitive impairment and normal aging. *International Journal of Geriatric Psychiatry, 18*, 1021–1028.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874.

Fischer, R., & Milfont, T. L. (2010). Standardization in psychological research. *International Journal of Psychological Research, 3*, 88–96.

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7,* 248–260.

Garske, J. P., & Anderson, T. (2003). Toward a science of psychotherapy research: Present status and evaluation. In S. O. Lilienfeld, S. J. Lynn & J. M. Lohr (Eds.), *Science and Pseudoscience in Clinical Psychology* (s. 145–175). New York: The Guilford Press.

Gelman, A., & Stern, H. (2006). The difference between „significant“ and „not significant“ is not itself statistically significant. *The American Statistician, 60,* 328–331.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The Psychology of intuitive Judgment*. New York: Cambridge University Press.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19–30.

Hänninen, T., Hallikainen, M., Tuomainen, S., Vanhanen, M., & Soininen, H. (2002). Prevalence of mild cognitive impairment: A population-based study in elderly subjects. *Acta Neurologica Scandinavica, 106*, 148-154.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*, 238**–**247.

Herlitz, A., Nilsson, L.-G., & Bäckman, L. (1997). Gender differences in episodic memory. *Memory & Cognition, 25*, 801–811.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*, 515–526.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Kazdin, A. E. (2008). Evidence-based treatment and practice. *American Psychologist, 63*, 146–159.

Knight, R. G., McMahon, J., Skeaff, C. M., & Green, T. J. (2007). Reliable change index scores for persons over the age of 65 tested on alternate forms of the Rey AVLT. *Archives of Clinical Neuropsychology, 22*, 513–518.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science, 2*, 53–70.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment*. New York: Oxford University Press.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement, 10*, 1–29.

Nunes, P. V., Diniz, B. S., Radanovic, M., Abreu, I. D., Borelli, D. T., Yassuda, M. S., & Forlenza, O. V. (2008). CAMCOG as a screening tool for diagnosis of mild cognitive impairment and dementia in a Brazilian clinical sample of moderate to high education. *International Journal of Geriatric Psychiatry, 23*, 1127–1133.

Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research, 16*, 6–17.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology, 56*, 303–308.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*, 33–65.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizibility theory. *American Psychologist, 44*, 922–932.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 1–25.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99–103.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1–26.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274–290.

Yamamoto, S., Mogi, N., Umegaki, H., Suzuki, Y., Ando, F., Shimokata, H., & Iguchi, A. (2004). The clock drawing test as a valid screening method for mild cognitive impairment. *Dementia and Geriatric Cognitive Disorders, 18*, 172–179.