Univerzita Karlova v Praze

Přírodovědecká fakulta

Analýza geneticko-epidemiologických dat

Petra Dušátková

Bakalářská práce

Praha 2006

Školitel: MUDr. Ondřej Cinek, Ph.D.

Garant: RNDr. Patrik Mottl, Ph.D.

OBSAH

1. Úvod 2
2. Epidemiologie 2
2.1 Genetická epidemiologie 2
2.2 Druhy epidemiologických studií 3
2.2.1 Observační studie 3
2.2.1.1 Deskriptivní studie 3
2.2.1.2 Analytická studie 4
2.2.2 Intervenční studie 6
2.2.3 Rozdělení studií z časového hlediska 6
2.2.3.1 Prospektivní studie 6
2.2.3.2 Longituální studie 6
2.2.3.3 Retrospektivní studie 6
3. Genetické varianty 7
3.1 Mutace 9
3.2 Polymorfismy 9
3.2.1 Jednonukleotidové polymorfismy 9
3.2.2 Ostatní polymorfismy 10
4. Testování asociace 11
4.1 Vazba, vazbová nerovnováha 11
4.2 Studium asociace 12
5. Studie případů a kontrol 14
5.1 Rodinné studie případů a kontrol 14
5.2 Populační studie případů a kontrol 15
5.3 Testování ve studiích případů a kontrol - TDT 15
5.4 Testování ve studiích případů a kontrol – populační kontroly 17
5.5 Korekce na mnohonásobné testování 17
5.6 Kontingenční tabulky 18
5.7 Kvantifikace rizika – odds ratio 18
5.8 Kvantifikace rizika – relative risk 20
5.9 Intervaly spolehlivosti 21
5.10 Chí kvadrát testy 21
5.10.1 McNemarův test 22
6. Závěr 23
7. Seznam literatury 24

PROHLÁŠENÍ

Prohlašuji, že jsem bakalářskou práci napsala samostatně. Veškeré literární prameny a informace, které jsem v práci využila, jsou uvedeny v seznamu literatury.

1. Úvod

Povědomí o základních typech epidemiologických studií je nezbytné pro vědecké pracovníky i kliniky, aby se mohli rozhodnout pro adekvátní design své studie, který by zaručil správné, relevantní a replikovatelné výsledky. Stejně jako výskyt, projevy nebo léčba nemoci je, především v posledních desetiletích, důležitá znalost jejího genetického pozadí. Díky ní máme možnost odhalit geny zodpovědné nebo spolu-zodpovědné za vznik onemocnění a následně tak předpovědět míru individuálního rizika onemocnění. Vše co jsem zmínila, by ovšem bylo zbytečné, pokud by nabyté poznatky nebyly správně statisticky zpracovány. Statistické metody zabývající se analýzou geneticko-epidemiologických dat spadají do kategorie studií vazby a asociace.

Cílem této bakalářské práce je zpracování literárních pramenů týkajících se geneticko-epidemiologických dat, zejména ve studiích případů a kontrol pro potřebu mé budoucí diplomové práce.

2. Epidemiologie

Epidemiologie je vědní obor, který se zabývá studiem výskytu infekčních chorob, jejich příčin a způsobů šíření v populaci. Slovo epidemiologie tvoří tři řecká slova: epi (nad), demos (lidé) a logos (nauka). Je to nauka o stavech populace z hlediska výskytu nejrůznějších onemocnění.

Epidemiologie hraje hlavní roli v hledání vhodných strategií ke zlepšení zdraví společnosti, a to především přes prevenci. Mezi hlavní úkoly epidemiologie tedy patří sledovat zdravotní stav populace (tzn. například měřit frekvenci výskytu nemoci, provádět dozor u vybraných nemocí a pod.), analyzovat zdravotní stav populace (tzn. zkoumat etiologii nemoci, měřit vztah – asociaci mezi nemocí a jejími příčinami a pod.) a v neposlední řadě zlepšovat zdravotní stav populace (tzn. vyhodnocovat diagnostické postupy, efektivitu nových léčiv, zavádět do lékařské praxe nové poznatky medicíny založené na výsledcích výzkumných studií, navrhovat a následně realizovat opatření vedoucí k omezení nebo úplnému vymýcení dané nemoci, a pod.).

2.1 Genetická epidemiologie

Genetická epidemiologie je vědní disciplínou zabývající se epidemiologickým hodnocením role dědičných příčin onemocnění v rodinách a populacích. Snaží se odhalit

model dědičnosti určité nemoci, lokalizovat gen nebo geny, které nemoc způsobují a najít marker vztahující se k náchylnosti k nemoci. Genetická epidemiologie rovněž studuje interakce gen – gen a gen – prostředí.

Genetická epidemiologie se jako samostatný vědní obor objevila v šedesátých letech minulého století. Vznikla spojením populační genetiky, statistiky, matematiky a molekulární epidemiologie. Mezi průkopníky genetické epidemiologie patří Newton Morton, Robert C. Elston, Elizabeth A. Thompson, Douglas Falconer a Neil Risch (Dorak, 2006).

Kroky, které zahrnuje geneticko-epidemiologický výzkum jsou: zjištění, zda se na vzniku poruchy podílí (mimo jiné) genetická složka, zjištění velikosti tohoto genetického efektu v porovnání s ostatními příčinami onemocnění a identifikace genu nebo genů zodpovědných za genetickou složku onemocnění.

2.2 Druhy epidemiologických studií

Epidemiologické studie se podle tradiční klasifikace dělí na dvě základní skupiny – studie observační a studie intervenční neboli experimentální.

2.2.1 Observační studie

V observační (pozorovací) studii výzkumníci nijak neovlivňují průběh pokusu, respektive neovlivňují studované populace, subjekty nebo jevy. V tomto typu studie se pouze pozorují, zaznamenávají a poté statisticky zpracovávají dané jevy. Observační studie poskytují jiný druh závěrů než intervenční studie – např. četnost výskytu choroby, u koho se objevuje nebo jaké jsou rizikové faktory. Lze i tvrdit, že observační studie přináší statisticky slabší výsledky než intervenční studie. U observačních studií je totiž větší pravděpodobnost omylu, a to v případě, že je přítomna nám neznámá asociace mezi zkoumaným faktorem a naším závěrem. Dalším důvodem je, že nelze přizpůsobit podmínky realizace potřebám výzkumu. Největším významem observačních studií je, že poskytují předběžné důkazy, které mohou být následně použity jako základ pro hypotézy zkoumané na základě silnějších intervenčních studií. Observační studie rozdělujeme na deskriptivní a analytické (Bencko, 2003).

2.2.1.1 Deskriptivní studie

Studie popisující rozložení výskytu onemocnění v populaci nazýváme deskriptivní. Deskriptivní studie je statisticky nejslabší epidemiologickou studií. Nepokouší se určit příčinu ani účinek onemocnění. I přesto je často právě deskriptivní studie prvním krokem ke

komplexní epidemiologické studii. Dovoluje definovat hypotézu, která je testována dalšími typy epidemiologických studií. Deskriptivní studie totiž zkoumá, jaké skupiny obyvatel jsou nemocí postiženy, ve které lokalitě se nemoc nejvíce či nejméně vyskytuje, jak se výskyt nemoci mění v čase atd. Na základě těchto otázek a odpovědí je možné vyslovit hypotézu a testovat ji. Deskriptivní studie lze rozdělit na kazuistiku, průřezovou studii a ekologickou studii (Bencko, 2003).

Kazuistika (case report) je základní typ deskriptivní studie prováděné na jednotlivci. Typicky popisuje projevy onemocnění u daného jedince, klinickou léčbu a prognózu (Gay, 1999). Kvůli velké variabilitě v těchto aspektech nemá kazuistika tak velký význam pro výzkum.

Kazuistika může být někdy rozšířena na popis série onemocění (case series). Jednotlivci v sérii jsou spolu určitým způsobem svázáni (Bencko, 2003). Popis série onemocnění je studie, která je užívána jako zdroj hypotézy testované silnějšími typy studií.

Dalším typem deskriptivní studie prováděné na skupinách jednotlivců je průřezová studie (cross-sectional study). Jedná se o studii vztahu mezi onemocněním a ostatními faktory v určitém čase v určité populaci. Průřezová studie je co se týče provedení relativně jednoduchá. Poskytuje užitečná data o počtu onemocnění v dané populaci a čase. Tento typ studie nám dovoluje odhadnout procento nemocných i procento osob s rizikovým faktorem. Rovněž můžeme testovat velikost asociace mezi expozicí a nemocí (Bencko, 2003). Hlavním problémem ale je, že expozice a onemocnění jsou měřeny současně (Pai, 2006). Nelze proto zpravidla určit, zda expozice předcházela nemoci či naopak. Průřezová studie je z tohoto důvodu některými autory řazena mezi deskriptivní a zároveň analytické studie.

Ekologická studie se někdy označuje i jako korelační studie (correlational study). V této studii jsou údaje (expozice rizikového faktoru, výskyt onemocnění) zjišťovány na úrovni populací. Agregovaná data z různých skupin v populaci jsou porovnávána za účelem objevit mezi nimi asociaci (Gay, 1999). Závěry z těchto dat však nemusí nutně platit na úrovni jednotlivců. Některé expozice jsou totiž takového druhu, že na individuální úrovni neexistují (hustota obyvatel nemá pro jednoho člověka žádný ekvivalent). Proto poskytuje ekologická studie empiricky slabý důkaz. Měla by se použít pouze k vytváření hypotéz, ne k jejich ověřování.

2.2.1.2 Analytická studie

Hlavní náplní analytických studií je formulovat a testovat epidemiologické hypotézy z popisu rozložení nemoci. Tímto způsobem lze ověřit předpokládané příčiny vzniku

onemocnění nebo faktory, které pomohou vzniku nemoci zabránit. V těchto studiích jde tedy hlavně o měření asociace mezi potencionálně rizikovým faktorem (expozicí) a zdravotním následkem (onemocněním nebo úmrtím).

Hranice mezi analytickými a deskriptivními studiemi není přesně dána. Odvíjí se spíš od cílů a účelů studie. Analytické studie se dělí na studie případů a kontrol a kohortové studie (Bencko, 2003). Jeden z hlavnívh rozdílů je v tom, že ve studiích případů a kontrol známe nemocné jedince před zahájením studie, kdežto na začátku kohortové studie žádný z respondentů nemocí netrpí.

Do studie případů a kontrol (case-control study) jsou bráni jednotlivci s danou nemocí (případy) a osoby netrpící danou nemocí (kontroly). Skupiny případů a kontrol jsou pak porovnávány vzhledem k četnosti výskytu nebo velikosti expozice. Expozice je určována zpětně užitím dokumentace, genetickým vyšetřením nebo formou dotazníku. Studie případů a kontrol je nejlepší metodou k zjištění etiologie vzácných onemocnění. Dovoluje současně objevit mnoho možných asociací s nemocí. Velikost skupin bývá nezřídka malá, a přesto může být asociace prokázána. Důvodem je, že kohorta by musela obsahovat velké množství respondentů a stejně by nemusel být žádný se zkoumanou vzácnou nemocí nalezen. Proto se výzkumní pracovníci vzdají přirozeného poměru mezi případy a kontrolami, aby získali statisticky silný výsledek. Na druhou stranu míry rizika asociovaného s expozicemi budou pouze odhadnutá a ne přesně stanovená, jak je to běžné u kohortové studie.

Zvláštním případem studie případů a kontrol je strukturálně vyvážená studie případů a kontrol. Kontroly jsou zde vybírány tak, aby se co nejvíce shodovaly s případy (např. stejné pohlaví, věk, prostředí). Nesmí se ale shodovat v expozicích, které by mohly mít vztah ke studované expozici. Potom se totiž vytrácí síla studie. Studii případů a kontrol se budu podrobněji věnovat níže, v kapitole 5.

Kohorta v epidemiologii označuje skupinu lidí, kteří sdílí společné charakteristiky a expozice. Kohortová studie se od studie případů a kontrol liší tím, že k ověření epidemiologické hypotézy dochází pozorováním skupiny osob vzhledem k přítomnosti či nepřítomnosti expozice. Na začátku studie je známá expozice všech a žádný z jedinců zkoumanou nemocí netrpí. Osoby jsou obvykle sledovány po dlouhý čas (follow up). Poté se porovnávají skupiny exponovaných a neexponovaných vzhledem k četnosti výskytu onemocnění. Kohortové studie jsou silnější než studie případů a kontrol, protože riziko vzniku onemocnění není třeba odhadovat, přímo ho zaznamenáváme. Nevýhodou kohortových studií je, že jsou drahé a časově náročné. Na druhou stranu u kohortových studií je možné sledovat vliv zaznamenávaných expozic na jakýkoli počet nemocí. Expozice máme zjištěné na počátku

studie a jejich vliv na další nemoc lze snadno získat. Na vztah více expozic – jedna nemoc, je vhodnější studie případů a kontrol.

2.2.2 Intervenční studie

Intervenční neboli experimentální studie jsou charakteristické tím, že probíhají pod řízením výzkumných pracovníků. Oni sami určují, kdo bude podroben jakému režimu expozice. Realizace je ale velmi limitována etickými i právními omezeními – nelze někoho vystavit určité expozici, když existuje lepší varianta. Intervenční studie má prospektivní charakter. Nejprve se přiřadí expozice, a poté se čeká na reakci. Je možné např. porovnávat na dvou skupinách dva různé léky a tak zjišťovat, který je účinnější. Intervenční studie je, jak jsem již výše uvedla, silnější než observační studie. Možnost randomizace (náhodného přidělení do té které skupiny) zajišťuje lepší základ pro statistické postupy. Intervenční studie lze rozdělit na terénní kontrolované studie a klinické studie. Jelikož se ale genetická epidemiologie tímto typem studie příliš nezabývá, nebudu ho dále rozvádět.

2.2.3 Rozdělení studií z časového hlediska

2.2.3.1 Prospektivní studie

Dříve se prospektivní studií rozuměla kohortová studie. Kohorty ovšem mohou být i retrospektivní (retrospective cohort). Později tak převládl názor, že termín prospektivní studie bude vyhrazen studii, kdy se expozice měří dříve než je zjištěno onemocnění.

2.2.3.2 Longituální studie

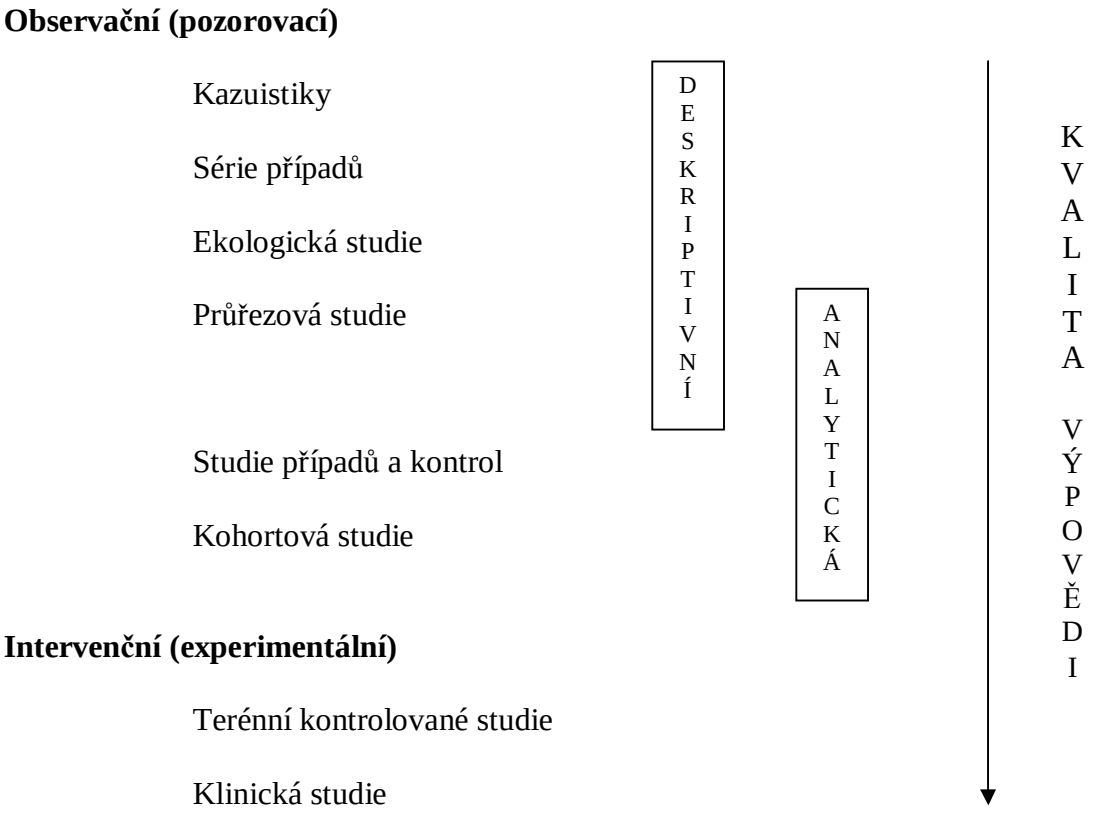
Kohortové či prospektivní studii se dříve říkalo i studie longituální. Dnes se ale pod tímto termínem spíše skrývá studie, kdy jsou jedinci zkoumáni a dotazováni opakovaně v dlouhém časovém intervalu (i několik let). Longituální studie jsou realizovány většinou prospektivně, ze záznamů lze získat data i retrospektivně.

Studie označované jako panel se zúčastňují soubory jednotek, nejčastěji rodiny. V dlouhém časovém intervalu zaznamenávají požadované údaje. Jedná se tedy o odnož longituální studie. V této studii hrozí nebezpečí, že jsou subjekty do výzkumu natolik vtaženy, že se nechovají přirozeně.

2.2.3.3 Retrospektivní studie

Podobně jako prospektivní studie označovala studii kohortovou, tak studie případů a kontrol byla synonymem pro retrospektivní studii. I zde však převládlo přesvědčení, že pojem retrospektivní studie by měl označovat obecnější metodu. Retrospektivní studie je tedy

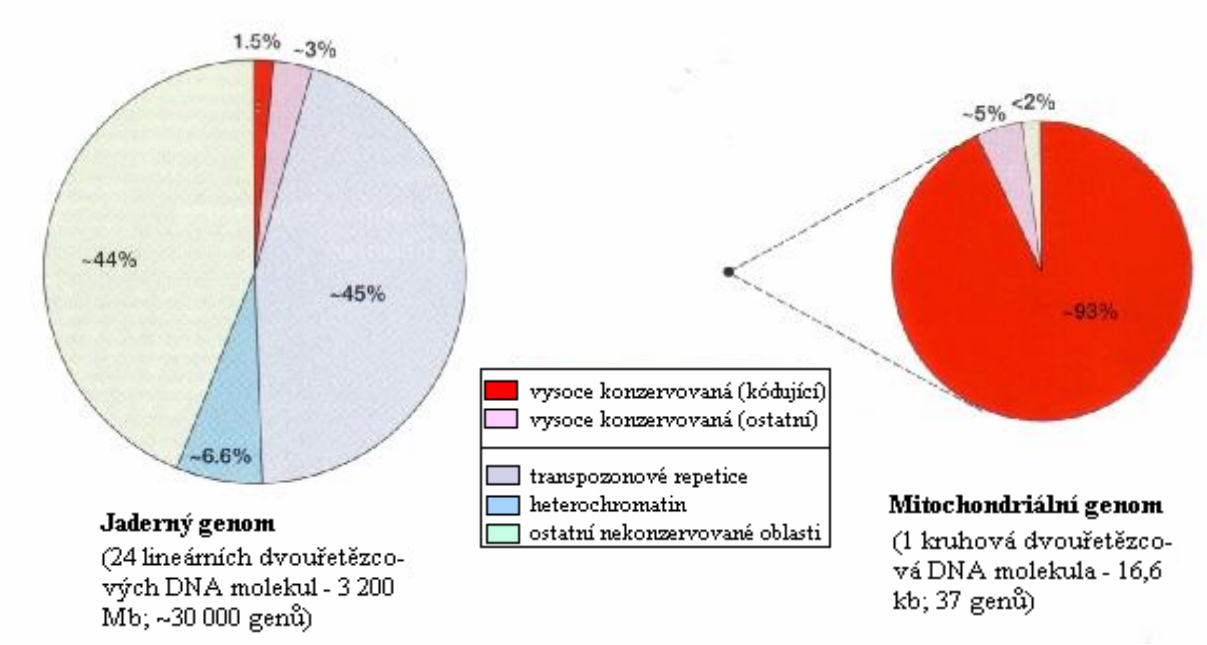
pojmenování pro situaci, kdy existuje onemocnění a hledá se jeho příčina. Expozice je měřena až po zjištění onemocnění.



Obrázek 1. Přehled epidemiologických studií (Bencko, 2003)

3. Genetické varianty

Pojem lidský genom vyjadřuje celkovou genetickou informaci uloženou v lidských buňkách. Ve skutečnosti lze lidský genom rozdělit na dva: jaderný a mitochondriální. Jaderný genom obsahuje zhruba 30 000 genů, mitochondriální pouze 37. Mitochondriální genom, jak je patrné i z grafu 1, je vysoce konzervovaný a nevytváří tedy velký prostor pro variabilitu, jako je to u jaderného genomu.



Graf 1. Organizace lidského genomu (upraveno dle Strachan, 2004)

Rozdíly ve fenotypu jedinců jsou dány genetickou variabilitou a vlivem prostředí. Člověk se od jiného, sobě nepříbuzného jedince, liší v průměru v 1 páru bazí (bp) na 1250 bp (Venter, 2001). Všichni příslušníci Homo sapiens jsou v 99,9% stejní. Pouze 0,1% lidského genomu dělá každého jedince geneticky jedinečným (Guttmacher, 2002 in Kansagra, 2004). Rozdíly v genomu se nacházejí ve všech 46 lidských chromozomech, ovšem ne se stejnou frekvencí. V některých částech genomu lze nalézt i stovky různých variací určité sekvence. Naopak v genomu existují i oblasti, v nichž je variabilita velmi vzácná. Většina genetických obměn je situována mimo kódující geny do extragenové části genomu. Mutace v této části totiž nejsou škodlivé pro organismus, a tak se mohou kumulovat bez větších problémů. Oblasti kódujících genů jsou stabilnější, protože mutace na ně působí většinou škodlivě a nevedou tak k přenosu do další generace.

Genetické varianty zahrnují mutace a polymorfismy. Lze říci, že tyto pojmy vyjadřují prakticky totéž (i když někteří je odlišují na základě četnosti výskytu v lidském genomu). Mutace je spíše vztahována k typu varianty, která způsobí nějaké onemocnění a rovněž je chápána jako pojmenování procesu, kterým nemoc vzniká.

3.1 Mutace

Typ genetické variace, která se vyskytuje u méně než 1% lidské populace, se nazývá mutace. Mutace mohou být rozděleny na základě toho, zda se týkají jedné DNA sekvence (jednoduché mutace) nebo zda zahrnují změny mezi dvěma alelickými nebo nealelickými sekvencemi (Strachan, 2004). Ovšem pro jiné než jednonukleotidové mutace neplatí, že jejich ekvivalentem je polymorfismus.

Nové mutace vznikají u jedinců v somatických i zárodečných liniích. Pokud mutace objevivší se v germinální linii neohrožuje schopnost jedince se rozmnožovat, může se rozšířit do dalších generací.

Mutace působí jako zdroj evolučních novinek. Ovšem také mohou být škodlivé. Obvyklá nízká hladina mutací se tak jeví jako rovnováha mezi evolučním vývojem a příčinou nemocí a smrtí příslušníků druhu.

3.2 Polymorfismy

Polymorfismus je typ genetické varianty, při níž se každá přítomná sekvence v lidské populaci vyskytuje s frekvencí větší než 1%.

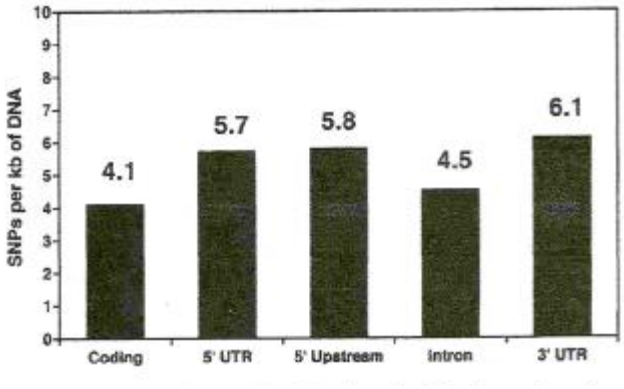
3.2.1 Jednonukleotidové polymorfismy

Jednonukleotidové polymorfismy (single-nucleotide polymorphisms – dále jen SNP) jsou nejfrekventovanějším typem variability lidského genomu. Nejčastěji je jeden nukleotid nahrazen jiným, ale vyskytují se i inzerce a delece jednoho nukleotidu. Mezinárodní skupina zabývající se mapováním SNP odhalila 1,42 mil. SNP, přičemž průměrně připadá jeden SNP na každých 1,9 kb genomu (Sachidanandam, 2001). Jednonukleotidových polymorfismů je mnohem více než výše zmíněných 1,42 milionů, jenže některé jsou nereplikované a jsou distribuovány s různou frekvencí. SNP mají nízkou hladinu opakujících se mutací, což je činí stabilními indikátory lidské historie (Sachidanandam, 2001).

Co se týče charakteristiky nukleotidů v rámci SNP, J. A. Schneider publikoval, že v 71,3% vznikají SNP tranzicí (Schneider, 2003). Tzn. výměnou purinové báze za purinovou (adenin (A) ↔ guanin (G)) nebo pyrimidinovou za pyrimidinovou (thymin (T) ↔ cytosin (C)), přičemž G je většinou frekventovanější než A, resp. C než T. Téměř 25% SNP vzniklých substitucí tvoří výměny T ↔ C (Wang, 1998).

Jak ukazuje graf 2, většina SNP se nachází v nekódujících oblastech lidské DNA. Nekódující DNA obsahuje sekvence důležité pro regulační funkce, a ty tudíž mohou být

v určitém rozsahu omezeny. Hustota SNP je u autosomů relativně konstantní (Sachidanandam, 2001). Nižší množství SNP převládá u pohlavních chromozomů. Chromozom Y čelí silnějšímu selekčnímu tlaku, protože se v populaci vyskytuje v menším množství než autosomy (Venter, 2001). Naopak jedním z chromozomů s vysokou heterogenitou byl označen chromozom 6, u kterého byla potvrzena přítomnost HLA lokusu (Horton, 1998 in Sachidanandam, 2001).



Graf 2. Distribuce SNP v genomových regionech – 1630 genů (Schneider, 2003)

Studium SNP může pomoci při zkoumání variability lidského genomu. SNP studie mohou shromáždit řadu haplotypů (tj. specifických sad alel na jednom chromozomu jedince, které jsou s vysokou pravděpodobností přenášeny spolu) zastoupených v různých etnických skupinách a poskytnout tak pohled na historii lidské populace a její migraci. V neposlední řadě mohou studie SNP vézt ke zlepšení predikce náchylnosti k určité nemoci, ke zlepšení porozumění nemoci, ke schopnosti ji diagnostikovat, a také k vývoji nových a efektivnějších metod léčby. Přes obrovský potenciál výzkumu SNP existují určité limity přístupu k této problematice. Např. některé SNP se mohou častěji vyskytovat u určitých etnických skupin, což by mohlo způsobit problémy při plošné aplikaci těchto genetických informací v medicínské léčbě (Phillips, 2001 in Kansagra, 2004).

3.2.2 Ostatní polymorfismy

SNP nejsou jediným typem polymorfismu, který lze nalézt v lidském genomu. VNTR polymorfismus (variable number of tandem repeat) vzniká kvůli nestabilitě souboru tandemových repetic a způsobuje tak změnu počtu repetic.VNTR je obecným termínem, který

zahrnuje mikrosatelitní VNTR a minisatelitní VNTR. Mikrosatelitní VNTR se rovněž nazývá krátký tandemový opakující se polymorfismus (short tandem repeat polymorphism – STRP) nebo jednoduchý sekvenční opakující se polymorfismus (simple sequence repeat polymorphism – SSRP). Jedná se o typ VNTR, který obsahuje malou opakující se jednotku – obvykle je dlouhá 1 až 4 nukleotidy. Minisatelitní VNTR obsahuje opakující se jednotku o délce od 9 do 65 bp. Některé z těchto polymorfismů se mohou objevit v blízkosti genů a mohou tak ovlivnit jejich expresi (Strachan, 2004).

Zhruba 45% lidského genomu tvoří transpozonové repetice. Většina z nich není aktivní, ale některé LINE1, Alu a LTR transpozony jsou stále aktivně transportovány. A jako důsledek v evoluci se opakujících transpozic jsou některé polymorfní.

Některé polymorfismy ovlivní tak velké oblasti nekódují DNA, že mohou být rozeznány cytogenetickými analýzami.

4. Testování asociace

Asociace je statisticky významná závislost mezi dvěma nebo více jevy. V genetické epidemiologii se jedná o výpověď o výskytu alel a fenotypů. Alela A je asociována s nemocí N, pokud lidé, kteří mají N, mají zároveň alelu A statisticky významně častěji (nebo i méně často), než by to odpovídalo předpovědi frekvencí A a N v populaci. Asociační metody testují, zda jsou frekvence alel či celých genotypů odlišné mezi 2 skupinami, obvykle případy a kontrolami. V posledních desetiletích se asociačním metodám dostává velké pozornosti (Risch, 1996).

4.1 Vazba, vazbová nerovnováha

Pokud jsou dva geny ve vazbě, pravděpodobnost jejich přenosu do stejné gamety je větší než 50%. Vazba je tedy specificky genetický vztah. Vazba sama o sobě neprodukuje asociaci v populaci. Vytváří asociace pouze uvnitř rodin, ale nikoliv mezi nepříbuznými jedinci. Z toho vyplývá, že studie vazby mohou pouze použít rodinná data, zatímco studie asociace mohou být založeny na rodinných i populačních datech. Díky širšímu poli aplikace jsou studie asociace považovány za citlivější testy (s větší statistickou sílou) než studie vazby.

Pojem vazbová nerovnováha (linkage disequilibrium, LD) znamená vztah mezi dvěma alelami, které se spolu na chromozomu vyskytují častěji, než by se dalo přisuzovat náhodě. Tyto alely si jsou na chromozomu fyzicky blízko a velmi zřídka jsou od sebe odděleny vlivem

rekombinace. Většinou je jedním z míst pozorovaný lokus s testovatelným markerem a druhým místem je lokus se zatím neznámým genem, který má vliv na výskyt nemoci. Spielman a Ewens definují vazbovou nerovnováhu jako přítomnost jak vazby tak asociace mezi markerem a nemocí (Spielman, 1996). Míra vazbové nerovnováhy je míra asociace alel na haplotypech. Míra LD se obvykle značí D` a pohybuje se v rozmezí 0 až 1. Pokud dva lokusy nejsou ve vazbě, například jsou na jiných chromozomech, D` = 0. V případě, že se hodnoty D` blíží 1, předpokládáme, že je lokus s markerem blízko lokusu, který má vliv na nemoc. Musíme ale vědět, kde se lokus s vlivem na vznik nemoci nachází. Statistické metody zahrnující LD mají významně vyšší pravděpodobnost nálezu genů s vlivem na onemocnění než studie vazby (Risch, 1996). Typy LD mohou významně kolísat uvnitř i mezi populacemi kvůli regionální variabilitě rekombinačních typů, genetickému driftu, etnické diverzitě, populačnímu promísení atd. (Cargill, 1999).

Jak jsem již výše zmínila, studie vazby, jako nástroj mapování míst v genomu se souvislostí s nemocí, mají limitovaný potenciál. Mohou uspět pouze pokud je genový efekt velký. Jednou z možností jak kvantifikovat tento efekt, je parametr γ. Risch a Merikangas definovali parametr γ následovně: předpokládejme, že riziko vzniku nemoci pro jednince např. s genotypem Aa (frekvence alely A + frekvence alely a = 1) je γ-krát větší než pro jednince s genotypem např. aa. Čím vyšší je parametr γ, tím vyšší je šance, že jedinec s určitým genotypem bude postižen studovanou nemocí (Risch, 1996). Risch a Merikangas ukázali, že studia vazby pravděpodobně budou úspěšná pouze při hodnotách γ rovných 4 a větším, ale ne pokud se $\gamma = 2$ a méně. Jako alternativu navrhli genomické asociační studie (Risch, 1996). Mezi některé praktické výhody studií asociace patří, že mají značnou statistickou sílu najít geny s malými efekty. Navíc tento typ studie nevyžaduje rodiny s více nemocnými jedinci, což je celkem vzácné. Nevýhodou asociačních studií je, že polymorfismy uvnitř genů, které mají vliv na nemoc, musí být pokusně identifikovány před testováním asociace (Risch, 1996).

4.2 Studium asociace

Genetické studie asociace hodnotí korelace mezi genetickými variantami a fenotypovými odlišnostmi na populační úrovni. Genetická stránka studie pracuje v nynější době s téměř dvěma miliony identifikovaných polymorfismů (viz kapitola 3.2). Fenotypy hodnocené ve studiích asociace zahrnují status nemoci (např. osteoporóza), míry určitých relevantních veličin (např. hustota kosti) a citlivost na stimuly vnějšího prostředí (např.

účinnost léků). Asociační studie mohou být rodinného nebo populačního charakteru. Rodinné studie vyžadují data případů a jejich rodičů nebo jiných příbuzných – sourozenců, bratranců a sestřenic, atd. Populační studie užívají jako kontroly nepříbuzné jedince.

Je nutné vyřešit otázku, jaké markery použít pro studium asociace. Bylo zjištěno, že mikrosatelitní markery nejsou ideálním typem markeru pro testování asociace (Allen-Brady, 2003). Za vhodné markery byly zvoleny jednonukleotidové polymorfismy (SNP) a to ze dvou důvodů. Narozdíl od mikrosatelitů jsou dostatečně početné, aby definovaly ostrovy LD. SNP rovněž prochází mutacemi méně často než mikrosatelity. Pokud totiž platí předpoklad, že náchylnost k onemocnění je hlavně určena běžnými, letitými DNA variantami, pak je nutné užít k identifikaci ancestrálních haplotypů markery, které jsou stabilní během dlouhé doby. Pravdou ovšem zůstává, že nikdo zatím nezná optimální strategii pro výběr DNA markerů. Zřejmě každá studie vyžaduje strategii, která jí bude „šita na míru“.

Příčin nálezu asociace může být hned několik. První je přímá příčinnost (kauzalita). Tzn. že přítomnost alely A zvyšuje pravděpodobnost výskytu nemoci N. Druhou příčinou může být přírodní selekce, což odpovídá stavu, kdy přítomnost nemoci N i alely A zvyšuje pravděpodobnost přežití. Dalším důvodem nálezu asociace může být populační rozvrstvení (stratifikace) a to v případě, že se populace skládá z několika geneticky vzdálených subpopulací a alela A s nemocí N jsou časté v jedné ze subpopulací. Problém populační stratifikace je kompletně eliminován užitím příbuzných jako kontrol, protože pocházejí ze stejné genetické skupiny. Čtvrtou příčinou může být chyba 1. druhu, tj. náhoda. Eliminujeme ji opakováním výsledků na nezávislém souboru dat. Asociační studie obyčejně testují vysoký počet markerů. Na hladině p = 0,05 bude 5% výsledků statisticky významných i bez jakéhokoli pravdivého efektu. Proto p hodnota potřebuje korekci v závislosti na počtu zkoumaných otázek (viz kapitola 5.5). V minulosti někteří výzkumní pracovníci aplikovali na p hodnotu špatné korekce a nálezy asociace nemohly být replikovány v následujících studiích.

Z výše vyjmenovaných možných příčin nálezu asociace je zřejmé, že existuje mnoho možností, jak asociaci nesprávně určit. Proto je nutné pokusit se dodržovat určitá pravidla. Pokud je to možné, mělo by být vybráno několik kontrolních populací, aby byla maximalizována stejnorodost vzhledem k populaci nemocných. Velikost vzorku by měla být dostatečně velká, aby pokryla ne úplně optimální podmínky, např. vzácně se vyskytující alely nebo neúplnou LD.

Studia asociace ovšem přináší i některá úskalí, na která je třeba upozornit. Běžnými chybami jsou (i) malá velikost výběru, (ii) mnohonásobné testování, (iii) náhodná chyba, (iv) špatně strukturálně vyvážená kontrolní skupina, (v) chyba v detekci LD se sousedním

lokusem, (vi) přecenění výsledků, (vii) nezaručená deklarace „kandidátního genu“ po identifikování asociace v libovolném genetickém regionu (Cardon, 2001). Přes všechny výše zmíněné problémy (a mnohé další) zůstávají studia asociace základním krokem k určení genetických variant zprostředkujících různá onemocnění.

5. Studie případů a kontrol

Tradiční studie případů a kontrol poskytují silnou a účinnou metodu, která hodnotí asociaci mezi kandidátními geny a onemocněním. Prvním krokem je stanovení hypotézy, následně definujeme případy a kontroly, stanovíme expozici a připravíme statistickou analýzu získaných dat. Případy jsou definovány jako jedinci se sledovanou nemocí či jiným znakem. Je nutné stanovit přesné podmínky, na základě kterých budou případy vybírány, protože to ovlivní míru zobecnitelnosti výsledků. Kontroly by měly být co možná nejpodobnější případům, kromě přítomnosti onemocnění nebo jiného sledovaného jevu. Rozlišujeme dvě základní varianty kontrol: náhodně vybrané z populace případů nebo rodinné příslušníky případů. Výběr kontrol je jednou z nejsložitějších částí studie a budu se mu podrobněji věnovat v kapitolách 5.1 a 5.2.

Pokud je počet případů příliš nízký a nelze jej nijak jednoduše zvýšit, řešením je zvýšit počet kontrol na jeden případ. Tak dosáhneme zvětšení statistické síly testu. V případě, že nejsme přesvědčeni o reprezentativnosti našeho souboru kontrol, můžeme použít více kontrolních skupin. Problém však může nastat, když nám analýzy založené na různých kontrolních skupinách přinesou rozdílné výsledky. Nemusí být zřejmé, která odpověď na hypotézu je správná.

5.1 Rodinné studie případů a kontrol

Rodinné studie případů a kontrol (family-based case-control studies) se staly populárními, protože jsou zdatné vůči potenciální populační genetické heterogenitě. Možná nejoblíbenější je užití jednoho nemocného dítěte (případu) a jeho rodičů. Kontrolami jsou v tomto případě možné genotypy, které nebyly přeneseny od rodičů potomkovi. Poměrem alel, které rodiče na potomka přenesli a které ne, se zabývá metoda TDT – transmission/disequilibrium test (viz kapitola 5.3). Místo genotypů rodičů mohou být zvoleni i sourozenci a bratranci a sestřenice nemocného. Rodinné studie vždy nejsou ideálním řešením. Počet kontrol v každé rodině je omezen, genotypy rodičů někdy nejsou informativní

a ne každý případ má k sobě vhodnou kontrolu. Příbuzní navíc mohou být podobnější případům ve vztahu k expozici, o kterou se zajímáme než náhodný vzorek populace.

5.2 Populační studie případů a kontrol

V populačních studiích případů a kontrol (population-based case-control studies) jsou kontroly vybírány z populace, ze které pochází případy. U těchto kontrol hrozí tzv. populační stratifikace, tj. situace, kdy genové frekvence na studovaném lokusu kolísají mezi subpopulacemi (viz také kapitola 4.2). Asociace se může rovněž objevit mezi geny, které spolu nejsou ve vazbě, a to v etnicky různých populacích, jednoduše jako výsledek smísení různých alelových frekvencí. Zkreslení je kontrolováno strukturálně vyváženými kontrolami (tj. kontroly splňující určité parametry – pohlaví, věk) a stratifikační analýzou. Tyto postupy ovšem vyžadují, aby byl zkreslující faktor znám. Kontrola zkreslení může být velmi složitá pro neznámé geny, které mohou mít silné gradienty mezi nebo uvnitř etnických skupin. V ideálním případě by kontroly měly reflektovat etnické a genetické složení vzorku případů.

5.3 Testování ve studiích případů a kontrol – TDT

Pokud jsou případy a kontroly vybírány do studie různým způsobem, může to způsobit tzv. výběrové zkreslení. V důsledku tohoto zkreslení lze pak pouze velmi obtížně zajistit srovnatelnost skupiny případů a skupiny kontrol, neboť kontrolní skupina nereprezentuje populaci, ze které pocházejí případy (Bencko, 2003). Výběr kontrol je velmi složitý, i protože mnoho populací je rozrůzněno do etnických podskupin, protože etnické pozadí je velmi těžko měřitelné a protože genetické markery velmi často kolísají mezi etnickými skupinami. Snaha vyhnout se těmto problémům vedla k předložení několika typů testu. Jedním z nich byl i R. S. Spielman, který roku 1993 představil TDT (transmission/disequilibrium test) jako test vazby mezi komplexní nemocí a genetickým markerem (Spielman, 1993). Původně měl TDT testovat vazbu s markerem lokalizovaným blízko kandidátního genu, a to v případě, kdy již byla nalezena asociace. Poslední dobou je ale TDT využíván jako test asociace v přítomnosti vazby nebo i přímo pro kandidátní polymorfismy. TDT používá data z rodin s nejméně jedním nemocným dítětem a hodnotí přenos asociované alely z heterozygotního rodiče na nemocného potomka. Procedura testování porovnává kolikrát heterozygotní rodič přenesl asociovaný marker na nemocného potomka a kolikrát rodič přenesl alternativní alelu (Spielman, 1993). Tento test není limitován přítomností stratifikace, protože jak případy tak

kontroly pochází jistě z jedné populace. TDT je platným testem vazby a asociace i v případě, že asociace je zapříčiněna rozčleněním nebo promísením populací (Ewens, 1995).

Ještě před Spielmanovou metodou byly představeny i jiné testy. Falk a Rubinstein navrhli užití rodičů případů jako alternativní typ kontroly v proceduře zvané „haplotype relative risk“ (HRR) (Falk, 1987 in Schaid, 1998). Dvě alely rodičů, které nebyly přeneseny na jejich nemocné dítě, byly zkombinovány a vytvořily tzv. pseudokontrolu. Frekvence jednotlivých markerů mezi případy a pseudokontrolami byly použity k získání poměru šancí (odds ratio – viz kapitola 5.7), který nazýváme HRR v případě pseudokonrol.

Zpět k TDT. Podoba statistiky TDT je docela jednoduchá. Předpokládáme existenci lokusu N, který má vliv na přítomnost nemoci. Lokus N obsahuje alelu N_1 s vlivem na nemoc a normální alelu N_2. Dále máme marker s alelami M_1 a M_2. V nejobecnější rovině jsou data, která analyzujeme počty „přenosů“. Tzn. že pro rodiče s genotypy M_1M_1, M_1M_2, M_2M_2 určíme, kolikrát byla alela M_1 nebo M_2 přenesena na nemocného potomka (Spielman, 1993). Tyto počty označíme následovně:

a – kolikrát genotyp M_1M_1 přenesl M_1 na nemocného potomka;

b – kolikrát genotyp M_1M_2 přenesl M_1 na nemocného potomka;

c – kolikrát genotyp M_1M_2 přenesl M_2 na nemocného potomka;

d – kolikrát genotyp M_2M_2 přenesl M_2 na nemocného potomka.

Nulová hypotéza zní, že marker a onemocnění nejsou ve vazbě. Již víme, že TDT používá pouze data od heterozygotních rodičů (M_1M_2), tzn. *b* a *c*. Statistika TDT zní: $(b - c)^2/(b + c)$. Pokud mezi markerem a nemocí existuje vazba, stejně jako alelická asociace, *b* a *c* se ve svých hodnotách budou lišit. V případě, že neexistuje vazba, alely M jsou rozdělovány nezávisle na nemoci, přítomnost asociace nevyústí v rozdílné hodnoty *b* a *c*. Statistická významnost TDT je hodnocena χ^2 (McNemarův test – viz 5.10.1) nebo přesným binomiálním testem (Spielman 1993, Spielman 1996).

Mezi výhody TDT patří jeho jednoduchá aplikace na různé skupiny, jeho necitlivost k populační stratifikaci, takže se vyhýbá nálezu falešné asociace. Metoda TDT je ve většině případů citlivější než jiné rodinné testy asociace, data z rodin s jedním nemocným mohou být lehce kombinovány s daty rodin s více nemocnými potomky. Vážnou nevýhodou TDT ale je, že vyžaduje informace o genotypu obou rodičů, což výrazně limituje její použití na nemoci, které se objevují v pozdějším věku života.

Jednou z možností, jak obejít problém chybějících rodičů je metoda zvaná sib TDT neboli S-TDT (Spielman a Ewens, 1998). S-TDT používá místo rodičovských dat data od zdravých sourozenců (siblings). Tato aplikace nebrání upotřebit metodu TDT. Data

z markerových genotypů sourozenců musí splňovat dva požadavky: je nutné, aby v příbuzenstvu byl nejméně jeden nemocný a jeden zdravý sourozenec a tito sourozenci nesmí mít stejný genotyp ve zkoumaném lokusu. Aplikace S-TDT je vhodná u nemocí, které se objevují později během života jako je např. diabetes 2. typu, kardiovaskulární nemoci, nebo jiné nemoci mající souvislost s věkem – např. Alzheimerova demence nebo Parkinsonova nemoc.

5.4 Testování ve studiích případů a kontrol – populační kontroly

Alternativou k TDT jsou populační kontroly, které v této době přicházejí opět do obliby. Populační studie případů a kontrol potřebují méně vzorků než TDT a jsou jednoduší u nemocí s pozdním propuknutím, kdy jsou data rodičů (ale i sourozenců) zřídka dostupná. Výběr kontrol tak, aby byly co nejbližší případům z hlediska etnického i genetického, je složitý, jak již jsem několikrát zmínila. Jedním z řešení může být shromáždění několika kontrolních skupin, které by reflektovaly různé substruktury zkoumané populace.

Obava z možnosti nálezu falešné asociace z důvodu populační stratifikace je, zdá se, příliš přehnaná. Např. česká populace je velmi homogenní. Navíc data mohou být zkontrolována porovnáním frekvencí alel v řadě lokusů případů a kontrol, které nejsou ve vazbě (Pritchard, 1999). Podle Rische a Tenga je optimálním designem studie asociace užití dvojice nemocných sourozenců jako případy a dvě nepříbuzné kontroly (Risch, 1998).

5.5 Korekce na mnohonásobné testování

Je nutné dávat si pozor na přílišné přikládání důležitosti jednomu statisticky významnému výsledku mezi množstvím statisticky nevýznamných výsledků (Bland, 1995). Pokud testujeme nulovou hypotézu na 5% hladině významnosti a ta hypotéza je ve skutečnosti pravdivá, s pravděpodobností 95% získáme statisticky nevýznamný, to jest správný, výsledek. Ovšem v případě, že takových hypotéz budeme testovat dvacet, pravděpodobnost statisticky nevýznamného závěru bude $0{,}95^{20} = 0{,}36$. To nám dává $((1-0{,}95^{20}) = 1-(1-0{,}05)^{20} = 1-0{,}36 = 0{,}64)$ 64% šanci, že získáme minimálně jeden statisticky „významný“ výsledek. Všimněme si, že máme větší pravděpodobnost nálezu signifikatního výsledku než správného nesignifikantního. Mnohonásobné testování významnosti tedy dává vysokou pravděpodobnost, že bude výrazný rozdíl nalezen náhodou.

Nejznámější metodou řešící korekci na mnohonásobné testování je Bonferroniho metoda. Cílem metody je udržet hladinu pravděpodobnosti chyby 1. druhu (většinou 5%) a v podstatě ji rozdělit na všechna porovnání (Zvárová, 1998). Bonferroniho metoda upravuje hladinu významnosti α. Tuto hladinu vydělíme celkovým počtem porovnání, které chceme provézt - např. 4 porovnání na 5% hladině: α = (5/4)% = 1,25%. Porovnání pak testujeme ne na 5% hladině, ale na hladině významnosti 1,25%.

Příkladem využití Bonferroniho korekce na mnohonásobné testování může být situace, kdy marker obsahuje více než dvě alely. Pro testování každé z nich je vypočítáno několik TDT a je nutné uplatnit korekci, abychom se vyhnuli zvyšující se pravděpodobnosti nálezu falešně pozitivního výsledku.

5.6 Kontingenční tabulky

Ve studiích případů a kontrol porovnáváme dvě skupiny. Jejich data jsou obvykle prezentována ve formě kontingenční tabulky (contingency table). Tato tabulka je vyžadována pro testování statistické významnosti, odhady rizik (poměr šancí a relativní riziko - viz kapitoly 5.7 a 5.8) a pro výpočet intervalu spolehlivosti (viz kapitola 5.9). V genetické epidemiologii řádky (resp. sloupce) popisují stav jedinců (případy nebo kontroly) a sloupce (popř. řádky), referují o specifických alelách, genotypech nebo haplotypech. Z tabulky 1 je zřejmé, že kontingenční tabulka může obsahovat více než 2 sloupce a 2 řádky.

	Genotyp			
Status	**AA**	**AB**	**BB**	**Celkem řádky**
Případ	23	47	30	100
Kontrola	12	40	48	100
Celkem sloupce	35	87	78	200

Tabulka 1. Hypotetická data v kontingenční tabulce pro případy a kontroly zkoumaných v dialelickém lokusovém markeru (upraveno dle Gordon, 2005)

5.7 Kvantifikace rizika – odds ratio

Poměr šancí (odds ratio, OR) je jednou ze statistik hodnotících nebezpečí výskytu určité nemoci, pokud je přítomný daný faktor či expozice. Odds ratio je poměr pravděpodobnosti, že se daná událost stane a pravděpodobnosti, že se nestane. To je obvykle

testováno jako poměr počtu případů, kdy se událost stala a počtu případů, kdy tomu bylo naopak. Odds ratio je nejpoužívanější statistickou metodou odhadující míru rizika ve studiích případů a kontrol. Data, která použijeme pro kalkulaci OR, mohou být shrnuty do 2x2 tabulky.

	N+ (případy)	**N– (kontroly)**	**Součet**
E+	a	b	a + b
E-	c	d	c + d
Součet	a + c	b + d	a + b + c + d

N+ = nemocný; N- = zdravý; E+ = exponovaný určitému faktoru; E- = neexponovaný určit. faktoru

Tabulka 2. 2x2 tabulka pro výpočet OR a RR

(upraveno dle Knapp, 1992)

Pro studii případů a kontrol vypadá výpočet OR takto (upraveno dle Knapp, 1992):

$$OR = \frac{šance,_ že_ nemocný_ byl_ vystaven_ určitému_ faktoru}{šance,_ že_ zdravý_ byl_ vystaven_ určitému_ faktoru}$$

$$OR = \frac{O_{E+|N+}}{O_{E+|N-}}$$

$$OR = \frac{P(E+ \mid N+) / P(E- \mid N+)}{P(E+ \mid N-) / P(E- \mid N-)}$$

$$OR = \frac{a/c}{b/d}$$

$$OR = ad/bc$$

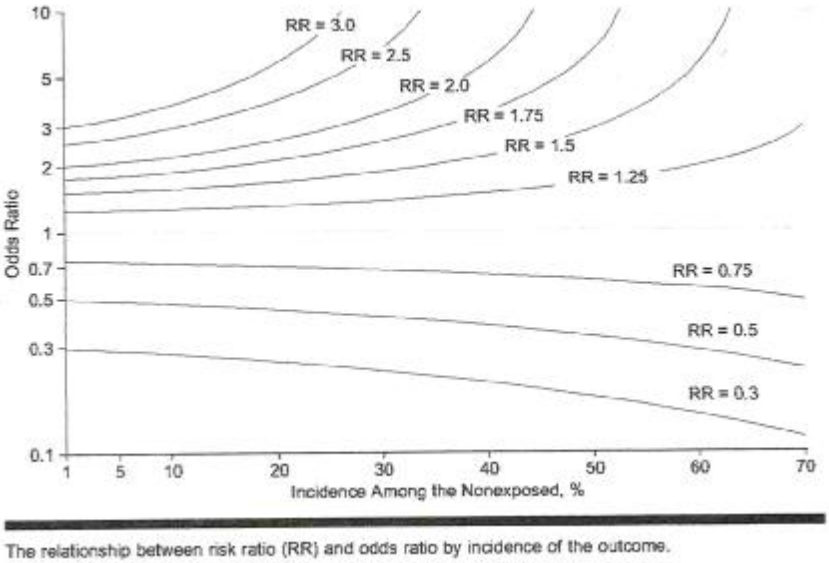
Hodnoty OR nemůžou nabývat záporných hodnot, nejsou ohraničeny shora. Čím vyšší je hodnota OR, tím silnější je asociace mezi onemocněním a zkoumaným faktorem. Pokud je OR okolo 1, pak spolu nemoc a zkoumaný faktor nesouvisí, tzn. že šance, že jedinec vystavený určitému faktoru onemocní, je stejná, jako u jedince, který stejnému faktoru vystaven nebyl. Hodnoty OR menší než 1 ukazují negativní asociaci, tzn. že sledovaný faktor má protektivní efekt (Grant, 1992). Zlogaritmováním hodnoty OR můžeme vypočítat standardní odchylku (SE) pro log OR a dále i intervaly spolehlivosti okolo odhadů rizika (viz kapitola 5.9).

$$SE(\log(or)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

OR je velmi užívané v lékařských výzkumech. Důvodem je, že poskytuje odhad (s intervaly spolehlivosti) vztahu mezi dvěma binárními (ano – ne) veličinami, dále umožní zkoušet efekty dalších proměnných na tento vztah užitím logistické regrese. Poskytuje rovněž vyhovující interpretaci ve studiích případů a kontrol (Bland, 2000). OR je navíc běžnou cestou prezentování výsledků meta-analýzy – statistické analýzy, která kombinuje výsledky několika studií (Crichton, 2001).

5.8 Kvantifikace rizika – relative risk

Relativní riziko (relative risk, RR) udává kolikrát větší je pravděpodobnost vzniku onemocnění ve skupině se sledovaným faktorem E, tzn. v populaci exponovaných, než ve skupině bez sledovaného faktoru (E- - populace neexponovaných). Ve studiích případů a kontrol nemůže být RR získáno přímo, proto se častěji využívá OR (viz kapitola 5.7). Obrázek 1 ukazuje, že pokud je výskyt onemocnění v populaci nízký (menší než 10%), hodnota OR je v blízkosti RR. Ale pokud je incidence nemoci častější, OR nadhodnocuje relativní riziko (v případě, že OR >1) nebo naopak podceňuje RR (v případě, že OR <1).



Obrázek 2. Vztah mezi RR a OR vzhledem k incidenci nemoci (Zhang, 1998)

K výpočtu RR mohu být použita data z 2x2 tabulky (viz tabulka 2). Poté je RR vypočítáno následovně (upraveno dle Knapp, 1992):

$$RR = \frac{P(N+ \mid E+)}{P(N+ \mid E-)}$$

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

Jelikož RR není ve studiích případů a kontrol využíván tolik jako OR (spíše v kohortových studiích), nebudu mu věnovat velkou pozornost. Pouze konstatuji, že u nemocí s incidencí větší než 10% nelze OR aproximovat na RR přímo. Vhodné metody, které z OR odhadnou RR jsou například stratifikační analýza a log-binominální model (McNutt, 2003).

5.9 Intervaly spolehlivosti

Interval spolehlivosti (confidence interval, CI) ohraničuje rozpětí, do něhož s určitou pravděpodobností (obvykle 95%) spadá výsledek testované hypotézy nebo odhad poměru šancí. Jinak řečeno, pokud je konfidenční interval spočítán na 95% hladině, znamená to, že 95% výběru z populace bude mít zkoumaný populační parametr uvnitř tohoto intervalu, neboli že 95% pokusů (opakovaných stále dokola) bude zahrnovat skutečnou populační hodnotu.

Obecný vzorec pro výpočet intervalu spolehlivosti na 95% hladině je: statistika ± 1,96 SE, kde SE vyjadřuje standardní odchylku statistiky a 1,96 je kritická hodnota (Z_C) pro p = 0,05 (pro p = 0,01 platí Z_C = 2,58). Kritická hodnota nezávisí na velikosti zkoumaného vzorku. Výpočet intervalu spolehlivosti pro OR nebo RR vyžaduje jejich přirozené zlogaritmování, protože bez toho nemají normální rozdělení. Výpočet CI se provádí Woolfovou metodou (Woolf, 1955 in Knapp, 1992):

95% interval spolehlivosti pro ln OR = ln OR ± 1,96 . $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$

Horní (UL) a dolní (LL) limit tohoto CI je následně spočítán jako antilogaritmus horního a dolního limitu pro ln OR: $UL_{OR} = e^{UL_{lnOR}}$; $LL_{OR} = e^{LL_{lnOR}}$.

5.10 Chí kvadrát testy

Jak jsem již výše zmínila, je statistická významnost asociačních testů vesměs hodnocena chí kvadrát testem (χ^2). Chí kvadrát test funguje na následujícím principu. Pro každou buňku v kontingenční tabulce (viz kapitola 5.6) najdeme frekvenci, kterou bychom očekávali v případě platnosti nulové hypotézy (tzn. žádný vztah mezi proměnnými).

Očekávanou (expected, E) frekvencí je míněna průměrná frekvence, kterou bychom získali při dlouho opakovaném měření. Obecně lze výpočet očekávaných frekvencí (E) definovat vzorcem:

$$E = \frac{\text{součet řádků x součet sloupců}}{\text{celkový součet}}$$

Následně porovnáme očekávané a skutečně získané (observed, O) frekvence. Pokud nejsou námi zkoumané veličiny ve vzájemném vztahu, oba druhy frekvencí by měly mít podobné hodnoty. Nejlepší statistikou, jak porovnat očekávané a získané frekvence je

$$c^2 = \Sigma \frac{(O - E)^2}{E}$$

Velikost výsledného rozdílu ale závisí i na počtu pacientů. χ^2 test je spolehlivý pro velké výběry. Čím menší budou očekávané frekvence, tím pochybnější bude výsledek testu. Statistik W. G. Cochran zavedl pravidlo, že χ^2 test je platný, pokud je nejméně 80% očekávaných frekvencí větších než 5 a všechny překračují hodnotu 1 (Bland, 1995). V případě, že naše výpočty očekávaných frekvencí toto pravidlo nesplňují, obvykle se kombinují nebo mažou sloupce a řádky kontingenční tabulky, tak aby hodnoty byly vyšší. Tato korekce samozřejmě nejde uplatnit u 2x2 tabulek. I když i tato opatření neposkytnou požadované hodnoty nebo pokud máme diskontinuitní charakter dat, přijde na řadu Yatesova korekce

$$c^2 = \sum \frac{\left(|O - E| - \frac{1}{2}\right)^2}{E}$$

nebo Fisherův exaktní test, který přímo počítá pravděpodobnost výskytu dané čtyřpolní tabulky. Exaktní testy jsou výhodné a často využívané, protože berou v úvahu všechny možné alternativy a výsledek bez počítače nespočítáme.

5.10.1 McNemarův test

χ^2 test porovnává shodnost dvou proměnných získaných ze dvou nezávislých výběrů. Pokud ale tyto proměnné pochází ze stejných jedinců, nemůže být χ^2 použit jako test nezávislosti dvou proměnných. Místo něj přichází v úvahu McNemarův test. Tento test je využíván v situacích, kdy je stejný výběr použit k zjištění např. shody mezi dvěma diagnostickými testy nebo rozdílu mezi dvěma druhy léku. V nulové hypotéze testujeme, zda sou dvě frekvence stejné. Očekávané frekvence jsou tedy $(f_a + f_b)/2$ a McNemarova statistika pak vypadá takto:

$$c^2 = \frac{(f_a - f_b)^2}{f_a + f_b}$$

McNemarův test se používá jako test statistické významnosti u TDT (viz kapitola 5.3).

6. Závěr

Z observačních epidemiologických studií poskytují statisticky nejsilnější výsledky studie případů a kontrol a kohortová studie, která je však časově i finančně náročnější.

Variabilita lidského genomu je v největší míře zapříčiněna existencí obrovského množství polymorfismů. Jejich lokalizace a charakterizace (zejména pak nejpočetnějších jednonukleotidových polymorfismů) nám poskytuje účinný nástroj k detekci genů způsobujících onemocnění, k předpovědi náchylnosti ke vzniku nemoci i k přípravě účinnějších léčiv.

Jednou z nejpoužívanějších metod k určení či vyvrácení příčinné asociace mezi genem a nemocí je studie případů a kontrol. Tento typ studie je vhodný pro vzácně se vyskytující nemoci, neklade přehnaně vysoké časové a finanční nároky a hodí se i v případech dlouhé časové prodlevy mezi expozicí a následkem. Pokud se vyhneme několika základním prohřeškům (zkreslení vyšetřovanou osobou, zkreslení způsobené experimentátorem, mnohonásobné testování apod.), poskytne nám tento typ studie statisticky významné zhodnocení dat.

Dle charakteru kontrol rozlišujeme rodinný a populační typ studie. Rodinná studie případů a kontrol je výhodnější, jestliže máme možnost získat genetická data obou rodičů. Vyhneme se tak riziku zkreslení vlivem populační stratifikace. Populační studie případů a kontrol vybírá své kontroly z populace případů. Snaží se, aby se co nejvíce shodovaly s případy (kromě přítomnosti onemocnění). Tento typ studie je vhodnější pro výzkum onemocnění, která se projevují později v lidském životě a nejsou tak již dostupná data rodičů nebo i jiných příbuzných. Riziko vzniku nemoci kvantifikujeme poměrem šancí, který je u studie případů a kontrol vhodnější než relativní riziko.

Optimální design studií asociace ale není plošně aplikovatelný na jakýkoli typ dat a způsob zkoumání. Závisí na vlastnostech výsledné proměnné, genové frekvenci, genetickém relativním riziku, důležitosti interakcí gen - gen a gen - prostředí (Thomas, 1999).

Tato bakalářská práce má být teoretickou přípravou na zpracování a vyhodnocení geneticko-epidemiologické studie v diplomové práci.

7. Seznam literatury

- Allen-Brady K., Farnham J.M., Weiler J., Camp N.C. (2003) A Cautionary Note on the Appropriateness of Using a Linkage Resource for Association Study. BioMed Central Genetics 4, 89 – 94
- Bencko V., Holcátová I., Králíková E., Kudlová E., Pikhart H., Rameš J., Slámová A., Wallenfels J., Zvárová J. (2004) Hygiene & Epidemiology. The Karolinum Press, Prague
- Bencko V., Hrach K., Malý M., Pikhart H., Reissigová J., Svačina Š., Tomečková M., Zvárová J. (2003) Biomedicínská statistika III – Statistické metody v epidemiologii. Nakladatelství Karolinum, Praha
- Beyene J., Yan J., Greenwood C.M.T. (2005) Linkage and Association Analysis in Pedigrees from Different Populations. BioMed Central Genetics 6, 59 – 64
- Bland J.M. (1995) An Introduction to Medical Statistics. Oxford University Press Inc., New York
- Bland J.M., Altman D.G. (1995) Multiple Significance Tests: the Bonferroni Method. British Medical Journal 310, 170
- Bland J.M., Altman D.G. (2000) Statistics Notes: The odds ratio. British Medical Journal 320, 1468
- Cardon L.R., Bell J.I. (2001) Association Study Designs for Complex Diseases. Nature Reviews – Genetics 2, 91 – 99
- Cargill M., Altshuler D., Ireland J., Sklar P., Ardlie K., Patil N., Lane C.R., Lim E.P., Kalyanaraman N., Nemesh J., Ziaugra L., Friedland L., Rolfe A., Warrington J., Lipshutz R., Daley G.Q., Lander E.S. (1999) Characterization of Single-Nucleotide Polymorphisms in Coding Regions of Human Genes. Nature Genetics 22, 231 – 238
- Clark A.G., Boerwinkle E., Hixson J., Sing C.F. (2005) Determinants of the Succes of Whole-Genome Association Testing. Genome Research 15, 1463 – 1467
- Crichton N. (2001) Information Point: Odds Ratio. Journal of Clinical Nursing 10, 268 – 269
- Dorak M.T. (2006) Genetic Epidemiology http://www.dorak.info/epi/genetepi.html
- Dorak M.T. (2006) Statistical Analysis in HLA and Disease Association Studies http://www.dorak.info/hla/stat.html
- Ewens W.J., Spielman R.S. (1995) The Transmission/Disequilibrium Test: History, Subdivision, and Admixture. American Journal of Human Genetics 57, 455 – 464

§ Gay J. (1999) Clinical Study Design and Methods Terminology. http://www.vetmed.wsu.edu/courses-jmgay/GlossClinStudy.htm

§ Gordon D., Finch S.J. (2005) Factors Affecting Statistical Power in the Detection of Genetic Association. Journal of Clinical Investigation 115, 1408 –1418

§ Hsu L., Zhao L.P., Aragaki C. (2000) A Note on a Conditional-Likelihood Approach for Family-Based Association Studies of Candidate Genes. Human Heredity 50, 194 – 200

§ Kansagra S, List J. (2004) The Clinical Relevance of Genomic Variation. Journal of the American Medical Association 291, 1641

§ Knapp R.G., Miller M.C. (1992) Clinical Epidemiology and Biostatistics. Williams & Wilkins, Baltimore, Maryland

§ Martin E.R., Kaplan N.L., Weir B.S. (1997) Tests for Linkage and Association in Nuclear Families. American Journal of Human Genetics 61, 439 – 448

§ McNutt L., Wu C., Xue X., Hafner J.P. (2003) Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. American Journal of Epidemiology 157, 940 – 943

§ Moore R.M., Pinel T, Zhao J.H., March R., Jawaid A. (2005) Selecting Cases from Nuclear Families for Case-Control Association Analysis. BioMed Central Genetics 6, 105 - 110

§ Pai M. (2006) Observational Study Design. http://www.sunmed.org/Obser.html

§ Pritchard J.K., Rosenberg N.A. (1999) Use of Unlinked Genetic Marjers to Detect Population Stratification in Association Studies. American Journal of Human Genetics 65, 220 – 228

§ Risch N., Merikangas K. (1996) The Future of Genetic Studies of Complex Human Diseases. Science 273, 1516 – 1517

§ Risch N., Teng J. (1998) The Relative Power of Family-Based and Case-Control Designs for Linkage Disequilibrium Studies of Complex Human Diseases I. DNA Pooling. Genome Research 8, 1273 – 1288

§ Sachidanandam R. *et al.* (2001) A Map of Human Genome Sequence Variation Containing 1.42 Million Single-Nucleotide Polymorphisms. Nature 409, 928 – 933

§ Schaid D.J. (1998) Transmission Disequilibrium, Family Controls, and Great Expectations. American Journal of Human Genetics 63, 935 – 941

§ Schneider J.A., Pungliya M.S., Jiang C.R., Sun X.J., Salisbury B.A., Stephens J.C. (2003) DNA Variability of Human Genes. Mechanisms of Ageing and Development 124, 17 – 25

§ Slager S.J., Schaid D.J. (2001) Evaluation of Candidate Genes in Case-Control Studies: A Statistical Method to Account for Related Subjects. American Journal of Human Genetics 68, 1457 – 1462

§ Spielman R.S., Ewens W.J. (1996) The TDT and Other Family-Based Tests for Linkage Disequilibrium and Association. American Journal of Human Genetics 59, 983 – 989

§ Spielman R.S., Ewens W.J. (1998) A Sibship Test for Linkage in the Presence of Association: The Sib Transmission/Disequilibrium Test. American Journal of Human Genetics 62, 450 – 458

§ Spielman R.S., McGinnis R.E., Ewens W.J. (1993) Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-dependent Diabetes Mellitus (IDDM). American Journal of Human Genetics 52, 506 – 516

§ Strachan T., Read A.P. (2004) Human Molecular Genetics 3. Garland Publishing, New York

§ Thomas D.C. (1999) Design of Gene Characterization Studies: an Overview. Journal of the National Cancer Institute Monographs 26, 17 – 23

§ Venter J.C. *et al.* (2001) The sequence of the Human Genome. Science 291, 1304 – 1351

§ Venter J.C. Institute. Genome News Network. http://www.genomenewsnetwork.org/

§ Wang D.G. *et al.* (1998) Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. Science 280, 1077 – 1082

§ Zhang J., Yu K.F. (1998) What´s the Relative Risk? Journal of the American Medical Association 280, 1690 – 1691

§ Zvárová J. (1998) Biomedicínská statistika I – Základy statistiky pro biomedicínské obory. Nakladatelství Karolinum, Praha